

**UNIVERSITE LILLE NORD DE FRANCE
ÉCOLE DOCTORALE BIOLOGIE SANTE
FACULTE DE MEDECINE HENRY WAREMBOURG**

**THESE SCIENTIFIQUE
POUR L'OBTENTION DU GRADE DE
DOCTEUR DE L'UNIVERSITE DE LILLE 2
BIostatISTIQUES**

**Soutenue le 09/02/2011
par Emmanuel Chazard**

**AUTOMATED DETECTION OF ADVERSE
DRUG EVENTS BY DATA MINING OF
ELECTRONIC HEALTH RECORDS**

Jury :

Pr. Elske Ammenwerth
Pr. Régis Beuscart
Pr. Paul Landais
Pr. Nicos Maglaveras
Pr. Christian Nøhr
Pr. Cristian Preda
Pr. Alain Venot

Examineur
Examineur
Rapporteur
Rapporteur
Examineur
Examineur
Examineur

SUMMARY

Automated Detection of Adverse Drug Events by Data Mining of Electronic Health Records

Introduction

Adverse Drug Events (ADE) are injuries due to medication management rather than the underlying condition of the patient. They endanger the patients and most of them could be avoided. The detection of ADEs usually relies on spontaneous reporting or medical chart reviews. The objective of the present work is to automatically detect cases of ADEs by means of Data Mining, which are a set of statistical methods particularly suitable for the discovery of rules in large datasets.

Material

A common data model is first defined to describe the available data extracted from the EHRs (electronic health records). More than 90,000 complete hospital stays are extracted from 5 French and Danish hospitals. Those complete records include diagnoses, lab results, drug administrations, administrative and demographic data as well as free-text reports. When the drugs are not available from any CPOE (Computerized Prescription Order Entry), they are extracted from the free-text reports by means of semantic mining. In addition, an exhaustive set of SPCs (Summaries of Product Characteristics) is provided by the Vidal Company.

Methods

We attempt to trace all the outcomes that are described in the SPCs in the dataset. By means of data mining, especially Decision Trees and Association Rules, the patterns of conditions that participate in the occurrence of ADEs are identified. Many ADE detection rules are generated; they are filtered and validated by an expert committee. Finally, the rules are described by means of XML files in a central rules repository, and are executed again for statistics computation and ADE detection.

Results

236 ADE-detection rules have been discovered. Those rules enable to detect 27 different kinds of outcomes. Several statistics are automatically computed for each rule in every medical department, such as the confidence or the relative risk. Those rules involve innovative conditions: for instance some of them describe the consequences of drug discontinuations.

In addition, two web tools are designed and are available through the web for the physicians of the departments: the *Scorecards* enable to display statistical and epidemiological information about ADEs in a given department and the *Expert Explorer* enables the physicians to review the potential ADE cases of their department.

Finally, a preliminary evaluation of the clinical impact of the potential ADEs is performed as well as a preliminary evaluation of the accuracy of the ADE detection.

RESUME

Détection automatisée d'Effets Indésirables liés aux Médicaments par fouille statistique de données issues du dossier patient électronique

Introduction

Les effets indésirables liés aux médicaments (EIM) sont des dommages liés au traitement médicamenteux plutôt qu'aux conditions sous-jacentes du patient. Ils mettent les patients en danger, et la plupart d'entre eux sont évitables. La détection des EIM repose habituellement sur les reports spontanés d'EIM et sur la revue de dossiers. L'objectif du présent travail est d'identifier automatiquement les cas d'EIM en utilisant des méthodes de *Data Mining* (fouille statistique de données). Le *Data Mining* est un ensemble de méthodes statistiques particulièrement adaptées à la découverte de règles dans de grandes bases de données.

Matériel

Un modèle de données commun est tout d'abord défini, dans le but de décrire les données qui peuvent être extraites des dossiers patient électroniques. Plus de 90 000 séjours hospitaliers complets sont extraits de 5 hôpitaux français et danois. Ces enregistrements incluent les diagnostics, les résultats de biologie, les médicaments administrés, des informations démographiques et administratives, et enfin du texte libre (courriers, comptes-rendus). Lorsque les médicaments ne peuvent être extraits d'un CPOE (système de prescription connectée), ils sont extraits des courriers par *Semantic Mining* (fouille de texte). De plus, la société Vidal fournit un ensemble exhaustif de RCP (Résumés des Caractéristiques du Produit).

Méthode

On tente de tracer dans les données tous les événements indésirables décrits dans les RCP. Puis en utilisant les méthodes de *Data Mining*, en particulier les arbres de décision et les règles d'association, on identifie les circonstances qui favorisent l'apparition d'EIM. Plusieurs règles de détection des EIM sont ainsi obtenues, elles sont ensuite filtrées et validées par un comité d'experts. Enfin, les règles sont décrites sous forme de fichiers XML et stockées dans une base. Elles sont exécutées afin de calculer certaines statistiques et de détecter les cas d'EIM.

Résultats

236 règles de détection des EIM sont ainsi découvertes. Elles permettent de détecter 27 types d'événements indésirables différents. Plusieurs statistiques sont calculées automatiquement pour chaque règle dans chaque service, comme la confiance ou le risque relatif. Ces règles impliquent des conditions innovantes : par exemple certaines règles décrivent les conséquences de l'arrêt d'un médicament.

De plus, deux outils Web sont développés et mis à la disposition des praticiens via Internet : les *Scorecards* permettent de présenter des informations statistiques et épidémiologiques sur les EIM propres à chaque service, tandis que l'*Expert Explorer* permet aux médecins d'examiner en détail les cas probables d'EIM de leur service.

Enfin, une évaluation préliminaire de l'impact clinique des EIM est menée, ainsi que l'évaluation de la précision de détection des EIM.

ACKNOWLEDGEMENTS

I would like to thank the members of the board of examiners, and especially the reviewers. I am very grateful to them for it.

I would like to thank all the contributors to this work, especially:

- The people who provided me with the data or helped me extracting the data:
 - o Mrs. Julie Niès & Mr. Bertrand Guillot from the Medasys Company
 - o Dr Michel Degroisse, Mrs. Nicole Radi & Mrs. Laurie Ferret from the Denain General Hospital
 - o Mrs. Sanne Jensen, Mr. Kenneth Skovhus Andersen & Mr. Preben Poul Grothe Jensen from the Capital Region Hovedstaden Hospitals (Denmark)
 - o Pr Stefan Darmoni, Dr Philippe Massari & Mr. Ivan Kergourlay from the Rouen University Hospital
 - o Mr. Mostafa Maazi from the Lille University Hospital
 - o Mr. Jean-Charles Sarfati from the Oracle Company
- The people who performed the anonymization and the semantic mining of the free-text records and provided me with information about the indexing tools:
 - o Pr Stefan Darmoni, Dr Philippe Massari & Mr. Ivan Kergourlay from the Rouen University Hospital
 - o Mrs. Suzanne Pereira from the Vidal Company
- The people who provided me with already existing structured rules, detailed description of those rules and helped me filtering and computing them:
 - o Mr. Ludovic Durand-Texte from the Vidal Company
 - o Dr Grégoire Ficheur from the Lille University Hospital
- The people who brought important reflections about the data-mining-based rule induction process:
 - o Mr. Cristian Preda from the Lille 1 University
 - o Pr Régis Beuscart, Dr Grégoire Ficheur & Dr Béatrice Merlin from the Lille University Hospital
 - o Dr Peter Mc Nair from the Kennedy Center (Denmark)
 - o Mrs. Jytte Brender Mc Nair from the Aalborg University (Denmark)
- The clinicians, pharmacologists, pharmacists, and experts who helped filtering the rules, and writing the labels:
 - o Mrs. Elisabeth Serrot & Mrs. Sophie Tessier from the Vidal Company
 - o Pr Jacques Caron, Dr Sophie Gauthier, Dr Béatrice Merlin & Mr. Pierre Fontana from the Lille University Hospital
- The computer scientists who implemented the Expert Explorer and the Scorecards and the psychologists and ergonomists who helped me improving those tools:

- Mr. Adrian Baceanu & Mr. Ionut Atasiei from the IDEEA Company (Romania)
- Mrs. Marie-Catherine Beuscart-Zéphir, Mr. Romaric Marcilly & Mr. Nicolas Leroy from EVALAB
- The people who brought important reflections about the output of the present work:
 - Mr. Vassilis Koutkias & Mr. Vassilis Kilintzis from the Aristotle University (Greece)
 - Mrs. Elske Ammenwerth & Mr. Werner Hackl from the UMIT University (Austria)
- The people who are currently involved in the review of the clinical cases:
 - Pr Régis Beuscart, Dr Matthieu Genty, Dr Jean-Baptiste Beuscart & Dr Julie Quentin
- The people who brought me significant help in the writing of this document:
 - Pr Régis Beuscart, Dr Grégoire Ficheur, Mrs. Stéphanie Bernonville & Mrs. Marie-Catherine Beuscart-Zéphir

I would like to offer my most sincere thanks to Pr Régis Beuscart for having involved me in such an interesting work, for having trusted and encouraged me all along the project.

Finally, I would like to thank my wife Paule, my brother Edouard, my sister Caroline and my parents for their unfailing support.

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.

TABLE OF CONTENTS

1. INTRODUCTION.....	15
1.1. DEFINITION OF ADVERSE DRUG EVENTS.....	15
1.2. STATE OF THE ART IN ADVERSE DRUG EVENTS DETECTION	16
1.3. STATE OF THE ART IN DATA MINING	18
1.3.1. <i>Introduction.....</i>	18
1.3.2. <i>Requirements of data mining</i>	19
1.3.3. <i>Review of the methods.....</i>	25
1.3.4. <i>Conclusion of the State of the Art in Data Mining.....</i>	42
1.4. THE PSIP PROJECT.....	43
1.4.1. <i>Project summary</i>	43
1.4.2. <i>Project objectives.....</i>	43
1.4.3. <i>Structure of the project</i>	44
1.5. OBJECTIVES OF THE PRESENT WORK.....	46
1.6. ARIANE’S THREAD	47
2. MATERIAL.....	49
2.1. OVERVIEW	49
2.2. DEFINITION OF A COMMON DATA MODEL	50
2.2.1. <i>Consideration about normalization</i>	50
2.2.2. <i>Available data</i>	51
2.2.3. <i>Terminologies</i>	51
2.2.4. <i>Description of the data model.....</i>	52
2.2.5. <i>Iterative quality control</i>	57
2.3. DATA EXTRACTION	58
2.4. EXTRACTION OF THE DRUG CODES FROM THE FREE-TEXT REPORTS	59
2.4.1. <i>Objectives.....</i>	59
2.4.2. <i>The F-MTI tool.....</i>	60
2.4.3. <i>Main use of Semantic Mining in the present work.....</i>	61
2.5. ADE DETECTION RULES EXTRACTED FROM THE SPCs.....	62
2.5.1. <i>General content.....</i>	62
2.5.2. <i>Format of the rules.....</i>	62
2.5.3. <i>Usability of the rules</i>	66
3. METHOD	67
3.1. INTRODUCTION.....	67
3.2. IDENTIFICATION OF THE OUTCOMES IN RELATION TO ADEs.....	68
3.2.1. <i>Principle.....</i>	68
3.2.2. <i>Description of Outcomes from the SPCs</i>	68

3.2.3. <i>Tracing of SPC Outcomes in this work</i>	70
3.3. EVALUATION OF THE CLINICAL IMPACT OF ADES	72
3.4. DATA MINING: A FIVE-STEP PROCEDURE	74
3.4.1. <i>Step 1: Transformation of the data into event (data aggregation)</i>	76
3.4.2. <i>Step 2: qualification of the events as “potential condition” or “potential outcome”</i>	82
3.4.3. <i>Step 3: statistical associations between potential conditions and outcomes</i> 83	
3.4.4. <i>Step 4: filtering of the associations</i>	84
3.4.5. <i>Step 5: validation of the rules</i>	84
3.5. CENTRAL RULE REPOSITORY	86
3.5.1. <i>Knowledge integrated in the Central rule repository</i>	86
3.5.2. <i>Rule description and storage in the central rule repository</i>	89
3.6. CONCLUSION.....	93
4. RESULTS	95
4.1. OVERVIEW OF THE CHAPTER	95
4.2. OVERVIEW OF DATA-MINING RESULTS	95
4.3. DECISION RULES INTEGRATED IN THE CENTRAL RULE REPOSITORY	102
4.3.1. <i>Validated rules</i>	102
4.3.2. <i>Detailed example of five rules</i>	102
4.3.3. <i>Classification / Overview of the rules</i>	108
4.4. EVALUATION OF THE ADE DETECTION: PRELIMINARY RESULTS	111
4.4.1. <i>Cases of Hyperkalemia ($K^+ > 5.3$ mmol/l)</i>	111
4.4.2. <i>Cases of VKA overdose (INR > 4.9)</i>	113
4.4.3. <i>All kinds of potential ADE</i>	116
4.5. PRESENTATION OF THE RESULTS: THE EXPERT EXPLORER AND THE SCORECARDS	117
4.5.1. <i>Description of the Expert Explorer</i>	118
4.5.2. <i>Description of the Scorecards</i>	118
4.5.3. <i>Use case example of the web tools for ADE discovery in databases</i>	118
5. DISCUSSION	131
5.1. CONTRIBUTION OF THE PRESENT WORK TO ADE DETECTION	131
5.2. DISCUSSION OF THE METHOD	139
5.3. PERSPECTIVES	142
5.3.1. <i>Reusability of the tools</i>	142
5.3.2. <i>Meta-rules for the implementation into a CDSS</i>	142
5.3.3. <i>Reusability of the rules described using XML files</i>	144
6. CONCLUSION	147
7. REFERENCES.....	149

8. ARTICLES	157
8.1. PUBMED REFERENCES	157
8.2. SCIENCE DIRECT REFERENCES.....	157
9. APPENDIX 1: TIME REQUIRED TO PERFORM THE DATA MINING TASK.....	159
9.1. OBJECTIVE OF THIS CHAPTER	159
9.2. DESCRIPTION OF THE TASKS AND NEEDED TIME	160
9.2.1. <i>Getting data from a hospital, first time.....</i>	<i>160</i>
9.2.2. <i>Getting data from a hospital, next times.....</i>	<i>160</i>
9.2.3. <i>Aggregating the data.....</i>	<i>160</i>
9.2.4. <i>Discovering new rules.....</i>	<i>160</i>
9.2.5. <i>Testing the rules of the central repository on an aggregated dataset</i>	<i>161</i>
9.2.6. <i>Publishing the datasets and the rules occurrences on the web tools.....</i>	<i>161</i>
9.3. “HOW MUCH TIME...”: USE CASE POINT OF VIEW	161
9.3.1. <i>Already known partner: computing statistics</i>	<i>161</i>
9.3.2. <i>New partner: computing statistics</i>	<i>162</i>
9.3.3. <i>Discovering some new rules</i>	<i>162</i>
10. APPENDIX 2: ODP DESCRIPTION OF THE EXPERT EXPLORER.....	163
10.1. USER REQUIREMENT AND ENTERPRISE VIEWPOINT	163
10.1.1. <i>Use case 1</i>	<i>163</i>
10.1.2. <i>Use case 2</i>	<i>165</i>
10.1.3. <i>Use case 3</i>	<i>165</i>
10.2. INFORMATION VIEWPOINT	167
10.3. COMPUTATIONAL VIEWPOINT.....	167
10.4. ENGINEERING VIEWPOINT	168
10.5. TECHNOLOGY VIEWPOINT	169
10.6. DETAILED DESCRIPTION OF THE INTERFACE AND USE FLOW	169
10.6.1. <i>Introduction.....</i>	<i>169</i>
10.6.2. <i>Implementation and availability</i>	<i>170</i>
10.6.3. <i>First Page.....</i>	<i>170</i>
10.6.4. <i>Data sets.....</i>	<i>171</i>
10.6.5. <i>Reports</i>	<i>173</i>
10.6.6. <i>Rules.....</i>	<i>175</i>
10.6.7. <i>Visualization of a stay.....</i>	<i>178</i>
10.6.8. <i>User accounts.....</i>	<i>181</i>
10.6.9. <i>Experts’ queries validation task.....</i>	<i>182</i>
10.6.10. <i>Administration.....</i>	<i>185</i>
11. APPENDIX 3: ODP DESCRIPTION OF THE SCORECARDS.....	186
11.1. USER REQUIREMENTS AND ENTERPRISE POINT OF VIEW.....	186

11.1.1. Use case	186
11.2. INFORMATION VIEW POINT	187
11.3. COMPUTATIONAL VIEWPOINT.....	188
11.4. ENGINEERING VIEWPOINT	188
11.5. TECHNOLOGY VIEWPOINT	189
11.6. DETAILED DESCRIPTION OF THE INTERFACE AND USE FLOW	189
11.6.1. Introduction.....	189
11.6.2. Interface design and development	189
11.6.3. First page of the Scorecards	190
11.6.4. “Synthesis and Edition of detailed statistics” Page	190
11.6.5. “Detailed statistics” Page	192
11.6.6. “Review cases” Page.....	195
11.6.7. “Review cases” questionnaire.....	196
12. APPENDIX 4: DESCRIPTION OF THE OUTPUT OF THIS WORK (USE OF THE XML FILES)	201
12.1. MAPPING XML FILES	201
12.1.1. Overview	201
12.1.2. Diagnosis mapping: <i>mapping_diag.xml</i>	201
12.1.3. Drugs mapping: <i>mapping_drug.xml</i>	202
12.1.4. Lab results mapping: <i>mapping_lab.xml</i>	203
12.2. RULES XML FILES	205
12.2.1. Lexicon: <i>lexique.xml</i>	205
12.2.2. Rules repository: <i>rules_yyyy-mm-dd.xml</i>	206
12.2.3. Pre-rules repository: <i>rules_root_yyyy-mm-dd.xml</i>	207
12.2.4. Rules contextualization: <i>rules_result_yyyy-mm-dd.xml</i>	208
12.2.5. Rules explanations: <i>rules_explanations_yyyy-mm-dd.xml</i>	212
12.3. HOW TO IMPLEMENT THE RULES FOR A PROSPECTIVE USE (TRANSACTIONAL USE OF THE CDSS)?.....	212
12.3.1. General considerations.....	212
12.3.2. “Cause” conditions of the rule	212
12.3.3. Outcome of the rule.....	213
12.3.4. How to manage several rules that predict the same outcome?	213
12.4. HOW TO IMPLEMENT THE RULES FOR A RETROSPECTIVE USE (RETROSPECTIVE USE OF THE CDSS, DASHBOARDS, CONFIDENCE COMPUTATION)?.....	214
12.4.1. General considerations.....	214
12.4.2. The “event” concept	214
12.4.3. From data to events	215
12.4.4. Using events to compute the confidence of a rule.....	217
13. APPENDIX 5: VALIDATION OF THE USE OF SEMANTIC MINING FOR ADE DETECTION	219
13.1. INTRODUCTION.....	219

13.1.1. Objective	219
13.1.2. Rationale in Semantic Mining evaluation.....	219
13.2. MATERIAL AND METHODS	220
13.2.1. Evaluation, Step 1: extraction of ATC codes from free-text documents: agreement between F-MTI and experts	220
13.2.2. Evaluation, Step 2- extraction of ATC & ICD10 codes from free text: agreements between F-MTI and EHR.....	221
13.2.3. Evaluation, Step 3- validation of the use of the semantic mining results for data-mining-based ADE detection	222
13.3. RESULTS	223
13.3.1. Evaluation Step 1- extraction of ATC codes from free-text documents: agreement between F-MTI and experts	223
13.3.2. Evaluation Step 2- extraction of ATC codes from free text: agreements between F-MTI and EHR.....	223
13.3.3. Evaluation Step 3- validation of the use of the Semantic Mining results for data-mining-based ADE detection	223
13.4. DISCUSSION.....	224
13.4.1. Ability of F-MTI to extract codes from free-text reports	224
13.4.2. Ability of F-MTI to provide ATC codes instead of a CPOE	225
13.4.3. Ability of F-MTI to be used for ADE detection.....	225
13.5. CONCLUSION	226
14. APPENDIX 6: VALIDATED RULES	227
14.1. ANEMIA (Hb<10G/DL)	227
14.2. HEPATIC CHOLESTASIS (ALKALIN PHOSPHATASE>240 UI/L OR BILIRUBINS>22 μMOL/L)	227
14.3. HEPATIC CYTOLYSIS (ALANINE TRANSAMINASE>110 UI/L OR ASPARTATE TRANSAMINASE>110 UI/L)	228
14.4. HIGH A CPK RATE (CPK>195 UI/L)	228
14.5. HEMORRHAGE HAZARD (INR>4.9)	229
14.6. LITHIUM OVERDOSE (TO HIGH A LITHIUM LEVEL).....	235
14.7. HEPARIN OVERDOSE (ACTIVATED PARTIAL THROMBOPLASTIN TIME>1.23)....	235
14.8. HYPEREOSINOPHILIA (ÉOSINOPHILIA>10 ⁹ /L).....	236
14.9. HYPERKALEMIA (K ⁺ >5.3)	236
14.10. HYPOCALCEMIA (CALCEMIA<2.2 MMOL/L)	243
14.11. HYPOKALEMIA (K ⁺ <3.0).....	243
14.12. HYPONATREMIA (NA ⁺ <130)	244
14.13. RENAL FAILURE (CREAT.>135 μMOL/L OR UREA>8.0 MMOL/L)	244
14.14. VKA UNDERDOSE (INR<1.6).....	245
14.15. NEUTROPENIA (PNN<1500/MM3).....	247
14.16. INCREASE OF PANCREATIC ENZYMES (AMYLASE>90 UI/L OR LIPASE>90 UI/L)	248
14.17. PANCYTOPENIA	248

14.18. THROMBOCYTOSIS (COUNT>600,000).....	249
14.19. THROMBOPENIA (COUNT<75,000)	249
14.20. DIARRHEA (PRESCRIPTION OF AN ANTI-DIARRHEAL)	252
14.21. DIARRHEA (PRESCRIPTION OF AN ANTIPROPULSIVE)	252
14.22. BACTERIAL INFECTION (DETECTED BY THE PRESCRIPTION OF ANTIBIOTIC) ..	253
14.23. PARACETAMOL OVERDOSE (DETECTED BY THE PRESCRIPTION OF ACETYL- CYSTEIN).....	253
14.24. FUNGAL INFECTION (DETECTED BY THE PRESCRIPTION OF LOCAL ANTIFUNGAL)	253
14.25. FUNGAL INFECTION (DETECTED BY THE PRESCRIPTION OF A SYSTEMIC ANTIFUNGAL).....	254
14.26. HEMORRHAGE (DETECTED BY THE PRESCRIPTION OF HEMOSTATIC).....	255
14.27. VKA OVERDOSE (DETECTED BY THE PRESCRIPTION OF VITAMIN K)	256
15. TABLE OF FIGURES.....	257
16. TABLE OF TABLES.....	261

ABBREVIATIONS AND ACRONYMS

ANR	Agence Nationale de la Recherche (French national research agency)
AFSSAPS	Agence Française de Sécurité Sanitaire des Produits de Santé (French drug administration)
ATC	Anatomical Therapeutic Chemical classification
BoW	Bag of words algorithm
CART	Classification And Regression Tree
CDSS	Clinical Decision Support System
CIS	Clinical Information System
CISMeF	Catalogue et Index des Sites Médicaux Francophones; Catalog and Index of French Medical Resources on the Internet
C-NPU	Nomenclature, Properties and Units
CPOE	Computerized Physician Order Entry
CRAN	Comprehensive R Archive Network
CxCDSS	Contextualized CDSS
DOW	Description Of Work (Annex 1 of PSIP Grant Agreement)
DRGs	Diagnoses Related Groups
EHR	Electronic Health Record
EMA	European Medicines Agency
F-MTI	French Multi-Terminology Indexer
FDA	Food and Drug Administration
HF	Human Factors
HAS	Haute Autorité en Santé (French health agency)
HIS	Hospital Information System
HMTS	Health Multi-Terminologies Server (English acronym); see also SMTS (French acronym) for the HMTS
HMTP	Health Multi-Terminologies Portal
HTML	HyperText Markup Language
ICD10	10 th revision of the International Classification of Disease and Related Health Problems
ICT	Information and Communication Technologies
ICU	Intensive Care Unit
INN	International Non proprietary Names
INR	International Normalized Ratio from prothrombin times
IUPAC	International Union of Pure and Applied Chemistry
KNN	K Nearest Neighbors
LOCF	Last Observation Carried Forward
LMWH	Low Molecular Weight Heparin
MDC	Major Disease Category
MeSH	Medical Subject Heading
NCCMERP	National Coordinating Council for Medication Error Reporting and Prevention
NLP	Natural Language Processing
ODP	Open Distributed Processing
PDF	Portable Document Format
PSIP	Patient Safety through Intelligent Procedure in medication
R-ODM	R Interface to Oracle Data Mining

RPART	Recursive Partitioning
SA	Société Anonyme
SMTS	Health Multi-Terminologies Server (French acronym)
SO	Sub-Objective
SPC	Summaries of Product Characteristics
TUV	French acronym, which stands for Vidal Unified Thesaurus
VKA	Vitamin K Antagonist
WHO	World Health Organization
WP	Work Package
XML	eXtensible Markup Language
XSL	eXtensible Stylesheet Language
XSLT	eXtensible Stylesheet Language Transformation
XSL-FO	eXtensible Stylesheet Language-Formatting Objects

1. INTRODUCTION

Adverse drug events (ADEs) are a public health issue. Each year, they are responsible for 98,000 deaths in the USA [Kohn 2000]. There is a need to improve the knowledge about ADEs. The objective of this work is to automatically detect cases of ADEs by mining electronic records of past hospitalizations, and to identify the causes that led to those ADEs.

The first part of this chapter aims at providing a definition of ADEs, which is not a trivial task (see section 1.1 on page 15). The second part of this chapter describes the different approaches that are commonly used for ADE detection: reporting systems, medical chart reviews, data mining and natural language processing (see section 1.2 on page 16)1.2. Then, as the objective of this work is to use data mining, a state of the art in data mining is performed (see section 1.3 on page 18): the requirements are enounced and 3 unsupervised and 7 supervised data mining techniques are reviewed and assessed. Finally, the PSIP Project is described (see section 1.4 on page 43). The PSIP Project aims at detecting and preventing ADEs, and the present work is performed as part of this project. Finally, the objectives of this work are more precisely detailed (section 1.5 on page 46).

1.1. Definition of Adverse Drug Events

The definition of Adverse Drug Events is not trivial but a common definition has been agreed on by researchers. Defining Adverse Drug Events (ADEs) first requires defining Adverse Drug Reactions (ADRs).

The World Health Organization (WHO) and the European Union share the same definition of an ADR [EC 2001]: *“A response to a medicinal product which is noxious and unintended and which occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease or for the restoration, correction or modification of physiological function”*. The Institute Of Medicine [IOM 2007, Kohn 1999, Handler 2006] define an ADE as *“an injury resulting from medical intervention related to a drug”* [Kohn 1999, Bates 1995]. This definition has been simplified to *“an injury resulting from the use of a drug”* [Gurwitz 2000]. According to Nebeker et al. [Nebeker 2004], in that definition ADEs include harm caused by the drug (ADRs and overdoses) and harm from the use of the drug, including dose reductions and discontinuations of drug therapy.

The Institute Of Medicine [IOM 2007] also gives another definition of ADEs that is interesting because it integrates the part that the diseases of the patient play in the outcomes: an ADE is *“an injury due to medication management rather than the underlying condition of the patient”*. A more complete definition can be retrieved from the *“Glossary of terms related to patient and medication safety”* elaborated by the Committee of Experts on Management of Safety and Quality in Health Care / Expert group on Safe Medication Practices, commissioned by the Council of Europe [SPSQS 2009]: *“An Adverse Drug Event is any injury occurring during the patient’s drug therapy and resulting either from appropriate care, or from unsuitable or suboptimal care. Adverse drug events include: the adverse drug reactions during normal use of the medicine, and any harm secondary to a medication error, both errors of omission or commission.”*

In accordance with those definitions, researchers also agree on dividing ADEs into two categories: preventable ADEs and non preventable ADEs. Preventable ADEs are assimilated to “medication errors” while non preventable ADEs are considered ADRs that could not be avoided [Murff 2003]. It is worth noting that medication errors do not necessarily harm the patients. Only a limited portion of medication errors turns into actual ADEs; all of them are preventable. Conversely, all preventable ADEs are considered medication errors.

Finally, the existence of an ADE implicitly requires the presence of clinical harm for the patient. The committee of Experts cited above [SPSQS 2009] adds that “*An adverse drug event can result in different outcomes, notably: in the worsening of an existing pathology, in the lack of any expected health status improvement, in the outbreak of a new or to be prevented pathology, in the change of an organic function, or in a noxious response due to the medicine taken.*”. The American Food and Drug Administration define an adverse event as “serious” when the patient encounters one of the following outcomes [FDA 2010]:

- Death
- Life-Threatening outcomes
- Hospitalization (initial or prolonged)
- Disability - significant, persistent, or permanent change, impairment, damage or disruption in the patient's body function/structure, physical activities or quality of life.
- Congenital Anomaly
- Requires Intervention to Prevent Permanent Impairment or Damage

From a pharmacological point of view, ADRs are well defined. Six types have been identified [Rawlins 1977, Aronson 2002]:

- Type A: Augmented pharmacologic effects - dose dependent and predictable: intolerance and side effects
- Type B: Bizarre effects (or idiosyncratic) - dose independent and unpredictable
- Type C: Chronic effects
- Type D: Delayed effects
- Type E: End-of-treatment effects
- Type F: Failure of therapy

In this work, we shall stick to those definitions of ADEs. Especially, we shall not look for abnormalities in the drug prescription that would not have any consequence on the patient. We shall first look for any traceable outcome (e.g. hyperkalemia), and try to limit those outcomes to those that could be explained by at least one drug administration or one drug discontinuation. This approach first requires that some outcomes are being identified in the data.

1.2. State of the Art in Adverse Drug Events detection

In former scientific works, different methods have been used to identify and characterize ADEs and medication errors. Several classifications of these systems are

available [Bates 2003, Morimoto 2004, Amalberti 2006]. To date, the most prominent systems are reporting systems and medical chart review.

Reporting systems

Reporting systems of medication errors or incidents are the most ancient methods, and they were imported in healthcare from other domains such as Transportation (aviation) or Industry. Reporting systems are usually documented by healthcare professionals spontaneously or after prompting, but some systems are designed to be documented by the patients themselves. Although ADE reporting is made mandatory by the law in certain cases, authors usually agree that all reporting systems suffer from important under-reporting biases [Morimoto 2004, Murff 2003]. Indeed if an ADE is frequent, predictable and not too severe, its declaration is deemed not to bring any new knowledge and might uselessly involve the practitioner's responsibility. On the other hand, those reports contain very exhaustive information, including narrative sections of the context and causative factors. They remain extremely useful for the analysis and characterization of contributing factors of ADEs.

Medical chart reviews

Retrospective medical chart reviews or Electronic Health Record (EHR) reviews constitute the main source of reliable epidemiological knowledge on ADE. At first these reviews were performed by trained experts, but despite its promising results, this method rapidly showed important limitations. Except when experts are intensively trained [Classen 2005, Kilbridge 2006, Morimoto 2004], the inter-experts agreement regarding the identification of ADEs is usually moderate to low ($40 < k < 60$) and even more so when experts are asked to validate the causative factors of the ADE ($k < .05$) [Amalberti 2006]. Moreover the method is extremely time and resource consuming. As a consequence, researchers have tried to take the opportunity of the increasing availability of EHRs to automate partly the reviewing process. Indeed it is possible to screen integrated data sources (Hospital Information Systems, Electronic Medical Records, lab results, administrative data, etc.). Except for systems targeting a circumscribed domain such as anesthesia [Benson 2000], to date no system has been able to reach complete automation.

Data Mining in the current approaches

Due to the exponential increase of the available computerized patient data, one would think that, as in banking industry, insurance companies or mass retail sector, data mining is more and more used to automatically screen large amount of medical records. Paradoxically, in the field of ADE detection, data mining is mainly used to analyze voluntary ADE reports [Almenoff 2005, Almenoff 2007, Bate 2006, Bennet 2007, Coulter 2001, Hauben 2005]. Those studies use data mining techniques in order to identify which drugs from those listed in the reports were more likely to be responsible for the outcome. It brings interesting knowledge for ADE report interpretation, but the statistical links that are discovered are valuable only knowing that an ADE has been declared and can hardly be extended to ADE detection or ADE prevention.

Natural Language Processing in the current approaches

Another way to detect ADEs is to mine free-text reports by means of Natural Language Processing (NLP) [Cantor 2007, Gysbers 2008, Melton 2005, Aramaki 2010]. Interesting results can be obtained; assuming that the physicians or the nurses have detected the ADEs and reported them in free-text observations or letters. In that use, NLP stands between ADE declaration screening and chart reviews as it enable to screen big amounts of patients records but requires that a sort of informal ADE declaration is present in the texts, for instance in the discharge letter.

Use of detection rules in CDSSs

Clinical Decision Support Systems use rule-based algorithms for ADE prevention during the medication process. Whatever their origin (chart reviews, reporting systems, summaries of product characteristics, etc.), the rules are always described as a set of Boolean conditions that could lead to an outcome. In CDSSs, the rules that have been implemented are considered to be always reliable and are applied in the same way to the medical departments all over the world. As a consequence in this classical approach the alerts are too numerous and of poor accuracy because they don't consider the variability in patients' characteristics, drug use, and monitoring procedures. The physicians often complain of over-alerting and their trust in the system decreases to such an extent that some of them deactivate the CDSS or systematically skip the alerts.

In this work, we shall not rely on any voluntary reporting of ADEs. This excludes reporting systems, data mining of voluntary ADE reports, and NLP of discharge summaries. We shall apply data mining techniques on routinely collected data in order to discover cases of ADEs. The ADE cases will hide in the data, without any flag and without any preliminary expert-operated review. Moreover, we shall try to bring innovative solutions based on statistics, in order to reduce over-alerting when the detection rules are implemented into a CDSS.

1.3. State of the Art in Data Mining

1.3.1. Introduction

A state of the art is realized before mining the data. The qualities and drawbacks of different data mining methods are evaluated with respect to the nature of the data and the results that are expected from those methods.

In this section, the requirements of the data mining are considered through several aspects: the available data, the two procedures we plan to apply, the expected results and the qualities of the rules we try to obtain from data mining. Those aspects help us to list the qualities we expect from the data mining methods.

Then, several data mining methods are reviewed and confronted with those expected qualities: non supervised methods and supervised methods. The description of those methods follows a uniform presentation.

1.3.2. Requirements of data mining

1.3.2.1. Available data

In this work, the aim of the data mining process is to identify adverse drug events (ADE) and causal conditions. We expect to have to explain about 50 different kinds of outcomes using about 500 potential conditions. All the outcomes will be described using binary variables (outcome present or absent). The potential conditions will be in the form of binary variables or quantitative variables (e.g. age of the patient).

Depending on the medical department, the number of rows of the dataset could vary from 500 to 5,000. In addition, the different datasets we use (one dataset per medical department) do not necessarily contain exactly the same fields. As the available data are not always the same from one hospital to another, as the laboratory parameters that are measured are not always the same, as the extraction process undergoes permanent improvements, the columns of the datasets may vary a little from one dataset to another

1.3.2.2. Two procedures for rule induction

Two procedures are going to be used for the ADE-detection-rule induction: procedure A & procedure B.

The aim in Procedure A is to identify ADE cases as groups of hospital stays that involve different patterns of outcomes, and then to try to explain their belonging to those groups using several conditions.

As shown on Figure 1 the approach in Procedure A will be:

- (step 1) to identify atypical groups using the outcome variables (“effect x”)
- (step 2) to explain those groups using the condition variables (“cause x”).

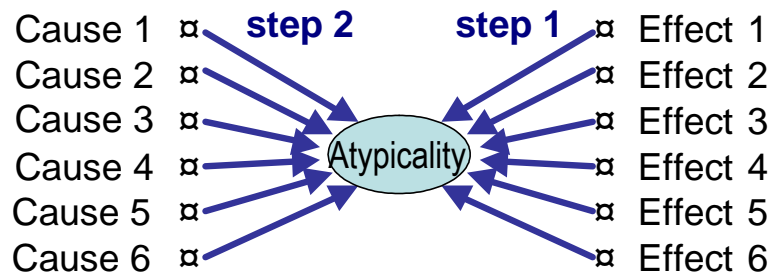


Figure 1. Decision rules induction in Procedure A

In Procedure B, results will be obtained using a simpler approach: outcome by outcome, we try to establish the link between the current outcome and all the available conditions (Figure 2).

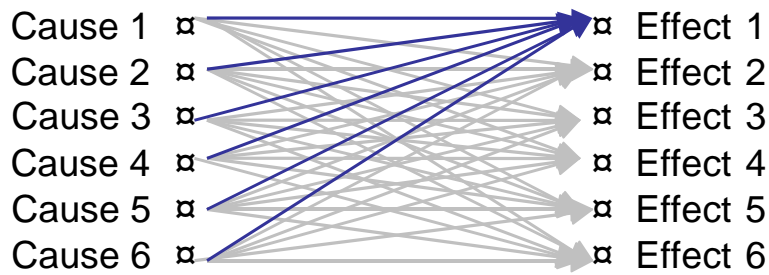


Figure 2. Approach for decision rules in Procedure B

The definition of the requirements is done with respect to the three different steps, as shown in Table 1.

Table 1. Definition of the requirements in respect with the procedure.

Procedure A		Procedure B
A step 1	A step 2	
To get some groups from outcome binary variables	To explain the groups using condition variables	To explain an outcome binary variable using condition variables

1.3.2.3. Expected qualities of the rules

1.3.2.3.1. Explanation of an outcome by a set of conditions

The objective of this work is to describe sets of conditions that could lead to an ADE. The term “effect” can be used:

- Either to describe the “1” value of a binary outcome variable (in procedure B)
- Or to describe a target value of a qualitative variable (a group, in procedure A)

Contrary to procedure B, in procedure A this explanation is split into two steps.

1.3.2.3.2. Simple formalization of cause-to-effect relationship

In procedures A & B, each effect is expected to be explained by a set of conditions. But the result presentation is preferred to look like a limited set of conditions leading to the realization of the effect:

$$C_1 \& \dots \& C_k \rightarrow E_1$$

with confidence = $P(E_1 | C_1 \cap \dots \cap C_k)$

This kind of result is strongly desirable for 3 main reasons:

- The rules have to be validated. The validation of a rule is only possible when a limited list of identified conditions lead to an identified effect. Even simple rules require a lot of time to get validated, the bibliographic validation of fuzzy conditions is not possible.
- Once validated, the rules have to be implemented into a CDSS. The CDSS module is not a statistical application, and methods that use too complex metrics are not usable. Even if the system had the ability to implement the metrics, this would suppose the availability of all the needed parameters, i.e. that all the datasets show exactly the same variables, and that this big amount of information is sent to the CDSS.
- Once validated and implemented, the rules must be explainable to the physicians who will use the CPOE. Medical reasoning can be summarized as decision rules where the combination of binary conditions linked by the AND operator leads to an effect.

1.3.2.3.3. Compatibility between the method and the kinds of variables

The methods have to be compatible with the available variables, with respect to the step of the procedure used (A1, A2 or B). Note that binary variables can in certain cases be considered either as quantitative variables or as qualitative variables.

1.3.2.3.4. Low number of predictors in the rule

A poor number of predictors are preferred in each rule because as mentioned above some variables might be missing from one dataset to another. Figure 3 displays the probability for a rule to be applicable in accordance with the number of missing variables among 500 variables, assuming that all the variables can be randomly used as predictors by the rules. Moreover, rules having more than 4 conditions are very difficult to explain according to the physicians in charge of rule validation.

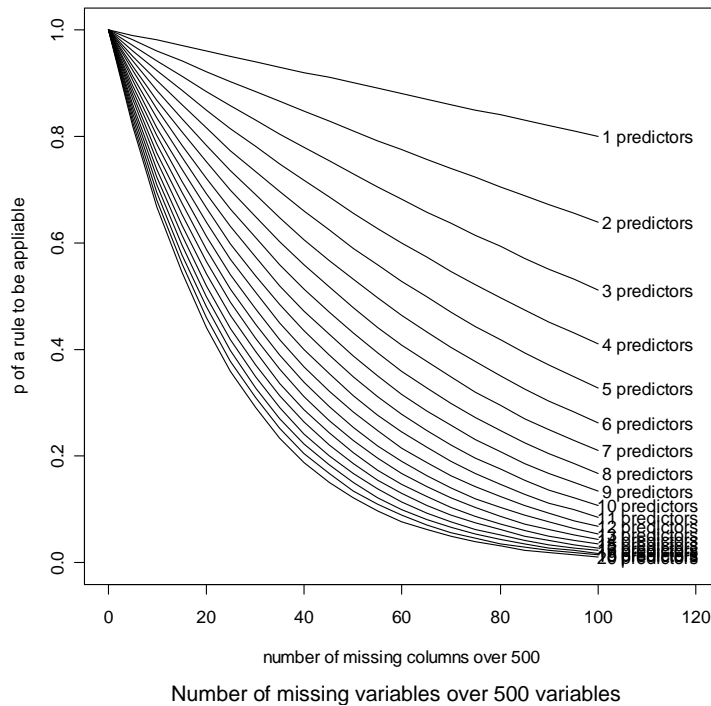


Figure 3. The probability of a rule to be applicable when some variables are missing strongly decreases when the number of predictors is high.

1.3.2.3.5. No assumption of risk additivity

We observed that, due to the high number of explanatory variables, additive models led to an over-adjustment risk. Most of the explanatory variables produce significant associations when used together in multivariate models although only a few variables are linked with the effect in univariate statistics. Moreover the additivity of the risks does not fit real situations in medicine: conditional probabilities seem to be more appropriate.

The following example shows the relative risks of larynx and hypopharynx cancers depending on alcohol and tobacco consumption: the hypothesis of risk additivity is clearly incongruous [Tuyns 1988]. As in many other examples, the risk additivity is not appropriate, and conditional probabilities are more reliable.

Alcohol increases the risk of cancer by 10.

$$P(C | A) = 10 * P(C)$$

Tobacco increases the risk of cancer by 5.

$$P(C | T) = 5 * P(C)$$

Tobacco and alcohol together increase the risk of cancer by 40.

$$P(C | A \cap T) = 40 * P(C)$$

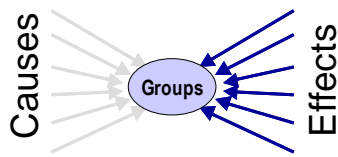
C=larynx or hypopharynx cancer

T=tobacco

A=alcohol

1.3.2.4. Expected qualities of the methods

1.3.2.4.1. In procedure A step 1 (N_2 effect variables \rightarrow K groups)

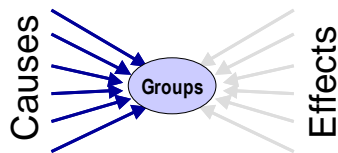


This step consists in discovering groups using all the outcome binary variables.

The following qualities are expected from the methods:

- Unsupervised method: discovery of groups from several outcomes
- Simple formalization of the conditions that enable to statute on the belonging to a group (pattern of outcomes)
- Compatibility between the method and the types of the variables:
 - o Several “outcome” binary variables as input
 - o One qualitative variable as output (normal, abnormal1, abnormal2...)
- Low number of predictors in the conditions of the rules

1.3.2.4.2. In procedure A step 2 (N_1 cause variables \rightarrow K groups)

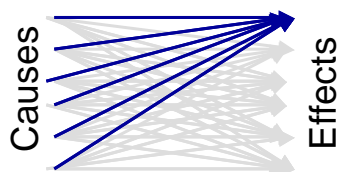


This step consists in predicting the groups from all the “condition” binary variables.

The following qualities are expected from the methods:

- Supervised method, explanation of a known group using a set of conditions
- Simple formalization of the relationship between conditions and group
- Compatibility between the method and the types of the variables:
 - o several binary or quantitative variables as conditions
 - o one qualitative variable as group (effect)
- Low number of predictors in the rule
- No hypothesis of risk additivity

1.3.2.4.3. In procedure B (N_1 condition variables \rightarrow 1 outcome variable)



This step consists in predicting a given outcome variable using all the “condition” binary variables.

The following qualities are expected from the methods:

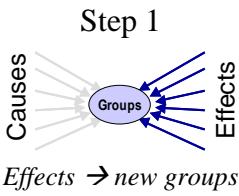
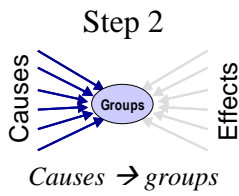
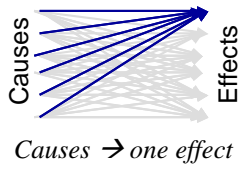
- Supervised method, explanation of an outcome by a set of conditions

- Simple formalization of the relationship between the conditions and the outcome
- Compatibility between the method and the types of the variables:
 - o several binary or quantitative variables as causes
 - o one binary variable as effect
- Low number of predictors in the rule
- No hypothesis of risk additivity

1.3.2.5. Descriptive summary of the data mining methods

At the end of the review of each data mining method, the following descriptive summary is checked. It makes it possible to quickly identify if the method is usable or not.

In each column, the first condition (first row) is mandatory: if the method doesn't fit this general requirement, there is no need to discuss its usability.

<i>Name of the method</i>		
Procedure A 	Procedure A 	Procedure B 
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Supervised: causes → one effect
<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Simple formalization of belonging conditions <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Low number of predictors	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> No risk additivity	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> No risk additivity
Conclusion: 😊😊😊	Conclusion: 😊😊😊	Conclusion: 😊😊😊

1.3.2.6. Methods to be reviewed

In this section, the following methods are reviewed:

- Non supervised methods:
 - o K means
 - o Agglomerative hierarchical clustering
 - o Latent class analysis
- Supervised methods:
 - o Logistic regression, multinomial logit model
 - o Cox proportional hazards model
 - o K-Nearest Neighbors algorithm
 - o Naïve Bayesian classification
 - o Neural networks

- Decision trees
- Association rules

1.3.3. Review of the methods

1.3.3.1. K Means

General principles

The k-means method is a clustering classification that enables to divide a data set into a certain number of partitions (*k homogeneous classes*) [MacQueen 1967].

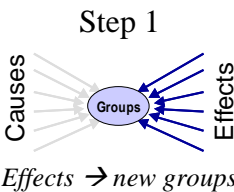
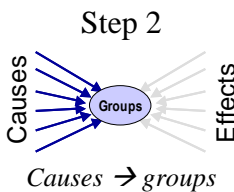
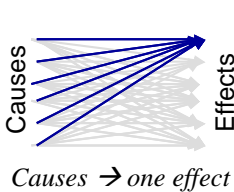
In each cluster the elements have to be as similar as possible, and clusters have to differ from the others as much as possible. Therefore, the number of clusters has to be defined (k) a priori. Then, k records are randomly chosen as “centroids”. Each record in the sample is checked by the algorithm [MacKay 2003] and is assigned to one of the k clusters, minimizing Euclidean distance to its centroid. Once this first loop is achieved, the n observations are now grouped into k clusters. For each cluster, the gravity center is computed and becomes the new centroid. Those steps are iterated. If one or more observations have moved to another cluster then centroids are computed again (for each cluster) and so on, otherwise the groups remain stable and the algorithm stops.

The algorithm of k-means aims at minimizing intra-class variance and maximizing interclass variance.

Assessment of the method

This method can be used when the means of the variables can be computed therefore only continuous variables are taken into account.

Conclusion

<i>K means</i>		
Procedure A		Procedure B
<p>Step 1</p>  <p>Effects → new groups</p>	<p>Step 2</p>  <p>Causes → groups</p>	 <p>Causes → one effect</p>
<input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> Supervised: causes → one effect
<input checked="" type="checkbox"/> Simple formalization of belonging conditions <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors		
Conclusion: ☹	Conclusion: ☹	Conclusion: ☹

1.3.3.2. Agglomerative hierarchical clustering

General principles

Agglomerative Hierarchical Clustering (AHC) aims at grouping n observations described by p explanative variables E_1, E_2, \dots, E_p in m groups as homogenous as possible according to a dissimilarity criterion. The hierarchical aspect can be explained by the fact that for a given precision level, two observations can be in the same group, while for a higher precision level, those observations would be assigned to two different groups. In other words, the method gradually aggregates observations according to their resemblance which is evaluated by a criterion.

How does it work?

The first step of the method consists of the choice of a dissimilarity index [Day 1984]. This index enables to evaluate the resemblance of two observations in relation to p explanative variables. In the case of continuous features, the Euclidean distance (Equation 1) can be used. Actually, the more important the distance is, the less similar the observations are. Several other distances can be quoted like squared Euclidian distance, Manhattan distance or Tchebychev distance. In the case of discrete explanative variables, chi-square distance (Equation 2) or percent disagreement distance can be used to evaluate similarity between observations.

$$d_{ii'}^2 = \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

Equation 1 Euclidian distance

$$d_{ii'}^2 = \sum_{j=1}^p \left(\frac{f_{ij}}{f_{.i} \sqrt{f_{.j}}} - \frac{f_{i'j}}{f_{.i'} \sqrt{f_{.j}}} \right)^2$$

Equation 2 Chi² distance

The second step consists of the choice of an aggregation index and therefore a classification algorithm (*i.e.* the algorithm is related to the distance computing). This index aims at evaluating distances between groups (or between an observation and a group). Several indexes can be quoted, such as:

- Single-linkage clustering: the distance between two groups is the smallest;
- Complete linkage clustering: the distance between two groups is the biggest;
- Distance between class' gravity centers;
- Ward's criterion, which is based on the intra-cluster variance minimization and the inter-cluster variance maximization;
- Sum of all intra-cluster variances: the higher this sum is, the less homogenous the clusters are.

First the method considers as many clusters as observations, therefore n clusters. Each observation is a proper cluster (the finest partitioning). Afterward, the algorithm forms the most similar couples of observations according to the aggregation index and iterates this logic at each step, decreasing the number of clusters until having one cluster, which contains all the observations.

The AHC method produces a classification binary tree, which is often named dendrogram. The root corresponds to the cluster that includes whole data and the leaves represent all single observations. The set of levels in the tree represents a hierarchy. Moreover, for each level the sample is divided into a certain number of clusters. In addition, the dendrogram shows, on the one hand, the order of aggregations and on the other hand the aggregation index value for each level. This value is important to choose a cut point therefore a certain number of clusters. Actually, it is generally relevant to select the cut point after the level of aggregations that shows low values of aggregation index and before the level that shows high values of index. Using this “gap” technique, we may expect to have a good quality partitioning therefore a reliable number of clusters.

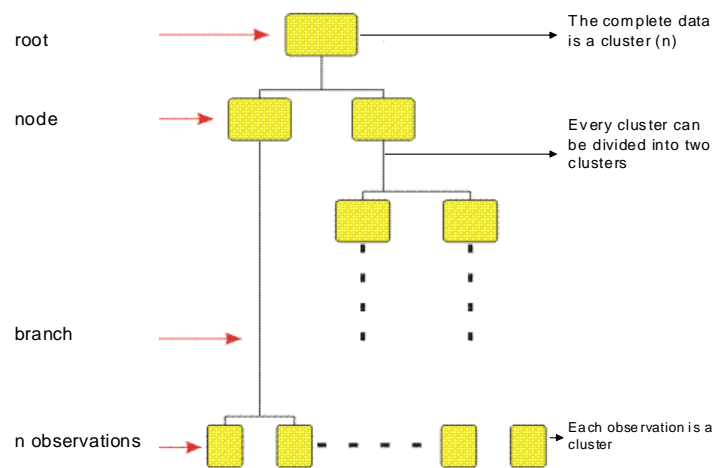


Figure 4. Partitioning with Agglomerative Hierarchical Clustering

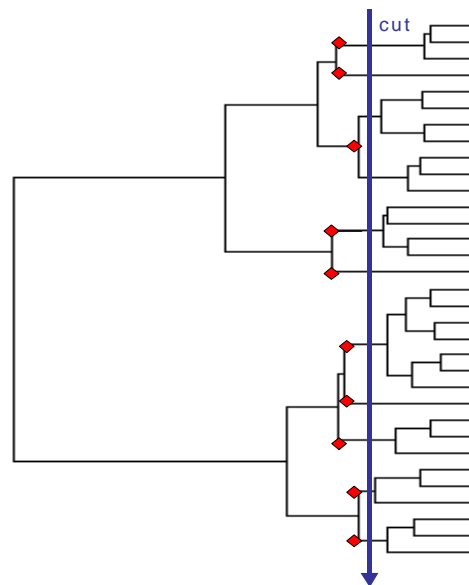


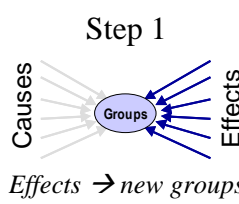
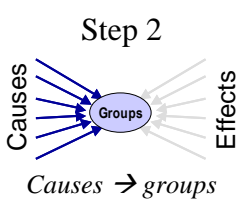
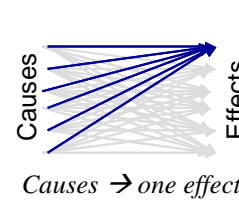
Figure. 5 Example of dendrogram

[Benzecri 1973, Didaye 1980, Cornuejol 2002].

Assessment of the method

The use of factorial analysis methods is advised before using the AHC method, in order to reduce noise in the data but these methods make the set of clusters more difficult to interpret and not necessary switchable in other medical departments.

Conclusion

Agglomerative hierarchical clustering		
Procedure A		Procedure B
<p>Step 1</p>  <p>Effects → new groups</p>	<p>Step 2</p>  <p>Causes → groups</p>	 <p>Causes → one effect</p>
<input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> Supervised: causes → one effect
<input checked="" type="checkbox"/> Simple formalization of belonging conditions <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors		
Conclusion: 😊	Conclusion: 😞	Conclusion: 😞

1.3.3.3. Logistic regression, multinomial logit model

General principles

Logistic regression is often used to predict the probability of the occurrence of an event by fitting data to a logistic curve [Amemiya 1985, Balakrishnan 1991]. It is the particular binomial use of a generalized linear model [McCullagh 1989]. It uses several numerical or categorical variables as predictors. It is widely used in the medical and social sciences and marketing applications. Being able to explain a binary effect justifies the popularity of the method:

$$P(E) = 1/(1+e^{-z}) \quad \text{where } z = a_0 + a_1*C_1 + a_2*C_2 + \dots + a_k*C_k$$

E is the variable to predict

C_i are the predictors

a₀ is the intercept

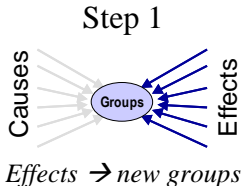
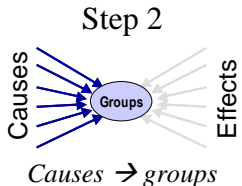
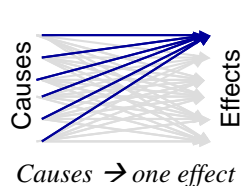
a_{i (i>0)} are the coefficients related to each predictor

Multinomial logit model is an extension of logistic regression that enables to use a categorical variable as dependant variable [Nakache 2003, Venables 2002]. We first used logistic regression as a simple and easy-to-perform way to identify multivariate associations.

Assessment of the method

This method is often used in the field of medicine. Actually, one of the project goals consists in obtaining rules that can be easily validated by physicians and implemented in a rule engine. However, the logistic regression does not provide this kind of rules. Actually, due to the affectation of a coefficient to each explanative variable, the model does not fulfill the previous rule requirements.

Conclusion

Logistic regression, multinomial logit model		
Procedure A		Procedure B
 <p>Step 1</p> <p>Effects → new groups</p>	 <p>Step 2</p> <p>Causes → groups</p>	 <p>Causes → one effect</p>
<input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> Supervised: causes → one effect
	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity
Conclusion: ☹	Conclusion: ☹	Conclusion: ☹

1.3.3.4. Cox proportional hazards model

General principles

The Cox proportional hazards model is a sub-class of survival models [Cox 1972]. The effects we trace can be considered as survival data: the effect might occur or not, when the effect occurs it occurs at a known date, when it doesn't occur (data are censored), the patient has been exposed to the predictors during a known duration. The popularity of the Cox model is due to its ability to take several predictors into account. But a strong assumption is that a given predictor multiplies hazard independently of the time.

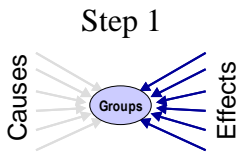
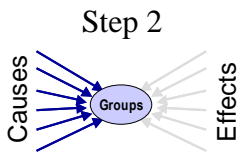
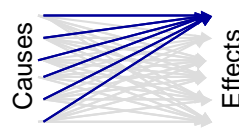
Assessment of the method

This method is appreciated because it is no hypothesis on the survival distribution. Furthermore, the main hypothesis of this method is the fact that the relative risks associated to predictors are constant over time. Then the additive relationship is still used to bring several predictors together. So the use of that model would lead to the same obstacles as generalized linear models:

- assertion of additivity of the effects

- over-adjustment due to a high number of predictors in respect with the number of records
- impossibility to validate and to output rules where predictors are linked together in a linear combination

Conclusion

<i>Cox proportional hazards model</i>		
Procedure A		Procedure B
 <p>Step 1</p> <p>Causes → Effects</p> <p>Effects → new groups</p>	 <p>Step 2</p> <p>Causes → Effects</p> <p>Causes → groups</p>	 <p>Causes → Effects</p> <p>Causes → one effect</p>
<input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> Supervised: causes → one effect
	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity
Conclusion: ☹	Conclusion: ☹	Conclusion: ☺

1.3.3.5. K-Nearest Neighbors algorithm (KNN)

General principles

KNN is a classification method based on distance computing. The main idea is that a new element is allocated to a class, which is the most common amongst the *k* nearest neighbors of the new element. *k* is the number of neighbors taken into account. It varies between 1 and *n*, where *n* is the number of observations.

The neighbors are identified by means of the computation of a distance. Actually the elements are represented by vectors in a multidimensional feature space, therefore Euclidean or Manhattan distances are usually used to identify the nearest neighbors of a new element.

This method makes a classification without venturing hypotheses on the function of classification $E=f(C_1, C_2, \dots, C_k)$ which links the dependent variable *E* to the predictors *C_i*. This method is non-parametric.

[Belur 1991, Shakhnarovich 2005, Garcia 2008]

How does it work?

For each new instance which has to be classified, the algorithm computes a distance between this instance and the other observations so as to find out the *k*-nearest observations which will be its *k*- nearest neighbors. The new instance is allocated to

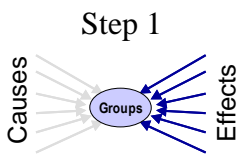
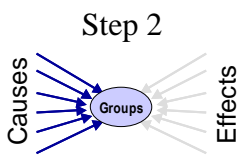
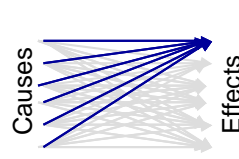
the class, which is the most common amongst the k -nearest neighbors. However, the Euclidean or Manhattan distances only work with continuous variables. In the case of discrete predictors another metric such as the overlap metric or Hamming distance can be used.

The k value is linked to the data. Generally, a high value of k enables a reduction of the noise in the classification but also involves less distinct boundaries between classes. An efficient value of k can be found thanks to various heuristics methods such as cross-validation.

Assessment of the method

The KNN method is well known for being one of the simplest methods in data mining, therefore the algorithm is easy to implement. However, this method is also well known for being time-consuming because of the computing of the k -nearest neighbors for each new instance to classify.

Conclusion

<i>K-Nearest Neighbors algorithm</i>		
Procedure A		Procedure B
 <p>Step 1</p> <p>Effects → new groups</p>	 <p>Step 2</p> <p>Causes → groups</p>	 <p>Causes → one effect</p>
<input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> Supervised: causes → one effect
	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity
Conclusion: ☹️	Conclusion: 😊	Conclusion: 😊

1.3.3.6. Naïve Bayesian classification

General principles

This type of classification is based on the Bayes' theorem with a naïve independence assumption for all the features.

In simple terms, this method supposes that the presence (or absence) of a specific cause of a class is unrelated to the presence (or absence) of any other cause. For instance, a fruit can be considered to be an orange if its color is orange, its shape is round and about 8cm in diameter. Although these features could be dependent on each other, the naïve Bayesian classification considers the opposite, that is to say that it

assumes that all the features contribute, in an independent way, to the probability that this fruit is an orange. [Domingos 1997, Hand 2001, Kotsiantis 2004, Minsky 1961, Morina 2004, Rish 2001].

How does it work?

This classification problem can be formulated using *a posteriori* probabilities:

$P(E/C_1 \cap C_2 \cap \dots \cap C_n)$ which corresponds to the probability that this instance (C_1, C_2, \dots, C_n) belongs to the E class. The main idea is to allocate a new instance to the class which maximizes the probability $P(E/C)$. Moreover, the method uses the Bayes' theorem which asserts that $P(E/C) = P(C/E) * P(E)/P(C)$, with $P(C)$ being considered constant for all classes and $P(E)$ the relative percentage of the E class.

Unfortunately, the probability $P(C/E)$ is not computable if the features are not independent.

The method assumes that all the features are independent and this hypothesis enables to easily compute the probability $P(C_1 \cap C_2 \cap \dots \cap C_n / E)$. Actually, under the independence hypothesis, $P(C_1 \cap C_2 \cap \dots \cap C_n / E) = P(C_1 / E) * \dots * P(C_k / E)$.

Thus, for all features C_i the probability $P(C_i / E)$ is computed. This probability is estimated thanks to the relative percentage of instances which have the value C_i , in the E class. Then, a new instance $C = (C_1, C_2, \dots, C_n)$ is allocated to the class which maximizes the probability $P(C_1 \cap C_2 \cap \dots \cap C_n / E)$ that is computed by the expression: $P(C_1 / E) * \dots * P(C_k / E)$.

Assessment of the method

This method is well known and often used because of its efficiency even with few data. Furthermore, it is based on probabilistic learning and provides good classification capacity.

However, the main hypothesis of this method is the independence of the predictors, which is not often verified in practice. Within the present work, the conditions that can lead to an adverse drug event are not uncorrelated. All the correlation coefficients between available predictor couples have been computed (about 10^5 different couples). Their distribution is displayed in Figure 6. A nullity test of the correlation coefficient has been computed too, the distribution of the p values is displayed in Figure 7 where the vertical line shows the 0.05 threshold. 23% of p values are under 0.05. In conclusion, 23% of the couples of variables are not independent.

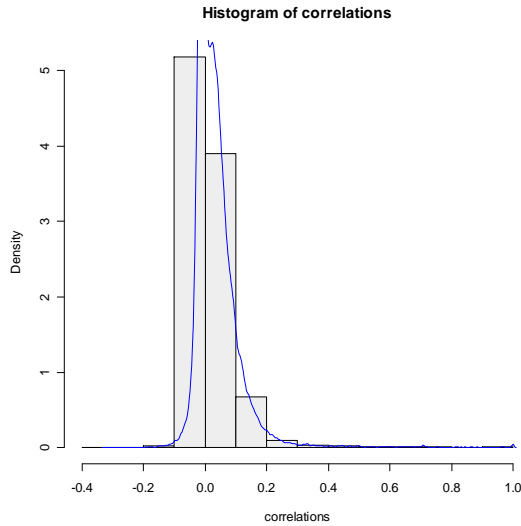


Figure 6. Distribution of the correlation coefficient between cause variables (about 10^5 values)

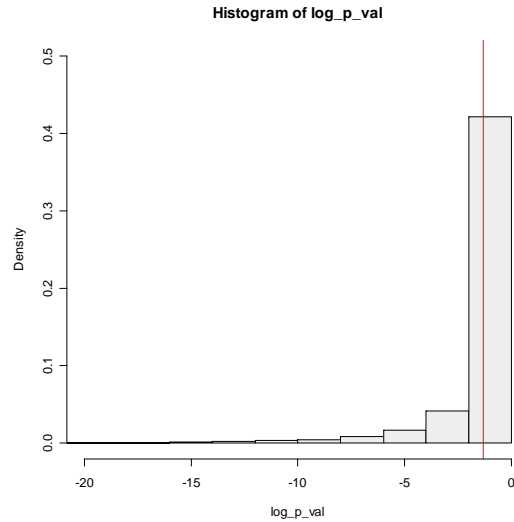


Figure 7. Distribution of $\text{Log}_{10}(p \text{ value of the correlation coefficient's nullity test})$ (about 10^5 values)

Considering the important number of predictors taken into account by the method so as to classify a new instance, it can be inferred that the naïve Bayesian classification is inappropriate in this work because we expect fairly short rules.

Conclusion

<i>Naïve Bayesian classification</i>		
Procedure A		Procedure B
<p>Step 1</p> <p><i>Effects → new groups</i></p>	<p>Step 2</p> <p><i>Causes → groups</i></p>	<p><i>Causes → one effect</i></p>
<input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> Supervised: causes → one effect
	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity
Conclusion: ☹️	Conclusion: 😊	Conclusion: 😊

1.3.3.7. Neural networks

General principles

The artificial neural networks, which are usually named neural networks, are mathematical models that are inspired by the biologic neural networks. These models are applicable when a relationship between the predictors (inputs) and the predicted variable(s) (output) can be demonstrated even if this relationship is complex and cannot be described by simple correlations or difference between groups.

An artificial neural network is generally composed of a series of layers, each one of them using in input the output of the previous one. Each layer i is composed of N_i neurons that use in input the output of the N_{i-1} neurons of the previous layer. A synaptic weight is associated with each synapse in such a way that the N_{i-1} neurons are multiplied by this weight and then added to the neurons of level i , which amounts to multiplying the input vector by a transformation matrix (combination function). The value given by the combination function is evaluated by the activation function using a threshold, and an activation value is computed. The activation function is, by definition, non-linear, therefore this method has interesting qualities of modeling.

In the present work, this method can be used so as to predict an ADE for a given stay. As it can be seen on Figure 8, the artificial neuron network model uses in input a set of causes and provides the presence (or absence) of an effect in output.

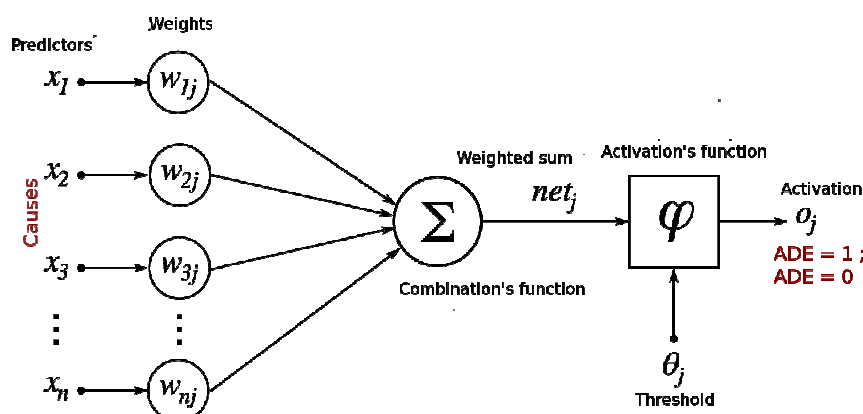


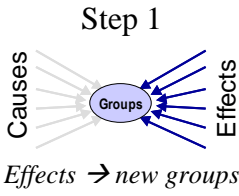
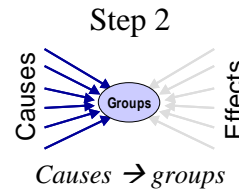
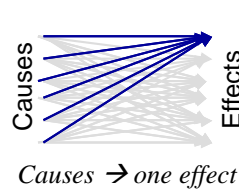
Figure 8. Scheme of an artificial neuron

Assessment of the method

This method is widely used in several fields because of its efficiency in modeling complex data and its ability to provide good quality of classification. Despite these obvious qualities, the neural networks are often considered as a “black-box” and the classification rules resulting from the model are difficult to interpret. Actually, the weight computing and the use of combination function involve the fact that the rules are abstract and meaningless, hence the “black-box” designation.

In this work, this element argues against the use of this method because the alert rules have to be easily validated and interpretable for the physicians. Furthermore, the alert rules have to be implemented in a simple fashion in engine rules.

Conclusion

<i>Neural networks</i>		
Procedure A		Procedure B
 <p>Step 1</p> <p>Effects → new groups</p>	 <p>Step 2</p> <p>Causes → groups</p>	 <p>Causes → one effect</p>
<input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> Supervised: causes → one effect
	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity
Conclusion: ☹	Conclusion: 😊	Conclusion: 😊

1.3.3.8. Decision trees

General principles

Decision trees consist of the construction of a tree by means of successive splits of the observations from a sample into two or more homogenous segments (called nodes) in relation to a dependent variable (binary, ordinal, discrete or continuous) and using the information of p explanative variables (binary, ordinal, discrete or continuous).

The inducted tree is represented by means of a reversed tree which has, as a root, the global sample to split. Furthermore the other nodes are either intermediate or final. The set of final nodes represents a partitioning of the sample into homogenous and distinct classifications in relation to the dependent variable. Branches symbolize conjunctions of predictors that lead to those classifications.

Decision tree models can be divided into two principal types: classification trees (discrete outcome) and regression trees (continuous outcome).

In addition, the different tree induction methods produce segmentation rules that describe each node (intermediate or final).

Decision trees (classification trees) are interesting in order to get some rules such as:

$$\text{Condition1} \ \& \ \text{Condition2} \ \& \ \dots \ \& \ \text{Conditionk} \ \rightarrow \ \text{Outcome}$$

$$C_1 \cap C_2 \cap \dots \cap C_k \rightarrow O$$

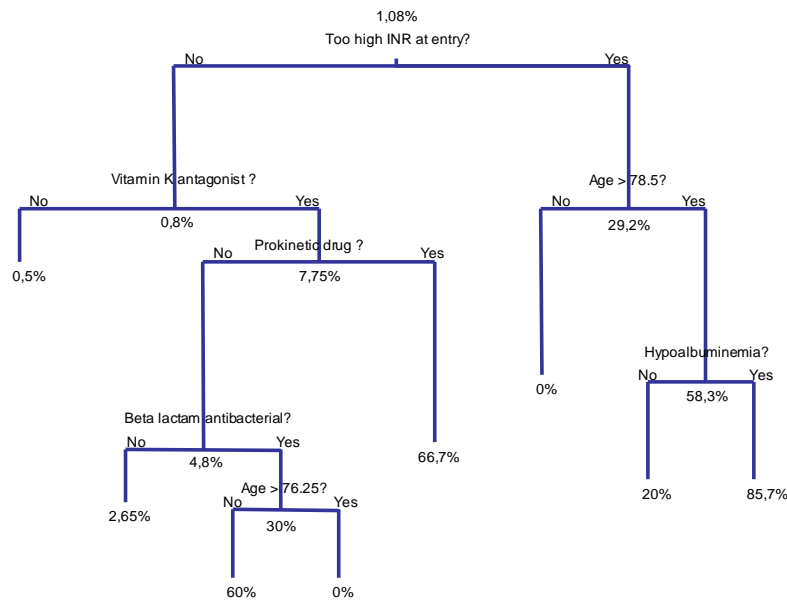


Figure 9. Example of decision tree

How does it work?

Several well-known methods produce decision trees, such as the Chi-squared Automatic Interaction Detector (CHAID) or Classification and Regression Trees (CART). Both methods are described below.

In both methods, the data are commonly divided into two sub-samples called growing set and validation set, respectively. The first method, which corresponds approximately to 70% of the data, is used for tree construction. The second method is used to carry out a validation of the model by means of the rest of the data, which can be considered as “unknown” data. This division aims at evaluating the ability of the model to provide reliable predictions on new data. In other words, the objective of this technique is to avoid overfitting.

Chi-squared Automatic Interaction Detector (CHAID)

CHAID is a non-parametric method, which produces either classification or regression trees depending on whether the dependent variable is discrete or continuous. In addition, this method produces N-ary trees. It is based on the independence chi-square test, therefore it needs an alpha risk that has to be chosen by the analyst. It is significant to add that produced trees are composed of rule set based on certain values of variables.

The algorithm follows several ordered points:

- Predictors are selected by means of the chi-square test. Actually, the predictor that maximizes the chi-square test statistic (this statistic has to be significant with respect to the alpha risk) is selected. That predictor provides the best split that differentiates observations based on the dependent variable.
- Once a predictor has been selected and has split a node into two or more “children” nodes, the same algorithm is applied to each child in a recursive way.
- The algorithm stops when the chi-square test is not significant for each predictor or in all terminal nodes, the number of observations is too low to

split these nodes again. These two elements define the pruning of the tree: as this pruning is performed during the tree expansion, a priori, it is called “pre pruning”.

Classification and Regression Tree (CART) method

This method is quite similar to CHAID. However, several differences can be highlighted. Actually, the split factor is based on the Gini coefficient instead of the chi-square test statistic. This coefficient measures the statistical dispersion of distributions. The higher the coefficient is, the better the split ability (notion of impurity reduction) of the considered predictor is. Moreover, this method only produces binary trees.

In addition, the method does not carry out the pruning of the tree during its growing stage (no pre-pruning). Actually, during the first stage, the biggest tree is produced even if certain branches are not really significant. The second stage consists of the pruning of the tree according to a main idea: finding the tree that optimizes its cost-complexity with the aim of minimizing the prediction error of the model. In order to carry out this pruning, a *pruning set* is randomly selected before the first stage. Thus, in CART method, the whole sample is divided into three parts: *growing set*, *pruning set* and *validation set*.

The third stage, similar to CHAID, consists of the validation of the model by means of the *validation set*. [Breiman 1984]

Comparison of the CHAID and CART tree induction methods

Table 1. Comparison between CHAID and CART tree induction methods

Features \ Method	CHAID	CART
Split factor	Independence chi-square test	Gini coefficient
Pooling	N-ary trees	Binary trees
Tree size determination	Pruning during the growing stage (pre pruning)	Pruning <i>a posteriori</i> by means of a pruning set (post pruning)
Advantages	Efficiency in an explanatory stage	Classification's efficiency Simplicity of use.
Drawbacks	Weak performances in classification Complex choice of alpha risk.	Weak efficiency with low number of observations.

Assessment of the method

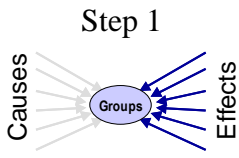
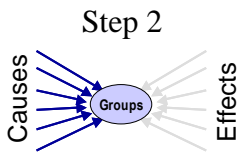
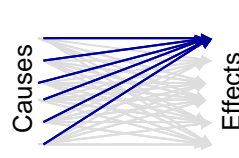
Decision trees are well-known and often used because they produce clear and understandable results even for neophyte users. Furthermore, they are commonly known as a “white box” model. Indeed, if the model provides a result, this result can be easily understood, validated and implemented (for replications) in a rule engine. However, decision trees are also well-known for their prediction unreliability when the size of the data set is too low, therefore the validation set use is absolutely necessary.

In this work, judging by these previous arguments, the use of decision trees is appropriate on one hand for the data nature and on the other hand for the formalism of

the output. Actually, decision tree methods provide sets of rules which, as said before, can be easily validated and implemented in a rule engine. Moreover, the CART method has been chosen because of its several qualities, compared to CHAID: the efficiency of the classification, the density of produced trees that involves the reliable prediction ability of the rules. Furthermore, in the case of the CHAID method, the determination of a good alpha risk is tricky. Indeed, if the risk is too high, the tree will be over-adjusted, otherwise it will be under-adjusted. This problem does not occur using the CART method because of its pruning stage that determines the optimal size of tree, therefore provides rules with the best prediction capacity in relation to data.

[Nakache 2005, Lebart 2000, Nakache 2003, Quilan 1993, Rakotomalala 1997, Rakotomalala 2005]

Conclusion

<i>Decision trees</i>		
Procedure A		Procedure B
 Step 1 <i>Effects → new groups</i>	 Step 2 <i>Causes → groups</i>	 <i>Causes → one effect</i>
<input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> Supervised: causes → one effect
	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity
Conclusion: ☹	Conclusion: ☺	Conclusion: ☺

1.3.3.9. Association rules

General principles

The association rule learning is a well-known and well-researched method, which aims at discovering relationships between variables in a large database using different measures of significance. This method originates in the researches of Piatetsky-Shapiro [Piatetsky-Shapiro 1991] who described strong rules discovered in database and the researches of Agrawal [Agrawal 1993] who introduced the association rules concept so as to discover relationships between sets of products in supermarket transaction data.

For instance the association rule learning method can provide rules such as:

$$\text{Condition1} \ \& \ \text{Condition2} \ \rightarrow \ \text{Outcome}$$

$$C_1 \cap C_2 \rightarrow O$$

How does it work?

First of all, some definitions have to be given. An itemset is an element set of a database such as (Condition1 ; Condition2).

Two principal indicators are important in the association rule mining: support and confidence. The support of an association rule corresponds to the probability of appearance of two itemsets at the same time $P(C_1 \cap C_2 \cap O)$. The confidence of an association rule is the probability of the outcome appearance knowing the set of causes $P(O/C_1 \cap C_2)$.

The main idea of the method is to find out association rules that meet user-defined thresholds: minimum support and confidence. First, the minimum support is used so as to find the “frequent itemsets” in the database. In a second time, the minimum confidence and the “frequent itemsets” are used to generate association rules. The first step is complex because it involves looking for all possible itemsets in the database (item combinations). Although the complexity of the method grows exponentially with the number of items, a downward-closure property of support can be used in order to provide efficiency for searching. This property is defined by the fact that if an itemset is frequent, then all its subsets are frequent, therefore if an itemset is not frequent, then all its subsets are not frequent. This property leads to efficient algorithms such as Apriori [Agrawal 1994].

Several additional indicators are to be used in order to decrease the number of rules on one hand, and to select the most statistically significant ones on the other hand. For instance, indicators such as lift, conviction [Brin 1997] are used. Moreover, we also use the hyperlift and hyperconfidence indicators that enable to test the independence between the antecedents and the consequences of a rule. The hyperconfidence indicator is linked to a unilateral exact Fisher’s test [Hasler 2007] therefore a probabilistic aspect has been introduced in the rule selection. In addition these indicators have the advantage of being robust in the case of infrequent items like the Adverse Drug Events.

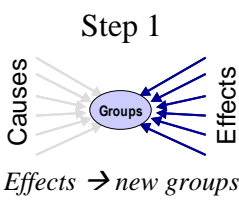
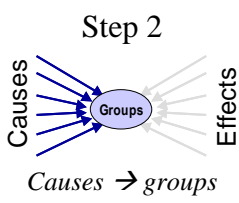
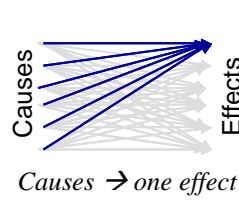
Assessment of the method

This method is well-known for its simplicity and its capacity to produce a large set of rules. Moreover, no data hypotheses are needed and the rules produced are easily interpretable.

However, this method is also well-known for its important time computing cost. Indeed, the search of the “frequent itemsets” is very time-consuming. Furthermore, this method has difficulties in working with an infrequent item because of the conception of the algorithm, therefore some rules can’t be significant.

In the present work, this last drawback has been avoided by means of the use of indicators such as hyperlift and hyperconfidence. Moreover, this method produces an exhaustive set of rules that can be easily validated by experts and integrated in a rule engine.

Conclusion

<i>Association rules</i>		
Procedure A		Procedure B
 <p>Step 1</p> <p>Effects → new groups</p>	 <p>Step 2</p> <p>Causes → groups</p>	 <p>Causes → one effect</p>
<input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> Supervised: causes → one effect
	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity	<input checked="" type="checkbox"/> Simple formalization <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Low number of predictors <input checked="" type="checkbox"/> No risk additivity
Conclusion: ☹	Conclusion: 😊	Conclusion: 😊

1.3.3.10. Latent class analysis

General principles

The latent variable models state that some unobservable variables exist, for which the effects can be observed through the covariates (observed variables). The principal components in multivariate factorial analysis are the best example of latent variables. The fundamental hypothesis is that the dependences between covariates could be explained by the dependence between latent variables and each one of the covariates. That involves the well-known principle of conditional independence: the covariates are independent conditionally to the latent variables.

A general introduction to latent variable models is given by [Lazarsfeld 1968, Everitt 1984, Bartholomew 1999]. The particular case of categorical latent variables yields naturally to the question of clustering addressed by [McCutcheon 1987, Hagenaars 2002]. If the classical techniques of clustering are in general developed for scalar covariates, the latent model based approach is particularly well adapted for categorical covariates, in particular for binary ones, as is the case in the present work. The theoretical developments for model estimation are based on the approach of likelihood maximization using the EM algorithm.

How does it work?

Let X_1, \dots, X_p be the covariates with values in $\{0, 1\}$ observed on n statistical units and Y be the latent variable with k levels (classes). We denote by p_{ij} the conditional probability that $X_i = 1$ for statistical unit belonging to the class j , i.e. $P(X_i = 1/Y = j)$, for

all $i=1\dots p$ and $j=1\dots k$. We also denote by π_j the a priori probability to belong to the class j .

If $f(x)$ is the density of the (X_1, \dots, X_p) then the principle of conditional independence yields to:

$$f(x) = \sum_{j=1}^K \pi_j \prod_{i=1}^p p_{ij}^{x_i} (1-p_{ij})^{1-x_i} .$$

The classification rule is based on the posterior conditional probability that a statistical unit with $x=(x_1, \dots, x_p)$ belongs to the class j :

$$h(j/x) = \pi_j \prod_{i=1}^p p_{ij}^{x_i} (1-p_{ij})^{1-x_i} .$$

The estimation of parameters p_{ij} and π_j is performed by the EM algorithm that maximizes the log-likelihood.

$$L = \sum_{h=1}^n \sum_{j=1}^K \pi_j \prod_{i=1}^p p_{ij}^{x_{hi}} (1-p_{ij})^{1-x_{hi}}$$

Thus, the classification rule giving the structure of the K clusters is given by the posterior probabilities

$$\hat{h}(j/x) = \pi_j \prod_{i=1}^p \hat{p}_{ij}^{x_i} (1-\hat{p}_{ij})^{1-x_i} ,$$

an observation (x) being assigned to the largest associated probability cluster.

Assessment of the method

This method is powerful and provides clusters characterized by patterns in data matching observations. It provides simple interpretation in terms of the characterization of classes as well as for the relationships between covariates. One of the main arguments in using it is that it represents one of the best alternatives to the classification on categorical variables after factorial analysis (multiple correspondence analysis) especially when some modalities are less frequent in the dataset. As with the K-means method, the number of clusters is fixed a priori and should be tuned.

Conclusion

<i>Latent class analysis</i>		
Procedure A <div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>Step 1</p> </div> <div style="text-align: center;"> <p>Step 2</p> </div> </div>		Procedure B
<input checked="" type="checkbox"/> Unsupervised, discovery of groups from several effects	<input checked="" type="checkbox"/> Supervised: causes → groups	<input checked="" type="checkbox"/> Supervised: causes → one effect
<input checked="" type="checkbox"/> Simple formalization of belonging conditions <input checked="" type="checkbox"/> Type compatibility <input checked="" type="checkbox"/> Interpretability		
Conclusion: 😊	Conclusion: 😞	Conclusion: 😞

1.3.4. Conclusion of the State of the Art in Data Mining

In this section, we formulated the characteristics we were expecting from the results of the data mining process (decision rules). After defining the different data mining steps (A₁, A₂ and B) we were able to formulate the required characteristics from the statistical methods to be chosen with respect to the data mining steps. Consequently after the review of several methods, only a few methods meet all our criteria:

- For step A₁: hierarchical clustering after factorial analysis and latent class analysis
- For steps A₂ and B: decision trees and association rules

Note: Unfortunately, after several months of research, it appeared that Procedure A led to results that were not interpretable according to the experts. As a consequence, the following “methods” section only uses the Procedure B for rule induction, i.e. decision trees and association rules.

1.4. The PSIP Project

1.4.1. Project summary

Adverse Drug Events (ADE) caused by product safety problems and medication errors caused by Human Factors are a major Public Health issue. They endanger the patients' safety and generate considerable extra hospital costs.

Healthcare Information and Communication Technologies (ICT) applications may help to reduce the incidence of preventable ADEs, by providing the healthcare professionals and patients with real-time relevant alerts, information and knowledge (guidelines, recommendations, etc.). But their efficiency is impeded by the lack of reliable knowledge about ADEs and the poor ability of ICT solutions to deliver contextualized knowledge focused on the problem at hand, aggravated by a poor consideration of causative human factors.

The project Patient Safety through Intelligent Procedures in medication (PSIP) [PSIP 2010] is a European project funded by the European Research Council [ERC 2010, FP7 2010]. It follows two objectives:

- 1- to facilitate the development of knowledge on ADE, and
- 2- to improve the medication cycle in hospital environments.

The first objective is to generate knowledge on ADE: to know, as accurately as possible, per investigated departments in hospital partners, their number, type, consequences and causes. Data mining of the structured hospital data bases, and semantic mining of the free-texts (letters and reports), enable to identify ADEs and to give them a context, with frequencies and probabilities, thus giving a better understanding of potential risks.

The second objective is to develop a set of innovative knowledge based on the mining results and to deliver a contextualized knowledge fitting the local risk parameters, in the form of alerts and decision support functions. The consortium will focus on the context handling to provide contextualized PSIP alerts to the medical staff. The design and development cycle of the PSIP solution will be HF oriented.

1.4.2. Project objectives

The overall PSIP project **general objectives** are to develop services (procedures, decision systems, prototypes) that:

- Identify, by State-of-the-Art Data and Semantic Mining techniques, Healthcare situations where patient safety is at risk
- Improve the decision support tool related to the medication cycle
- Deliver usable, efficient and contextualized alerts and just-in-time relevant information to healthcare professionals and patients
- Demonstrate a significant reduction of patient risk for a subset of diseases and practices in the hospital setting
- Implement standardized knowledge based tools.

More in detail, **scientific objectives** are:

- To get a better knowledge of the prevalence of ADEs and of their characteristics, per hospital, per region, per country

- To develop concepts and methods to achieve the contextualization of CDSS functions
- To model the architecture enabling both the independency and the interrelation between the knowledge (CDSS modules) and the connecting applications (HIS, CPOE).

Technical objectives are:

To run semi-automatically data and semantic mining techniques on existing healthcare data repositories

- To develop a platform incorporating the CDSS modules and easy to connect to healthcare IT applications primarily CPOE
- To design and develop a support system for healthcare professionals prototype integrated in the medication workflow
- To design and develop a support system prototype to help patients monitor their medication process.

The project PSIP addresses these objectives as follows:

1. The PSIP project will apply innovative concepts and methods to improve the existing knowledge on adverse events by mining the data repositories of Hospital Information Systems (HIS) or CPOES. This will enable to tackle the actual ADEs occurring in a given healthcare environment and to provide the healthcare professionals with a context-dependent information.
2. The PSIP project will develop contextualized-CDSS modules (CxCDSS) for all four main actors in the medication cycle: the physician in charge of the decision making and the ordering stage, the pharmacist in charge of the validation and the dispensing phase, the nurse in charge of the verification and administration phase, and the patient who finally receives and ingests (or not!) the drug.
3. The project intends to provide a proof of concept of an independent CDSS platform to support and secure the medication cycle equally accessible to various healthcare IT applications such as CPOE, e-prescribing, HIS, Clinical Information System (CIS), Electronic Health Record (EHR), etc. whether they are commercially available or locally (“home-grown”) developed.
4. The PSIP project will adopt a Human Factors Engineering approach to the design of the CDSS system and of its Human Computer Interface. This particular approach will encourage healthcare professionals to use the system and minimize the emergence of unexpected technology-induced negative effects.

1.4.3. Structure of the project

The PSIP project is divided into 3 stages:

- Stage 1 is dedicated to “improving knowledge on adverse events”
- Stage 2 aims at developing “clinical decision support systems”
- Stage 3, for “integration in existing IT solutions and usage”.

Each Stage is divided into Workpackages (WPs) described on Figure 10: WP1-3 for Stage 1, WP4-6 for Stage 2 and WP7-10 for Stage 3. Two WP (WP11 and WP12) handle the human factors engineering and evaluation all along the duration of the project. In addition, a supplementary WPO is dedicated to the overall management and

the quality management of the project, and a last WP13 will handle the dissemination and exploitation tasks.

The activity within the project starts with working on existing material resources brought in by all the partners. These material resources are data repositories of healthcare records and of Adverse Events, existing ICT solutions running in the participating hospitals and/or developed by the industrial partners, existing data and semantic mining techniques mastered by the academic teams, existing pharmaceutical knowledge on work processes and healthcare organizations brought in by Human Factors teams.

The present work is mainly performed as a part of the first stage of the PSIP Project: “improve knowledge on adverse events”. It aims at gathering routinely collected data from several hospitals’ EHRs into a common repository, analyzing those data in order to detect ADE cases, and producing knowledge about ADEs, mainly in the form of ADE detection rules.

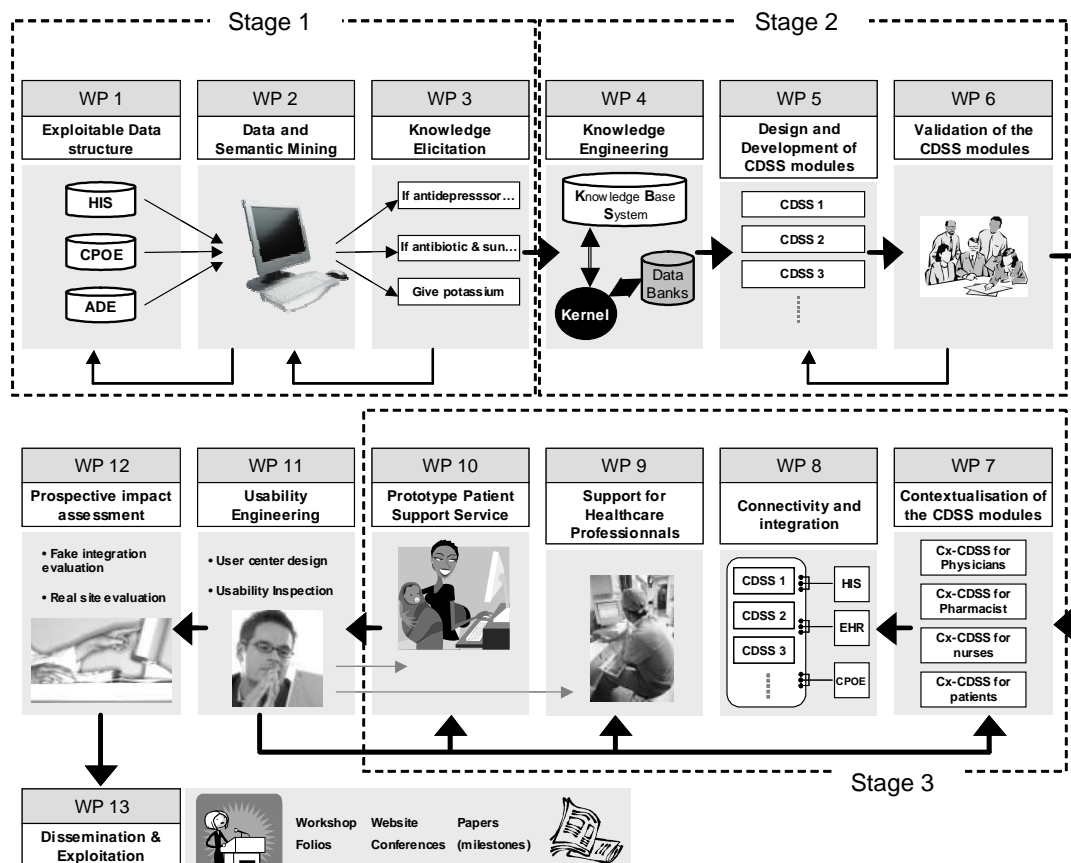


Figure 10. Description of the architecture of the PSIP project

1.5. Objectives of the present work

General objective

The general objective of the present work is to identify, in databases and data collections, expected or unexpected Adverse Drug Events, and their link with the demographic, medical and therapeutic data of patients. This objective has to be reached in the absence of any ADE declaration, and by mining routinely-collected data. Data Mining [Adriaans 1996] of the large medical databases of HIS or CPOEs, and Semantic Mining of the collections of free-texts documents will be used for that purpose.

That general objective consists of 2 sub-objectives:

- 1- to automatically detect cases of Adverse Drug Events and compute statistics to provide epidemiological knowledge about those cases
- 2- to identify decision rules that are able to be used for 2 different purposes:
 - a. for ADE detection: the rules have to provide the ability to be used to mine thousands of past hospitalizations in order to identify potential ADE cases
 - b. for ADE prevention: the rules have to provide the ability to be used by a Clinical Decision Support System to prevent ADEs by identifying risky situations during the prescription process and by alerting the prescriber

Operational objectives

To explore large databases we will use statistical methods and Data Mining techniques. The Data Mining techniques will be preferred to identify adverse events in the Hospital databases as they tend to be more robust when applied to “messy” real world data.

A rule-induction process will try to link some outcomes with the characteristics of the patient (age, gender, type of disease, drug intake, laboratory parameters) in order to identify at the same time ADE detection or prevention rules, and potential ADE cases. Several statistics such as the probabilities of occurrence will be computed. In some cases, when no data is available from the CPOEs, the data mining task will be made possible by the use of semantic mining to retrieve drug names from the free-text collections.

The rules will have to be filtered and validated by experts to make sense and to limit their number. At the same time, the rules will be accompanied by validated explanations.

The potential ADE cases will have to be verified. For that purpose, additional tools will be designed for the review of the cases. Indeed, the case-validation process will rely on the complete analysis of the patients’ records to assess the real occurrence of adverse events. This review will be performed by medical experts.

Close contacts with the people involved in the next steps of the PSIP project will be necessary to provide a usable output. Indeed, the output will be defined and fully structured in the form of XML files that will be loaded on-the-fly by the different software programs that will use them, such as a CDSS.

The final activity of this task is to develop semi-automatic Data Mining techniques that will explore regularly the patients databases of hospitals in order to detect automatically if adverse events continue to occur or not, and if new adverse events can be detected, when a new drug is prescribed, or when a new technique is used. By doing that, this work will guarantee the regular update of the system and will also prepare the impact assessment of the tools.

1.6. Ariane’s thread

In order to make the reading of the material, method and results chapters easier, an “Ariane’s thread” is displayed at the beginning of each chapter (Figure 11).

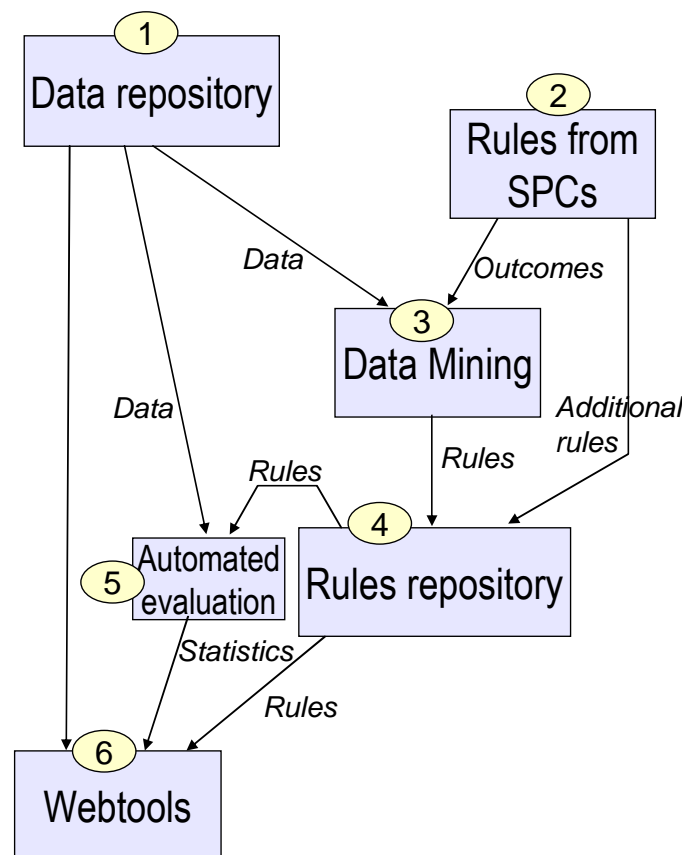


Figure 11. Ariane’s thread

The present study consists of several steps. Data from several hospitals are uploaded onto a data repository that fits a defined data model (label 1 on Figure 11). Some ADE detection rules are extracted from the Summaries of Product Characteristics (SPCs) (label 2 on Figure 11). An ADE detection rule induction is performed by means of data mining (label 3 on Figure 11). That step uses the data from the repository and a list of outcomes extracted from the SPCs. The rules generated by data mining are described and loaded into a rule repository (label 4 on Figure 11). Some other rules from SPCs are added. The rules from the repository are automatically evaluated against all the available data from the data repository (label 5 on Figure 11). Finally, the rules, their statistics and the original data are loaded into several web tools that allow for ADE detection and potential ADE cases exploration (label 6 on Figure 11).

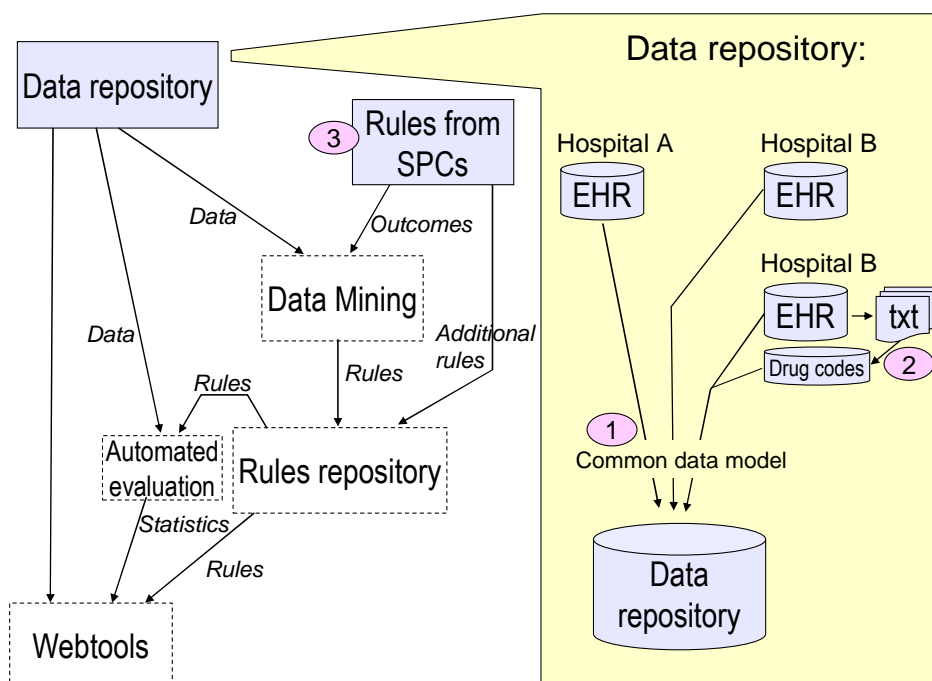
2. MATERIAL

2.1. Overview

The “Material” chapter consists of two main parts (Figure 12).

First, data are extracted from several hospitals and are used to feed a data repository. For that purpose, a data model is designed (label 1 on the right part of Figure 12, section 2.2 of this chapter). Unfortunately, in some hospitals there is no CPOE: in such cases, the drug codes are extracted from free-text reports (label 2 on Figure 12, section 2.4 of this chapter).

Then, summaries of product characteristics are loaded (label 3 on Figure 12). They are used to extract a list of outcomes and some additional rules.



**Figure 12. Ariane's thread – Available data and SPCs
(right part: details about the data repository)**

The different sections of the present chapter will present the steps described above:

- Section 2.2: the definition of a common data model
- Section 2.3: the extraction of the data into a common data repository that fits the data model
- Section 2.4: the enrichment of the data by extracting drug codes from free-text reports in some specific datasets
- Section 2.5: the extraction of ADE detection rules from the summaries of product characteristics.

2.2. Definition of a common data model

In order to perform data mining in past hospitalizations from different hospitals, it is mandatory to design a common way to describe the stays, whatever their origin. For that purpose, a data model is first defined in order to group together all the records from the various hospitals involved in this work and to structure the data into a common data repository.

2.2.1. Consideration about normalization

In order to define the data scheme, a compromise has to be reached from all those considerations (Figure 13). That compromise is reached by means of exchanges with physicians, computer scientists and statisticians concerning two main axes:

- Cardinality of the scheme (number of tables):
 - o If the data scheme contains fewer tables and relationships, the data quality control is made easier, as well as the data mining.
 - o If the data scheme contains more tables and relationships, the data are closer from the native scheme, the extraction is easier and with fewer errors, and the scheme is more stable over time.
- Number of columns (fields):
 - o If there are more columns, more data are available as well as calculated fields that allow for multivariate quality control.
 - o If there are fewer columns, the extraction is faster, with fewer errors, and the various datasets provided by the different partners are more likely to be compatible with each other and over time.

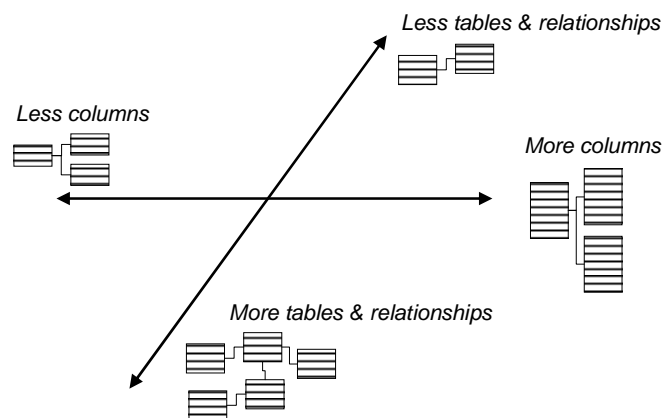


Figure 13. Compromise to reach in building a common data mode

When computer scientists are in charge of designing the data model of a production database, which is used in the daily processing of transactions, there is no doubt that the data model must be as normalized as possible. Normalization is a systematic way of ensuring that a database structure does not allow for any data redundancy, those redundancies being able to lead to a loss of data integrity in the case of insertion, update or deletion queries [Codd 1990]. Usually [Date 1999], a relational database is described as “normalized” when it fits the Third Normal Form as defined by Edgar F. Codd [Codd 1972].

However, in some cases, selective denormalization can be performed for performance reasons [Date 2005]. This is mainly encouraged in the field of data warehouse design [Kimball 2002] in case the database is not designed for transactional use, especially when the data are never partially updated. The present repository is supposed to be used only for reading purposes. The update of the data is performed by deleting all the records from a hospital and writing them again using a complete data export. There is no partial update of the data and, as a consequence, no potential loss of data integrity even if the data model is not fully normalized.

In addition, denormalized data models that allow for several redundancies in the data provide important features for the data quality management. This point will be detailed hereafter.

2.2.2. Available data

EHRs seem to be the best data source in the field of ADE detection [Gurwitz 2003, Jalloh 2006]. The aim is here to define a data model allowing for simple import of data from EHRs. A review of the available data is performed to answer the following questions:

- What structured data are available in the EHR of each partner?
- What part of those data is mandatory in most countries 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 to answer the following questions:

- What would be the simplest way to store laboratory results, drug administrations, diagnoses, demographic and administrative information?
- Are the available data schemes of the partners able to feed such a relational data scheme?

2.2.3. Terminologies

Terminologies are a standard way to label or designate concepts. In the medical field, terminologies allow for describing the non-quantitative knowledge that is available for a given patient. For instance, diabetic cataracts can be described using the code “H28.0” from the ICD10 instead of various free-text synonyms. In the present work, terminologies are useful to standardize the description of medical diagnoses, medical procedures, laboratory results and drug prescriptions. A review of the encoding systems enables us to choose common terminologies.

Diagnoses are encoded using the ICD10 [WHO 2010 (3)] (International statistical Classification of Diseases and related health problems, 10th Revision) of the World Health Organization.

Diagnoses related groups (DRGs) are encoded using the French or Danish national terminologies. 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: expected death frequency, average length of stay, expected ICU frequency, etc.

Drugs are encoded using the ATC [ATC 2009] (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 [IUPAC 2010]. 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 laboratory values.

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

2.2.4. Description of the data model

Figure 14 shows a simplified representation of the final data scheme. In the figure, fields are replaced using groups of fields.

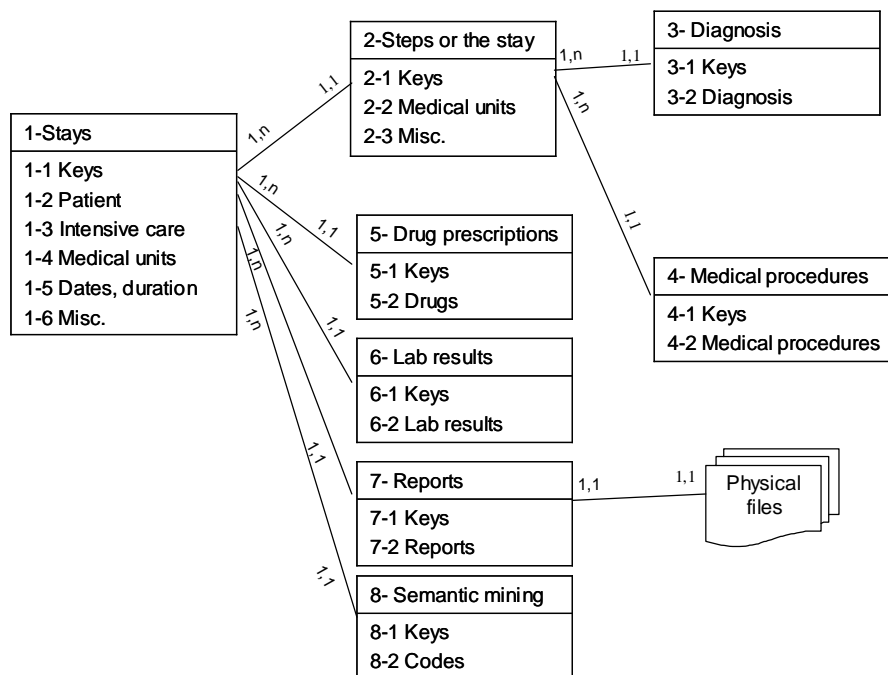


Figure 14. Simplified representation of the data scheme

The “1- Stays” table contains one row per stay (hospitalization) and mainly describes administrative and demographic information. One stay can be made up of one or several steps (emergency room, intensive car unit, cardiology department, etc.). The “2- Steps of the stay” table contains one row per step. Diagnoses and medical procedures are linked to the steps of the stays.

The data mining is performed stay-wise and details within the day level are ignored. For a given stay, drug prescriptions must be summed up 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 into the table “5- Drug prescriptions”.

In the “6-Lab results” table, one row corresponds to one measurement of one laboratory parameter at a given time. Each record should contain the normality range (upper bound and lower bound) as provided by the laboratory for each measure when this range is available. For a given parameter in a given hospital, those bounds may vary from one measure to another according to the robot that is used and to its calibration period.

Every free-text report is stored as a physical file linked to its stay by means of a specific table named “7-Reports”. In some hospitals, semantic mining is used to extract ICD10 and ATC codes from the reports, such as the discharge summary. A specific table enables to register those codes: “8-Semantic mining”.

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

The data scheme describes 8 tables from which 2 are dedicated to free-text reports and 89 fields from which 60 are not identifiers. The field list is shown in Table 2, Table 3, Table 4, Table 5, Table 6, Table 7, Table 8 & Table 9. The original version of the scheme description is complemented by detailed description of each field, but those detailed descriptions are not printed in the present document. That data scheme is complemented by physical files for the free-text reports.

This data model has been published [Chazard 2009 (1)] and is already used in another research project named AKENATON [Akenaton 2010]. This project is funded by the ANR (the French national research agency) and deals with managing the alerts of the devices that are used in telecardiology.

Table 2. The hospital stay table

Group	Field	Field (long name)	origin	kind
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 is a childbirth, the ID of the mother's stay	database	ID number
	id_stay_newborn	If it is 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 the current DRG	proportion, float between 0 and 1
	geo_state_01	Does the patient usually live in the hospital's country (state)?	constant	0/1
	geo_region_01	Does the patient usually live in the hospital's region?	geographic reference	0/1
	geo_dpt_01	Does the patient usually live in 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, through_icu_exp)	the number of stays computed in the whole hospital for the current 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 or resuscitation for this DRG	the proportion computed in the whole hospital for the current DRG	proportion, float between 0 and 1
	duration_icu	Duration in an intensive care or resuscitation unit	database	days (integer)
	duration_icu_exp	Expected duration in an intensive care or resuscitation unit	the average duration computed in the whole hospital for the current DRG	days (float)
	saps	Simplified Acute Physiological Score, 2 nd version (Gravity score)	database	integer
	duration_icu_sd	Standard deviation of the duration in an intensive care or resuscitation unit	the std dev of the duration computed in the whole hospital for the current DRG	days (float)
	delay_icu	Delay before ICU or 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 the current DRG	the average duration computed in the whole hospital for the current 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 the current DRG	the std dev of the duration computed in the whole hospital for the current 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 3. The steps of the hospital stays table

Group	Field	Field (long name)	origin	kind
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	Simplified Acute Physiological Score, 2 nd version (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 4. The diagnoses table

Group	Field	Field (long name)	origin	kind
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 5. The medical procedures table

Group	Field	Field (long name)	origin	kind
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 6. The drug table

Group	Field	Field (long name)	origin	kind
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 admission and the administration	database	Days (integer)
	dose	Total drug dose administered during the current day	database	number
	unit	Unit used for the total dose	database	name
	route	Route	database	name

Table 7. The lab results table

Group	Field	Field (long name)	origin	kind
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 admission 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 8. The reports table

Group	Field	Field (long name)	origin	kind
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 9. The semantic mining table

Group	Field	Field (long name)	origin	kind
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

2.2.5. Iterative quality control

Each table and field of the data model is complemented by a quality control procedure. The quality control is iteratively performed at each data extraction. Its goal is not to improve the datasets themselves, but to detect abnormalities in order to improve the extraction mechanisms of each hospital. The quality control consists of several checks that depend on the nature of the fields.

Unary checks apply to each value apart from the others, e.g. each cell of a table. Two unary checks are applied:

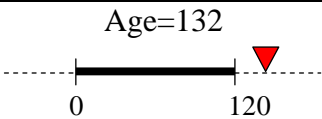
- Concerning the nature of the value:
 - o Is each value compatible with the expected type? The data types can be either numeric (binary, integer, float) or string (characterless or matching a specific pattern such as ICD10 codes)
- Concerning the value itself:
 - o If the value is a number, does it match the authorized range?
 - o If the value is a string, is it in the list of authorized values?

Some other checks apply to whole vectors, e.g. to one or several columns of a table. Several checks are applied:

- Univariate distribution:
 - o If the variable is numeric, is its histogram plausible?
 - o If the variable is a qualitative variable, is the proportion of the categories plausible?
- Multivariate distribution:
 - o Are conditional distributions plausible? Let Y be a qualitative variable or a set of classes of a quantitative variable, $Y \in \{Y_1, \dots, Y_k\}$. Is the distribution of X knowing that $Y=Y_i$ consistent with the value Y_i ?

Table 10 shows an example of quality control applied to a numeric variable, which is the age of the patients. Table 11 shows an example of quality control applied to a string variable, which is the principal diagnosis of the patients.

Table 10. Example of quality control and abnormal values for a numeric variable such as the age

Quality control	Example of incorrect result
Incorrect nature of the value <i>The value is not a number</i>	Age="old"
Value out of range <i>The number is too low or too high</i>	Age=132 

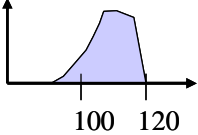
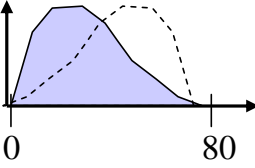
<p>Incorrect univariate distribution</p> <p><i>Each value is valid but the distribution is not plausible</i></p>	
<p>Incorrect multivariate distribution</p> <p><i>The total distribution is plausible, but the conditional distribution is not correct.</i></p>	 <p style="text-align: center;"><i>dashed line: department=pediatrics plain line: department≠pediatrics</i></p>

Table 11. Example of quality control and abnormal values for a string variable such as the principal diagnosis, encoded in ICD10

Quality control	Example of incorrect result
<p>Incorrect nature of the value</p> <p><i>The value doesn't match a given pattern or is not a string</i></p>	Diagnosis="AF35Z"
<p>Value out of range</p> <p><i>The value looks legal but the code doesn't exist</i></p>	Diagnosis="D80.9"
<p>Incorrect univariate distribution</p> <p><i>The proportion of a given value is abnormal</i></p>	$P("I46.9") = 60\%$ <i>I46.9 = cardiac arrest</i>
<p>Incorrect multivariate distribution</p> <p><i>The proportion of a given value looks plausible but is not compatible with a chosen subgroup</i></p>	$P("O80.0" \text{age} > 80) = 10\%$ <i>G80.0 = Spontaneous vertex delivery</i>

In addition, the redundancy of the denormalized data model allows for supplementary quality control checks. For instance, the fields “going through ICU” and “ICU duration” are strongly linked:

- if “going through ICU”=FALSE, then “ICU duration”=0
- if “ICU duration”>0 then “going through ICU”=TRUE

2.3. Data extraction

Data extractions are performed to feed a common repository. The extracted data have to fit the data model defined above. An important point is that no data have to be specifically recorded for the project as only routinely collected data are used. Those data include:

- Medical and administrative information, e.g. age, gender, admission date, and medical department.
- Diagnosis encoded using the 10th revision of the International Classification of Diseases, ICD10 [ICD 2009]
- Medical procedures encoded using national classifications, including therapeutic procedures (such as surgery) as well as diagnostic procedures (such as Magnetic Resonance Imaging)

- Drugs administered to the patient, encoded using the Anatomical Therapeutic Chemical (ATC) classification [ATC 2009]
- Laboratory results encoded using the IUPAC classification (International Union of Pure and Applied Chemistry) [IUPAC 2010]
- Free-text records, such as the discharge letter

Data extracted from EHRs are provided by the hospitals involved in the project. An iterative quality control of the extracted data is performed in order to improve the extraction mechanisms. The present work is performed using **92,686 complete records** from six hospitals. The records always include the six previously defined kinds of data. In Lille and Rouen, drug-related information is extracted from the reports and not from any CPOE. The dataset mostly concern cardiologic or geriatric units:

- Frederiksberg Hospital (RegionH, Denmark): 21,331 records
- Nordsjaelland Hospital (RegionH, Denmark): 23,067 records
- Denain General Hospital (France): 39,010 records
- Lille University Hospital (France): 7,711 records
- Rouen University Hospital (France): 1,367 records

In the Denain Hospital, the data extraction is automated in order to refresh the repository every month. The complete dataset allows for a 4-year follow up.

2.4. Extraction of the drug codes from the free-text reports

2.4.1. Objectives

In this work, ADE detection mainly uses administrative information, diagnoses and lab results from the EHR, and information about administered drug extracted from the CPOE (left part of Figure 15). Unfortunately, many hospitals still don't have any CPOE. In the datasets extracted from the Lille and Rouen university hospitals, there are no structured data about drug administration. In such a situation, a semantic mining tool called F-MTI is used in order to extract drug names from the free-text records (right part of Figure 15). Then, the drug codes are used as if they were retrieved from the CPOE.

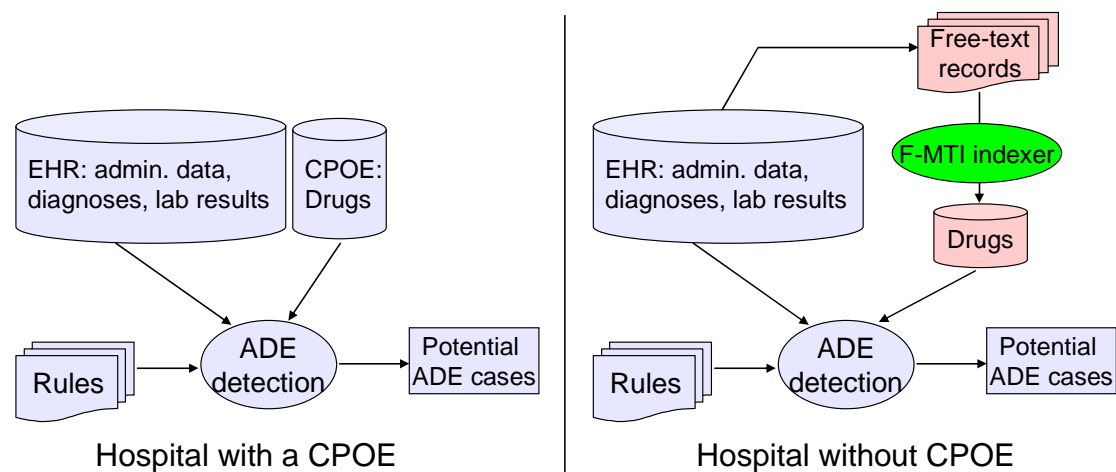


Figure 15. Rule-based detection of ADEs with a CPOE (left) or without CPOE (right)

2.4.2. The F-MTI tool

2.4.2.1. Presentation

The French Multi-Terminology Indexer (F-MTI) [Pereira 2008, Pereira 2009] is an automatic indexing tool based on a multi-terminology context. This is the first tool which is based on a multi-terminology context developed in another language than English. In English, the best-known automatic indexing tool is MTI from the US NLM. F-MTI was developed during collaboration between CISMef and the Vidal Company. It is the intellectual property of the Vidal Company.

The F-MTI tool has recently been enriched using four new terminologies devoted to drugs: the Anatomical Therapeutic Chemical (ATC) Classification (N=5,514), drug names with international non-proprietary names (INN) and brand names (N=11,353), the Orphanet thesaurus for rare diseases (N=7,424) and the chemical substances and pharmacological action terms of the MeSH Supplementary Concepts translated into French by the CISMef team (N=6,505 out of over 180,672). The terminologies and classifications around drugs and medical devices have been provided by the Vidal Company.

2.4.2.2. Functioning

F-MTI is called via a Web API, where the terminologies and the languages used for indexing can be selected. The results are returned in HTML, TXT or XML.

For each discharge letter, the document is first broken into sentences. Then each sentence is normalized (accents are removed, all words are switched to lower case and stemmed...) and stop words are removed to form a bag of words containing all the content words. The “bag” thus obtained is matched independently of the order of the words against all the terminology terms that have been processed in the same way. All terms formed with at least one word of the sentence are retrieved. Longer matches are preferred to shorter ones. All these candidate indexing terms are restricted to the semantically closest ATC and ICD10 terms using mappings. F-MTI can index a discharge letter in less than 1 second.

A phonemization algorithm has been developed to replace stemming, to improve the system and solve some mistakes found in the previous evaluations. Indeed, F-MTI encountered difficulties to recognize brand names due to incorrect spelling of the names in the discharge summaries. Some brand names are written improperly with dash ("-") or underscore ("_") or with an incorrect space " " (e.g. *di-antalvic*, *diffu k*, *di hydan*, *cacit D*, *calcidose vit D*, *co renitec*). Conversely, some brand names were written without dash ("-") or underscore ("_") or space (" "), as they normally should have to (e.g. *chibroproscar* instead of *chibro-proscar*; *bipreterax* instead of *bipreterax*).

Brand names are particular as they don't follow the French syntax. Brand Names can have English or Latin origins and in the spelling they don't mean anything at all. That's why using stemming (removing suffixes from the word) is logical for terms like those from ICD10 but not for brand names. Another alternative to using the exact word is the use of a phonemization algorithm. Phonemization [Brouard 2004] is based on the pronunciation of a word dependent on the language. The word is transformed into the corresponding list of phonemes. This method enables to take into account some incorrect spellings like “Eupanthyl” and “Eupantil”.

The algorithm was significantly changed to address the specific purpose of brand names' recognition. Specific English, Latin and Roman pronunciations were taken into account ("free" equals "fri", "chol" equals "kol", "I" equals "un", etc.). Some abbreviations were added ("vit" equals "vitamin"), single letters ("PH") were kept. The problems of dash, underscore and space were solved via alternative spellings. Some common words were removed (the brand name "PAR" equals "par"). And finally some items of the discharge summary were not taken into account: "way of life" and "exams" that can produce some noise in the indexing.

2.4.2.3. Terminologies and languages

F-MTI includes 22 terminologies:

- Drug-related terminologies:
 - o DCI (fr) for international names
 - o NC (fr) for commercial names
 - o WHO-ATC (fr/eng/dk): Anatomical Therapeutic Chemical Classification System (n=5,592)
- Patient safety-related terminologies:
 - o MedDRA (en) (n=19,862) [Northrop Grumman 2010]
 - o NCCMERP (eng) (n=468)
 - o PSIP taxonomy (eng) (n=320)
 - o WHO-ART (fr, eng): Adverse Reactions Terminology (n=3,483) [WHO 2010 (1)]
 - o WHO-ICPC (fr, dk, eng, du) International Classification for Patient Safety (n=647 in English and n=392 in French),
- Other terminologies: ACTS, CCAM, CISMef thesaurus, Cladimed, DRC, ICD9, ICD10, ICPC2, IUPAC, LPP, MedlinePlus, MeSH, MeSHSC, SNOMED3.5, TUV, VCM, & WHO-ICF

F-MTI also includes 27 equivalent mappings. The mappings are links between terms having the same meaning in different terminologies. F-MTI manages 6 different languages: French, English, Spanish, Portuguese, Danish and Dutch.

2.4.3. Main use of Semantic Mining in the present work

The semantic mining methods employed here are applied to unstructured documents and reports for information retrieval purposes: retrieving ATC codes from free-text records when no CPOE is available. Then, the approach adopted in this work is to use the retrieved information as complementary to the one available from structured hospital records and to inject it in the phase of ADE detection. In this regard, semantic mining (as a knowledge source) is therefore embedded in the Data Mining source. As a consequence, apart from the F-MTI evaluation that is performed hereafter, the semantic mining is not visible in the results of the present work.

The interest and the reliability of Semantic Mining is evaluated in the section 13 (*Appendix 5: validation of the use of Semantic Mining for ADE detection*) on page 219.

2.5. ADE detection Rules extracted from the SPCs

2.5.1. General content

In this stage, ADE detection rules are extracted from summaries of product characteristics. This will be useful (1) in order to get an exhaustive base of “official” ADE detection rules and (2) in order to get a list of outcomes that might result from ADEs.

The aim is to incorporate knowledge on ADEs that is available in existing databases maintained by pharmaceutical companies or related agencies. Such a database is made available by the Vidal Company, a partner involved in the PSIP Project. This database has been maintained for many years and contains thousands of possible risks related to drugs. The main sources of information come from international publications, health agencies (AFSSAPS, HAS), workshop groups, guidelines (AFSSAPS’s interaction thesaurus), and summaries of product characteristics (SPCs) provided by AFSSAPS, EMEA, FDA and others. These sources produce particularly reliable information, allowing the Vidal Company to produce *state of the art* information about drugs.

The database contains many types of interactions for a very large number of drugs, which results in a huge number of combinations among them and therefore corresponds to a huge number of possible events. For instance, 125,000 rules deal with drug-to-drug interactions and enables to generate 2,500 different alerts.

Those rules indicate the danger of interactions between drugs or contra-indications for drugs in specific circumstances. The rules may link two drugs regarding the interactions between them, or may link one drug with a particular diagnosis, allergy, laboratory value, or patient characteristic. The result of the rule is a description of the possible adverse effect and, depending on the type of rule, can also include a degree of severity or a suggestion for action.

2.5.2. Format of the rules

The knowledge content provided by the Vidal Company consists of a small number of general types of rules. Each type of rule consists of two conditions, and each condition is expressed on the basis of a specific parameter (e.g. drug, expressed as ATC code). By assigning values to these parameters (e.g. ATC= A02AB03), a large number of rules are created as instances of each rule type. The values of the rule parameters are stored in some tables and the content of these tables can be considered as the actual knowledge.

All the rules provided by the Vidal Company can be described through 9 different generic types. These 9 types consist of drug to drug interactions, drug to diagnosis interaction, drug to medical information (i.e. allergy) interaction, drug to laboratory result interaction, etc. The rules always associate together 2 conditions and a description of an interaction or an outcome. The syntax of those rules and an example are presented hereafter:

Kind of rule

r-atc-atc

The conditions are two ATC codes. When those drugs are prescribed together, the rule fires: there is a drug-to-drug interaction.

Conditions

$\text{Drug}(\text{ATC1})=1 \ \& \ \text{Drug}(\text{ATC2})=1 \ \& \ \text{ATC-ATC}([\text{ATC1}], [\text{ATC2}])=1$

The first drug matches ATC1 code. The second drug matches ATC2 code. An interaction is registered in a lookup table including pairs of drug ATC codes.

Effect

$\text{IAM}(\text{ATC-ATC}([\text{ATC1}], [\text{ATC2}]) \ [\text{id-IAM}])$

The rule detects a drug to drug interaction. In this case, the effect is a contra-indication (IAM). The description and the “conduct to take into account” is found in a lookup table from the effect code (id-IAM). The code of the interaction results as a function of the ATC codes of two drugs. This function is implemented as a lookup table associating pairs of ATC codes and effect codes. All possible effects are listed and coded.

Example

$\text{A02AB03} \ \& \ \text{B01AC30} \ \Rightarrow 194$

If a drug with ATC code A02AB03 is prescribed to a patient, then the prescription of drug with ATC code B01AC30 triggers id-IAM contraindication with code 194.

The rule type in the example above is instantiated to executable rules by assigning values to its parameters from a table, such as Table 12. Each row of this table associates the combination of two drugs (with ATC codes ATC1 and ATC2) with a specific interaction (with ID code id-IAM).

Table 12. Example of ATC-to-ATC interaction table. For each pair of ATC codes, the identifier of the interaction is provided if an interaction is described

ATC1	ATC2	id-IAM
A01AB09	B01AA03	181
A01AB09	B01AA07	181
A01AB17	B01AA03	342
A01AB17	B01AA07	342
A02AB01	B01AC06	194
A02AB01	B01AC30	194
A02AB01	C10BX02	194
A02AB01	N02BA01	194
A02AB01	N02BA51	194
A02AB03	B01AC06	194
...

An additional Table is used to list all the interactions found in the rules of this type and contains information related to each interaction. In the specific example of ATC-to-ATC rule type, the outcomes table (IAM) contains for each coded outcome a description of the risk, a suggestion and a code indicating the severity of the possible outcome.

Other types of rules correspond to different associations, such as a drug with a diagnosis. Since the rules contain different parameters and logic, they are described using other tables and mechanisms.

The types of rules that are stored in the VIDAL database are listed hereafter:

Kind of rule: r-atc-atc
Conditions: Drug(ATC1)=1 & Drug(ATC2)=1 & ATC-ATC([ATC1],[ATC2])=1
Effect: IAM(ATC-ATC([ATC1], [ATC2]) [id-IAM])
Description: Rules for alerts involving drug prescription with another pre-prescribed drug.
Example: IF a drug A02AB03 and a drug B01AC30 are prescribed to a patient
THEN caution of use (Reduced digestive absorption of acetylsalicylic acid and should take gastrointestinal topical agents and antacids some time (2 hours) after acetylsalicylic acid)

Kind of rule: r-atc-icd
Conditions: Drug(ATC) = 1 & Diag(ICD10) = 1 & ATC-ICD([ATC], [ICD10]) = 1
Effect: CI-PU(ATC-ICD([ATC], [ICD10])[id-CI-PU])
Description: IF a drug with ATC code X is prescribed to a patient who has a disease with ICD10 code Y,
THEN Contra-Indication or Precaution of Use.
Example: IF a drug B01AA03 is prescribed to a patient who suffers from disease D59.3
THEN Contraindication (Severe renal failure: creatinine clearance < 20 ml/min)

Kind of rule: r-atc-allergy
Conditions: Drug(ATC) = 1 & MedInfo(AllergyClass) = 1 & ATC-AllergyClass([ATC], [AllergyClass]) = 1
Effect: AllergyClass((ATC-AllergyClass([ATC], [AllergyClass])[id-AllergyClass])
Description: IF a drug with ATC code X is prescribed to a patient who suffers from an allergy id-AllergyClass
THEN Contraindication (AllergyClass).
Example: IF a drug A01AA51 is prescribed to a patient who suffers from Hypersensitivity to non-steroidal anti-inflammatory drugs (NSAID)
THEN Contraindication (Hypersensitivity to NSAID)

Kind of rule: r-atc-bio (1)
Conditions: Drug(ATC) = 1 & ATC-Bio([ATC]) = 1 & Bio(ATC-Bio([ATC])[bio-variable]) >= ATC-Bio([ATC])[min] & Bio(ATC-Bio([ATC])[bio-variable]) < ATC-Bio([ATC])[max]
Effect: CI-PU(ATC-Bio([ATC])[id-CI-PU])
Description: IF a drug with ATC code X is prescribed to a patient who has a lab value for idBio in the interval [min;max]
THEN Contra-Indication or Precaution for Use (id-CI-PU)
Example: IF a drug B01AA03 is prescribed to a patient who has a creatinine value between 0 and 20
THEN Contraindication (Severe renal failure)

Kind of rule: r-atc-bio (2)

Conditions: Drug(ATC) = 1 & ATC-Bio([ATC]) = 1
& Bio(ATC-Bio([ATC])[bio-variable]) >= ATC-Bio([ATC])[min]
& ATC-Bio([ATC])[max] = 0

Effect: CI-PU(ATC-Bio([ATC])[id-CI-PU])

Description: IF patient's biology measurement value for idBio >min
and MAX of this biology (abnormal value)=0
(probably max boundary does not make any difference)
and patient is going to be prescribed drug with ATC code x
THEN Contra-Indication or Precaution for Use (id-CI-PU).

Example: IF patient's creatinine value <0
and creatinine maximal abnormal value=0
and drug B01AB06 is prescribed
THEN Precaution for use (Chronic Renal Failure)

Kind of rule: r-atc-bio (3)

Conditions: Drug(ATC) = 1 & ATC-Bio([ATC]) = 1 & ATC-Bio([ATC])[min] = 0
& Bio(ATC-Bio([ATC])[bio-variable]) <= ATC-Bio([ATC])[max]

Effect: CI-PU(ATC-Bio([ATC])[id-CI-PU])

Description: This type of rules is for alerts involving drug prescription with a laboratory
result under a given boundary

Example: IF a drug B01AB06 is prescribed to a patient
who has a creatinine value under 60ml/min
THEN Contraindication (moderate renal failure)

Kind of rule: r-atc-medinfo (1)

Conditions: Drug(ATC) = 1 & ATC-MedInfo([ATC]) = 1
& Bio(ATC-MedInfo([ATC])[bio-variable]) >= ATC-MedInfo([ATC])[min]
& Bio(ATC-MedInfo([ATC])[bio-variable]) < ATC-MedInfo([ATC])[max]

Effect: CI-PU(ATC-MedInfo([ATC])[id-CI-PU])

Description: IF a drug with ATC code X is prescribed to a patient
who has a lab value Y above MAX or below MIN
THEN Contra-indication or Precaution for Use.

Example: IF drug B01AA03 is prescribed to a patient
whose age is greater than 780 months (65 years)
THEN Precaution for use ("patient is an elderly patient")

Kind of rule: r-atc-medinfo (2)

Conditions: Drug(ATC) = 1 & ATC-MedInfo([ATC]) = 1
& Bio(ATC-MedInfo([ATC]) [bio-variable]) >= ATC-MedInfo([ATC])[min]
& ATC-MedInfo([ATC])[max] = 0

Effect: CI-PU(ATC-MedInfo([ATC])[id-CI-PU])

Description: IF patient's lab value for bio-id >= MIN from the ATC-MedInfo table AND
MAX from the ATC-MedInfo table = 0 AND patient is going to be prescribed
drug with ATC code
THEN Contra-indication (id-CI-PU from ATC-Medinfo table)

Example: IF the drug N02BA01 is prescribed to a patient having a creatinine value
greater than 20 with max value=0
THEN Contraindication (Last 4 months of pregnancy)

Kind of rule: r-atc-medinfo (3)

Conditions: Drug(ATC) = 1 & ATC-MedInfo([ATC]) = 1
& ATC-MedInfo([ATC])[min] = 0
& Bio(ATC-MedInfo([ATC])[bio-variable]) <= ATC-MedInfo([ATC])[max]

Effect: CI-PU(ATC-MedInfo([ATC])[id-CI-PU])

Description: This type of rules is for alerts involving drug prescription with a bio variable (history/personal data) less than one limit.

Example: IF the drug B01AB06 is prescribed to a patient
whose age is smaller than 15 years
THEN Precaution of use (the patient is a child <15years old)

2.5.3. Usability of the rules

The main characteristics of this knowledge source are outlined below:

- The knowledge is provided in the form of rules. Every rule is made up of 2 conditions that lead to an effect.
- The parameters used in all rule conditions are
 - o drugs, encoded in ATC
 - o lab values, using a proprietary encoding
 - o or diagnoses, encoded in ICD10
- The result of the rules is of different kinds according to the type of rule. Unfortunately, the effect is described in free-text and, most of the time the effect is a contra-indication (example: “drugs A and B shouldn’t be prescribed together”), a warning or a use caution, but does not necessarily describe what could happen to the patient (e.g. “hemorrhage hazard”). In their form, it is not possible to use those rules to detect whether an outcome occurred or not because the output does not comply with any standard or taxonomy. The resulting effect is complex information and, according to the type of rule, may imply in addition a suggestion or a degree of severity.

However, by means of a manual examination and mapping of the rules, in several cases an outcome is present and can be described by a structured description. Such a work is performed systematically. From all the available rules, 228 unstructured descriptions of potential outcomes can be retrieved. Those outcomes are of two kinds:

- 149 clinical outcomes, e.g. “seizure hazard” or “hemorrhage”
- 79 paraclinical outcomes, e.g. “hyperkalemia” or “increase of pancreatic enzymes”

Section 3.2 (*Identification of the outcomes*) page 68 describes more precisely those outcomes and how they are used in this work.

3. METHOD

3.1. Introduction

The material consists of a data repository containing about 90,000 complete records. Each record contains demographic information, diagnoses, lab results and drug-related information (extracted from a CPOE or from free-text discharge letters). The objective is here to mine those records so as to produce ADE detection rules. The material also consists of rules from SPCs, which can be added to the rules discovered by data mining. Moreover, data-mining-based rule induction requires that some outcomes can be traced in the data. For that purpose, the SPCs presented in the material are also used to get an exhaustive list of possible outcomes in relation to ADEs.

In this chapter, the rules from the SPCs are analyzed so as to get a list of outcomes to trace for ADE detection purposes. This list is restricted to the outcomes that are traceable in the current datasets (label 1 of Figure 16, section 3.2 of the present chapter). Then, the data mining is performed using the available data and the list of outcomes (label 2 of Figure 16, section 3.4 of the present chapter). The data mining generates several ADE detection rules. The rules generated by data mining and some additional rules from the SPCs are integrated into a common rule repository (label 3 of Figure 16, section 3.5 of the present chapter). Then it is possible to enforce the execution of all the rules using the data of the repository: this will enable to calculate several statistics (label 4 of Figure 16, section 3.4 of the present chapter).

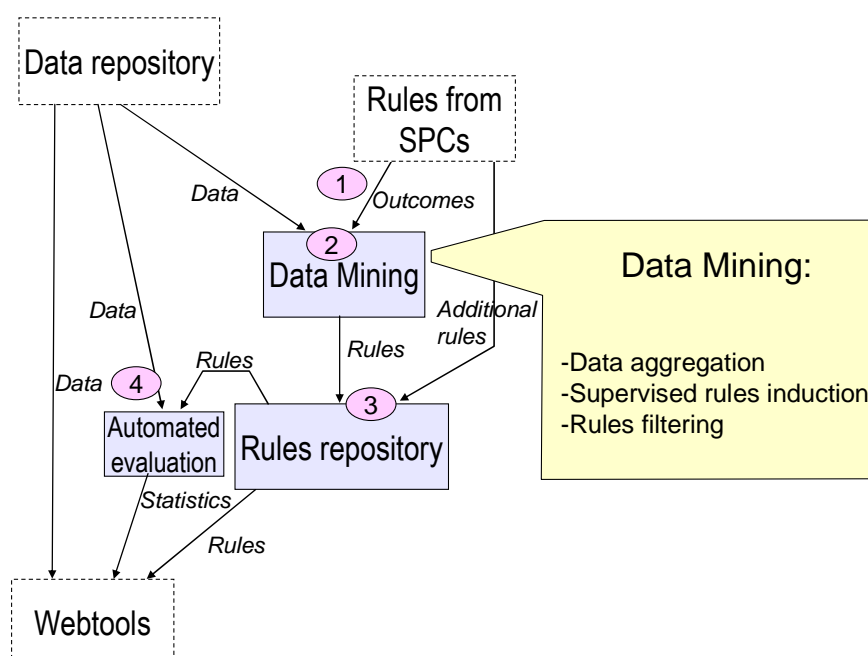


Figure 16. Ariane's thread – Rules induction, storage into a repository, and machine evaluation

The next sections follow the stages described above:

- Section 3.2: the rules extracted from the SPCs are used in order to get an exhaustive list of possible outcomes in relation with ADEs.
- Section 3.3: by means of a case review, several variables are identified in order to assess the clinical impact of ADEs on patients.
- Section 3.4: data mining is applied to the stays that are described in the data repository in order to identify (1) potential cases of ADEs and (2) situations at risk of ADEs in the form of ADE detection rules.
- Section 3.5: the ADE detection rules that are discovered by means of data mining are described into a common rules repository. This repository is complemented using ADE detection rules from other sources. All those rules are applied on the available data in order to compute various statistics.
- Section 3.6 provides the conclusion of the present chapter.

3.2. Identification of the outcomes in relation to ADEs

3.2.1. Principle

The present section describes how the list of traced outcomes is defined. The ADEs can be defined as injuries *due to medication management rather than the underlying condition of the patient* [IOM 2007]. Those injuries may consist of clinical or paraclinical signs that result from a drug intake, a dose modification, or a drug discontinuation. Those conditions are commonly described as *the use of a drug* [Gurwitz 2000]. All the ADEs that have been observed for a given drug are described in the related SPC (Summary of Product Characteristics). It is reasonable to think that the union of all the SPCs provides a list of all the potential outcomes related to drugs, according to the current knowledge.

In order to prepare data-mining-based rule induction, a list of outcomes is necessary. The drugs that are known to lead to those outcomes are not considered: only the list of the outcomes is used. The main source consists of the knowledge from SPCs validated by the AFSSAPS, the French national drug administration. The knowledge is collected by the Vidal Company and maintained in their knowledge database under the form of rules. From those rules, 228 different kinds of outcomes are precisely described. Those outcomes are clinical and paraclinical signs of ADEs. All of those 228 outcomes are examined in order to find *if* and *how* they can be traced in the EHRs.

3.2.2. Description of Outcomes from the SPCs

The outcomes from the SPCs described in the Vidal database belong to two major categories:

SPC clinical outcomes:

- **Definition:** these outcomes are observed only by means of clinical examination of the patient
- **Example:** the onset of urticarial reaction following the introduction of a drug during the hospitalization
- **Number:** 149 different clinical outcomes are described in the SPCs

SPC laboratory-related outcomes:

- **Definition:** these outcomes can be objectified by measuring a laboratory parameter
- **Example:** the occurrence of an acute renal failure during the stay
- **Number:** 79 different paraclinical outcomes are described in the SPCs

A significant number of outcomes have clinical and paraclinical definitions; every outcome is nevertheless assigned to a single category in the previous description.

The list of the outcomes is presented hereafter. As many outcomes are very similar to one another, they are grouped into the same label and the initial number of different outcomes is displayed in brackets (Figure 17).

Abortion (1)	Abruptio placentae (1)	Abscess (1)
Acidosis (1)	Acnea (1)	Addictive behavior (2)
Adrenal Insufficiency (1)	Agranulocytosis (1)	Alkalosis (1)
Amenorrhea (1)	Anaphylactic choc (1)	Anemia (2)
Angina pectoris (1)	Angioedema (1)	Angor (1)
Anorexia (1)	Antibodies depletion (1)	Antiphospholipid syndrome (1)
Anuria (1)	Anxiety (1)	Aplasia (3)
Asthma (2)	Ataxia (1)	Atherosclerosis (1)
Bacterial infection (2)	Calcium deficiency (1)	Cardiac insufficiency (1)
Cardiac rythm troubles (14)	Cardiopathia (4)	Colitis (1)
Coma (1)	Confusion (1)	Connective tissue disease (1)
Constipation (2)	Cough (1)	Cushing syndrome (1)
Dehydration (2)	Dental alteration (1)	Depression (1)
Dermatitis (4)	Diabetes (1)	Diarrhea (3)
Digestive pain (1)	Digitalic intoxication (1)	Diplopia (1)
Disseminated intravascular coagulation (1)		Drowsiness (2)
Dyskinesia (1)	Dysmenorrhea (1)	Dyspepsia (1)
Dyspnea (1)	Eczema (2)	Encephalopathy (1)
Endometriosis (1)	Enuresis (2)	Eruption (1)
Esophageal achalasia (1)	Extrapyramidal syndrome (1)	Fecal incontinence (1)
Fibrinolysis (1)	Fiever (1)	Fungal infection (2)
Gastric ulcar (1)	Gastritis (5)	General physical deterioration (1)
Glaucoma (1)	Gout (1)	Haemorrhage (3)
Haemorrhage hazard (9)	Headache (1)	Hematuria (1)
Hemolysis (1)	Hepatic cholestasis (1)	Hepatic cytolysis (2)
Hepatic fibrosis (1)	Hepatic insufficiency (2)	Hepatitis (2)
Hyperaldosteronism (1)	Hyperammonemia (1)	Hyperbilirubinemia (3)
Hypercalcemia (1)	Hypercalciuria (1)	Hypercapnia (1)
Hyperchloremia (1)	Hypercholesterolemia (1)	Hyper eosinophilia (1)
Hyperglycemia (1)	Hyperhydration (1)	Hyperkalemia (1)
Hyperlactatemia (1)	Hyperlipidemia (2)	Hypermagnesemia (1)
Hypernatremia (1)	Hyperparathyroidism (1)	Hyperphosphatemia (1)
Hyperprolactinemia (1)	Hyperuricemia (1)	Hypoalbuminemia (1)
Hypocalcemia (1)	Hypochloremia (1)	Hypodermatitis (1)
Hypoglycemia (1)	Hypokalemia (1)	Hypomagnesemia (1)
Hyponatremia (1)	Hypophosphatemia (1)	Hypopituitarism (1)
Hypoprotidemia (1)	Hypotension (2)	Hypoxemia (2)
Inflammation (1)	Intestinal obstruction (1)	Ionic disorder (1)
Lacrimal apparatus disease (1)		Leucopenia (2)
Loss in visual acuity (1)	Macrophage Activation Syndrome (1)	
Malaria attack (1)	Manic episode (1)	Metabolic trouble (1)
Myasthenia (1)	Myelodysplastic syndrome (1)	
Myocardial infarction (2)	Myopathia (1)	Nasal polyps (1)
Nausea (1)	Nephrotic syndrome (1)	Neuroleptic malignant syndrome (1)

Neuromuscular trouble (1)	Neutropenia (1)	Oliguria (1)
Osteoporosis (1)	Pancreatitis (2)	Pericarditis (2)
Pleuresia (1)	Polycythemia (1)	Psychotic disorders (1)
Pulmonary edema (1)	Raynaud syndrome (1)	Renal failure (4)
Respiratory distress (2)	Rhabdomyolysis (1)	Rhinitis (1)
Sarcoidosis (1)	Seizure (1)	Specific antibody (1)
Tendinopathy (1)	Thrombocytosis (1)	Thrombopenia (1)
Thrombosis (2)	Thrombosis hazard (3)	Thyroid disease (5)
Urticaria (1)	Uveitis (1)	Vasculitis (1)
Vomiting (1)	Water-electrolyte imbalance (1)	

Figure 17. List of outcomes extracted from the SPCs

3.2.3. Tracing of SPC Outcomes in this work

The appearance of the outcomes listed above has to be traced in the data. This is possible through different ways with respect to the category defined before.

SPC clinical outcomes:

- **Properties:**
 - o Most of them can be defined by means of ICD10 codes.
 - o Some of them can indirectly be traced by means of laboratory results. For instance, a hemorrhage under Vitamin K Antagonists is strongly linked to an increase of the INR, and might be followed by a decrease of the Hemoglobin blood level.
 - o Some of them can be traced as drug administrations. For instance, a hemorrhage under VKA is often followed by a Vitamin K administration.
- **Implementation:**
 - o ICD10-related outcomes cannot be used *as is*. Indeed, the diagnoses are encoded retrospectively at the discharge of the patient, and the ICD10 codes are provided without any date. As a consequence, it is not possible to know if a diagnosis such as “hemorrhage” is the reason of admission or a side effect that occurred during the hospitalization. Most of the time, it is the reason of admission.
 - o As much as possible, clinical outcomes are translated into laboratory-related signals or into drug-prescription-related signals.

SPC laboratory-related outcomes:

- **Properties:** these outcomes can be objectified by a deviation in laboratory results
- **Implementation:** all the paraclinical outcomes that are traceable in the laboratory results of the dataset are taken into account. Some of these outcomes are traced by means of combinations of several parameters, *e.g.* pancytopenia is the combination of anemia, leucopenia and thrombopenia.

The use of outcomes that are traced through abnormalities in lab results (*e.g.* hyperkalemia) or specific drug prescriptions (*e.g.* specific antidote) presents multiple advantages:

- The values are measured several times and the detailed chronology is available

- Many laboratory parameters are nearly systematically measured at least once, especially in some diseases or drug intakes, *e.g.* the INR when a patient is under Vitamin K Antagonists
- They are sensitive to changes induced by drugs
- They are objective

On the other hand, thresholds have to be defined for laboratory results. For instance, in this work a hyperkalemia is defined as a potassium value higher than 5.3 mmol/l. Too low a threshold would result in many false positive (i.e. cases that are not considered as ADEs by experts), and too high a threshold would result in too few cases, so that ADE detection rules cannot be discovered by data mining. The thresholds are defined empirically, reaching a consensus within a committee of 3 physicians of the project and 2 pharmacologists who are not part of the project. The thresholds are embedded in the name of the effect.

Finally, 83 outcomes out of the 228 initial outcomes can be traced within the present work. Those 83 outcomes are traced through 56 different variables. Several distinct outcomes in the SPC database can be traced by means of only one variable in the present work. Two examples are given below:

- The variable “*Hyperbilirubinemia*” takes into account 3 different outcomes in the SPC database: “*bilirubinemia higher than twice the normal upper bound*”, “*hyperbilirubinemia*” and “*jaundice*”.
- The variable “*Hypoxemia*” takes into account 2 different outcomes in the SPC database: “*arterial hypoxemia*” and “*PO₂ < 60 mm Hg at rest*”.

As a result, 37% of the SPC outcomes can be traced within the present work. The variables that are used are listed in Table 13. This limitation is not due to the method that is used, but is due to the available data. For instance, in the future, as soon as the electrocardiographic results are routinely available in a database, we will be able to trace 14 additional outcomes.

Table 13. List of outcomes that are traced in the data

Anemia (Hb<10g/dl)
Bacterial infection (detected by prescription of antibiotic)
Diarrhea (detected by prescription of an anti-diarrheal)
Diarrhea (detected by prescription of an antipropulsive)
Fungal infection (detected by prescription of a systemic antifungal)
Fungal infection (detected by prescription of griseofulvin)
Fungal infection (detected by the prescription of local antifungal)
Hemorrhage (detected by a prescription of haemostatic)
Heparin overdose (APTT>1.23)
Hepatic cholestasis (alkal. phosphatase>240 UI/l or bilirubins>22 µmol/l)
Hepatic cytolysis (alanine transa.>110 UI/l or aspartate transa.>110 UI/l)
High a CPK rate (CPK>195 UI/l)
Hypereosinophilia (eosinophilocytes>109/l)
Hyperkalemia (K+>5.3 mmol/l)
Hypernatremia (Na+>150 mmol/l)
Hypocalcemia (Ca++<2.2 mmol/l)
Hypokalemia (K+<3.0 mmol/l)
Hyponatremia (Na+<130 mmol/l)

Increase of pancreatic enzymes (amylase>90 UI/l or lipase>90 UI/l)
 Neutropenia (count<1500/mm³)
 Renal failure (creat.>135 μmol/L or urea>8.0 mmol/l)
 Thrombocytosis (count>600,000)
 Thrombopenia (count<75,000)
 VKA overdose (detected by a prescription of vitamin K)
 VKA overdose (INR>4.9)
 VKA underdose (INR<1.6)
 Acetaminophen overdose (detected by prescription of N-acetyl-cystein)
 Digitalis overdose (detected by the prescription of antidote)
 Digitalis overdose (digoxinemia>2.6 nmol/l)
 Drug overdose leading to methemoglobin formation (detected by the prescription of antidote)
 Drug overdose leading to sulfhemoglobin formation (detected by the prescription of antidote)
 Glaucoma (detected by the prescription of antiglaucoma miotic)
 Hyperalbuminemia (albuminemia>60 g/l)
 Hypercalcemia (calcemia>2.6 mmol/l)
 Hypocapnia
 Lithium overdose (to high a lithium rate)
 Opioids overdose (detected by the prescription of antidote)
 Pancytopenia
 Acidosis (pH<7.35)
 Alkalosis (pH>7.45)
 Cardiac failure (detected by prescription of cardiotoxic agent)
 Delirium (detected by the prescription of an antipsychotic)
 Gastric ulcer (detected by the prescription of antiH2)
 Hypercapnia
 Hyperglycemia (detected by the prescription of insulin analogue)
 Hyperglycemia (glycemia>15 mmol/l)
 Hyperthyroidism (T4>160 nmol/l or fT4>22 pmol/l or T3>3 nmol/l)
 Hypoalbuminemia (albuminemia<30 g/l)
 Hypoglycemia (glycemia<2.8 mmol/l)
 Hypothyroidism (T4<60 nmol/l or fT4<12 pmol/l or T3<1 nmol/l)
 Hypoxemia
 Inflammation (CRP>12 mg/l or VS1>50)
 Leukocytosis (leukocytes>15.10⁹/l)
 Leucopenia (leukocytes<3.10⁹/l)
 OEdema (detected by the prescription of diuretic)
 Seizure (detected by the prescription of intravenous anticonvulsive)

3.3. Evaluation of the clinical impact of ADEs

Adverse Drug Events cannot consist only of biological abnormalities because the definition describes effects *which [are] noxious* to the patients [EC 2001]. For that reason, it is necessary to find some ways to assess the clinical impact of the ADE we might be able to detect, such as death, higher length of stay, change in medication management, etc.

The main idea is that some variables could help to automatically assess the clinical impact of adverse drug events. In order to identify such variables, 90 hospitalizations with adverse events are reviewed by physicians assisted by a computer scientist. The experts are asked to review carefully the stays and to describe in what ways the various observed outcomes could have consequences on the patient or on his stay, such as “death”, or “longer hospitalization”.

Table 16. Third example of ADE case and inferred variable

Clinical case (ADE)	Mrs. Y encountered a vitamin K antagonist (VKA) overdose due to a pharmacokinetic interaction. The INR rose up to 7. The VKA was discontinued and a dose of vitamin K was administered.
What anomaly is visible in the EHR?	A dose of vitamin K is administered to a patient who is under vitamin K antagonist.
What variable(s) could be useful?	<i>vitaminK</i> {0;1}: = 1 if the patient is under VKA and is administered vitamin K with date of administration > 2 days from the admission = 0 in other cases
Examples of possible uses of the new variable(s)	Binary use: use the variable as is

The main output of this task, which is presented through three examples, consists of a list of variables that can be used to assess the clinical impact of ADEs. The variables that are constructed this way are automatically traced for each potential ADE case. Some examples of such variables are detailed in the Result chapter, in section 4.4 (*Evaluation of the ADE detection: preliminary results*) on page 111.

3.4. Data mining: a five-step procedure

A five-step procedure is followed (Figure 18) [Chazard 2009 (2-3)]:

1. Data are transformed into events: the native data are complex (thousands of codes, repeated measurement of various laboratory parameters, etc.). They are transformed into binary events. This point is discussed in section 5.2 (*Discussion of the method*) on page 139. Those events can happen or not. If they happen, they have a starting date and a stopping date.
2. The events are qualified as “potential cause of ADEs” or “potential events of ADEs”.
3. Statistical associations between conditions and outcomes are automatically discovered by means of decision trees and association rules. At this stage, associations do not necessary mean ADE. For instance we could find “age>90 & renal insufficiency => too long stay”.
4. The associations are filtered and only the associations that contain drugs in their list of causes are kept.
5. The rules obtained in this way are validated against academic knowledge and are tuned during validation sessions with experts.

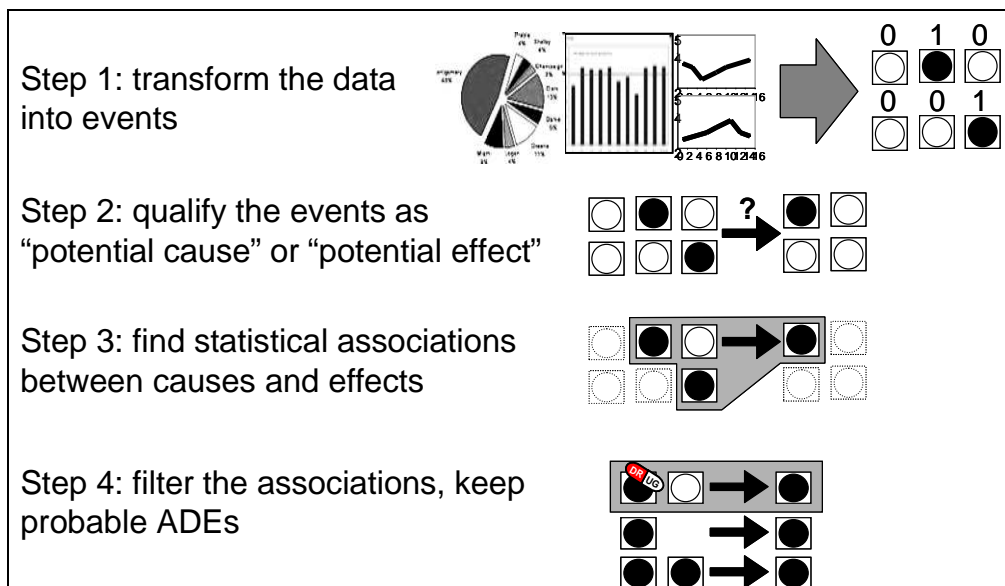


Figure 18. The 4 first steps of the procedure to "fish" ADEs

There are important differences between the present procedure and usual supervised rule induction. Supervised data mining methods often rely on a simple idea:

- The observed effect is known: the group with outcome (ADE=1) and the group without outcome (ADE=0) have already been identified.
- A large set of potential conditions is available; and a statistical method is used to find the most relevant conditions. The appropriate methods are used to explain the outcome by some of the conditions.

The approach is different in the present work, and the classical rule induction procedure is not usable, mostly because *the outcome is not identified*: nobody flagged the cases as "normal" (ADE=0) nor "abnormal" (ADE=1), and our objective is to avoid a time-consuming staff operated review. Our procedure can be described as follows:

- The data are transformed into events. Some of the events are flagged as "potential ADE manifestation", for example the occurrence of hyperkalemia.
- Supervised data mining methods are used to link events to potential ADE signs.
- The results of the rule induction are sets of conditions that lead to the outcome, but not sets of conditions that lead to an ADE. The patterns that are discovered require a lot of care in the interpretation, and the rule filtering operations are very important. The cases must be reviewed in order to identify the participation of the drugs in the observed outcomes, in relation to the underlying conditions of the patients. For instance, a drug might be the consequence of a chronic disease, and the outcome might be the consequence of that disease and not of the drug itself. However, in that example, statistical methods will highlight a participation of the drug in the outcome: there is a statistical link but there is no causal relationship.

Each one of those steps is described in a dedicated section hereafter (see sections 3.4.1, 3.4.2, 3.4.3, 3.4.4 & 3.4.5).

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 throughout the hospitalization) 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-day-long stay). The engines transform data into events. For a given hospital stay, events can have one of the two following values (Figure 20):

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

This transformation enables to describe five patterns as shown in Figure 20.

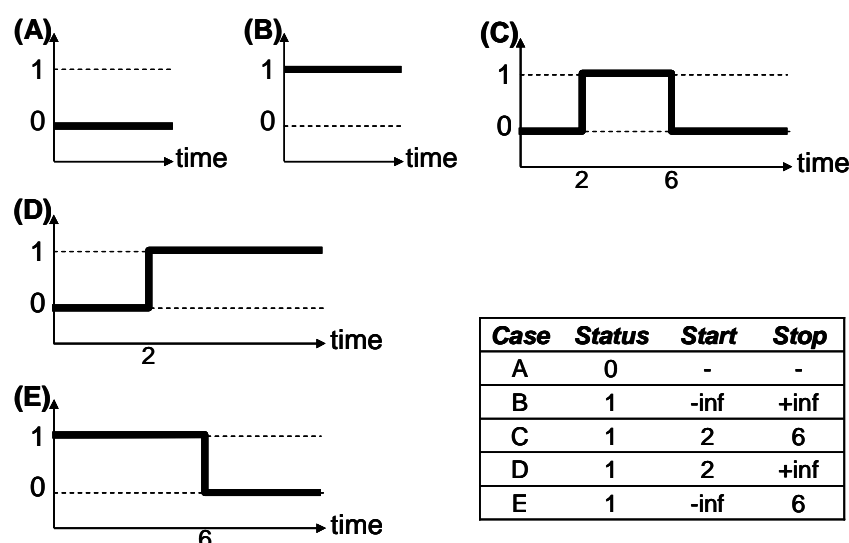


Figure 20. Different possible shapes of the events

3.4.1.2. Aggregation of the diagnoses

The mapping policy of the Diagnoses consists of grouping together similar codes. It mainly takes into account chronic diseases such as chronic renal insufficiency, hepatic insufficiency, some chronic infections, some neurodegenerative diseases, etc.

The referral diagnosis would be an interesting data, but in some countries such as in France, it is not formally encoded, although it is available as free text.

The complete mapping process is described in the following sections of *Appendix 4: Description of the output of this work (use of the XML files)*:

- Section 12.3 (*How to implement the rules for a prospective use (transactional use of the CDSS)?*) on page 212
- Section 12.4 (*How to implement the rules for a retrospective use (retrospective use of the CDSS, dashboards, confidence computation)?*) on page 214

3.4.1.3. Aggregation of the administrative information

The available administrative information is also mapped into events. The complete mapping process is described in the following sections of *Appendix 4: Description of the output of this work (use of the XML files)*:

- Section 12.3 (*How to implement the rules for a prospective use (transactional use of the CDSS)?*) on page 212
- Section 12.4 (*How to implement the rules for a retrospective use (retrospective use of the CDSS, dashboards, confidence computation)?*) on page 214

3.4.1.4. Aggregation of the information about drug administration

The aggregation of the drugs administered is first explained through an example, and then general principles are detailed.

Example of data aggregation

As an example, let's examine the aggregation of three different drugs:

- Heparin, that matches the category "Heparin"
- Warfarin and Phenindione, that both match the category "Vitamin K Antagonist (VKA)"

The drugs are taken into account since they are administered at least once.

The drug aggregation engine generates in this example 2 binary variables: *heparin* and *vit_K_antagonist*. Those variables are accompanied by starting dates and stopping dates when their values are set to 1 (Figure 21). In this example, as Warfarin and Phenindione both match the same category, they are grouped together.

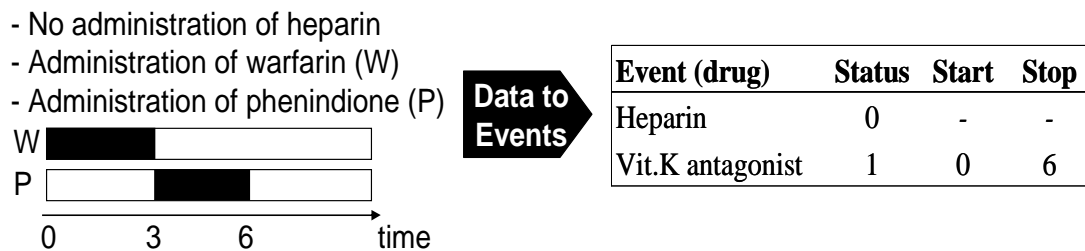


Figure 21. Example of administered drug aggregation: anticoagulant drugs

General principles

The definition of the drug mapping policy is an important work. In the present work, the ATC classification is used to map the drugs. But the ATC classification also offers a way to rank the drugs into a tree. Like many drug classifications, this tree is mainly based on the *therapeutic indication* of the drugs. It couldn't be used for the drug mapping: the ADEs are not related with the intention of the prescriber, but with the intrinsic properties of the drugs, those properties being either pharmacokinetic or pharmacodynamic.

The following example (Figure 22) shows that the acetyl salicylic acid (e.g. in aspirin) is ranked into 8 different ATC codes, those codes being ranked into 5 different chapters of the ATC. But whatever the therapeutic indication, the pharmacodynamic properties of the Aspirin are always the same: anti-inflammatory, pain killer, antipyretic, platelet aggregation inhibitor; and the pharmacokinetic properties still remain the same, due to its acidity. As a consequence some ADEs are shared by all

those available forms of Aspirin, despite therapeutic indications (and thus the ATC chapters) being different: gastric ulcer, hemorrhage, renal failure, overdoses of other drugs, etc.

A alimentary tract and metabolism <ul style="list-style-type: none">↳ A01AD other agents for local treatment<ul style="list-style-type: none">↳ <u>A01AD05</u> ...aspirin...
B blood and blood forming organs <ul style="list-style-type: none">↳ B01AC platelet aggregation inhibitors<ul style="list-style-type: none">↳ <u>B01AC06</u> ...aspirin...
C cardiovascular system <ul style="list-style-type: none">↳ C10BX (...) other combinations<ul style="list-style-type: none">↳ <u>C10BX01</u> & <u>C10BX02</u> ...aspirin...
M musculo-skeletal system <ul style="list-style-type: none">↳ M01BA anti-inflammatory<ul style="list-style-type: none">↳ <u>M01BA03</u> ...aspirin...
N nervous system <ul style="list-style-type: none">↳ N02BA salicylic acid and derivatives<ul style="list-style-type: none">↳ <u>N02BA01</u>, <u>N02BA51</u>, <u>N02BA71</u> ...aspirin...

Figure 22. Different positions of the Acetyl-Salicylic acid (Aspirin) in the ATC classification

Conversely, Rifampicin and Isoniazid could be ranked into the same category as they are both antibiotics used for the treatment of tuberculosis (Figure 23). But the first is a liver enzyme activator, and the second is a liver enzyme inhibitor. As a consequence, they may have opposite contributions to various ADEs.

J antiinfectives for systemic use <ul style="list-style-type: none">↳ J04A drugs for treatment of tuberculosis<ul style="list-style-type: none">↳ J04AB Antibiotics<ul style="list-style-type: none">↳ J04AB02 Rifampicin↳ J04AC Hydrazides<ul style="list-style-type: none">↳ J04AC01 Isoniazid

Figure 23. Places of Rifampicin and Isoniazid in the ATC classification

For all those reasons, it is necessary to design a customized drug classification. This classification aims at describing both pharmacodynamic and pharmacokinetic properties of drugs, whatever their therapeutic indication. Taking into account those properties leads to a redundant classification: a given drug can match several properties at the same time (Figure 24a). Moreover, a hierarchical redundancy is also useful (Figure 24b), making it possible for the statistical data-mining methods to automatically select the group or subgroup that maximizes the statistical link between causes and probability of ADE.

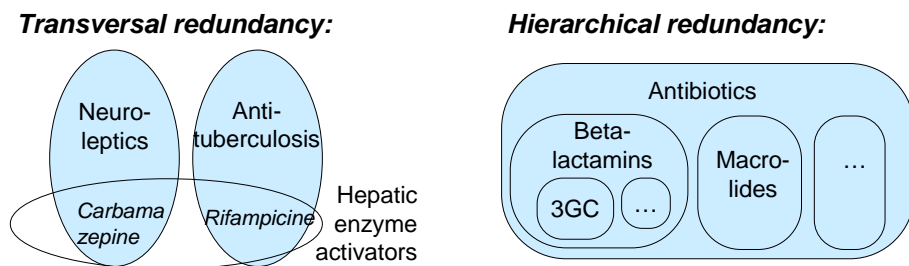


Figure 24. Transversal and hierarchical redundancy of the classification

The complete mapping process is described in the following sections of *Appendix 4: Description of the output of this work (use of the XML files)*:

- Section 12.3 (*How to implement the rules for a prospective use (transactional use of the CDSS?)*) on page 212
- Section 12.4 (*How to implement the rules for a retrospective use (retrospective use of the CDSS, dashboards, confidence computation?)*) on page 214

3.4.1.5. Aggregation of the laboratory Results

The aggregation of laboratory results is first explained through an example, and then general principles are detailed.

Example of data aggregation

As an example, let's examine the aggregation of two laboratory parameters:

- The digoxin level in the blood: Digoxin is a drug that is used for the treatment of cardiac insufficiency. Unfortunately overdoses are frequent and dangerous, that is why its level is frequently measured:
 - o it is considered as "too high" over 3 ng/ml (hyperdigoxinemia)
- The potassium level in the blood: potassium is an electrolyte. In some circumstances its concentration in the blood may vary out of the normality range. Unfortunately, those variations may lead to lethal cardiac arrhythmias:
 - o it is considered as "too low" under 3 mmol/l (hypokalemia)
 - o it is considered as "too high" over 5.3 mmol/l (hyperkalemia)

Thanks to boundaries, those 2 parameters enable to define 3 different anomalies. Those boundaries can be wider than the laboratory usual reference intervals: the thresholds have been defined in order to catch laboratory value anomalies that have a high probability to lead to clinical symptoms, or situations that are assumed to endanger the patient.

In that example the aggregation engine generates 3 binary variables: *hyperdigoxinemia*, *hypokalemia* and *hyperkalemia*. Those variables are accompanied by starting dates and stopping dates when their values are set to 1 (Figure 25). LOCF (Last Observation Carried Forward) is used to interpolate the available values.

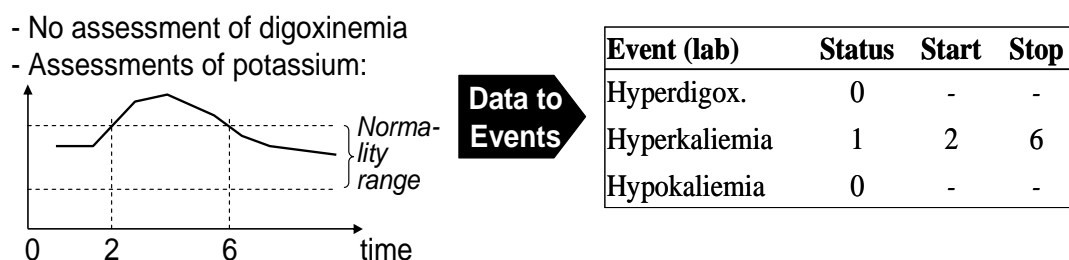


Figure 25. Example of laboratory result aggregation: the Digoxin and Potassium values of a stay

General principles

All the different laboratory parameters that are available in the data of the various hospital partners are examined. The list of the main syndromes that those parameters could help to detect is defined. Finally, a mapping table is designed, comparing:

- the reference interval that is declared in the data by each laboratory
- the reference interval of each parameter that can be found in the literature, taking into account the threshold beyond/above which clinical symptoms may appear
- the observed distribution of the values contained in the datasets (with respect to the observed diseases of the patients)

Finally, for a given syndrome, several theoretical laboratory parameters can be involved. E.g. the “pH” and the “base excess” can both be used to diagnose acidosis. Moreover, for a given laboratory parameter, several different real-data parameters can be used. E.g. either NPU12474 or NPU12491 IUPAC codes can be used to measure the “pH”. For a given parameter, several IUPAC codes might exist to render the measurement method that is used but also the various units that can be used. The ranges that are defined must take into account the different units that are used.

The complete mapping process is described in the following sections of *Appendix 4: Description of the output of this work (use of the XML files)*:

- Section 12.3 (*How to implement the rules for a prospective use (transactional use of the CDSS)?*) on page 212
- Section 12.4 (*How to implement the rules for a retrospective use (retrospective use of the CDSS, dashboards, confidence computation)?*) on page 214

3.4.1.6. Synthesis of the data aggregation

Diagnoses:

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

Administrative information:

The various administrative and demographic fields are aggregated into 15 different variables.

Drug administrations:

The 5,400 ATC codes are aggregated into 250 drug categories. Those categories are designed to be redundant: they enable 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 discontinuation is also traced as a potential ADE cause, the number of variables is then 500.

Lab results:

The laboratory results are aggregated into 35 laboratory signals.

The data aggregation produces one dataset per medical department. In each dataset, 588 variables can be used as *conditions* to explain 56 different outcomes.

3.4.2. Step 2: qualification of the events as “potential condition” or “potential outcome”

An informal analysis is first performed: in the available data, some can be identified as “potential condition of an ADE” and some others as “potential outcome of an ADE”. Table 17 shows examples of classifications.

Table 17. Examples of information classification: potential condition / potential outcome

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

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

Example of the Laboratory Results

In the previous example (Figure 25 on page 81), a hyperkalemia (too high potassium level) occurs from the 2nd day (included) to the 6th day (included). As the potassium level is assessed, two binary variables are generated in the dataset: *hyperkalemia* (which is true in this case) and *hypokalemia* (which is false in this case). They both can be used as conditions and as outcomes:

- *Hyperkalemia* (in this case: equals 1 from day 2 to day 6):
 1. *Hyperkalemia* is able to be an outcome with value=1. All the other events that occur before day 2 will be candidate to explain that outcome.
 2. *Hyperkalemia* is able to be a condition for every outcome that occur between day 2 and day 6.
- *Hypokalemia* (in this case: equals 0 all along the stay):
 1. *Hypokalemia* is able to be an outcome with value=0. All the other events will be candidate to explain the absence of outcome, whatever their date.
 2. *Hypokalemia* is able to be a condition with value=0 for every outcome, whenever it occurs.

This approach has two important advantages:

- A statistical association doesn't have any direction. But taking the dates into account prevents cause-to-effect relationship inversion. Events that are posterior to the outcomes cannot be interpreted as conditions. Events that are anterior but too far from the outcome are not taken into account.
- Outcomes can become conditions in their turn. That approach makes it possible to consider an **ADE domino effect**. A domino effect is observed when a first reaction leads to an outcome, and that outcome (alone or in combination with other factors) induces another outcome. For instance:
first drug A & age>70 → acute renal failure
then acute renal failure & drug B → hemorrhage

3.4.3. Step 3: statistical associations between potential conditions and outcomes

The previous steps enable to identify potential ADE conditions and potential ADE outcomes. The aim of statistical analysis is then to identify some links between (the combination of) potential conditions and potential outcomes. A result of the section 1.3 (*State of the Art in Data Mining*) on page 18 is the choice of decision trees and association rules for that purpose. This choice is done with respect to the type of the available data, and the expected output, which is a set of rules.

Decision trees are an important part of medical reasoning [Dzeroski 1996]. Fortunately many statistical methods enable to produce them. Decision trees and the CART method [Breiman 1984, Fayyad 1996, Lavrac 1999, Quinlan 1986, Ripley 1996, Zhang 2001] are used by means of the RPART package of R [Therneau 2007, R 2008]. Association rules [Agrawal 1993] are also used in addition to decision trees.

Decision trees and association rules enable to identify 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). In addition, the relative risk is computed (3: probability of the outcome knowing that the conditions are met, divided by the probability of the outcome knowing that the conditions are not met) and a Fisher's exact test for independency between the set of conditions and the outcome is performed.

Those statistics are not only provided by the methods, but recalculated afterwards. Indeed, those methods do not natively take the time constraints into account. Decision trees could conclude "drug A & lab result B → outcome C" although the 2 conditions and the outcome are not compatible with respect to the chronology. But once the rules are available, we are able to compute again those statistics by integrating time constraints.

$$Confidence = P(outcome | condition_1 \cap \dots \cap condition_K) \quad (1)$$

$$Support = P(outcome \cap condition_1 \cap \dots \cap condition_K) \quad (2)$$

$$Relative\ risk = \frac{P(outcome | condition_1 \cap \dots \cap condition_K)}{P(outcome | \overline{condition_1} \cap \dots \cap \overline{condition_K})} \quad (3)$$

Decision trees and association rules are automatically launched for each outcome in each hospital and each medical department. Thousands of rules are provided by both methods.

3.4.4. Step 4: filtering of the associations

Thousands of rules are generated, and reliable statistics are computed afterwards due to the complexity of time constraint integration. In addition, hopefully the most probable reasons why outcomes occur are in relation with the patient's underlying conditions rather than with the medication management. As a consequence, most of the rules that are discovered deal with the effect of diseases rather than the effects of medication. For those reasons, the rules require to be automatically filtered according to the following criteria:

- The rule must contain at least one of the following event types as a condition:
 1. one drug,
 2. or one drug discontinuation,
 3. or one laboratory parameter 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 outcome:
relative risk > 1
- The rule must lead to a significant Fisher's exact test for independency between the set of conditions and the outcome in at least one place (the rules are discovered separately in every medical department):
p value < 5%

In addition, an automatic modification of conditions related to the age of the patient is performed: the thresholds used for the age are automatically rounded off to a multiple of 5 years. This simple modification enables to remove duplicate rules.

3.4.5. Step 5: validation of the rules

Several technical meetings are organized with experts: physicians, pharmacologists, pharmacists and statisticians. The rules are examined by the experts and validated against summaries of product characteristics and bibliography. That review uses several drug-related web information portals [Pharmacorama 2009, BDAM 2009, Thériaque 2009], Pubmed referenced papers [Pubmed 2009], and French summaries of product characteristics provided by the Vidal Company.

During this review, the experts may suggest different modifications: queries to check what the pathological context of the cases is, and what the drugs involved precisely are, or tuning of the rules, to test different classes or subclasses of drugs. Sometimes, rules are manually enforced in agreement with academic knowledge in order to test some hypothesis.

Finally, the rules are reorganized: the conditions of each rule are characterized using one of the following types:

- **Subgroup conditions** help defining a subgroup that is used as the reference group for some statistic computations, such as the relative risk.
- **Cause conditions** are those that are explained in the comments of the rule.

- **Segmentation conditions** do not explain why an outcome occurs, but have an impact on the statistics that are computed.

Those kinds of conditions are explained hereafter.

The cause conditions are the most common conditions. A rule has at least one “cause condition”, quite often there are only “cause conditions”.

The subgroup conditions are used when it makes no sense to consider all the records at the same time: doing so could lead to overestimating the value of the relative risk for a given rule.

Let’s imagine a rule and its basic statistics:

VKA & drug_X → too high INR (VKA=vitamin K antagonist)

confidence = P(too high INR | VKA ∩ drug_X) = 10%

prevalence = P(too high INR) = 1%

relative risk = confidence / prevalence = 10

Fisher’s exact test p value = 0

A fast interpretation of those statistics could lead to the following conclusion: “the risk of having too high an INR is multiplied by 10 when *drug_X* is administered”. This is false because it makes no sense to use the control group that is implicitly used here: the control group is “no (*VKA and drug_X*)”, which means “no *VKA* or no *drug_X*”. But it is well known that the probability of encountering too high an INR without intake of *VKA* is very weak and has no interest in the field of ADE detection. The solution of this problem is to consider that, in order to discover causes of INR deviations, only the patients who are under *VKA* treatment have to be used to discover and evaluate the rules. This is what we called “subgroup conditions”.

Considering that “*VKA*” is a subgroup condition, the statistics become (changes appear in bold):

VKA & drug_X → too high INR (VKA=vitamin K antagonist)

confidence = P(too high INR | VKA ∩ drug_X) = 10%

*prevalence = P(too high INR | **VKA**) = **10%***

*relative risk = confidence / prevalence = **1***

*Fisher’s exact test p value = **1***

In that example, the rules cannot be validated anymore as the Fisher’s exact test p value is not significant anymore.

In the rules generated in this work, the following subgroups are systematically used to correct the statistics:

- “*VKA*” is used as “subgroup condition” (and is manually added if necessary) as soon as the outcome is one of the following:
 - “too high an INR” (*VKA* overdose, risk of hemorrhage)
 - “too low an INR” (*VKA* underdose, risk of thrombosis)
 - “vitamin K administration” (probable sign of *VKA* overdose)

- “Heparin” is used as “subgroup condition” for the outcome “too long APTT” (Heparin overdose, risk of hemorrhage)
- When the outcome is “hyperkalemia”, two subgroups are systematically mined separately: patients with renal failure and patients without renal failure.
- Finally, in most cases, when the outcome is a decrease of blood cell count (thrombopenia, anemia, neutropenia), the only tested subgroup is made up of patients without any ICD10 code of malignant disease.

In all those cases, the “subgroup” approach prevents the false discovery of rules.

The segmentation conditions are conditions that can modify the probability of the outcome. They are systematically explored. Their definition helped to improve rule management by reducing overlap and void spaces. Implementing such conditions (when justified) also helps to reduce over-alerting. Here is a simple example of 2 rules:

drug_X & age \geq 70 => renal failure
confidence=15%
explanation: drug_X can induce renal failure

drug_X & age<70 => renal failure
confidence=3%
explanation: drug_X can induce renal failure

The explanation of the rules is exactly the same, but for a given stay zero or one of those rules will fire, providing a more reliable confidence.

3.5. Central rule repository

A central rule repository is built. The aim is to group together rules from various origins through a common format, and to automatically execute and test all the rules in all the available datasets.

3.5.1. Knowledge integrated in the Central rule repository

The central rule repository is fed by different sources:

- Automatic rule production from the Denain hospital (F) and the RegionH hospital (Dk), using data mining (decision trees and association rules) [Chazard 2009 (2)].
- Manual transformation of rules coming from foreign sources: academic knowledge from the SPCs (provided by the Vidal Company) and scientific articles (presently Jah et al.)

The different sources of rules are explained hereafter. How the rules are stored and what they are used for are also described.

Each of the various sources of ADE rules has its own characteristics. A comparison is provided in Table 18. It seems that an efficient rule repository should incorporate rules from various sources because of the advantages and drawbacks of each method.

Table 18. General considerations about various sources of rules

Question	Academic knowledge	Staff operated record reviews	Data mining
Number of rules	Very high	Low	Medium
Need for validation	No, already performed before	Yes, performed during the review	Yes, must be performed, but sometimes difficult
Confidence of the rules	Not available	Computed by experts	Automatically computed
Number of the conditions	Few (1 or 2)	Depends on the review, often few	Variable, potentially high
Population segmentation, confidence optimization	No	No	Yes
Ability to propose rules when conditions never occur (e.g. absolute contra-indications)	Yes	No	No
Ability to describe very rare events	Yes	Sometimes possible	No
Ability to find all the interesting rules of a dataset	Yes	Yes but limited by the size of the review	Yes depending on the methods (association rules better than decision trees)
Time needed to find rules	Already available	Very time-consuming	Quite fast
Time needed to update confidence over space and times	Not possible	Very time-consuming	Very fast

3.5.1.1. Rules discovered by data mining (decision trees and association rules)

As described previously, data mining is used in the present work to discover ADE detection rules. Decision trees [Breiman 1984, Fayyad 1996, Lavarc 1999, Piatetsky-Shapiro 1991, Quinlan 1986, Ripley 1996, Zhang 2001, Therneau 2007] and association rules [Agrawal 1993] are used. The rules are computed separately for each medical department and are tuned and organized by a committee of experts. The obtained rules associate a variable number of conditions to a traceable outcome and take chronology into account. The conditions can have various natures:

- a drug prescription
- the presence of a group of ICD10 diagnoses
- an acute or chronic laboratory result anomaly
- data about the patient (e.g. gender, age)
- data about the organizational conditions of the hospital stay (e.g. admission by emergency, with a too high INR, on Saturday, etc.)

Advantages:

- The rules can be automatically implemented.
- Confidence (positive predictive value) and support are provided.
- Each rule can consider a variable number of causes, from various natures (lab, drugs, diagnoses, demographic and administrative variables, organizational causes).
- The population is segmented in order to optimize the confidence of the rules and to decrease over-alerting.
- The rules are contextualized: their confidence is computed separately on each medical department.

Drawbacks:

- Only events that are not too rare can be observed because a strong statistical link is required. Rare or unobserved events are not treated.
- Only conditions that are observed can be considered: absolute contra-indications should never appear in the output, however their implementation in a CDSS is mandatory. For instance, if two drugs are contraindicated in association because of a high probability of outcome, we hope we'll never be able to observe those drugs in association. As a consequence we won't be able to discover such a rule by means of data mining. However, the integration of such a rule is mandatory in case it would happen.
- Trees are known for their instability in relation with sampling and for the risk of omitting interesting rules in case an interesting condition always ranks second in the procedure.

Integration of this knowledge in the repository:

- The rules are implemented without any change in the rule repository. Only the rules that can be validated by the expert in agreement with SPCs, drug-related web information portals or Pubmed referenced articles are used [Morimoto 2004, Chazard 2009, Vidal 2009, Pubmed 2009].

3.5.1.2. Rules from academic knowledge

Vidal S.A. [Vidal 2009] is a French company that provides information about drugs and therapeutics. That information is used by almost all French physicians and is also available in different languages in other countries. Vidal's knowledge comes from official summaries of product characteristics, recent studies, official recommendations, literature, and experts' advice. Vidal's products and services include drug information, therapeutic guidelines and decision support modules. Being a partner of the project, the Vidal Company provides formalized association rules that are deduced from the summaries of product characteristics. Those SPCs are initially produced by the French national drug agency, the AFSSAPS (Agence Française de Sécurité Sanitaire des Produits de Santé) [AFSSAPS 2009].

The rules provided by Vidal S.A. describe four alert levels:

- absolute contra-indication,
- relative contra-indication,
- use caution, and
- notice.

The rules are always built as follows: two causes brought together are linked to one effect. The effect is expressed according to a proprietary thesaurus. Both causes can be of several kinds (but at least one of the conditions is a drug or a class of drugs):

- drugs or classes of drugs
- classes of diagnoses
- creatinine clearance lower than a given threshold
- age, gender, pregnancy, allergies, breast feeding...

Advantages:

- That source provides all the "official" information and is exhaustive.

- Even rare outcomes are mentioned.
- Even conditions that might never occur together are described (exhaustive list of absolute contra-indications).

Drawbacks:

- The number of rules is very high.
- Support and confidence (positive predictive values) of the rules are not provided (academic knowledge relies on clinical trials and spontaneous declarations that do not reflect the prevalence of ADEs [Morimoto 2004, Murff 2003]).
- Some outcomes described by the rules are not encoded in the EHRs except in unstructured free text (e.g. clinical events). Only some of the outcomes are usable in this work. For that purpose, a mapping has to be defined.
- There are only and always two conditions per rule, no segmentation.
- There is no contextualization: the knowledge is supposed to be valuable all over the hospitals and medical units.

Integration of this knowledge in the repository:

- The rules are first restricted to absolute contra-indications.
- The rules are then limited to those where the effect is an outcome that can be traced in the database. Some approximations are done to trace some outcomes, e.g. “*drug_A => rhabdomyolysis*” is transformed into “*drug_A => hyperkalemia & elevation of muscle enzymes & renal insufficiency*”.
- Only rules that were not already discovered by the Data Mining procedure are added into the repository.

3.5.1.3. Rules from scientific articles

In a recent paper, Jah et al. published a list of 30 alerts from the VigiLanz commercial application [Jha 2008]. Those alerts are rules composed of a drug as the cause, and a laboratory alert. Ten of those alerts are validated as ADEs or potential ADEs.

Advantages:

- The rules are easy to implement, the outcome is traceable.
- Rules have been validated by a staff operated record review.
- Confidence (positive predictive value) and support have been computed.

Drawbacks:

- The number of events is low.
- There is only and always one condition per rule, there is no segmentation.
- The rules are not contextualized.

Integration of this knowledge in the repository:

- The rules are implemented without any change in the rule repository.

3.5.2. Rule description and storage in the central rule repository

An XML [XML 2009] scheme has been conceived to depict the rules. XML is chosen because of the following characteristics:

- XML enables to build semi-structured databases: a complex data scheme with much cardinality can be defined in a much simpler way than by using relational databases. Any update of the scheme is easy too.
- XML can be easily produced by many programs. R scripts are able to automatically generate XML in addition to standard output (Figure 26). During the test phase, it was easier to edit the data with only a simple text editor and to get preliminary results.
- XSL and XSL-FO transformation enable to quickly design many kinds of outputs (e.g. text files, HTML, PDF, and other XML files). All the programming languages are able to load XML data, which is useful when the computations are too complex for XSLT.
- A unique central repository can then be used to store all our knowledge about ADEs, including free text comments and bibliographic references.

The XML data scheme contains two main parts: (1) the rule description (2) the last available data about rule occurrences on every place (Figure 27). A more detailed description of those files (rules, roots of rules, occurrences, lexicon, explanations...) is provided in the Appendix in the section 12.2 (*Rules XML files*).

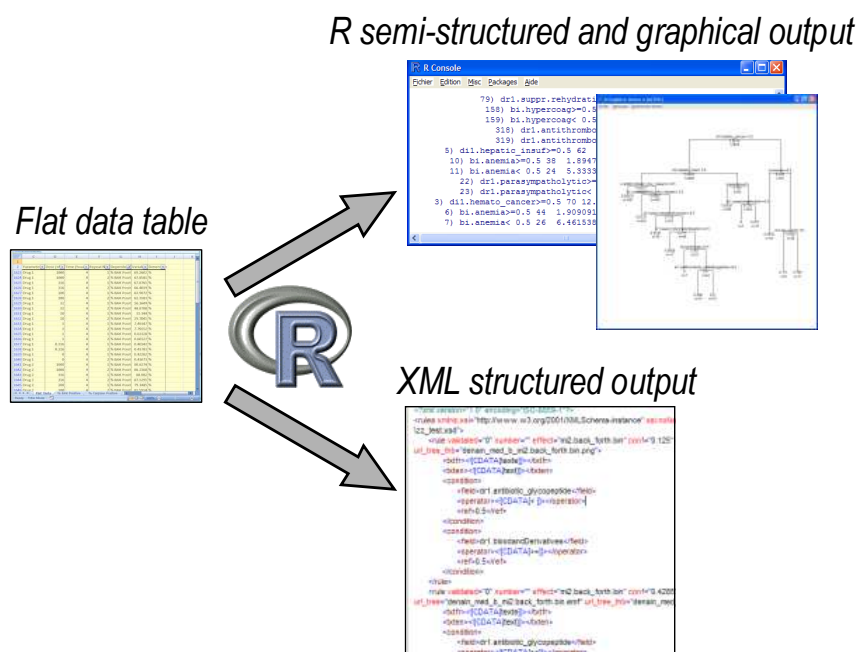


Figure 26. Automatic XML output of R scripts

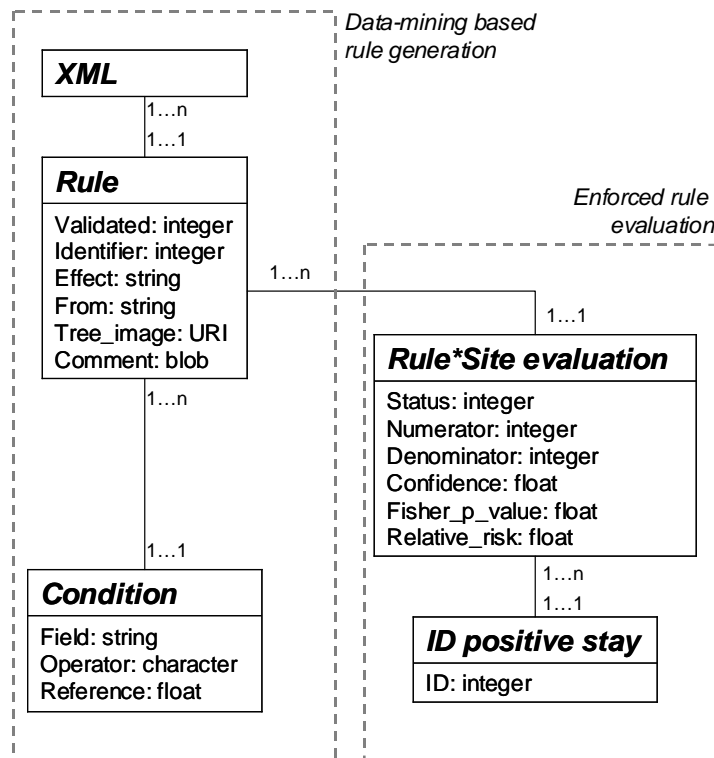


Figure 27. XML data scheme

All the rules are included in the central repository (Figure 28) and, whatever their origin, in a few minutes all the validated rules can be automatically evaluated in every medical department (Figure 29) [Chazard 2009 (4)]. That evaluation is important for several reasons:

- A rule might have been discovered in only one medical department or directly imported from academic knowledge. It must be evaluated in every other department.
- The hospitals that don't have any CPOE are not used for rule discovery, but can be used for automated machine evaluation of the rules and then for potential ADE detection.
- Some additional statistics are interesting and have to be computed. They are detailed below.
- Time consideration is complex and is partially taken into account in the data-mining discovery of rules. Conversely, the automated machine evaluation of the rules fully respects time constraints.

The automated machine evaluation of the rules uses the 90,000 available stays from all the hospital partners.

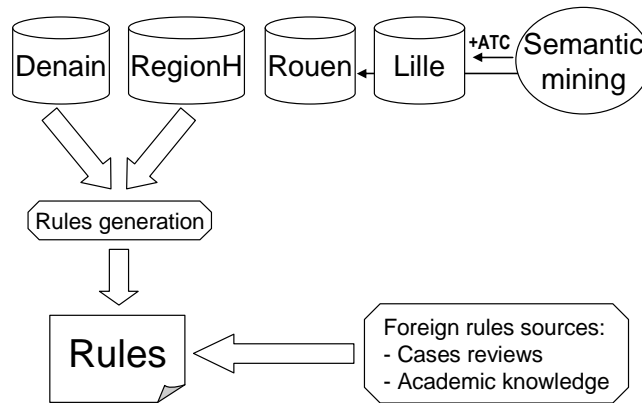


Figure 28. Rules inclusion in the repository

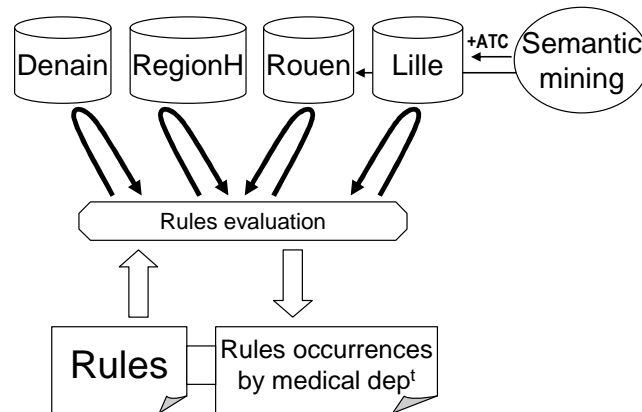


Figure 29. Automated machine evaluation of the rules in every department

The automated machine evaluation of the rules enables to add knowledge into the database: rule occurrences. It is possible to answer several questions for each rule, separately in each medical department. A rule is a set of conditions leading to an outcome, such as $C_1 \& \dots \& C_k \Rightarrow O$. The questions are:

- Do some hospital stays match the conditions?
number of stays = $\#(C_1 \cap \dots \cap C_k)$
- Among those stays, do some stays encounter the expected outcome?
number of stays = $\#(O \cap C_1 \cap \dots \cap C_k)$
support = $P(O \cap C_1 \cap \dots \cap C_k)$
confidence = $P(O | C_1 \cap \dots \cap C_k)$
- What are the identifiers of the stays that match the complete rule? They will be used in the case review.
- Is it possible to quantify the strength of the association?

$$\text{relative risk } RR = \frac{P(O | C_1 \cap \dots \cap C_k)}{P(O | (C_1 \cap \dots \cap C_k))}$$

p value of the Fisher's exact test for independency between the outcome (O) and the set of conditions (C₁ ∩ ... ∩ C_k)

NB: the computation is less optimistic in some cases, as discussed in the section 3.4.5 (Step 5: validation of the rules)

- When the outcome occurs, what is the delay between the conditions and the outcome?

*median delay between $t1$ and $t2$ where
 $t1$: time when all the conditions are met
 $t2$: time when the outcome occurs*

- Are those patients similar to others? (Descriptive statistics only)
on the subset $O \cap C_1 \cap \dots \cap C_k$, compute some descriptive statistics: sex ratio, average age, proportion of alcoholism, proportion of renal failure, proportion of hepatic insufficiency, etc.
- What happens then to those patients? (Descriptive statistics only)
on the subset $O \cap C_1 \cap \dots \cap C_k$, compute some descriptive statistics: proportion of death, average length of stay, etc.

3.6. Conclusion

In the *Method* chapter, data extracted from hospitals are mined by means of decision trees and association rules in order to identify ADE detection rules. Those rules are filtered, tuned and validated by Experts. They are then described into a central rule repository. This rule repository is completed using rules from the summaries of product characteristics, and rules from a commercial ADE detection system. All the rules are described by using the same formalism. Then the rules are automatically evaluated in all the datasets. This enables to automatically compute statistics for each rule in each medical department, and to detect potential ADE cases.

4. RESULTS

4.1. Overview of the chapter

A first part of the chapter gives an overview of the data mining results, through some examples of decision trees (label 1 of Figure 30, section 4.2 of the present chapter). Then the decision rules that are present in the rule repository are described, as well as the statistics that are computed by means of machine automated rule evaluation (label 2 of Figure 30, section 4.3 of the present chapter). Preliminary results of the evaluation of the clinical impact of ADEs are provided in section 4.4 (as this evaluation is not automated, it is not displayed on Figure 30). Finally, the web tools that enable ADE-related knowledge display and case review are presented through a use case (label 3 of Figure 30, section 4.5 of the present chapter). Those tools are fed by the data of the repository, the rules of the repository, and the statistics that have been computed.

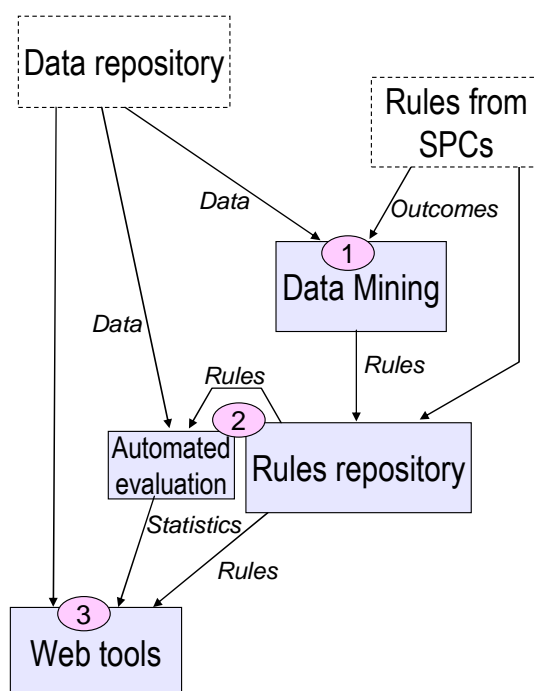


Figure 30. Breadcrumb trail – Results of data mining, rules of the repository and related statistics, web tools

4.2. Overview of data-mining results

Decision trees and association rules are systematically computed in order to explain each outcome by all the available potential conditions, in every hospital, and then in every medical departments. Thousands of rules are automatically generated.

Example of decision tree and rule generation

In the following example the outcome “VKA underdose (INR<1.6)” is followed-up. When patients are under vitamin K antagonist (VKA) treatment, the international normalized ratio from prothrombin times (INR) is traced in order to evaluate the treatment. In the case of too high INR, there is a VKA overdose; the patient could present a hemorrhage. In the 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 outcome is mostly associated with the admission of a patient with too high an INR (risk of bleeding, Figure 31). When a patient is admitted into the department with too high an INR, there might be an over-correction of the treatment and a risk of thrombosis in 29 % of cases. Elderly patients admitted with too high an 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: admission with too high an INR

The patients who are admitted with a normal INR value and receive at the same time VKA and a digestive prokinetic drug have too low an INR in 67 % cases (Figure 32). Digestive prokinetic drugs decrease the bio-availability of VKA.

The patients aged less than 76 who are given VKA and beta lactam antibiotics have too low an INR in 60 % cases (Figure 32). 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.

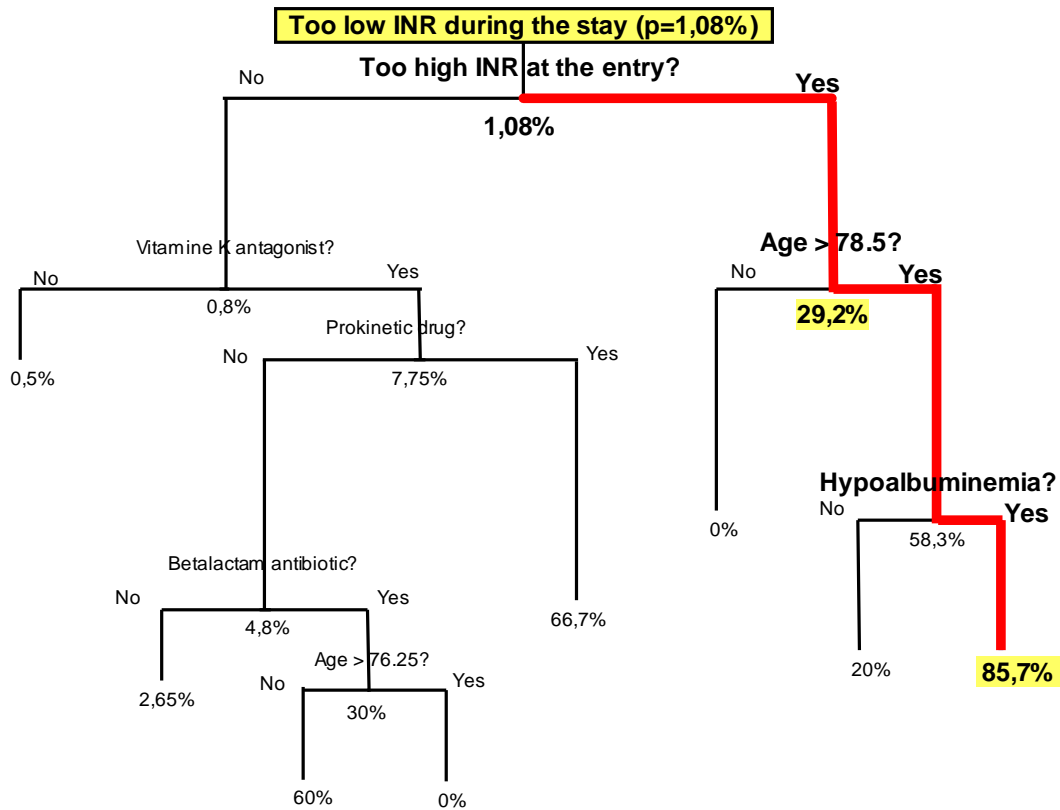


Figure 31. First rule gives $p(\text{too low INR during stay})=86\%$ instead of 1%

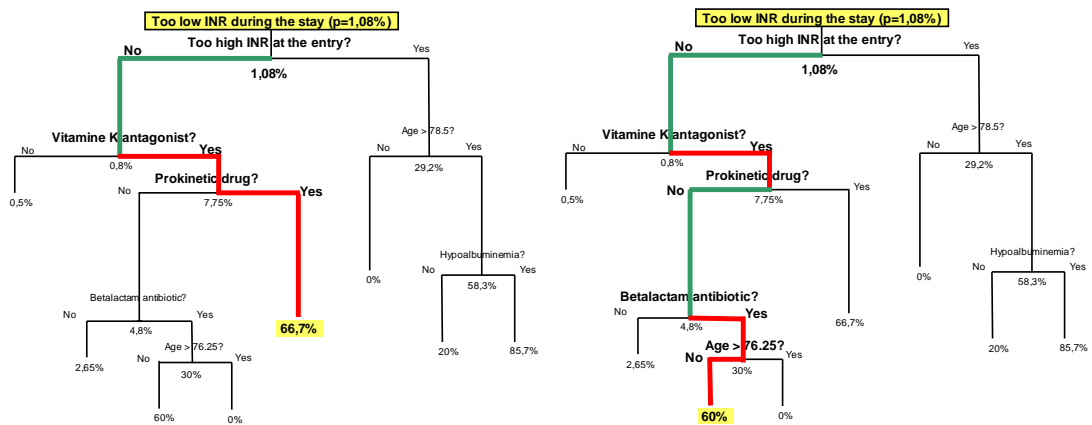


Figure 32. Second and third rules give $p(\text{too low INR during stay})=67\%$ and 80%

Overview of the rules discovered by means of data mining

For many kinds of outcomes, several cases are observed in the database and, by means of data mining, reproducible causes are identified for those outcomes. Some of those causes are drug administrations or drug discontinuations. This is detailed in the following sections. Those outcomes are presented in Table 19.

Table 19 Traceable outcomes that enabled to discover ADE detection rules

Anemia (Hb<10g/dl)
Bacterial infection (detected by prescription of antibiotic)
Diarrhea (detected by prescription of an anti-diarrheal)
Diarrhea (detected by prescription of an antipropulsive)
Fungal infection (detected by prescription of a systemic antifungal)
Fungal infection (detected by prescription of griseofulvin)
Fungal infection (detected by the prescription of local antifungal)
Hemorrhage (detected by a prescription of haemostatic)
Heparin overdose (APTT>1.23)
Hepatic cholestasis (alkal. phosphatase>240 UI/l or bilirubins>22 µmol/l)
Hepatic cytolysis (alanine transa.>110 UI/l or aspartate transa.>110 UI/l)
High a CPK rate (CPK>195 UI/l)
Hypereosinophilia (eosinophilocytes>10 ⁹ /l)
Hyperkalemia (K ⁺ >5.3 mmol/l)
Hypernatremia (Na ⁺ >150 mmol/l)
Hypocalcemia (Ca ⁺⁺ <2.2 mmol/l)
Hypokalemia (K ⁺ <3.0 mmol/l)
Hyponatremia (Na ⁺ <130 mmol/l)
Increase of pancreatic enzymes (amylase>90 UI/l or lipase>90 UI/l)
Neutropenia (count<1500/mm ³)
Renal failure (creat.>135 µmol/L or urea>8.0 mmol/l)
Thrombocytosis (count>600,000)
Thrombopenia (count<75,000)
VKA overdose (detected by a prescription of vitamin K)
VKA overdose (INR>4.9)
VKA underdose (INR<1.6)

In the dataset that is used in this work, some other outcomes are quite rare, so that they do not allow for data-mining-based rule discovery. For example, too high digoxinemia blood levels couldn't be observed. Anti-convulsive drugs were observed (it might be a signal of seizure) but they were nearly always administered from the day of admission: in the dataset there was probably not any first seizure during hospitalization. The outcomes that are too rare for rule induction in the current dataset are presented in Table 20.

Table 20 Traceable outcomes for which no case or too few cases can be observed in the dataset

Outcome	Number	Incidence rate
Acetaminophen overdose (detected by prescription of N-acetyl-cystein)	1	0.00% [0;0]
Digitalis overdose (detected by the prescription of antidote)	0	0.00% [0;0]
Digitalis overdose (digoxinemia>2.6 nmol/l)	17	0.03% [0.03;0.03]
Drug overdose leading to methemoglobin formation (detected by the prescription of antidote)	0	0.00% [0;0]
Drug overdose leading to sulfhemoglobin formation (detected by the prescription of antidote)	0	0.00% [0;0]
Glaucoma (detected by the prescription of antiglaucoma miotic)	58	0.11% [0.11;0.11]
Hyperalbuminemia (albuminemia>60 g/l)	0	0.00% [0;0]
Hypercalcemia (calcemia>2.6 mmol/l)	92	0.15% [0.15;0.15]
Hypocapnia	73	0.33% [0.32;0.33]
Lithium overdose (to high a lithium rate)	2	0.00% [0;0]

Opioids overdose (detected by the prescription of antidote)	11	0.02%	[0.02;0.02]
Pancytopenia	71	0.11%	[0.11;0.11]

Such outcomes should probably be explored again in bigger datasets. However, this kind of situation justifies the import of ADE detection rules from the summaries of product characteristics.

Finally, for some specific outcomes, several cases are observed but cannot be linked to any drug reproducible context in the current dataset. Those outcomes seem to be more in relation with the patients' underlying conditions than with the medication management. This is the direct opposite of the ADE definition we use. It is comforting to observe that a great number of outcomes are due to the diseases of the patients rather than the drugs they are administered. No rule could be found in the present work, but perhaps another dataset could give more interesting results

Those outcomes are presented in Table 21.

Table 21 Traceable outcomes that don't allow for ADE detection rule discovery

Outcome	Number	Incidence rate	
Acidosis (pH<7.35)	263	1.17%	[1.16;1.19]
Alkalosis (pH>7.45)	451	2.01%	[1.99;2.04]
Cardiac failure (detected by prescription of cardiotoxic agent)	258	0.48%	[0.48;0.49]
Delirium (detected by the prescription of an antipsychotic)	983	1.83%	[1.82;1.85]
Gastric ulcer (detected by the prescription of antiH2)	233	0.43%	[0.43;0.44]
Hypercapnia	297	1.33%	[1.31;1.34]
Hyperglycemia (detected by the prescription of insulin analogue)	296	0.55%	[0.55;0.56]
Hyperglycemia (glycemia>15 mmol/l)	246	0.39%	[0.39;0.4]
Hyperthyroidism (T4>160 nmol/l or fT4>22 pmol/l or T3>3 nmol/l)	145	0.24%	[0.23;0.24]
Hypoalbuminemia (albuminemia<30 g/l)	1712	2.75%	[2.73;2.77]
Hypoglycemia (glycemia<2.8 mmol/l)	141	0.23%	[0.22;0.23]
Hypothyroidism (T4<60 nmol/l or fT4<12 pmol/l or T3<1 nmol/l)	427	0.69%	[0.69;0.7]
Hypoxemia	379	1.69%	[1.67;1.71]
Inflammation (CRP>12 mg/l or VS1>50)	3145	5.05%	[5.01;5.08]
Leukocytosis (leukocytes>15.109/l)	862	1.38%	[1.37;1.39]
Leucopenia (leukocytes<3.109/l)	195	0.31%	[0.31;0.32]
Edema (detected by the prescription of diuretic)	1173	2.19%	[2.17;2.21]
Seizure (detected by the prescription of intravenous anticonvulsive)	725	1.35%	[1.34;1.36]

A representative example of such outcomes is the occurrence of a leucopenia, defined as a leukocyte count under 3×10^9 per liter of blood appearing at least 2 days after the admission. The decision tree that tries to explain such an outcome is printed in Figure 33.

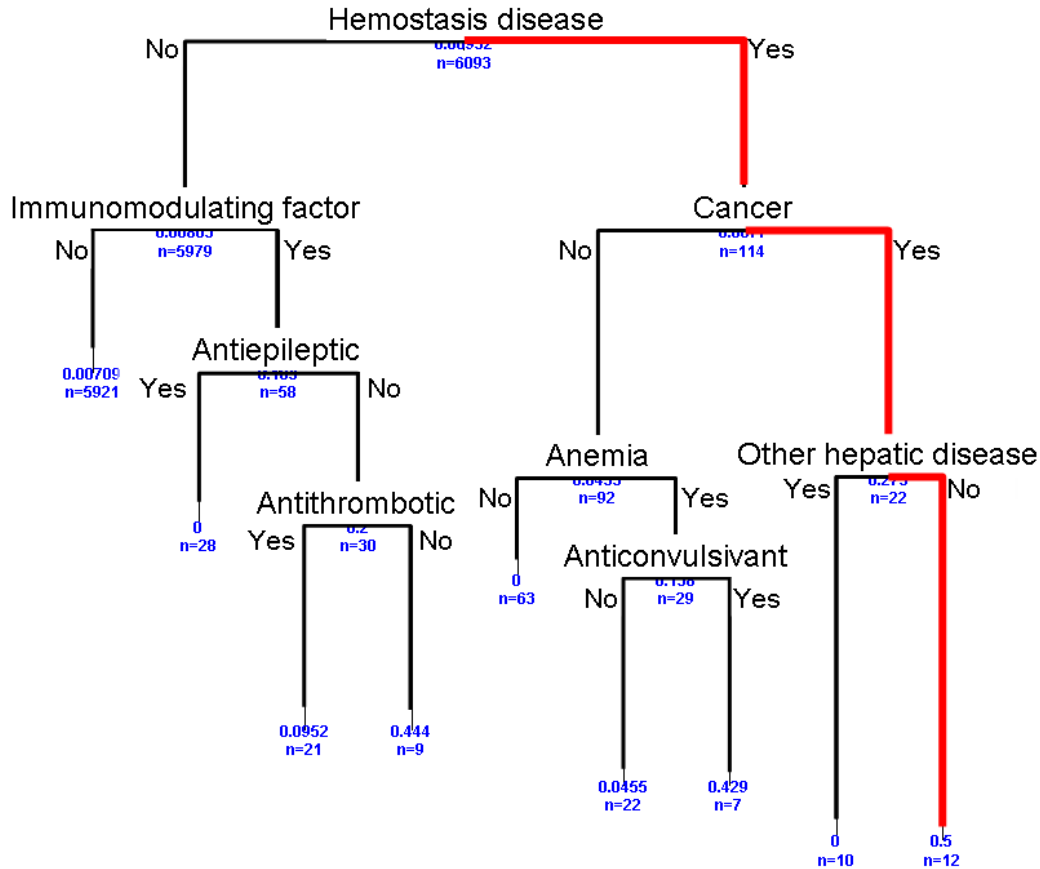


Figure 33. Decision tree: circumstances that lead to a leukopenia (leukocyte count < 3^{E09/l})

The branch that is drawn in red provides us with the following rule:

*Hemostasis disease & cancer & other hepatic disease → leukopenia
with 12 stays matching the conditions*

(Hemostasis disease ∩ cancer ∩ other hepatic disease) = 12

and 6 stays matching the conditions and the outcome

(Hemostasis disease ∩ cancer ∩ other hepatic disease ∩ leukopenia) = 6

the confidence of the rule is then 6/12 = 50%

Such a rule does not contain any drug-related condition. In that case, the outcome seems to be more linked with the patient's underlying conditions than with the medication management.

In the other rules of the present tree, some drugs appear but they are always involved knowing that a severe disease is already observed:

- The antiepileptic and antithrombotic drugs are involved knowing that the patient already receives immunomodulating factors, but such drugs are administered only in severe hematologic diseases. In addition, their presence is linked with the absence of outcome.
- The anticonvulsivant drugs only appear knowing that the patient already has an anemia.

The list below reproduces all the rules that can be read in the tree, from the right to the left. Each branch provides the probability (p) that a leucopenia occurs. Drugs are underlined:

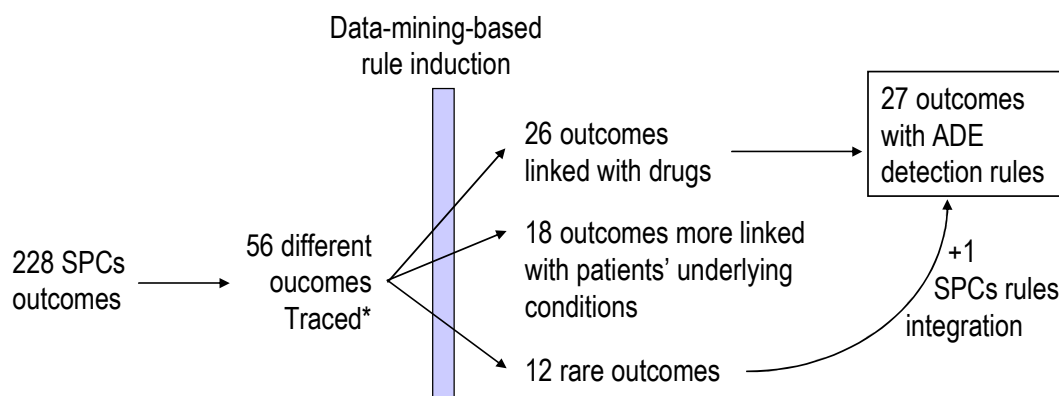
1. Hemostasis disease & cancer & other hepatic disease → p=0.5
2. Hemostasis disease & cancer & NO other hepatic disease → p=0
3. Hemostasis disease & NO cancer & anemia & anticonvulsivant → p=0.429
4. Hemostasis disease & NO cancer & anemia & NO anticonvulsivant → p=0.0455
5. Hemostasis disease & NO cancer & NO anemia → p=0
6. NO Hemostasis disease & immunomodulating factor & NO antiepileptic & NO antithrombotic → p=0.444
7. NO Hemostasis disease & immunomodulating factor & NO antiepileptic & antithrombotic → p=0.0952
8. NO Hemostasis disease & immunomodulating factor & antiepileptic → p=0
9. NO Hemostasis disease & NO immunomodulating factor → p=0.007

How many outcomes are mined, how many of them are detected by the rules?

Figure 34 is a flowchart of the different outcomes that are explored during the data-mining process. The review of the SPCs enables to identify 228 different outcomes: 83 of them can be traced in the current data by means of 56 specific variables (some of the 83 initial outcomes are very similar to one another):

- 12 of those outcomes are so rare that no rule can be discovered by means of data mining
- 18 of those outcomes frequently occur but seem to be more in relation to the patients' underlying conditions than to medication management
- Finally 26 of those outcomes appear to occur in relation to the medication, and allow for rule discovery.

Finally 27 outcomes are taken on board by the ADE detection rules that are described in the rule repository, as 40 additional rules from the Vidal's knowledge base are added in the rule collection as detailed in section 4.3.3.3 (*Origin of the rules (data mining, SPCs)*) on page 110.



(*) Those outcomes cover 83 of the initial SPC outcomes

Figure 34. Flowchart of the different outcomes explored in this work

4.3. Decision rules integrated in the central rule repository

4.3.1. Validated rules

Two hundred and thirty-six validated rules are now integrated in a central rule repository. This list is probably not exhaustive: other datasets would probably enable to detect other rules. The complete list is provided in chapter 14 (*Appendix 6: Validated rules*) on page 227. The following statistics are automatically computed for each rule in every medical department:

- number of cases, confidence and support
- relative risk, p value of a Fisher's exact test
- median delay of appearance of the outcome, other temporal quantiles
- descriptive statistics of the concerned cases

A complete output is available in English, French and Danish.

The following section displays the complete example of five decision rules. Such outputs are computed for all the 236 rules and cannot be displayed here.

4.3.2. Detailed example of five rules

In the five following examples, a rule is first described as a set of conditions that lead to an outcome. The link between causes and consequences is illustrated using a right arrow "→". When a condition is the absence of an event, the condition appears in gray.

For each rule, a table is provided. In the table, five hospitals (H1, H2, H3, H4 & H5) are displayed. For each hospital, the first line is called "Hx_all" and shows the various statistics when computed in every medical department at the same time. Then, when available, the statistics are computed department by department.

Below the table, some free-text is made available. Those explanations are to be used in the Scorecards (a web tool presented in section 4.5 (*Presentation of the results: the Expert Explorer and the Scorecards*) on page 117) and for the design of a contextualized CDSS. There are 3 kinds of text:

- A short text that can be used to display the explanations briefly.
- A long text that explains the rule in details. It sometimes includes bibliographic references.
- A text that can be used as recommendation of action.

The different lines of the table help to notice that the confidence of the rules varies a lot from one department to another. It is probably due to differences in the patient population (age and disease patterns), in the treatment, and in monitoring policies. Taking into account those various statistics will prevent the CDSS from over alerting.

Example of rule b003-0

Rule: VKA & selective serotonin recapture inhibitor
& NO respiratory obstruction → VKA overdose (INR>4.9)

Department	Confidence (PPV)	Support (frequency)	Median delay	Relative risk	Fisher's test P value
H1_all	2/21=9.5%	2/6110=0.3‰	5j	6.59	0.0376
H1_chir	<i>No stay</i>				
H1_geriatrics	1/2=50%	1/358=2.8‰	6j	12.71	0.0822
H1_gynobs	<i>No stay</i>				
H1_med_a	1/7=14.3%	1/1337=0.7‰	5j	8.64	0.1146
H1_med_b	0/7=0%	0/1026=0‰		0	1
H1_pneumo	1/7=14.3%	1/881=1.1‰	4j	4.31	0.216
H2_all	4/39=10.3%	4/11923=0.3‰	7.5j	36.95	0
H2_apoplexy	2/5=40%	2/369=5.4‰	13.5j	145.6	0.0004
H2_cardio_endocrino	1/16=6.3%	1/1967=0.5‰	3j	6.42	0.1514
H2_geriatrics	1/8=12.5%	1/493=2‰	11j	60.62	0.0322
H2_gynecology	<i>No stay</i>				
H2_icu	<i>No stay</i>				
H2_internal_med	0/1=0%	0/1514=0‰		0	1
H2_obstetrics	<i>No stay</i>				
H2_orthopedic	<i>No stay</i>				
H2_rheumatology	0/6=0%	0/446=0‰		0	1
H2_urology	0/1=0%	0/1107=0‰		0	1
H3_all	0/1=0%	0/1022=0‰		0	1
H4_all	0/5=0%	0/7685=0‰		0	1
H5_all	0/8=0%	0/1816=0‰		0	1

Comment:

- *Increased effect of the oral anticoagulant and hemorrhagic risk by SSRI.*
- *Selective serotonin reuptake inhibitors inhibit the oral anticoagulant metabolism and increase the risk of hemorrhage.*
- *When receiving SSRI, the dose of vitamin K antagonist will be adapted and clinical and laboratory monitoring will be enhanced.*

Example of rule b004-0

Rule: VKA & proton pump inhibitor & NO benzamide neuroleptic
 → VKA overdose (INR>4.9)

Department	Confidence (PPV)	Support (frequency)	Median delay	Relative risk	Fisher's test P value
H1_all	9/33=27.3%	9/6110=1.5‰	4j	20.46	0
H1_chir	<i>No stay</i>				
H1_geriatrics	<i>No stay</i>				
H1_gynobs	<i>No stay</i>				
H1_med_a	3/12=25%	3/1337=2.2‰	4j	16.56	0.0009
H1_med_b	3/9=33.3%	3/1026=2.9‰	7j	21.19	0.0004
H1_pneumo	3/12=25%	3/881=3.4‰	3j	8.05	0.0064
H2_all	0/101=0%	0/11923=0‰		0	1
H2_apoplexy	0/7=0%	0/369=0‰		0	1
H2_cardio_endocrino	0/25=0%	0/1967=0‰		0	1
H2_geriatrics	0/15=0%	0/493=0‰		0	1
H2_gynecology	<i>No stay</i>				
H2_icu	0/3=0%	0/357=0‰		0	0
H2_internal_med	0/4=0%	0/1514=0‰		0	1
H2_obstetrics	<i>No stay</i>				
H2_orthopedic	0/34=0%	0/1132=0‰		0	0
H2_rheumatology	0/9=0%	0/446=0‰		0	1
H2_urology	0/4=0%	0/1107=0‰		0	1
H3_all	0/4=0%	0/1022=0‰		0	1
H4_all	1/35=2.9%	1/7685=0.1‰	4j	3.08	0.2812
H5_all	0/4=0%	0/1816=0‰		0	1

Comment:

- Proton pump inhibitors may induce VKA overdose.
- The frequent intake of proton pump inhibitors increases the effect of the vitamin K antagonist and therefore the risk of bleeding.
Ref. : A prospective case-control from an emergency department. G. Cadioua La Revue de Médecine Interne Volume 27, Supplement 3, December 2006, Page S304 54ème Congrès de la Société nationale française de médecine interne, Congrès SNFMI.
- When receiving a PPI, the dose of VKA will be adjusted and clinical and laboratory monitoring will be enhanced.

Example of rule b012-1

Rule: VKA & amoxicillin and clavulanic acid & age ≥ 70
→ VKA overdose (INR>4.9)

Department	Confidence (PPV)	Support (frequency)	Median delay	Relative risk	Fisher's test P value
H1_all	15/73=20.6%	15/6110=2.5‰	6j	16.54	0
H1_chir	0/3=0%	0/1150=0‰		0	1
H1_geriatrics	5/12=41.7%	5/358=14‰	5j	14.42	0
H1_gynobs	<i>No stay</i>				
H1_med_a	2/13=15.4%	2/1337=1.5‰	4j	9.7	0.0197
H1_med_b	3/17=17.7%	3/1026=2.9‰	3j	11.13	0.0031
H1_pneumo	6/32=18.8%	6/881=6.8‰	9.5j	6.63	0.0005
H2_all	1/10=10%	1/11923=0.1‰	6j	33.09	0.0306
H2_apoplexy	<i>No stay</i>				
H2_cardio_endocrino	1/2=50%	1/1967=0.5‰	6j	51.71	0.0202
H2_geriatrics	0/2=0%	0/493=0‰		0	1
H2_gynecology	<i>No stay</i>				
H2_icu	<i>No stay</i>				
H2_internal_med	0/5=0%	0/1514=0‰		0	1
H2_obstetrics	<i>No stay</i>				
H2_orthopedic	<i>No stay</i>				
H2_rheumatology	<i>No stay</i>				
H2_urology	<i>No stay</i>				
H3_all	0/1=0%	0/1022=0‰		0	1
H4_all	0/8=0%	0/7685=0‰		0	1
H5_all	<i>No stay</i>				

Comment:

- Hemorrhage risk due to increased effect of VKAs under penicillin.
- Penicillin antibiotics reduce the hepatic metabolism of the vitamin K antagonist and increase the available fraction. Hemorrhagic risk is thus increased.
Ref: thésaurus IAM-AFSSAPS June 2009.
- When receiving penicillin antibiotics, the dose of the vitamin K antagonist will be adjusted and clinical and laboratory monitoring will be enhanced.

Example of rule b038-0

Rule: VKA & anti-diarrheal → VKA overdose (INR>4.9)

Department	Confidence (PPV)	Support (frequency)	Median delay	Relative risk	Fisher's test P value
H1_all	9/41=22%	9/6110=1.5‰	3j	16.45	0
H1_chir	0/4=0%	0/1150=0‰		0	1
H1_geriatrics	1/2=50%	1/358=2.8‰	2j	12.71	0.0822
H1_gynobs	<i>No stay</i>				
H1_med_a	3/9=33.3%	3/1337=2.2‰	3j	22.13	0.0003
H1_med_b	3/22=13.6%	3/1026=2.9‰	4j	8.56	0.0066
H1_pneumo	3/11=27.3%	3/881=3.4‰	3j	8.79	0.0049
H2_all	2/9=22.2%	2/11923=0.2‰	2j	75.64	0.0003
H2_apoplexy	<i>No stay</i>				
H2_cardio_endocrino	1/4=25%	1/1967=0.5‰	1j	25.83	0.0401
H2_geriatrics	1/4=25%	1/493=2‰	3j	122.2	0.0162
H2_gynecology	<i>No stay</i>				
H2_icu	<i>No stay</i>				
H2_internal_med	0/1=0%	0/1514=0‰		0	1
H2_obstetrics	<i>No stay</i>				
H2_orthopedic	<i>No stay</i>				
H2_rheumatology	<i>No stay</i>				
H2_urology	<i>No stay</i>				
H3_all	0/2=0%	0/1022=0‰		0	1
H4_all	0/8=0%	0/7685=0‰		0	1
H5_all	0/1=0%	0/1816=0‰		0	1

Comment:

- *Increased effect of VKAs with anti-diarrheal treatment.*
- *Under anti-diarrheal treatment, the effect of vitamin K antagonist anticoagulant treatment can be increased.*
- *In the case of an anti-diarrheal treatment, the dosage has to be adapted and the clinical and biological monitoring has to be increased.*

Example of rule b134-0

Rule: other beta lactam
 → fungal infection (detected by the prescription of a systemic antifungal)

Department	Confidence (PPV)	Support (frequency)	Median delay	Relative risk	Fisher's test P value
H1_all	5/23=21.7%	5/6110=0.8‰	3j	13.93	0
H1_chir	<i>No stay</i>				
H1_geriatrics	0/1=0%	0/358=0‰		0	1
H1_gynobs	<i>No stay</i>				
H1_med_a	2/2=100%	2/1337=1.5‰	3j	102.7	0.0001
H1_med_b	4/21=19.1%	4/1026=3.9‰	2.5j	6.6	0.0036
H1_pneumo	1/2=50%	1/881=1.1‰	6j	9.55	0.1039
H2_all	18/126=14.3%	18/11923=1.5‰	4j	7.73	0
H2_apoplexy	<i>No stay</i>				
H2_cardio_endocrino	2/15=13.3%	2/1967=1‰	5j	7.89	0.0279
H2_geriatrics	1/5=20%	1/493=2‰	28j	3.75	0.2463
H2_gynecology	<i>No stay</i>				
H2_icu	2/17=11.8%	2/357=5.6‰	1.5j	4.44	0.0912
H2_internal_med	11/75=14.7%	11/1514=7.3‰	5j	2.78	0.0028
H2_obstretics	0/1=0%	0/1969=0‰		0	0
H2_orthopedic	0/2=0%	0/1132=0‰		0	1
H2_rheumatology	1/2=50%	1/446=2.2‰	2j	7.93	0.126
H2_urology	0/5=0%	0/1107=0‰		0	1
H3_all	0/3=0%	0/1022=0‰		0	0
H4_all	0/101=0%	0/7685=0‰		0	0
H5_all	1/3=33.3%	1/1816=0.6‰	8j	15.11	0.0662

Comment:

- Occurrence of fungal infection during treatment with some cephalosporin related drugs.
- In some patients receiving cephalosporin related drugs (such as carboxypenicillins), fungal infection might occur, requiring the use of a systemic antifungal drug.
- Monitor the occurrence of fungal infection during treatment with cephalosporin related drugs.

4.3.3. Classification / Overview of the rules

The 236 rules can be classified according to several points of view. The following sections describe three kinds of classifications.

4.3.3.1. Modules

A module corresponds to a group of rules that lead to the same outcome. The modules and number of rules are detailed in Table 22.

Table 22. Modules and corresponding number of rules

Module	# rules
Coagulation disorders	
hemorrhage (detected by the prescription of haemostatic)	7
heparin overdose (activated partial thromboplastin time>1.23)	5
VKA overdose (INR>4.9)	57
VKA overdose (detected by the prescription of vitamin K)	2
thrombocytosis (count>600,000)	5
thrombopenia (count<75,000)	24
VKA underdose (INR<1.6)	18
Nosocomial infections	
bacterial infection (detected by the prescription of antibiotic)	4
fungal infection (detected by the prescription of a systemic antifungal)	8
fungal infection (detected by the prescription of local antifungal)	2
Ionic and renal disorders	
hyperkalemia ($K^+ > 5.3$ mmol/l)	63
hypocalcemia ($Ca^{++} < 2.2$ mmol/l)	1
hypokalemia ($K^+ < 3.0$ mmol/l)	1
hyponatremia ($Na^+ < 130$ mmol/l)	2
renal failure (creat.>135 μ mol/l or urea>8 mmol/l)	8
Others	
anemia (Hb<10g/dl)	2
diarrhea (detected by the prescription of an anti-diarrheal)	1
diarrhea (detected by the prescription of an antipropulsive)	1
hepatic cholestasis (alkaline phosphatase>240 UI/l or bilirubins>22 μ mol/l)	3
hepatic cytolysis (alanine transaminase>110 UI/l or aspartate transaminase>110 UI/l)	4
high a CPK rate (CPK>195 UI/l)	2
hypereosinophilia (eosinophilocytes>10 ⁹ /l)	4
increase of pancreatic enzymes (amylase>90 UI/l or lipase>90 UI/l)	7
lithium overdose (to high a lithium rate)	1
neutropenia (count<1,500/mm ³)	2
pancytopenia	1
paracetamol overdose (detected by the prescription of acetyl-cystein)	1
Total	236

4.3.3.2. Clinical niches

The rules can also be classified according to clinical niches. The modules focus on the different circumstances that could lead to given outcomes. Conversely, the clinical niches try to explore what could happen to specific groups of patients. Formally, clinical niches look more like a classification according to the causes of the rules.

In the scope of the present work, as proof of the concept, it was decided to focus the data mining on specific clinical niches:

- the consequences of anticoagulation
- the consequences of proton pump inhibitors

The discovery of rules in those niches is exhaustive (according to the available datasets), but it is not in the other niches. As a rule consists of one or several conditions, it can belong to several niches: a rule containing vitamin K antagonists and proton pump inhibitors belong to two different niches. The rules can be classified as follows:

- the consequences of anticoagulation: 111 rules
- the consequences of proton pump inhibitors: 16 rules
- rules out of any clinical niche: 111 rules

The detailed results are displayed in Table 23, Table 24 & Table 25.

Table 23 Anticoagulation niche

Module	# rules
hemorrhage (detected by the prescription of haemostatic)	7
heparin overdose (activated partial thromboplastin time>1.23)	5
VKA overdose (INR>4.9)	57
hyperkalemia (K+>5.3 mmol/l)	20
thrombopenia (count<75,000)	4
VKA underdose (INR<1.6)	18
Total	111

Table 24 Proton pump inhibitor niche

Module	# rules
anemia (Hb<10g/dl)	1
diarrhea (detected by the prescription of an anti-diarrheal)	1
diarrhea (detected by the prescription of an antipropulsive)	1
fungal infection (detected by the prescription of a systemic antifungal)	2
hemorrhage hazard (INR>4.9)	2
hepatic cytolysis (alanine transaminase>110 or aspartate transaminase>110)	1
hypocalcemia (Ca ⁺⁺ <2.2 mmol/l)	1
hyponatremia (Na+<130 mmol/l)	2
increase of pancreatic enzymes (amylase>90 UI/l or lipase>90 UI/l)	2
neutropenia (PNN <1500/mm ³)	1
thrombopenia (count<75,000)	2
Total	16

Table 25 Rules out of the niches

Module	# rules
anemia (Hb<10g/dl)	1
bacterial infection (detected by the prescription of antibiotic)	4
fungal infection (detected by the prescription of a systemic antifungal)	6
fungal infection (detected by the prescription of local antifungal)	2
hepatic cholestasis (alkal. phosphatase>240 UI/l or bilirubins>22 µmol/l)	3
hepatic cytolysis (alanine transaminase>110 UI/l or aspartate transaminase>110 UI/l)	3
high a CPK rate (CPK>195 UI/l)	2
hypereosinophilia (eosinophilocytes>10 ⁹ /l)	4
hyperkalemia (K+>5.3 mmol/l)	43
hypokalemia (K+<3.0 mmol/l)	1
increase of pancreatic enzymes (amylase>90 UI/l or lipase>90 UI/l)	5
lithium overdose (too high a lithium level)	1
neutropenia (PNN <1,500/mm ³)	1
Pancytopenia	1
paracetamol overdose (detected by the prescription of acetyl-cystein)	1
VKA overdose (detected by the prescription of vitamin K)	2
renal failure (creat.>135 µmol/L or urea>8 mmol/L)	8
thrombocytosis (count>600,000)	5
thrombopenia (count<75,000)	18
Total	111

4.3.3.3. Origin of the rules (data mining, SPCs)

The 236 rules can also be classified into several categories (Table 26):

- **SPCs only:** rules that only come from the Vidal knowledge base and that do not produce significant statistics. The contribution of the present work is to compute the confidence of those rules, and to quantify their usefulness: **40 rules**
- **DM & SPCs:** rules that have been found by data mining and confirmed by the Vidal knowledge base: **25 rules**
- **DM+:** rules found by data mining, that are confirmed by Vidal and bring new knowledge (additional segmentation conditions, significant reorganizing of the knowledge): **127 rules**
- **DM++:** rules found by data mining and that cannot be found in Vidal's knowledge base but can be indirectly explained by other information coming from that base (e.g. effect of a drug discontinuation, new context variables, calcemia instead of albuminemia...): **44 rules**

Table 26 Count of rules per origin

Source	# rules
SPCs only	40
DM & SPCs	25
DM+	127
DM++	44
Total	236

4.4. Evaluation of the ADE detection: preliminary results

The main operational objective of this work is to discover ADE detection rules and to execute them in order to detect potential ADE cases in past patients' records. A further step is to ensure that the cases are really ADEs. Another further step is to evaluate the impact of those ADEs on the patients in hospital. For that purpose, an evaluation is currently being performed in one of the hospitals. Preliminary results are detailed here concerning two sets of rules (modules):

- 57 rules leading to a VKA overdoses, detected by an $\text{INR} > 4.9$
- 63 rules leading to hyperkalemia, $\text{K}^+ > 5.3 \text{ mmol/l}$

Then, a global description of all the potential ADE cases is provided, whatever the kind of ADE.

The figures presented in the following subsections (4.4.1, 4.4.2, & 4.4.3) are computed by using stays from Denain, Frederiksberg and Nordsjaelland hospitals where CPOEs are available, in order to get reliable chronological data about the drugs. As for the rules discovery, only the stays that lasted at least 2 days are used.

In the following subsections, the "filtering of the rules" refers to the filtering settings that have been empirically decided in order to filter the rules in the Scorecards and in the CDSS. For a given hospital, only the rules that meet the 3 following conditions are applied (those statistics are described in section 3.5.2 *Rule description and storage in the central rule repository*):

- confidence $\geq 10\%$
- relative risk > 1
- p value of the Fisher's exact test $< 5\%$

The next subsections provide the preliminary results of the evaluation of several rules:

- rules that deal with the appearance of a hyperkalemia (section 4.4.1),
- rules that deal with the appearance of a VKA overdose (section 4.4.2), and
- all the rules grouped together (4.4.3).

4.4.1. Cases of Hyperkalemia ($\text{K}^+ > 5.3 \text{ mmol/l}$)

4.4.1.1. Definition

In this work, hyperkalemia is defined as the occurrence of a potassium blood level strictly over 5.3 mmol/l. This outcome is important because such a value could lead to severe and potentially lethal cardiac rhythm troubles. Patients suffering from renal insufficiency are more likely to have a hyperkalemia, spontaneously or due to the

action of drugs. Patients who don't suffer from renal insufficiency can also have hyperkalemia, mainly as part of adverse drug events.

4.4.1.2. Rule-based detection

Number of cases:

In the test hospital, 604 cases of hyperkalemia have occurred during hospitalization (second day or later), over 36,210 stays. The incidence rate amounts to 1.67% [1.65;1.69] of the stays

Note: Most of those cases are probably more in relation to the patient's underlying conditions than to drug intakes.

Number of cases within validated rules:

Without filtering the rules, 588 cases of hyperkalemia have occurred within the rules. The incidence rate amounts to 1.62% [1.61;1.64] of the stays.

Filtering the rules, 322 cases of hyperkalemia have occurred within the rules. The incidence rate amounts to 0.89% [0.88;0.90] of the stays.

Number of rules:

The cases of hyperkalemia are detected by means of 63 rules: 29 of them concern patients without renal failure and 34 of them concern patients suffering from renal failure.

Estimate of the accuracy of the detection rules:

29 different cases from the test hospital have been reviewed by experts. According to them, there was an ADE as described in the rules in 22 cases. The estimate of the accuracy of the rules in the field of hyperkalemia is 75.9% [69.0;82.7].

4.4.1.3. Clinical impact of ADEs leading to hyperkalemia

In the test hospital, 175 cases of hyperkalemia are detected using the ADE detection rules, over 21,737 stays. Two groups are defined:

- the 175 stays that involve a hyperkalemia within the filtered rules, that are called "potential ADE group"
- the 21,562 other stays, that are called "control group"

Both groups are compared with respect to several variables. Several differences are found and displayed here. Comparisons of quantitative variables are computed by means of Student's tests. Comparison of binary variables are computed by means of Fisher's exact tests (as the proportion of ADE is low, Chi-square tests are not always feasible). Confidence intervals are provided with a 95% confidence.

A cause-to-effect relationship would require a complete case review that is not performed here.

Higher length of stay in potential ADE group

The average length of stay in control group is 7.96 days

The average length of stay in potential ADE group is 17.8 days

The confidence interval for the difference is [8.13;11.63] days. It is significantly different from 0 ($p < 2.2e-16$)

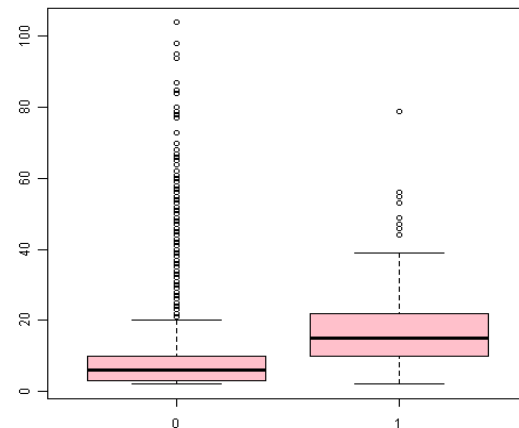


Figure 35. Length of stay in control group (left) and potential ADE group (right)

Higher death rate in potential ADE group

The proportion of death in control group is 3.15%.

The proportion of death in potential ADE group is 27.4%.

The odds ratio is 12 [8.08;16.5], which is significantly different from 1 ($p < 2.2e-16$).

Higher multimorbidity rate in potential ADE group

The “multimorbidity” variable is set to 1 if the different principal diagnoses of the stay belong to several medical specialties; otherwise it is set to 0.

The proportion of multimorbidity in control group is 3.15%.

The proportion of multimorbidity in potential ADE group is 10.9%.

The odds ratio is 3.9 [2.25;6.30], which is significantly different from 1 ($p = 2.44e-06$).

4.4.2. Cases of VKA overdose (INR>4.9)

4.4.2.1. Definition

Patients who are at risk of thrombosis or myocardial infarction are most often treated using vitamin K antagonists (VKAs). In that case, the dose must be regularly adapted. For that purpose, the International Normalized Ratio from prothrombin times (INR) is systematically assessed. In case of too high an INR, there is a VKA overdose: the patient is at risk of encountering a hemorrhage. In case of too high an INR, the therapeutic choice of the practitioner depends on the INR level. In this work an INR over 4.9 is considered as “too high”. For very high values, the VKA can be stopped, and vitamin K can be administered. If the patient encounters a hemorrhage, a traceable consequence can be the occurrence of an anemia.

4.4.2.2. Rule-based detection

Number of cases:

292 cases of VKA overdoses (INR>4.9, second day or later) have occurred during hospitalization over 35,442 stays. The incidence rate amounts to 0.82% [0.82;0.83] of the stays.

Note: Most of those cases are more probably linked with the patient's diet or genetic background than with drug interactions.

Number of cases within the validated rules:

Without filtering the rules, 248 cases occurred within the rules. The incidence rate amounts to 0.70% [0.69;0.71] of the stays. The number of cases and the incidence rate are the same when the rules are filtered.

Number of rules:

57 rules enable to detecting VKA overdoses. 17 of them involve VKA, antibiotics and sometimes other conditions. 44 other rules involve VKA, other drugs than antibiotics and sometimes other conditions. Finally, 6 other rules involve VKA and other conditions that are not related to drugs.

Estimate of the accuracy of the detection rules:

11 different cases from the test hospital have been reviewed by experts. According to them, there was an ADE as described in the rules in 5 cases. The estimate of the accuracy of the rules in the field of VKA overdoses is 45.4% [30.5;60.4].

4.4.2.3. Clinical impact of ADEs leading to VKA overdoses

In the test hospital, 232 cases of VKA overdoses (detected by an INR over 4.9) are detected using the ADE detection rules, over 21,737 stays. Two groups are defined:

- the 232 stays that have an INR over 4.9 within the filtered rules, that are called "potential ADE group"
- the 21,505 other stays, that are called "control group"

Both groups are compared with respect to several variables. Several differences are found and are displayed here. Comparisons of quantitative variables are computed by means of Student's tests. Comparison of binary variables are computed by means of Fisher's exact tests (as the proportion of ADE is low, Chi-square tests are not always feasible). Confidence intervals are provided with a 95% confidence.

A cause-to-effect relationship would require a complete case review that is not performed here.

Higher length of stay in potential ADE group

The average length of stay in control group is 7.93 days

The average length of stay in potential ADE group is 17.8 days

The confidence interval for the difference is [8.39;11.3] days. It is significantly different from 0 ($p < 2.2e-16$)

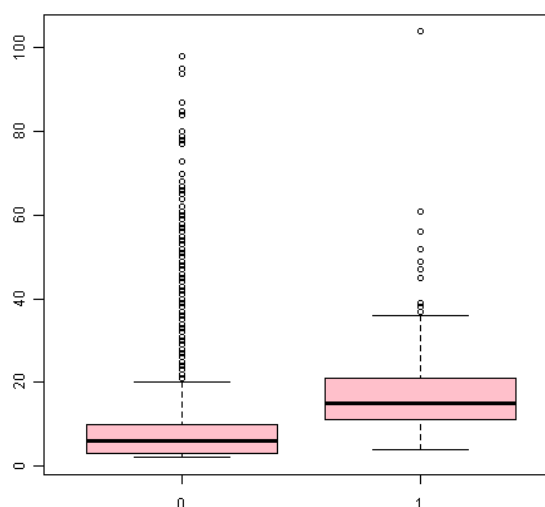


Figure 36. Length of stay in control group (left) and potential ADE group (right)

Higher death rate in potential ADE group

The proportion of death in control group is 3.28%.

The proportion of death in potential ADE group is 9.05%.

The odds ratio is 2.9 [1.77;4.63]. It is significantly different from 1 ($p = 3.74e-05$).

Higher multimorbidity rate in potential ADE group

The “multimorbidity” binary variable is set to 1 if the different principal diagnoses of the stay belong to several medical specialties; otherwise it is set to 0.

The proportion of multimorbidity in control group is 3.07%.

The proportion of multimorbidity in potential ADE group is 6.90%.

The odds ratio is 2.3 [1.30;3.91]. It is significantly different from 0 ($p=0.00323$).

Higher vitamin K administration rate in potential ADE group

This binary variable is set to 1 if vitamin K is administered in the absence of too high an INR, or if vitamin K is administered *quickly after* the occurrence of too high an INR; otherwise it is set to 0.

The proportion of vitamin K administration in control group is 1.06%.

The proportion of vitamin K administration in potential ADE group is 23.3%.

The odds ratio is 28 [19.9;39.7]. It is significantly different from 1 ($p < 2.2e-16$).

Lower anemia rate in potential ADE group

This binary variable is set to 1 if an anemia occurs in the absence of too high an INR, or if an anemia occurs *quickly after* the occurrence of too high an INR; otherwise it is set to 0.

The proportion of anemia in control group is 15.9%.

The proportion of anemia in potential ADE group is 7.76%.

The odds ratio is 0.44 [0.26;0.72]. It is significantly different from 1 ($p0.000379$).

Higher VKA discontinuation rate in potential ADE group

VKA are considered as “discontinued” after a given date if the VKA is administered before the given date and never administered after that date until the discharge. The binary variable “VKa discontinuation” is set to 1 if a VKA is discontinued before the discharge in the absence of too high an INR, or if VKA is discontinued before the discharge and *quickly after* the occurrence of too high an INR; otherwise it is set to 0.

The proportion of vitamin K administration in control group is 7.16%.

The proportion of vitamin K administration in potential ADE group is 32.8%.

The odds ratio is 6.3 [4.72;8.41] . It is significantly different from 1 ($p < 2.2e-16$).

4.4.3. All kinds of potential ADE

4.4.3.1. Rule-based detection

All the stays that match any of the validated rules are considered as “potential ADEs”. This groups together very various outcomes, from hemorrhage hazard to fungal infections. It is useful in order to get global statistics on ADEs.

Number of cases within validated rules:

Without filtering the rules, 3,333 cases of potential ADE have occurred. The incidence rate amounts to 9.40% [9.31;9.49] of the stays.

Filtering the rules, 1,431 cases of potential ADE have occurred. The incidence rate amounts to 4.04% [4.00;4.08] of the stays.

As the review is still in progress and has been first focused on VKA overdoses and cases of hyperkalemia, it is still not possible to provide global figures about the accuracy of the rules.

4.4.3.2. Clinical impact of potential ADEs (all together)

In the test hospital, 451 cases are detected by using the complete set of ADE detection rules, over 21,737 stays. Two groups are defined:

- the 451 stays for which at least one of the filtered rules fires, that are called “potential ADE group”
- the 21,286 other stays, that are called “control group”

Both groups are compared with respect to several variables. Several differences are found and displayed here. Comparisons of quantitative variables are computed by means of Student’s tests. Comparison of binary variables are computed by means of Fisher’s exact tests (as the proportion of ADE is low, Chi-square tests are not always feasible). Confidence intervals are provided with a 95% confidence.

Higher length of stay in potential ADE group

The average length of stay in control group is 7.83 days

The average length of stay in potential ADE group is 17.8 days

The confidence interval for the difference is [8.86;11.07] days. It is significantly different from 0 ($p < 2.2e-16$)

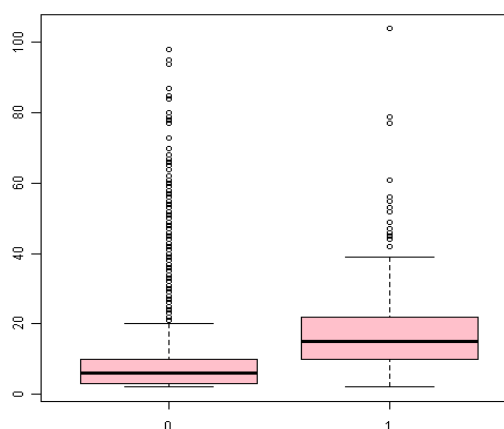


Figure 37. Length of stay in control group (left) and potential ADE group (right)

Higher death rate in potential ADE group

The proportion of death in control group is 3.07%.

The proportion of death in potential ADE group is 16.4%.

The odds ratio is 6.2 [4.71;8.08], which is significantly different from 1 ($p < 2.2e-16$).

Higher multimorbidity rate in potential ADE group

The “multimorbidity” variable is set to 1 if the different principal diagnoses of the stay belong to several medical specialties; otherwise it is set to 0.

The proportion of multimorbidity in control group is 2.99%.

The proportion of multimorbidity in potential ADE group is 9.09%.

The odds ratio is 3.2 [2.27;4.53], which is significantly different from 1 ($p = 9.89e-10$).

4.5. Presentation of the results: the Expert Explorer and the Scorecards

This section shows two web tools that are designed as part of this work. Both tools enable to display information to physicians in the medical departments. The Expert Explorer is a tool that enables to display all the available information about a given stay (hospitalization): diagnoses, lab results, drugs, etc. The Scorecards is a tool that displays the rules and the statistics that have been computed within the present work.

An overview of those tools is described, and completed by an ODP-compliant description that is provided in the appendix. The tools are presented through a use case that will give a more comprehensive view of how they can be used.

4.5.1. Description of the Expert Explorer

The Expert Explorer is a web tool that enables to display all the details of a given stay in an easy-to-use navigation interface. The data are loaded by using the same data model as defined in section 2.2 (*Definition of a common data model*) on page 50. Once logged in, everyone can simply visualize all the available information about a given stay.

The Expert Explorer has been totally specified as part of the present work. It has been implemented by IDEEA Advertising, a Romanian software engineering company that is involved in the PSIP Project.

A complete ODP-compliant description of the tool is available in chapter 10 (*Appendix 2: ODP description of the Expert Explorer*) on page 163. However, the section 4.5.3 (*Use case example of the web tools for ADE discovery in databases*) on page 118 shows a complete use case that helps to understand more easily the main features of the tool.

4.5.2. Description of the Scorecards

The Scorecards are a web tool that enables to detect ADEs in the past stays of a given medical department or hospital. The tool is connected to the Expert Explorer described above. The tool is directly fed with the results of the present work, in the XML format defined in chapter 12 (*Appendix 4: Description of the output of this work (use of the XML files)*) on page 201. Once logged in, the user can see the potential ADEs that are detected by the data mining procedure. In addition, the rules that help to detect several cases in the medical department or hospital are displayed, as well as information about ADEs. Finally, it is possible to visualize the real-data cases that match the rules by means of the Expert Explorer. This tool is very convenient for physicians because it enables them to learn which ADEs may occur in their department, and to be informed about probabilities.

The Scorecards have been totally specified as part of the present work. In addition, this visualization tool is fed by the XML files produced in the present work. It has been implemented by IDEEA Advertising too.

A complete ODP-compliant description of the tool is available in chapter 11 (*Appendix 3: ODP description of the Scorecards*) on page 186. However, the section 4.5.3 (*Use case example of the web tools for ADE discovery in databases*) on page 118 shows a complete use case that helps to understand more easily the main features of the tool.

4.5.3. Use case example of the web tools for ADE discovery in databases

This section shows a sequence of commented screenshots that correspond to the following possible scenarios:

“A physician working in a hospital, from which the web tools are available, uses those tools for 3 purposes.

Firstly, he wants to have a comprehensive overview of the ADEs that have been detected in his medical department during the last 6 months (scenario 1).

Secondly, he wants to explore one of those probable ADE cases to form his own opinion (scenario 2).

Thirdly, he has to participate in the review of some cases (scenario 3).”

This use case is developed in the next three sections.

4.5.3.1. Scenario 1: Comprehensive overview of potential ADEs in a given medical department - the Scorecards

As the Scorecards are a web tool, the user just has to use a computer connected to the web and equipped with a web browser. He first logs in to the Scorecards (Figure 38).



Figure 38. Login page of the Scorecards

Then, he has access to the synthesis page (Figure 39). As the tool is multilingual, the user is able to get it in French, English or Danish using the language toolbox on the top of the page.

The synthesis page (Figure 39) consists of 3 zones:

1. The big blue table displays the number of detected ADEs month per month. Each line of the table is a kind of ADE; each column is a month of the current year.
2. The line chart displays the same information as the table using a chart
3. In the yellow zone, the user can change the period in order to explore the years 2007, 2008 or 2009 instead of 2010. He is also able to choose some kinds of ADEs and validate the form in order to generate the scorecards per kind of ADE.

In this example, the user checks 5 different kinds of ADEs that are of interest for him and he validates the form.

Synthesis

Number of stays with adverse events

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
<input checked="" type="checkbox"/> VKA overdose (INR>4.9)	5	6	3	6	3	6	4					
<input checked="" type="checkbox"/> VKA overdose (detected by a prescription of vitamin K)	2	0	3	2	2	5	3					
<input checked="" type="checkbox"/> Anemia (Hb<10g/dl)	1	2	4	3	4	3	1					
<input checked="" type="checkbox"/> Bacterial infection (detected by prescription of antibiotic)	6	6	2	3	5	3	1					
<input checked="" type="checkbox"/> Diarrhoea (detected by prescription of an anti-diarrhoeal)	1	0	1	1	0	0	0					
<input checked="" type="checkbox"/> Diarrhoea (detected by prescription of an antipropulsive)	0	0	1	1	0	0	0					
<input checked="" type="checkbox"/> Renal failure (creat>135 micromol/L or urea>16.6 mmol/L)	9	10	12	10	18	13	4					
<input checked="" type="checkbox"/> Thrombocytosis (count>600,000)	2	4	1	3	2	2	1					
<input checked="" type="checkbox"/> Thrombopenia (count<75,000)	5	3	4	5	3	4	2					

Number of detected cases by effect and by month

Edit detailed statistics

Analysis period: jan-jul 2010

Detected effects:

- VKA overdose (INR>4.9) (33)
- VKA overdose (detected by a prescription of vitamin K) (17)
- anemia (Hb<10g/dl) (18)
- bacterial infection (detected by prescription of antibiotic) (26)
- diarrhoea (detected by prescription of an anti-diarrhoeal) (3)
- diarrhoea (detected by prescription of an antipropulsive) (2)
- fungal infection (detected by prescription of a systemic antifungal) (15)
- fungal infection (detected by the prescription of local antifungal) (13)
- hemorrhage (detected by a prescription of hemostatic) (25)
- heparin overdose (APTT>1.23) (2)
- hepatic cholestasis (alkal. phosphatase>240 UM or bilirubins>22 μmol/L) (22)
- hepatic cytolysis (alanine transa.>110 or aspartate transa.>110) (9)
- high a CPK rate (CPK>195 UM) (5)
- hypereosinophilia (eosinophilia>10⁹/l) (7)
- hyperkalemia (K<+>5.3) (65)
- hypocalcemia (ca<+>2.2 mmol/L) (7)
- hypokalemia (K<+>3) (2)
- hyponatremia (Na<+>130) (4)
- increase of pancreatic enzymes (amylase>90 UM or lipase>90 UM) (5)
- neutropenia (count<1500/mm³) (1)
- renal failure (creat>135 micromol/L or urea>16.6 mmol/L) (76)
- thrombocytosis (count>600,000) (15)
- thrombopenia (count<75,000) (26)

Generate Scorecards

Figure 39. Synthesis page of the Scorecards

Once the user has validated the form, a new page (Figure 40) displays 5 links to the 5 different scorecards he has asked for.

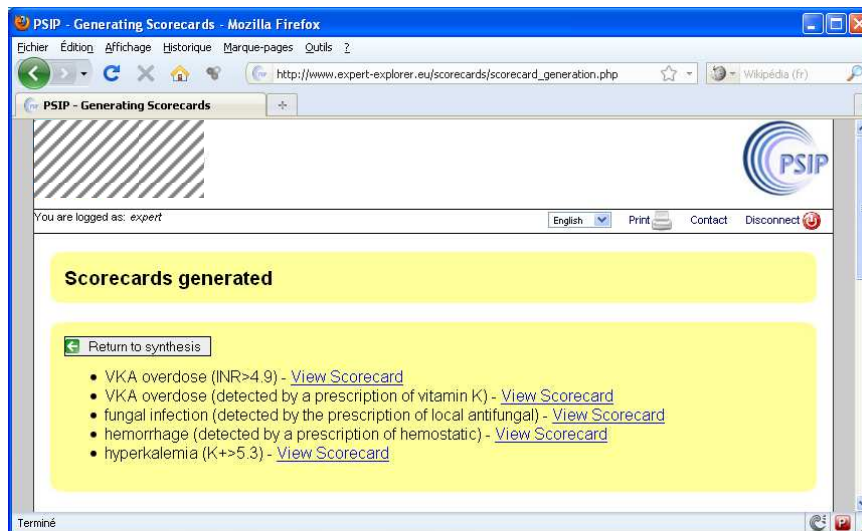


Figure 40. List of generated scorecards

In this use case, the user focuses on the latest outcome, that describes cases of hyperkalemia ($K^+ > 5.3$): this kind of abnormal lab value endangers the patient, as it could lead to lethal cardiac rhythm troubles. The user clicks (Figure 40) on the hypertext link “View Scorecard” to reach the page. The page opens (Figure 41).

The complete scorecard is displayed (Figure 41). It is conceived to be either explored on the screen or printed on paper. The page contains 4 zones:

1. At the top of the page, the user can read the period, the place, and the outcome that is traced
2. In the yellow area, descriptive statistics are computed. They describe all the stays that have been detected within all the rules.
3. In the blue zone, all the rules that enable to detect potential ADE cases in the current department are displayed. For instance, the user can read that Low Molecular Weight Heparins (LMWH) can induce hyperkalemia especially for patients suffering from renal insufficiency (rule N°1). In the current department, 17% of patients with LMWH and renal failure encountered a hyperkalemia in a median delay of 4.5 days (2 cases).
4. At the bottom of the page, more detailed explanations are provided for each rule. They can be reached by clicking on the internal hypertext links placed on the number of each rule.

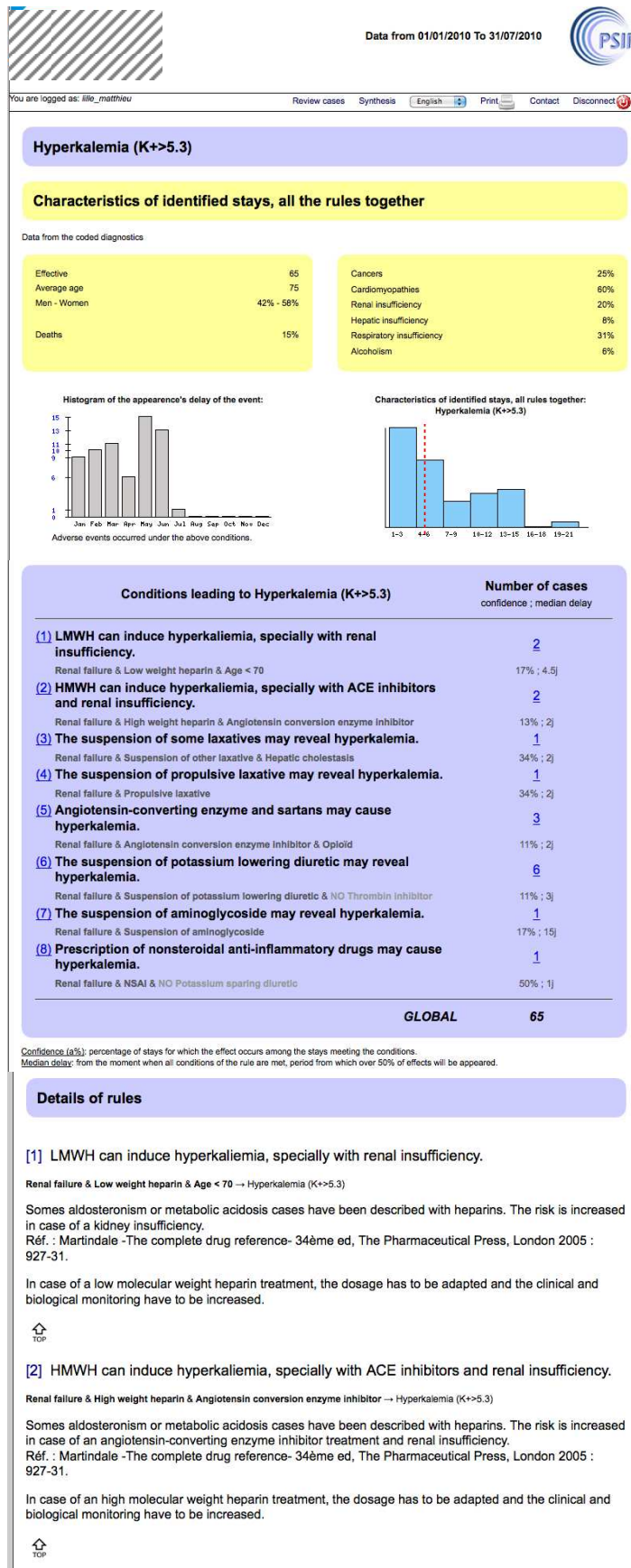


Figure 41. Scorecard of hyperkalemia (K⁺>5.3)

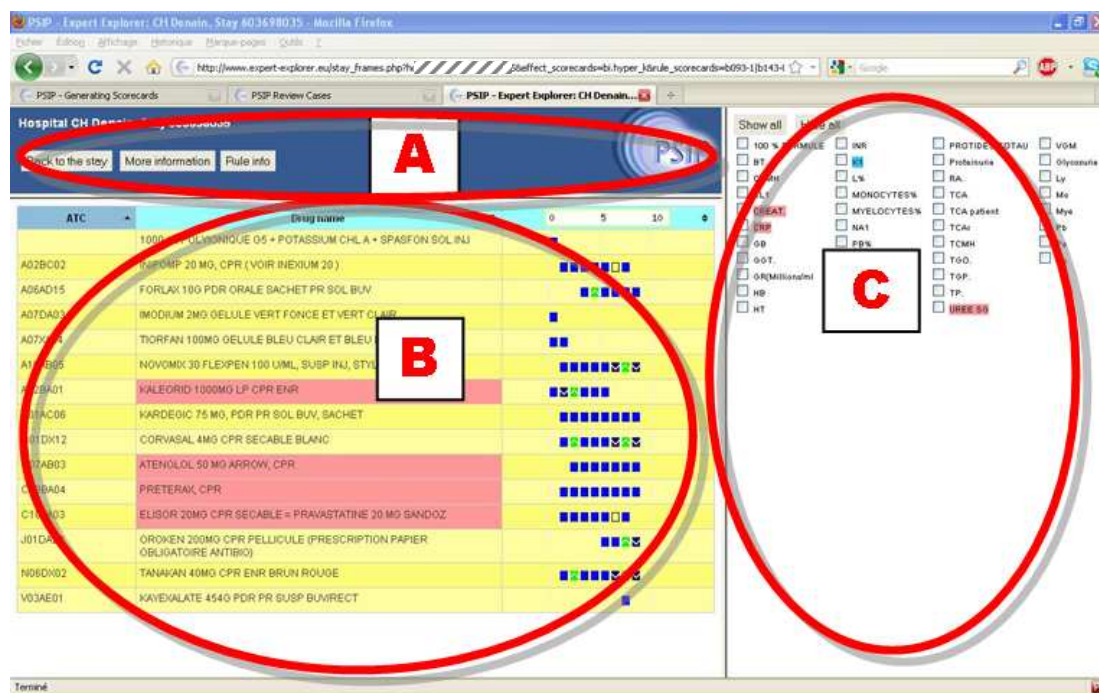


Figure 44. The 3 zones of the Expert Explorer main page

The screen is made up of 3 main parts (Figure 44 & Figure 43):

- A: the header contains several buttons that will be described later
- B: the drug panel helps reviewing all the drugs that were administered to the patient
- C: the lab panel helps reviewing all the laboratory results

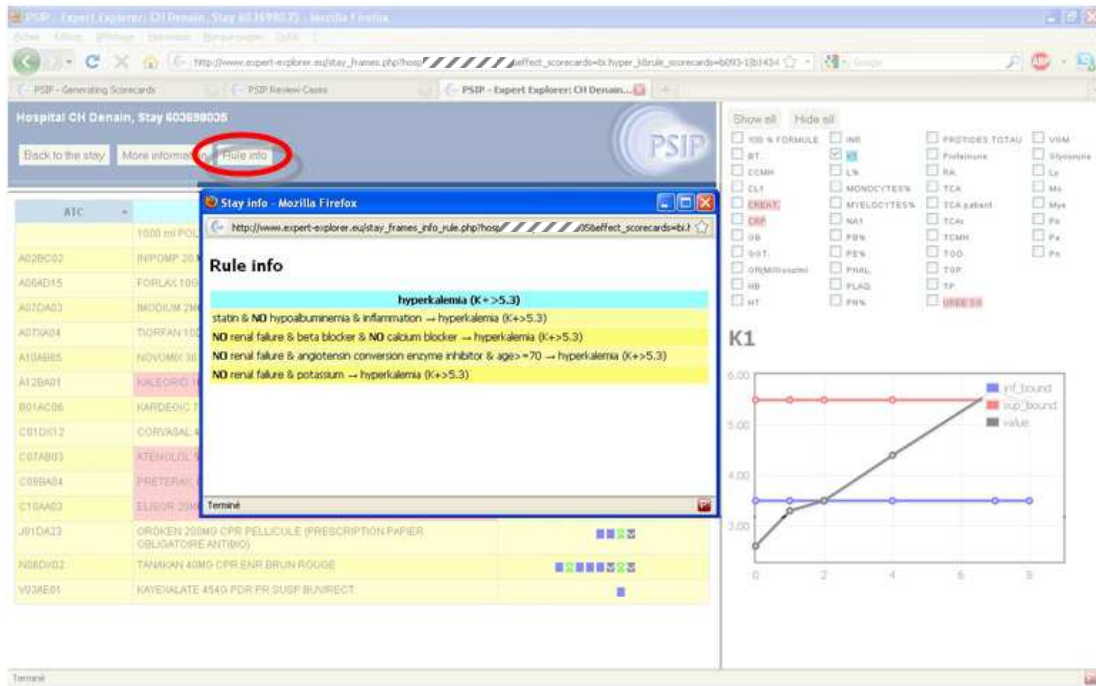


Figure 46. When required, a popup displays information about the rules that fire on the current stay

If the user wants to see the rules that fire for that stay, he just has to click on the button “rules info” in the head panel. A popup appears as displayed in Figure 46. In the present case, according to the rules, the drugs involved are statins, beta blockers, angiotensin conversion enzyme inhibitor, and potassium.

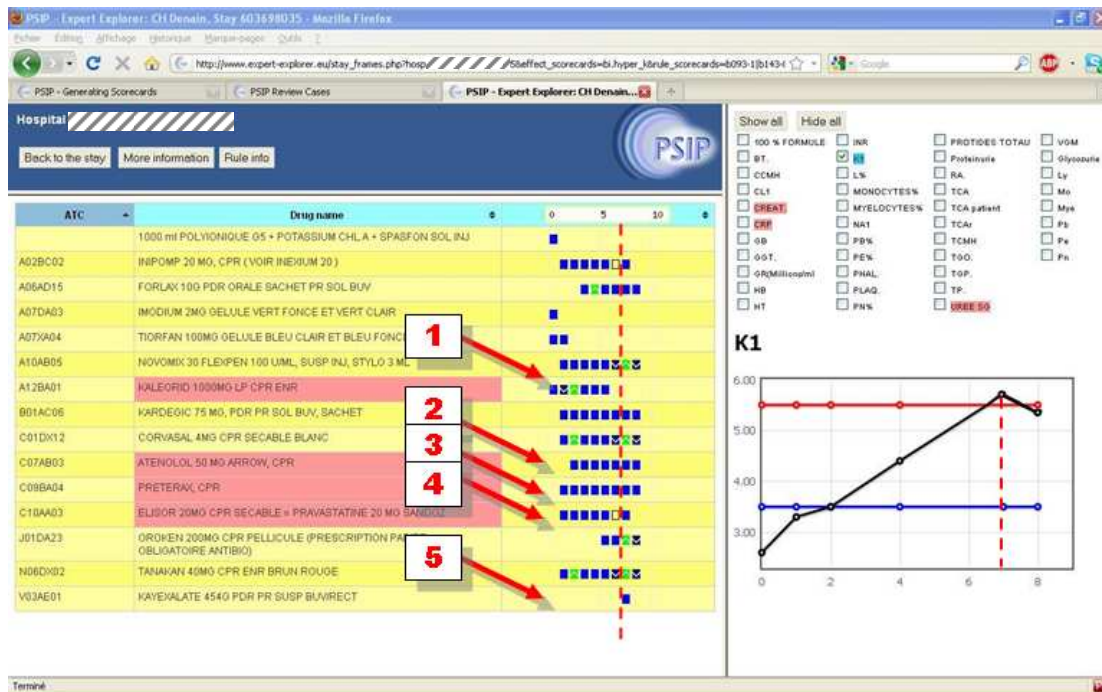


Figure 47. Analysis of the administered drugs

On the drug panel (Figure 47 & Figure 43), the user can now review the drugs. The drugs that correspond to the various rules appear on a pink background so that they are easy to localize (labels 1-4). The user can check that the potassium (label 1), the beta blocker (label 2), the association of the angiotensin conversion enzyme inhibitor and potassium sparing diuretic (label 3) and the statin (label 4) were administered before day 7, the date of the outcome. All those drugs are known to increase the potassium blood level. On Figure 47, two red dotted lines have been manually added for didactic purposes; they show the seventh day in both lab chart and drug chart.

In the present example, the user can also notice the reactions of the physicians. Hopefully the potassium is suspended before the hyperkalemia occurs (label 1). But as the potassium level reaches a very high level, a potassium lowering drug is administered during the seventh day (label 5).

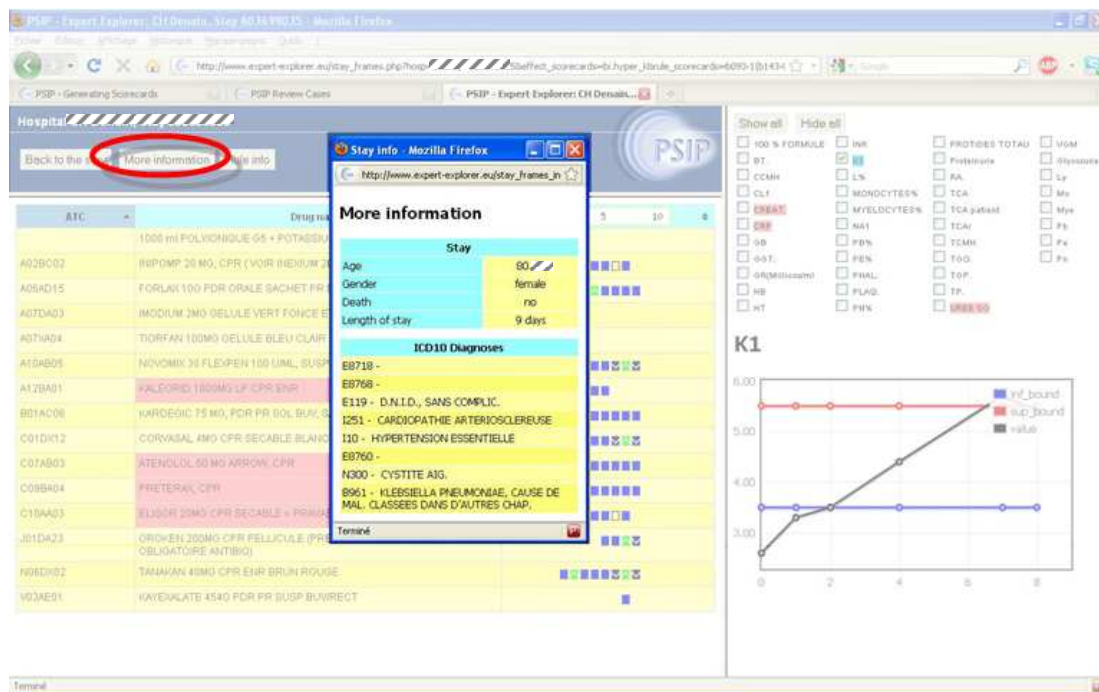


Figure 48. The "more information" popup

The user can also access additional information by clicking on the “more information” button of the head panel. A popup appears as displayed on Figure 48. It enables to review the ICD10 diagnoses that have been encoded by the physicians. In the present case, the hypokalemia is encoded, but not the hyperkalemia.

Finally, the tool also enables the user to read the anonymized letters and reports. This feature is not shown here and is detailed in the ODP description of the tool in chapter 10. In that precise case, the hypokalemia is mentioned in the report but not the hyperkalemia. The physician concludes “woman admitted for a hypokalemia in relation to a gastro-enteritis (...) after correction, the potassium level is normal (...)”.

4.5.3.3. Scenario 3: Review of a stay

Finally, some users are also designated as Experts and are allowed to give their advice on a given stay to validate it as a *real* case of ADE. If the user is an expert, he can review a stay for which one or several rules fired. He can then fill in a form in order to validate the ADE status of the stay, and to quote the cause-to-effect relationship described by the rules. This information will be used in order to assess the *real* positive predictive value of the set of rules, this is very important for the assessment of the system.

The forms are not detailed here. Only the main review page is displayed (Figure 49).

For each stay, the following features are available:

- A picture showing whether the stay has already been reviewed or not (top, middle)
- A link to the Expert Explorer for the present stay (top, left)
- A link to the forms that are used to validate the stay (top, right)

- Then, the list of the rules that fire for the present stay is provided. Beside each rule there is a link to the details of the rules.

Indeed, for a given stay, several rules might fire and predict the same outcome at the same time.

[View the stay](#) // // // 955 - Reviewed by an other user [Assess this stay](#)

[b060-1](#) NO Renal failure & Low weight heparin & Age ≥ 70 → Hyperkalemia (K+>5.3) (3%)
LMWH can induce hyperkalemia.

[View the stay](#) // // // 203 - Reviewed by an other user [Assess this stay](#)

[b139-0](#) NO Renal failure & Suspension of other laxative → Hyperkalemia (K+>5.3) (4%)
The suspension of some laxatives may reveal hyperkalemia.

[b149-0](#) NO Renal failure & Potassium → Hyperkalemia (K+>5.3) (3%)
Prescription of potassium may cause hyperkalemia.

[View the stay](#) // // // 035 - Reviewed by an other user [Assess this stay](#)

[b093-1](#) Proton pump inhibitor → Hyperkalemia (K+>5.3) (1%)

[b143-0](#) NO Renal failure & Beta blocker & NO Calcium blocker → Hyperkalemia (K+>5.3) (1%)
Beta blocker may cause hyperkalemia.

[b146-1](#) NO Renal failure & Angiotensin conversion enzyme inhibitor & Age ≥ 70 → Hyperkalemia (K+>5.3) (1%)
Angiotensin-converting enzyme and sartans may cause hyperkalemia.

[b149-0](#) NO Renal failure & Potassium → Hyperkalemia (K+>5.3) (3%)
Prescription of potassium may cause hyperkalemia.

Figure 49. List of the potential ADE cases to review

5. DISCUSSION

5.1. Contribution of the present work to ADE detection

Summary

In this work, a common data model is first defined (see section 2.1 on page 49). More than 90,000 complete stays are loaded into a repository that fits the data model. Those complete records include diagnoses, lab results, drug administrations, administrative and demographic data as well as free-text reports. When the drugs are not available from any CPOE, they are extracted from the free-text reports by means of semantic mining (see section 2.4 on page 59). Then, ADE detection rules are discovered by means of data mining, in particular decision trees and association rules (see section 3.4 on page 74). Those rules are completed by using some ADE detection rules extracted from the SPCs. All in all, 236 rules are described: they enable to trace 27 different outcomes (see section 4.1 on page 95). All those rules are described by using a common formalism and are loaded into a common rule repository (see section 3.5 on page 86). There, all the 236 rules are automatically assessed in every hospital and every medical department. Several statistics are automatically computed, such as the confidence. In addition, two web tools are designed in order to show epidemiological information about the potential ADE cases and to explore those cases (see section 4.5 on page 117). Finally, a preliminary evaluation of the clinical impact of the potential ADEs is performed as well as a preliminary evaluation of the accuracy of the ADE detection (see section 4.4 on page 111).

Data sources

Many other scientific papers deal with rule-based detection of adverse drug events. Those papers concern prospective ADE prevention within a CDSS [Paterno 2009, Jha 2008] or retrospective ADE detection in past stays [Honigman 2001, Kaushal 2003]. The fact that the rules are used for ADE detection or ADE prevention doesn't have any impact on the formalism or the content of the rules. The present work deals with retrospective ADE detection, but also proposes to use the same detection rules in a CDSS for ADE prevention: the rules are versatile.

As in the present work, the various contributions use the same information source for ADE detection:

- data from the EHRs: diagnostic codes (ICD10 or ICD9), drug allergies, patients' characteristics (demographic data, pregnancy), and laboratory results [Kaushal 2003],
- data from the CPOEs: drugs with or without dose [Kaushal 2003], and
- free-text reports [Honigman 2001].

Formalism and content of the rules

In all the papers that have been reviewed, each rule consists of one or two conditions that lead to an effect. Those two conditions can be of different types but, as diagnostic codes and free-text are not available when a CDSS is used, they are not involved in

the rules usable for ADE prevention. The conditions used in the ADE detection rules can be of different types:

- a drug and another drug [Bates 1994, Jha 1998, Del Fiol 2000, Gandhi 2005, Judge 2006, Paterno 2009, Schedelbauer 2009, Teich 1999]
- a drug and a lab result [Bates 1994, Kuperman 1996, Jha 1998, Honigman 2001, Field 2004, Morimoto 2004, Schedelbauer 2009]
- a drug alone with its dose [Bates 1994, Morimoto 2004, Gandhi 2005, Judge 2006]
- a drug and a patient characteristic [Bates 1994, Morimoto 2004]
- a drug with its dose and a patient characteristic [Schedelbauer 2009]
- a drug and a drug allergy [Bates 1994, Honigman 2001, Schedelbauer 2009]
- a drug and a chronic disease [Schedelbauer 2009]

The present work uses nearly all the previous kinds of conditions combined together. Only drug doses and drug allergies are not taken into account. In the PSIP project, those features are supposed to be already implemented in the CPOEs instead of being supported by the CDSS itself. In addition, the present work considers drug discontinuations as conditions of an ADE; this feature is not described in other researches.

In several works, a traceable outcome is available. In [Handler 2007], Handler et al. perform a systematic review of 12 studies describing 36 unique ADE signals. Over the 36 different signals that are detected, 7 are administrations of antidotes, and 19 are abnormal laboratory test results; 10 are suprathereapeutic medication levels which formally peaking are not an outcome but a suspicious circumstance. In the present work, an outcome is systematically traced and can be of abnormal laboratory results (38 kinds of outcomes, from which 19 are involved in ADE detection rules), or drug prescriptions (18 kinds of outcomes, from which 7 are involved in ADE detection rules). As the doses are not used, we are not able to detect supreatherapeutic medication levels.

Finally, despite a low number of conditions involved in the rules, some projects use segmentation conditions, i.e. conditions that do not explain the outcome but modulate its confidence. Those conditions are the age, the renal function, the hepatic function and the patient's weight [Bates 1994, Schedelbauer 2009]. In the present work, such conditions exist but are not systematic: they are used only if the Data Mining shows that they significantly modulate the confidence of the rules. Unfortunately, the patient's weight is not used as it is not sufficiently documented in the data. But the method used here would enable to take the weight into account in another dataset.

Schedlbauer et al. describe a typology of ADE prevention rules [Schedlbauer 2009]. According to the authors, the rules can be classified as follows (Table 27).

Table 27. Categories of drug alerts [Shedlbauer 2009]

<ul style="list-style-type: none">□ BA. Basic Drug Alerts <i>Provision of basic clinical decision support. Clinical information systems should first have basic alerting systems in place before moving on to more enhanced alerting features.</i><ul style="list-style-type: none">○ BA1. Drug allergy warnings <i>Alert is generated in orders of medication to which the patient has an electronically</i>

documented allergy

- **BA2. Drug-drug interactions**
Alert is generated when the mode of action of one drug is known to be affected by the simultaneous prescribing of another drug.
- **BA3. Duplicate medication or therapeutic duplication alerts**
Alert is generated when the patient is already receiving the medication just ordered or a different drug in the same therapeutic category.
- **BA4. Basic medication order guidance**
Alerts providing dosing strings with default dosing being the most appropriate initial dosing

- **AA. Advanced Drug Alerts**
Provision of enhanced clinical decision support in CPOE system.
 - **AA1. Drug-lab alerts**
Alerts are generated when administration of drug requires close monitoring of laboratory parameters before or/and after administration.
 - **AA2. Drug-condition alerts**
Raise awareness of specific prescribing for certain conditions.
 - **AA2.1 Drug-disease contraindication alerts**
alerts are generated to warn against prescribing of a certain drug in a specific disease.
 - **AA2.2 Drug-condition alerts aiming at appropriate prescribing**
alerts are generated to encourage prescribing of a certain drug in a specific disease or condition.
 - **AA2.3 Drug-age alerts**
Alerts are generated to discourage prescribing of a certain drug in the elderly.
 - **AA3. Drug-formulary alerts**
alert provided when a particular brand or drug is not included or not recommended in the formulary of the prescribing location.
 - **AA4. Dosing guidelines**
Advanced medication dosing alerts that take into account complex patient characteristics such as age, weight, height, renal function, liver function, and fluid status; co-morbidities, other medications the patient may be currently taking and indication for the drug.
 - **AA4.1 Dosing guidelines based on renal function**
 - **AA4.2 Dosing guidelines based on age**
 - **AA4.3 Dosing guidelines based on pregnancy/female of childbearing potential**
 - **AA4.4 Dosing guidelines based on pediatric patients/weight based dosing**
 - **AA4.5 Dosing guidelines based on drug utilization restriction**
 - **AA4.6 Dosing guidelines based on indications**
 - **AA5. Complex prescribing alerts**
Combined features of basic and advanced alerts

Rules of type *BA1* describe drug-allergy warnings. This feature is very important, but such rules can probably not be found by using data mining as those events are never

supposed to occur. This is typically the reason why such rules must be imported from an SPC knowledge base such as provided by the Vidal Company. However, this feature is not so simple to implement, as the drug allergies are not precisely described in the ICD10 classification. As a consequence, in many EHRs, drug allergies are described by using free text and are difficult to take into account automatically.

Rules of type *BA2* are generated in the present work.

Rules of type *BA3* and *BA4* are out of the scope of the present work, but are mandatory in any CPOE. As part of the PSIP Project, as this kind of alert is quite classical, it was decided that it was taken care of by the CPOE itself and not by the CDSS developed in the project.

Rules of type *AA1* and *AA2* are typically what is done in the present work. In addition to all the *AA2.x* types, the present work adds a significant contribution by identifying other kinds of contexts such as laboratory results, drug discontinuations, and potentially administrative or organizational conditions.

Rules of type *AA3* are very dependent on the local cultures. Such rules are not provided by the present work and probably mainly depend on the scientific and economic policies of the managers of a hospital.

Dosing guidelines -as described in the rules of type *AA4*- are not directly provided in the present work. For technical reasons, the doses are not taken into account in the data mining step. However, most of our VKA-related rules tend to provide dose-adaptation recommendations.

Finally, most of the rules we provide are “complex prescribing alerts”, which correspond to the *AA5* type.

Origin of the rules

In many papers that deal with ADE detection or prevention, the rules have been written by experts.

In quite numerous studies, data mining is used to analyze ADE reports (post-marketing data) [Almenoff 2005, Almenoff 2007, Bate 2006, Bennet 2007, Coulter 2001, Hauben 2005]. As opposed to the present work, those approaches are based on classical supervised rule induction. To our mind, the interest of such an approach is limited:

- It is currently admitted that less than 5% of the ADEs are reported. As no data is available to know whether the 95% other reports are missing at random or not, it is difficult to conclude about the representativeness of such reports.
- The statistical associations that are discovered provide some results such as “knowing that an ADE *Y* has been reported, the drug *X* has a probability of $P(X|Y)$ to be responsible for the ADE”. But the question that is of real interest to the physicians is another one: “Knowing that the drug *X* is prescribed, what is the probability $P(Y|X)$ that the ADE *Y* occurs?”

If such studies might bring interesting knowledge, this knowledge can’t directly be used for ADE detection or ADE prevention.

The interest of data mining of the EHRs is discussed in some papers [Berlin 2008, Cerrito 2001] but the “data mining” term often conceals classical multivariate methods such as logistic regressions, using a limited number of explicative factors that are preselected using expert knowledge [Ingram 2008].

In the present work, we mine hundreds of variables extracted from the EHRs, taking time into account. Data-mining-induced rules and rules extracted from the SPCs are pooled together. 17% of the rules only come from the SPCs, and 72% of the rules only come from the data mining.

Contextualization of the rules

In this work, all the rules are automatically tested in every hospital and medical department in order to compute statistics such as the confidence. Those statistics vary with respect to the site. As a consequence, we recommend filtering the rules using thresholds applied to the local statistics. In the literature, it seems that the behaviors of the CDSSs are never contextualized. Most of the time, no statistic is computed and when statistics are computed, only one site is used and no recommendation is made to filter the rules: the rules are supposed to have the same interest everywhere.

Meta-rules and over-alerting prevention

In the literature, meta-rules are not explicitly described. Kuperman et al. introduce a feature that is very close to meta-rules [Kuperman 1994]: for instance a patient receiving both potassium and potassium sparing diuretics would only have the interaction reported if one of the three most recent serum potassium measurements is above 4.7 or if the serum potassium has not been measured in the last 3 days. This kind of feature is interesting as it may reduce over-alerting.

In most papers, this kind of feature is not possible, as the outcome is described in free text but cannot be automatically traced, as in the Vidal's rules describing contraindications. Ignoring the appearance of the outcome is the most classical approach in the CDSSs, but it induces over-alerting: it is not possible to know whether the outcome is traced or not, is to occur or has already occurred.

Meta-rules for the prospective implementation are proposed later in the present discussion.

Chronology consideration

In the literature, the time constraints seem not to be taken into account. It is not a critical issue in the case of prospective implementation (CDSS). However, even for a prospective use when the outcome is traced, there is no certitude that the outcome occurred after the drugs were administered. In the third meta-rule we propose in the present discussion, the problem is solved: we distinguish the case when the conditions may be responsible from the outcome and the case when the conditions may worsen the outcome. In addition, in the present work we empirically used a 5-day delay: if a drug is discontinued on day d , then it can still be involved in an outcome that occurs till day $d+5$. In further works, this 5-day delay should be replaced by using the half-life of the drug.

Positive predictive values of the set of rules

There are two different ways to compute the positive predictive value of a set of rules (Figure 50).

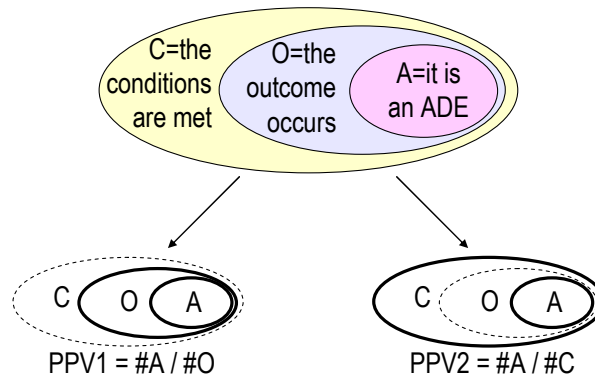


Figure 50. Two different ways to compute the positive predictive value of a set of rules

The first way is to consider the probability that a case is a real ADE, knowing that the case matches all the conditions of at least one rule and leads to the outcome. This corresponds to the use of rules for ADE detection in past records. That figure is provided in many articles, we'll call it "PPV1" (left part of Figure 50).

The second way is to consider the probability that a real ADE occurs knowing that the conditions are met. In that case, we don't know if the outcome occurs or not. This figure would be useful to qualify the warnings of a CDSS that does not wait until the outcome occurs to alert. That figure is provided in a few articles, we'll call it "PPV2" (right part of Figure 50). It should necessarily be lower than PPV1. If there were only one rule, PPV2 would result from the multiplication of PPV1 and the confidence of the rule $P(outcome/conditions) = \#O/\#C$.

Values of PPV1:

- In Kuperman et al.'s work, PPV1 = 5% [Kuperman 1994].
- In Jha et al.'s work, PPV1 = 10.5% [Jha 1998] and PPV1 = 1.6% [Jha 2008].
- In the review performed by Handler et al. [Handler 2007], several outcomes are considered. Cases of hyperkalemia are traced in 5 different studies and lead to PPV1 comprised between 0 and 67%. Cases of INR increase are traced in 4 different studies and lead to PPV1 comprised between 5% and 100%. Cases of Vitamin K administration are traced in 3 different studies and lead to PPV1 comprised between 2% and 30%.
- In the present work, for rules enabling to detect VKA overdoses we get PPV1 = 45.4% [30.5;60.4] (see section 4.4.2.2 on page 114) and for rules enabling to detect cases of hyperkalemia we get PPV1 = 75.86% [69.0;82.7] (see section 4.4.1.2 on page 112). Those very good results are due to the fact that the rules are well segmented. In addition, in those modules we expected to get good results. Other more complex modules were not included in the review.

Values of PPV2:

- In Field et al.'s work, PPV2 = 6.4% [Field 2004].
- In Honigman et al.'s work, PPV2 = 3.3% for drug-lab alerts [Honigman 2001].

Incidence rate of ADEs

According to the literature, ADEs are common and occur in 2.4 to 5.2 per 100 hospitalized adult patients [Bates 1994, Bates 1995, Classen 1997, Nebeker 2005,

Senst 2001]. In [Jha 1998], 2.8 ADEs occur for 100 patients*days, this could correspond to 5-10% of the stays.

In the present work, we identify 4% of potential ADEs in the datasets (see section 4.4.3.1 on page 116). One can assume that on the one hand only half of those potential ADEs are real cases of ADEs (the accuracy is 45% for VKA overdoses and 76% for cases of hyperkalemia). On the other hand, we are missing several purely clinical outcomes (such as rashes). All in all, those results are compatible with the commonly admitted figure of 5% of ADEs during hospitalizations. However, only a complete review of the cases and the stays that do not match any rule could enable to provide an ADE frequency.

Clinical impact of ADEs

The clinical impact of the ADEs is not systematically measured in the literature. Though, the clinical impact is a part of the definition of ADEs. In [Kuperman 1994], 3% of the detected interactions with outcome have led to a change in drug prescriptions. In [Bates 1994], 64% of the ADEs are considered as severe. Petersen et al. [Petersen 1992] define severe ADEs as “ADEs resulting in death, at least 1 month of disability or a minimum of 4 added hospitalization days”. According to [Bates 1997], each ADE is estimated to increase the length of hospital stay by 2.2 days and to increase the hospital cost by \$3,244.15.

In the present work, several statistics are computed. The interpretation of the following figures must be done very carefully: firstly the figures concern *potential ADE cases* and not *validated ADE cases*, secondly a cause-to-effect relationship cannot be established without a complete case review.

The patients who are qualified as *hyperkalemia as part of a potential ADE* have a 9.8 day higher length of stay, and the death rate is 8.7 times higher (see section 4.4.1.3 on page 112).

The patients who are qualified as *VKA overdose as part of a potential ADE* have a 9.9 day higher length of stay, and the death rate is 2.7 times higher (see section 4.4.2.3 on page 114). In addition, the outcome leads to a prescription of vitamin K in 23.3% cases and to a definitive VKA discontinuation in 32.8% of cases.

The patients who are qualified as *potential ADE (whatever the outcome)* have a 10.0 day higher length of stay, and the death rate is 5.3 times higher (see section 4.4.3.2 on page 116).

According to those figures, the potential ADEs would have an important clinical impact. But it cannot be excluded that the patients have ADEs because of a worse health status, which could explain the higher mortality rate. However, diagnoses of severe diseases didn't appear as segmentation conditions in the decision trees. Similarly, the longer the stay is, the higher the probability of detecting outcomes is: that could explain the higher length of stay. On the other hand, specific impact measures such as “VKAs discontinuation” or “vitamin K administration” seem to be quite reliable.

The grand challenges of clinical decision support system

In [Sittig 2008], Sittig et al identify the grand challenges of clinical decision support system. According to the authors, the CDSS of the future will have to address several issues that are listed in Table 28.

Table 28. Summary of the grand challenges of clinical decision support system [Sittig 2008]

#	Grand Challenge Description
1	Improve the human-computer interface
2	Disseminate best practices in CDS design, development, and implementation
3	Summarize patient-level information
4	Prioritize and filter recommendations to the user
5	Create an architecture for sharing executable CDS modules and services
6	Combine recommendations for patients with co-morbidities
7	Prioritize CDS content development and implementation
8	Create internet-accessible clinical decision support repositories
9	Use free-text information to drive clinical decision support
10	Mine large clinical databases to create new CDS

The present work and, more generally speaking, the PSIP Project, answer or try to answer several points.

Points #1 and #2 are well defined in the objectives of the PSIP Project. Ergonomists, psychologists and human-factor engineers involved in the project are aware of those challenges and are currently trying to bring innovative answers to those issues.

Point #3 highlights that most of the time only drug-prescription-related conditions are used in the CDSSs. Concretely, all the CDSSs are rule-based, and very often the rules only take into account drug-related conditions. The present work is innovative as discussed before, as the data-mining step enables to integrate diseases, lab results, drug-related information, demographic information (sex, gender...) and administrative description of the stays (admission season or day of week, admission by emergency or ICU, etc.). Theoretically, in the present work we would have been able to discover even organizational causes. Unfortunately, the 236 rules we have discovered “only” take into account the age, chronic diseases, lab results, drug prescriptions and drug discontinuations, because according to statistical results the other conditions don’t appear to be involved in the appearance of the outcomes.

Point #4 highlights that in many existing CDSSs the alerts are not sorted nor filtered. The CDSS that is under development in the PSIP Project enables to sort and filter the alerts. Those features are directly based on another great contribution of the present work: we are able to automatically compute several statistics that can be used to sort and filter the rules, such as the confidence, the relative risk and the Fisher’s p value. In addition, what is not described by Sittig et al., is the possibility that the sorting and filtering of the rules vary with respect to the place where the CDSS is used: this is contextualization. That feature is also made possible by the present work, as all the statistics are computed separately in each hospital and in each medical unit. Contrary to the accepted wisdom, we demonstrated that the confidence of the rules is deeply dependent on the place where they are used. As a consequence, defining a contextualized CDSS seems to be a good way to increase the accuracy of the alerts.

Point #5 is addressed by this work, as we have defined and proposed a common data model, and a sharable set of XML files to describe the behavior of the CDSS through a set of rules and related mappings. In addition, in the PSIP Project, the CDSS is not a standalone application but is reachable through a Connectivity Platform. As a

consequence, it is very easy to load the same knowledge into a new CDSS, or to connect any new tool to the existing PSIP CDSS through the connectivity platform.

Point #6 is only a focus on a specific kind of variables, and can be generalized through point #3. This issue is fully addressed by the present work. However, it is interesting to notice that, according to data mining results, the laboratory sign of co-morbidities seems more reliable than the ICD10 encoding of those morbidities.

Point #7: this challenge is commonly shared by the PSIP project.

Point #8 is now possible as the behavior of a CDSS can be described by using the XML files that are provided by the present work. Intellectual property issues have to be solved first.

Point #9 is partially addressed in the present work. Indeed, free-text records are used for the retrospective data mining that enables to discover past ADE cases. But till now free texts have not been used to discover situations at risk of ADE within the daily drug prescriptions. However it would be an interesting feature, as when a physician prescribes a drug, the diagnoses of the patients are still not available. But on the other hand, in the present work the diagnoses of the patient are not deeply involved in ADEs, except chronic diseases that can already be retrieved from the previous stays of the patient.

Point #10 is exactly the objective of the present work.

5.2. Discussion of the method

Weaknesses

Despite its advantages, the present procedure suffers from two main weaknesses.

First, **only the data that are recorded can be mined**. Some clinical events might occur and might be present in the data of the EHR as unstructured free-text observation but not encoded as structured information. The main advantage of non automated methods is to be able to take into account every kind of information, even unstructured informations.

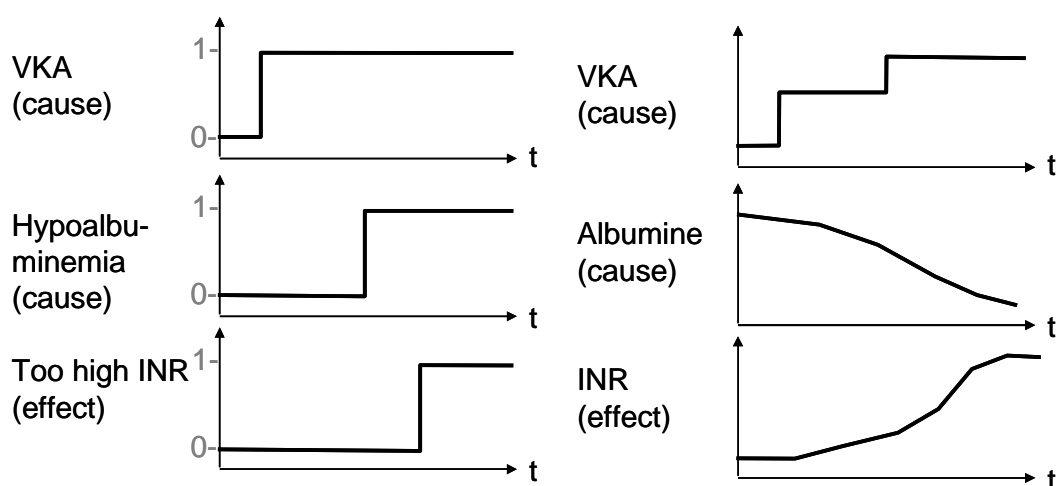
Secondly, **diagnostic codes** are important to describe acute and chronic diseases. Till now we have only been able to take into account chronic diseases and acute diseases that are necessarily the admission ground. For instance, the diagnosis “polytrauma” can only be the referral diagnosis, as it is not probable that a polytrauma occurs during the hospitalization. Conversely, “hemorrhage” could be either the admission ground or an adverse event occurring during the hospitalization. It is simpler for chronic diseases: a disease labeled as “chronic renal insufficiency” is necessarily an admission ground. In Danish hospitals, the referral diagnosis is encoded in the EHR at the admission, while in France the admission ground is registered using free-text and is later encoded as the “principal diagnosis”. But legally the physicians are not allowed to encode it since the patient hasn’t been discharged. As a consequence, in a CDSS context, only the chronic diseases are available, if the patient has already been admitted in the past.

Other important data haven’t been integrated as they were not sufficiently available: the patient’s **weight** and the patient’s **drug allergies**.

Finally, the **binary approach** for laboratory results and drug administration aggregation can be discussed. This kind of transformation has the inconvenient of an information loss. But it enables to date the events. An example is shown in Figure 51. The case of this example matches the following rule:

vitamin K antagonist & hypoalbuminemia → *risk of hemorrhage (too high INR)*

After a binary transformation of VKA doses, albumin blood level and INR (left part of Figure 51), it appears clearly that the VKA administration and the hypoalbuminemia have predated the too high INR and are candidates to explain it. This kind of representation seems to be the only way to discover statistical associations while taking time into account in a sample of thousands of cases. If we look at the quantitative variables without any binary transformation (right part of Figure 51), depending on the threshold that is chosen, it is not clear whether the VKA administration and the albumin level can explain or not the too high INR values. As far as we can mine the data, a quantitative approach seems not to be compatible with the time acknowledgment. In some cases we were able to generate some quantitative variables for data that are measured once and only once per hospitalization: this is the case of the patient's age.



**Figure 51. Example of ADE case matching the rule *VKA & hypoalbuminemia* → *too high INR*.
 Left: with binary transformation. Right: without binary transformation.**

Advantages

The present procedure also has some advantages.

The first advantage is to be able to confirm already known rules, to complete existing knowledge and to discover some new rules, as detailed in Chapter 4.3.3.3 (*Origin of the rules (data mining, SPCs)*) on page 110.

The main new kinds of knowledge are:

- *The consequences of drug discontinuations:* those kinds of conditions are rarely described in the summaries of product characteristics. However, particularly concerning pharmacokinetic drug interactions, it seems easy to accept that the importance of their effect might be the same as for the introduction of new drugs.
- *Organizational causes:* the rules from the present work can take into account practical causes. For instance the condition “the patient has been admitted with too low an INR” increases the probability that the patient has too high an INR during the hospitalization. This kind of conditions is typically useful but absent from academic knowledge. Organizational circumstances are probably not enough considered.

- *Segmentation conditions*: the aim of such conditions is to identify factors that are not directly responsible for an ADE but that significantly change the confidence of the rules. Most of segmentation conditions are related to the age, and the threshold that spontaneously appears is very close to 70 years old. In many rules, the segmentation condition is the absence of a drug or a disease, and this absence increases the confidence of the rule. In such cases, it can be supposed that, in case the disease or the drug is present, the patient benefits from more frequent monitoring of many laboratory parameters. Paradoxically, the absence of the drug or the disease may induce a higher rate of ADEs.

As opposed to academic knowledge, the results of the present work enable to sort the knowledge according to the probabilities of the outcomes. For instance the “contraindication” and “use caution” sections of the French summaries of product characteristics of current VKAs are 3,300 word 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 by the huge amount of information. Conversely, by means of the Scorecards and the automated computation of the confidences of the rules, it is possible for the physicians to get comprehensive and ordered information about the ADE cases that really happened in their medical department. By means of the link that enables to review the cases in the Expert Explorer, they can learn while reviewing the cases of patients they remember. The fact that the ADE statistics are linked with real cases has an impact on the perception of the tool by the practitioners of a medical department.

Statistics such as the confidence are automatically computed for each rule in each medical department. The output presented in chapter 14 (*Appendix 6: Validated rules*) on page 227 demonstrates that the confidence of the rules varies a lot from one department to another. It is probably due to differences in the patients, in treatments, and mostly in monitoring policies. Taking into account those various statistics will prevent the CDSS from over-alerting: by means of meta-rules presented in section 5.3.2 (*Meta-rules for the implementation into a CDSS*) on page 142, it is possible for the CDSS to have a behavior adapted to the final user.

The fact that the confidences of the rules are evaluated on-the-fly on fresher datasets will help to monitor the changes in therapeutic choices and consequences in several medical departments. This is particularly important for the Prospective Impact Assessment that will be performed as part of the PSIP Project (latest workpackage). It is possible that, if some specific ADEs are highlighted, the medication practices are improved and the number of cases decreases. A decreasing of the confidences of the rules could be a good piece of evidence of positive impact. But on the other hand, in order to avoid alert-fatigue, it will be very important to take into account the change in the confidence and, by means of meta-rule, to deactivate some rules. This will be automatically done by means of the actual process. Naturally, it will be done only for departments where it is appropriate: the statistics are computed separately in every medical department.

Those results are encouraging and announce a new approach to the ADE studies, current approaches being essentially based on staff operated case reviews [Bates 2003] or database queries [Honigman 2001, Honigman 2003, Seger 2007].

5.3. Perspectives

5.3.1. Reusability of the tools

All the programs and methods that have been used within this work can easily be reused in industrial context or in further researches.

The Expert Explorer and the Scorecards consist of a set of PHP Scripts and a MySQL database, making it easy to install the tools as it is on any new web server. A few hours are required to install duplicate versions of both tools, assuming that data compliant to the data model are provided as well as a description of users and medical units. In addition, the SQL scripts are limited to basic data management operations and a database abstraction library is used (PDO: PHP Data Objects). As a consequence, an Oracle database can very easily be used instead of MySQL without modifying the PHP scripts.

The data-mining scripts use the R language. This warrants that those scripts are widely reusable and can easily be interfaced with any existing database system. R is a system for statistical computation. It consists of a language plus a run-time environment. The R language is very similar in appearance to Chambers & Wilks' S language. R is free software distributed under a GNU-style copyleft, and an official part of the GNU project ("GNU S"). Since 1997, the R source code is updated by a core group (the "R Core Team"). Besides this core group, many R users contribute application codes: nearly 1,500 publicly-available packages are distributed through the Comprehensive R Archive Network (CRAN). Dozens of thousands people currently use R worldwide.

5.3.2. Meta-rules for the implementation into a CDSS

The knowledge rules provided by this work are not properly speaking operational rules for a CDSS: they describe knowledge and can be used for ADE detection in past stays, but they do not explain how a CDSS has to react in a prospective use during the medication process. Three meta-rules enable to transform the knowledge rules into operational rules which described the behaviors of the CDSS. They are proposed on Figure 52, Figure 53 and Figure 54.

For a given medical department, the knowledge rules can be ranked into 3 categories:

- the rule provides no stay (denominator=0) → activate the rule
- the rule provides some stays (denominator>0)
but $p\ value > 0.05$ **or** confidence < threshold* **or** relative risk ≤ 1
→ deactivate the rule
- the rule provides some stays (denominator>0)
and $p\ value < 0.05$ **and** confidence > threshold* **and** relative risk > 1
→ activate the rule

*The threshold has to be defined according to the severity of the outcome. For example, a 10 % threshold seems appropriate.

The same static filter is currently applied in the Scorecards.

Figure 52. Meta-rule n°1: static filter

The rules are provided with a description of the delay between (t_1) all the conditions are met and (t_2) the outcome occurs. Let $p80$ be the 80th percentile of the delay. As a consequence, when the outcome occurs, in only 20% of cases the outcome occurs with a delay greater than $p80$ days. In other words, when conditions have been met for a delay of $p80$ days, the confidence of the rule is divided by 5.

- For a given rule in a given medical department, use the $p80$ threshold of the delay (it is an expiration delay): $p80$
- For the given stay, compute the delay between causes and now (the prescription day): $delay$
- If $delay > p80$, do not use the rule anymore

Figure 53. Meta-rule n°2: temporal filter

This meta-rule is to be used when a rule involves a lab-related anomaly as the outcome:

$A+B \rightarrow C$ (C being an abnormal value of a laboratory parameter)

- IF the conditions A&B of the rule are present & the outcome C is not present & the outcome C is monitored \rightarrow don't alert
- IF the conditions A&B of the rule are present & the outcome C is not present & the outcome C is not monitored \rightarrow request monitoring C
- IF the conditions A&B of the rule are present & the outcome C is present \rightarrow alert by showing values of C and explain **according to the chronology of events**:
 - IF A&B started **before** the outcome C: A&B are **potential initial conditions of the outcome**.
 - IF A or B started **after** the outcome C: A&B are **potential conditions of worsening the outcome but not the initial causes of the outcome**.

This meta-rule requires that secondary thresholds are defined for each laboratory parameter.

Example:

Initial rule: VKA & hypoalbuminemia \rightarrow too low an INR

If both conditions are present, the behavior of the CDSS depends on the INR:

- IF INR is monitored & INR is not too low \rightarrow don't alert
- IF INR is not monitored \rightarrow request monitoring of the INR
e.g. *no value of NPU01685 is readable in the dataset during the last 5 days*
- IF INR is monitored & INR is too low \rightarrow alert
 - IF VKA and hypoalbuminemia were both present before the too low INR value, present conditions as "initial conditions"
 - IF VKA or hypoalbuminemia was not present before the too low INR value, present conditions as "worsening conditions"

Figure 54. Meta-rule n°3: rules with laboratory-related outcomes

5.3.3. Reusability of the rules described using XML files

The main outputs of this work are:

- **Clean and structured datasets** from various hospitals and departments
- **Structured XML files**, described in chapter 12 (*Appendix 4: Description of the output of this work (use of the XML files)*), on page 201, including:
 - **Mapping policies**: those files enable to aggregate the data into events (for diagnoses, drug administrations, and laboratory results).
 - **Rules**, identified as set of conditions linked to outcomes
 - **Occurrences of the rules**, i.e. statistics computed for each rule in each medical department or hospital
 - **Lexicon** to automatically make the conditions and outcomes of the rules readable for humans. It provides labels in English, French and Danish.
 - **Explanations** to understand how a rule can be argued and what the prescriber could do when an alert fires. It provides texts in several languages (English, French and Danish) for several uses (short text, long text, action) and for several users (physicians, nurses and patients).
- **Automated HTML output**, making it possible to present all the rules, statistics and explanations at the same time, as shown in section 4.3.2 (*Detailed example of five rules*) on page 102, and in chapter 14 (*Appendix 6: Validated rules*) on page 227

The format of the structured XML files has been agreed on by all the different users, who are the persons involved in the next workpackages of the PSIP Project. Those XML files are frequently updated and loaded on-the-fly by several partners for several uses:

- **Retrospective use**: the files are used to produce information about stays from which the patients have already been discharged (Figure 55). For that kind of use, the present output is used as it is and is linked to all the available stays. Those stays are stored in a central repository, describing all the extractions of all the available hospitals. This kind of architecture is used by several processes:
 - **The automated HTML outputs** described above
 - **The periodic ADE Scorecards**: this tool enables to visualize the rules and the related statistics in a given medical department
 - **The Questionnaire of the Expert Explorer**: it is loaded in order to review and validate or invalidate the rules
- **Prospective use: the CDSS**. For that kind of use, the present output is used as it is too. Each time a drug is prescribed, the CPOE provides the Connectivity Platform developed in PSIP with the data concerning the current stay. The CDSS is queried and uses the XML files to answer (Figure 56). This kind of architecture will be used by several software applications:
 - **The IBM prototype**, a tool connected to a CPOE
 - **The Medasys Prototype**, a tool connected to a CPOE

- **The Web prototype**, a web tool enabling drug prescription simulations, for education purposes.
- **The Patient component**, a tool designed to provide patients with information in relation to their own treatment

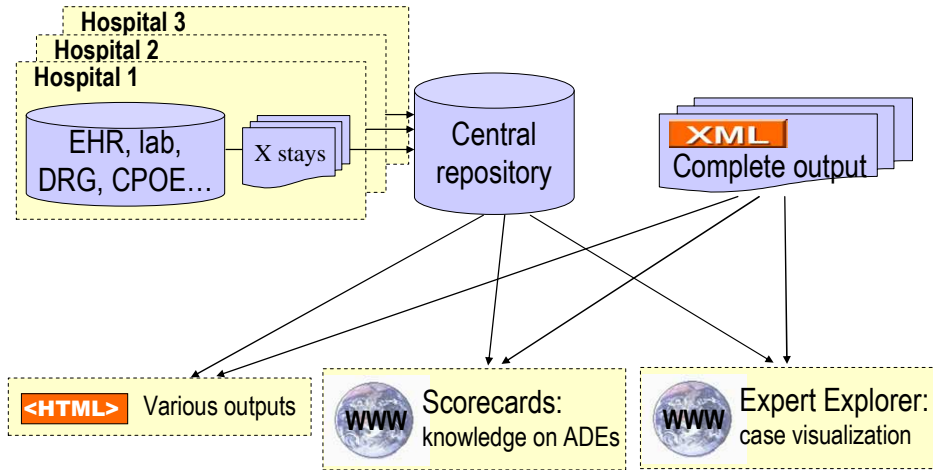


Figure 55. Retrospective use of the set of XML files

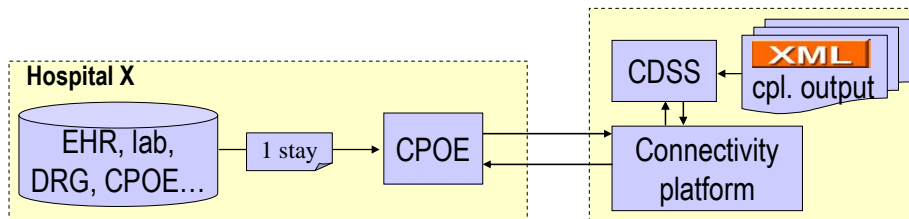


Figure 56. Prospective use of the set of XML files

6. CONCLUSION

In this work, we have defined a simple data model to group together data extracted from EHRs and CPOEs. This data model has been used to collect more than 90,000 hospital stays. Mainly by means of data mining, 236 ADE detection rules have been described in a rule repository through a set of easily reusable XML files. Those rules enable to detect and prevent 27 different kinds of outcomes.

The execution of those rules on large datasets describing past hospitalizations enable to detect potential ADE cases with a good accuracy and to compute automatically interesting statistics. In order to exploit such results, two web tools have been developed and can be easily implemented: the Scorecards to display epidemiological information about ADEs, and the Expert Explorer to review the potential ADE cases. The same rules can also be used for ADE prevention in a CDSS. The definition of several statistics and related thresholds enables to reduce the false positive rates, and would participate in reducing the over-alerting of CDSSs.

The present work contributes to the challenge of ADE detection. It seems that this work is the first successful experiment of data-mining-based rule induction for ADE detection. One output of this work is a set of 236 validated and commented rules. Those rules seem to have a good accuracy thanks to the segmentation and contextualization of the rules. In addition, this work highlights the contribution of factors that have always been ignored: drug discontinuations and organizational causes.

This work must be continued in order to detect less frequent outcomes on larger datasets, to evaluate more precisely the accuracy of the ADE detection, and to quantify more precisely the clinical impact of ADEs.

Finally, the method used here can incorporate other kinds of data as soon as they are available in the EHRs, such as the structured results of any paraclinical exam (e.g. electrocardiograms). In addition, the method showed an unexpected ability to detect some cases of nosocomial infection. This field deserves to be more deeply explored.

7. REFERENCES

- [Adriaans 1996] Adriaans P, Zantinge D, Syllogic (Firm). *Data mining*. Harlow, England ; Reading, Mass.: Addison-Wesley; 1996.
- [Agrawal 1993] 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.
- [Agrawal 1994] Agrawal R, Srikant R. *Fast algorithms for mining association rules in large databases*. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.
- [AFSSAPS 2009] *Agence Française de Sécurité Sanitaire des Produits de Santé* [cited October 2010]; Available from: <http://www.afssaps.fr/>
- [Akenaton 2010] *Akenaton* [Cited 2010 September 28] Available from: <http://resmed.univ-rennes1.fr/testjps/akenaton/>
- [Almenoff 2005] Almenoff J, Tonning JM, Gould AL, Szarfman A, Hauben M, Ouellet-Hellstrom R, et al. *Perspectives on the use of data mining in pharmaco-vigilance*. Drug Saf. 2005;28(11):981-1007.
- [Almenoff 2007] Almenoff JS, Pattishall EN, Gibbs TG, DuMouchel W, Evans SJ, Yuen N. *Novel statistical tools for monitoring the safety of marketed drugs*. Clin Pharmacol Ther. 2007 Aug;82(2):157-66.
- [Amalberti 2006] Amalberti R, Gremion C., Auroy Y, Michel P, Salmi R, Parneix P, et al. *Typologie et méthode d'évaluation des systèmes de signalement des accidents médicaux et des événements indésirables*. Report, 2006.
- [Amemiya 1985] Amemiya T. (1985). *Advanced Econometrics*. Harvard University Press. ISBN 0-674-00560-0.
- [Aramaki 2010] Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, Ohe K. *Extraction of adverse drug effects from clinical records*. Stud Health Technol Inform. 2010;160(Pt 1):739-43.
- [Aronson 2002] Aronson JK. Drug therapy. In: Haslett C, Chilvers ER, Boon NA, Colledge NR, Hunter JAA, eds. *Davidson's principles and practice of medicine 19th ed*. Edinburgh: Elsevier Science, 2002:147-
- [ATC 2009] *Anatomical and Therapeutical Classification*. [cited 2009 february 24]; Available from: <http://www.whocc.no/atcddd>.
- [Băceanu 2009] Băceanu A, Atasiei I, Chazard E, Leroy N, *The Expert Explorer: A Tool for Hospital Data Visualization and Adverse Event Rules Validation*. Studies in Health Technology and Informatics 148(2009), 85-94.
- [Balakrishnan 1991] Balakrishnan N. (1991). *Handbook of the Logistic Distribution*. Marcel Dekker, Inc. ISBN 978-0824785871.
- [Bartholomew 1999] Bartholomew DJ, Knott M. *Latent Variable Models and Factor Analysis*. 2nd edition, Arnold, 1999.
- [Bate 2006] Bate A, Edwards IR. *Data mining in spontaneous reports*. Basic Clin Pharmacol Toxicol. 2006 Mar;98(3):324-30.
- [Bates 1994] Bates, D.W., et al., *Potential identifiability and preventability of adverse events using information systems*. J Am Med Inform Assoc, 1994. 1(5): p. 404-11.
- [Bates 1995] Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. *Incidence of adverse drug events and potential adverse drug events. Implications for prevention*. ADE Prevention Study Group. JAMA. 1995;274:29-34.
- [Bates 1997] Bates D, Spell N, Cullen DJ, et al. *The costs of adverse drug events in hospitalized patients*. JAMA 1997.
- [Bates 2003] 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.

- [BDAM 2009] *Banque de Données Automatisée sur les Médicaments*. [cited 2009 february 24]; Available from: <http://www.biam2.org/accueil.html>.
- [Belur 1991] Belur V. Dasarathy, editor (1991) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*.
- [Bennet 2007] Bennett CL, Nebeker JR, Yarnold PR, Tigue CC, Dorr DA, McKoy JM, et al. *Evaluation of serious adverse drug reactions: a proactive pharmacovigilance program (RADAR) vs safety activities conducted by the Food and Drug Administration and pharmaceutical manufacturers*. Arch Intern Med. 2007 May 28;167(10):1041-9.
- [Benson 2000] Benson M, Junger A, Michel A, Sciuk G, Quinzio L, Marquardt K, et al. *Comparison of manual and automated documentation of adverse events with an Anesthesia Information Management System (AIMS)*. Stud Health Technol Inform 2000;77:925-9.
- [Benzecri 1973] Benzécri JP. *L'Analyse de Données : la Taxinomie*. Dunod, Paris, 1973.
- [Berlin 2008] Berlin JA, Glasser SC, Ellenberg SS. *Adverse event detection in drug development: recommendations and obligations beyond phase 3*. Am J Public Health. 2008 Aug;98(8):1366-71.
- [Breiman 1984] Breiman L. *Classification and regression trees*. Belmont, Calif.: Wadsworth International Group; 1984.
- [Brin 1997] Brin S, Motwani R, Ullman JD, Tsur S. *Dynamic itemset counting and implication rules for market basket data*. Proceedings ACM SIGMOD International Conference on Management of Data, pages 255-264, Tucson, Arizona, USA, May 1997.
- [Brouard 2004] Brouard F. *L'art des Soundex*. [cited 2010 september, 27] Available from: <http://sqlpro.developpez.com/cours/soundex>
- [Cantor 2007] Cantor MN, Feldman HJ, Triola MM. *Using trigger phrases to detect adverse drug reactions in ambulatory care notes*. Qual Saf Health Care. 2007 Apr;16(2):132-4.
- [Cerrito 2001] Cerrito P. *Application of data mining for examining polypharmacy and adverse effects in cardiology patients*. Cardiovasc Toxicol. 2001;1(3):177-9.
- [Chazard 2009 (1)] Chazard E, Merlin B, Ficheur G, Sarfati JC; PSIP Consortium, Beuscart R. *Detection of adverse drug events: proposal of a data model*. Stud Health Technol Inform. 2009;148:63-74.
- [Chazard 2009 (2)] Chazard E, Preda C, Merlin B, Ficheur G, Beuscart R. *Data-mining-based detection of adverse drug events*. Stud Health Technol Inform. 2009;150:552-6..
- [Chazard 2009 (3)] Chazard E, Ficheur G, Merlin B, Genin M, Preda C; PSIP consortium, Beuscart R. *Detection of adverse drug events detection: data aggregation and data mining*. Stud Health Technol Inform. 2009;148:75-84.
- [Chazard 2009 (4)] Chazard E, Ficheur G, Merlin B, Serrot E; PSIP Consortium, Beuscart R. *Adverse drug events prevention rules: multi-site evaluation of rules from various sources*. Stud Health Technol Inform. 2009;148:102-11.
- [Classen 1997] Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. *Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality*. JAMA 1997.
- [Classen 2005] Classen DC, Pestotnik SL, Evans RS, Burke JP. *Computerized surveillance of adverse drug events in hospital patients*. 1991. Qual Saf Health Care 2005 Jun;14(3):221-5.
- [Codd 1972] Codd E.F. *Further Normalization of the Data Base Relational Model*. IBM Research Report RJ909. Republished in Randall J. Rustin (ed.), Data Base Systems: Courant Computer Science Symposia Series 6. Prentice-Hall, 1972.
- [Codd 1990] Codd, E.F. *The Relational Model for Database Management: Version 2*. Addison-Wesley (1990).
- [Cornuejols 2002] Cornuéjols A, Miclet L. *Apprentissage artificiel concepts et algorithmes*. Eyrolles, 2002.
- [Coulter 2001] Coulter DM, Bate A, Meyboom RH, Lindquist M, Edwards IR. *Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study*. BMJ. 2001 May 19;322(7296):1207-9.

- [Cox 1972] Cox DR (1972). *Regression Models and Life Tables*. Journal of the Royal Statistical Society Series B 34 (2): 187–220. JSTOR: 2985181. MR0341758.
- [Date 1999] Date C.J. *An Introduction to Database Systems*. Addison-Wesley (1999).
- [Date 2005] Date C.J. *Database in Depth: Relational Theory for Practitioners*. O'Reilly (2005).
- [Day 1984] Day WHE, Doucette CR (1984). *Expected Behaviour of Quartet Distances Between Undirected Phylogenetic Trees*. Eighteenth International Numerical Taxonomy Conference, Cornell University, Ithaca, New York, 5-8 October 1984.
- [Del Fiol 2000] Del Fiol G, Rocha BH, Kuperman GJ, Bates DW, Nohama P. *Comparison of two knowledge bases on the detection of drug-drug interactions*. Proc AMIA Symp. 2000:171-5.
- [Didaye 1980] Diday E. *Optimisation en Classification Automatique, tomes 1 et 2*. INRIA, 1980
- [Domingos 1997] Domingos P, Pazzani M (1997) *On the optimality of the simple Bayesian classifier under zero-one loss*. Machine Learning, 29:103–137.
- [Dzeroski 1996] Dzeroski S, Lavrac N. *Rule induction and instance-based learning applied in medical diagnosis*. Technol Health Care. 1996 Aug;4(2):203-21.
- [EC 2001] *Directive 2001/83/EC on the Community code relating to medicinal products for human use*. [Cited october 2010]; Available from: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:2001L0083:20090720:EN:PDF>
- [ERC 2010] *European Research Council*. [cited 2010 October]; Available from: <http://erc.europa.eu>
- [Evans, 1996] Evans DA, Brownlow ND, Hersh WR, Campbell EM. *Automating Concept Identification in the Electronic Medical Record: An Experiment in Extracting Dosage Information*. AMIA 1996 Symposium Proceedings. 1996:388-92.
- [Everitt 1984] Everitt BS. *An Introduction to Latent Variable Models*. Chapman & Hall, London, 1984.
- [Fayyad 1996] 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.
- [FDA 2010] *MedWatch - What Is A Serious Adverse Event?* [cited 2010 October]; Available from: <http://www.fda.gov/Safety/MedWatch/HowToReport/ucm053087.htm>
- [Field 2004] Field, T.S., et al., *Strategies for detecting adverse drug events among older persons in the ambulatory setting*. J Am Med Inform Assoc, 2004. 11(6): p. 492-8.
- [FP7 2010] *Seventh Framework programme*. [cited 2010 February, 10]; Available from: http://cordis.europa.eu/fp7/home_en.html
- [Gandhi 2005] Gandhi, T.K., et al., *Outpatient prescribing errors and the impact of computerized prescribing*. J Gen Intern Med, 2005. 20(9): p. 837-41.
- [Garcia 2008] Garcia V, Debreuve E, Barlaud M. *Fast k nearest neighbor search using GPU*. Proceedings of the CVPR Workshop on Computer Vision on GPU, Anchorage, Alaska, USA, June 2008.
- [Gold, 2008] Gold S, Elhadad N, Zhu X, Cimino J J., George Hripsak, *Extracting Structured Medication Event Information from Discharge Summaries* AMIA 2008 Symposium Proceedings 237-241
- [Gurwitz 2000] Gurwitz JH, Field TS, Avorn J, McCormick D, Jain S, Eckler M, et al. *Incidence and preventability of adverse drug events in nursing homes*. Am J Med. 2000;109:87-94.
- [Gurwitz 2003] 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.
- [Gysbers 2008] Gysbers M, Reichley R, Kilbridge PM, Noirot L, Nagarajan R, Dunagan WC, et al. *Natural language processing to identify adverse drug events*. AMIA Annu Symp Proc. 2008:961.
- [Hagenaars 2002] Hagenaars JA, McCutcheon AL (eds.). *Applied latent class analysis*. Cambridge University Press, 2002.

- [Hand 2001] Hand DJ, Yu K. (2001). *Idiot's Bayes - not so stupid after all?* International Statistical Review. Vol 69 part 3, pages 385-399. ISSN 0306-7734.
- [Handler 2006] Handler SM, Wright RM, Ruby CM, Hanlon JT. *Epidemiology of medication-related adverse events in nursing homes.* Am J Ger Pharma 2006
- [Handler 2007] Handler SM, Altman RL, Perera S, Hanlon JT, Studenski SA, Bost JE, et al. *A systematic review of the performance characteristics of clinical event monitor signals used to detect adverse drug events in the hospital setting.* J Am Med Inform Assoc. 2007 Jul-Aug;14(4):451-8.
- [Hasler 2007] Hasler M, Hornik K. (2007). *New probabilistic interest measures for association rules.* Intelligent Data Analysis, pages 437 - 455.
- [Hauben 2005] Hauben M, Patadia V, Gerrits C, Walsh L, Reich L. *Data mining in pharmacovigilance: the need for a balanced perspective.* Drug Saf. 2005;28(10):835-42.
- [Honigman 2001] Honigman B., et al., *Using computerized data to identify adverse drug events in outpatients.* J Am Med Inform Assoc, 2001. 8(3): p. 254-66.
- [Honigman 2003] 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.
- [ICD 2009] *International Classification of Diseases.* [cited 2009 february 24]; Available from: <http://www.who.int/classifications/icd/en>.
- [Ingram 2008] Ingram PR, Lye DC, Tambyah PA, Goh WP, Tam VH, Fisher DA. *Risk factors for nephrotoxicity associated with continuous vancomycin infusion in outpatient parenteral antibiotic therapy.* J Antimicrob Chemother. 2008 Jul;62(1):168-71.
- [IOM 2007] Institute Of Medicine. *Preventing Medication Errors.* Washington, DC: The National Academic Press; 2007.
- [ISO 2009] ISO 13407:1999, *Human-centred design processes for interactive systems.*
- [IUPAC 2010] *International Union of Pure and Applied Chemistry.* [cited 2009 february 24]; Available from: <http://www.iupac.org>.
- [Jalloh 2006] 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.
- [Jha 1998] Jha, A.K., et al., *Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report.* J Am Med Inform Assoc, 1998. 5(3): p. 305-14.
- [Jha 2008] Jha AK, Laguette J, Seger A, Bates DW. *Can surveillance systems identify and avert adverse drug events? A prospective evaluation of a commercial application.* J Am Med Inform Assoc. 2008 Sep-Oct;15(5):647-53.
- [Judge 2006] Judge, J., et al., *Prescribers' responses to alerts during medication ordering in the long term care setting.* J Am Med Inform Assoc, 2006. 13(4): p. 385-90.
- [Kaushal 2003] Kaushal R, Shojania KG, Bates DW. *Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review.* Arch Intern Med. 2003 Jun 23;163(12):1409-16.
- [Kilbridge 2006] Kilbridge PM, Campbell UC, Cozart HB, Mojarrad MG. *Automated surveillance for adverse drug events at a community hospital and an academic medical center.* J Am Med Inform Assoc 2006 Jul;13(4):372-7.
- [Kimball, 2002] Kimball R. *The Data Warehouse Toolkit, 2nd Ed.*. Wiley Computer Publishing (2002).
- [Kohn 1999] Kohn LT, Corrigan JM, Donaldson MS. *To Err Is Human: Building a Safer Health System.* Washington, DC: National Academy Pr; 1999.
- [Kotsiantis 2004] Kotsiantis S, Pintelas P. *Increasing the Classification Accuracy of Simple Bayesian Classifier.* Lecture Notes in Artificial Intelligence, AIMS 2004, Springer-Verlag Vol 3192, pp. 198-207, 2004
- [Kuperman 1994] Kuperman, G.J., et al., *A new knowledge structure for drug-drug interactions.* Proc Annu Symp Comput Appl Med Care, 1994: p. 836-40.

- [Kuperman 1996] Kuperman, G.J., et al., *Detecting alerts, notifying the physician, and offering action items: a comprehensive alerting system*. Proc AMIA Annu Fall Symp, 1996: p. 704-8.
- [Kuperman 1999] Kuperman, G.J., et al., *Improving response to critical laboratory results with automation: results of a randomized controlled trial*. J Am Med Inform Assoc, 1999. 6(6): p. 512-22.
- [Lavrac 1999] Lavrac N. *Selected techniques for data mining in medicine*. Artif Intell Med. 1999 May;16(1):3-23.
- [Lazarsfeld 1968] Lazarsfeld PF, Henri NW. *Latent Structure Analysis*. Houghton Mifflin, Boston, 1968.
- [Lebart 2000] Lebart L, Morineau A, Piron M. (2000). *Statistique exploratoire multidimensionnelle*. Dunod.
- [MacKay 2003] MacKay D. (2003). Chapter 20. *An Example Inference Task: Clustering*. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. pp. 284–292. MR2012999. ISBN 0-521-64298-1.
- [MacQueen 1967] MacQueen JB. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1: 281–297, University of California Press.
- [McCullagh 1989] McCullagh P, Nelder J. (1989). *Generalized Linear Models*. London: Chapman and Hall. ISBN 0-412-31760-5. Chapter 1.
- [McCutcheon 1987] McCutcheon AL. *Latent Class Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences n°64, SAGE publications, 1987.
- [Makhoul 1999] Makhoul, John; Francis Kubala; Richard Schwartz; Ralph Weischedel: *Performance measures for information extraction*. In: Proceedings of DARPA Broadcast News Workshop, Herndon, VA, February 1999.
- [Marcilly 2010] Marcilly R, Chazard E, Beuscart-Zephir MC, Hackl W, Baceanu A, Kushniruk A, Borycki E, *Design of Adverse Drug Events-Scorecards*
- [Melton 2005] Melton GB, Hripcsak G. *Automated detection of adverse events using natural language processing of discharge summaries*. J Am Med Inform Assoc. 2005 Jul-Aug;12(4):448-57.
- [Minsky 1961] Minsky M. (1961). *Steps toward Artificial Intelligence*. Proceedings of the IRE 49(1):8-30.
- [Morimoto 2004] 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.
- [Morina 2004] Mozina M, Demsar J, Kattan M, & Zupan B. (2004). *Nomograms for Visualization of Naive Bayesian Classifier*. Proc. of PKDD-2004, pages 337-348.
- [Murff 2003] 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.
- [Nakache 2003] Nakache J. & Confiais J. (2003). *Statistique explicative appliquée*. Technip.
- [Nakache 2005] Nakache J. (2005). *Approche pragmatique de la classification*. Technip.
- [Nakache 2005] Nakache D., Metais E., Timsit J. *Evaluation and NLP*. proceedings of DEXA Database and Expert System Application, 626–632, 2005.
- [Nebeker 2004] Nebeker Jonathan R. *Clarifying Adverse Drug Events: A Clinician's Guide to Terminology, Documentation, and Reporting*. Ann Intern Med. 2004;140:795-801.
- [Nebeker 2005] Nebeker JR, Hoffman JM, Weir CR, Bennett CL, Hurdle JF. *High rates of adverse drug events in a highly computerized hospital*. Arch Int Med 2005.
- [Névéol, 2006] Névéol, A; Rogozan, Aé & Darmoni, SJ. *Automatic indexing of online health resources for a French quality controlled gateway*. Information Processing & Management , May, Volume 42, Number 3, 695-709, 2006.
- [Névéol, 2007] Névéol, A.; Pereira, S.; Kerdelhué, G.; Dahamna, B.; Joubert, M. & Darmoni, SJ. *Evaluation of a Simple Method for the Automatic Assignment of MeSH Descriptors to Health Resources in a French Online Catalogue*. Stud Health Technol Inform,

U.S. National Library of Medicine, National Institutes of Health, Bethesda, USA., Volume 129, 407-411, 2007.

- [Northrop Grumman 2010] Northrop Grumman. *MedDRA and the Maintenance and Support Services Organization*. [cited 2010 May 5] Available from: <http://www.meddrasso.com/MSSOWeb/index.htm>
- [Paterno 2009] Paterno, M.D., et al., *Tiering drug-drug interaction alerts by severity increases compliance rates*. J Am Med Inform Assoc, 2009. 16(1): p. 40-6.
- [Pereira 2008] Pereira, S; Névéol, A; Kerdelhué, G; Serrot, E; Joubert, M & Darmoni, SJ. *Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue*. AMIA symp., 586-590, 2008.
- [Pereira 2009] Pereira S, Sakji S, Névéol A, Kergourlay I, Kerdelhué G; Serrot E; Joubert M, Darmoni SJ. *Multi-terminology indexing for the assignment of MeSH descriptors to medical abstracts in French*. AMIA2009, Biomedical and Health Informatics: From foundations to Applications to policy, San Francisco, nov 2009.
- [Petersen 1992] Petersen LA, Lee TH, O'Neil AC, Cook EF, Brennan TA. *Potentially preventable adverse events identified by physician self-report and medical record review: demographic and resource utilization data*. Clin Res. 1992;40:588A.
- [Pharmacorama 2009] *Pharmacorama*. [cited 2009 february 24]; Available from: <http://www.pharmacorama.com>.
- [Piatetsky-Shapiro 1991] Piatetsky-Shapiro G, Frawley W. *Knowledge discovery in databases*. Menlo Park, Calif.: AAAI Press: MIT Press; 1991.
- [PSIP 2010] *Patient Safety by Intelligent Procedures in medication*. [cited 2010 october]; Available from: <http://www.psip-project.eu>.
- [Pubmed 2009] *Pubmed*. [cited 2009 february 24]; Available from: <http://www.ncbi.nlm.nih.gov/pubmed>.
- [Quinlan 1986] Quinlan JR. *Introduction of Decision Trees*. Machine Learning. 1986;1:81-106.
- [Quilan 1993] Quilan R. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufman.
- [R 2008] R_Development_Core_Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008.
- [Rakotomalala 1997] Rakotomalala R. (1997). *Graphes d'induction*. PhD thesis, Université Claude Bernard Lyon 1.
- [Rakotomalala 2005] Rakotomalala R. (2005). *Arbre de décision*. Revue Modulad, 33: 163–187.
- [Rawlins 1977] Rawlins MD, Thompson JW. *Pathogenesis of adverse drug reactions*. In: Davies DM, ed. Textbook of adverse drug reactions. Oxford: Oxford University Press, 1977:10.
- [Ripley 1996] Ripley BD. *Pattern recognition and neural networks*. Cambridge ; New York: Cambridge University Press; 1996.
- [Rish 2001] Rish I. (2001). *An empirical study of the naive Bayes classifier*. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.
- [Schedlbauer 2009] Schedlbauer, A., et al., *What evidence supports the use of computerized alerts and prompts to improve clinicians' prescribing behavior?* J Am Med Inform Assoc, 2009. 16(4): p. 531-8.
- [Seger 2007] 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.
- [Senst 2001] Senst BL, Achusim LE, Genest RP, et al. *Practical approach to determining costs and frequency of adverse drug events in a health care network*. Am J Health-Sys Pharm 2001.
- [Shakhnarovich 2005] Shakhnarovich G, Darrell T & Indyk P. *Nearest-Neighbor Methods in Learning and Vision*. Edited by Shakhnarovich, Darrell, and Indyk, The MIT Press, 2005.
- [Sirohi, 2005] Sirohi E, Peissig P. *Study of Effect of Drug Lexicons on Medication Extraction from Electronic Medical Records*. Pacific Symposium on Biocomputing; 10:308-18. 2005

- [Sittig 2008] Sittig, D.F., et al., *Grand challenges in clinical decision support*. J Biomed Inform, 2008. 41(2): p. 387-92.
- [SPSQS 2009] Committee of experts on management of safety and quality in health care (SPSQS)/ Expert group on safe medication practices. *Glossary of items related to patient and medication safety*. [cited 2010 october] available from <http://www.bvs.org.ar/pdf/seguridadpaciente.pdf>
- [Teich 1999] Teich, J.M., et al., *The Brigham integrated computing system (BICS): advanced clinical systems in an academic hospital environment*. Int J Med Inform, 1999. 54(3): p. 197-208.
- [Thériaque 2009] *Theriaque*. [cited 2009 february 24]; Available from: <http://www.theriaque.org/InfoMedicaments>.
- [Therneau 2007] Therneau TM, Atkinson B, Ripley B. *Rpart: Recursive Partitioning*. 2007.
- [Tuyns 1988] Tuyns AJ et al. *Cancer of the larynx/hypopharynx, tobacco and alcohol: IARC international case-control study in Turin and Varese (Italy), Zaragoza and Navarra (Spain), Geneva (Switzerland) and Calvados (France)*. 1988.
- [Van Rijsbergen 1979] Van Rijsbergen, C.V.: *Information Retrieval*. London; Boston. Butterworth, 2nd Edition 1979.
- [Venables 2002] Venables WN & Ripley BD. (2002). *Modern Applied Statistics with S*. Fourth edition. Springer.
- [Vidal 2009] Vidal S.A. [cited 2009 April, 23]; Available from: <http://www.vidal.fr/societe/vidal>.
- [WHO 2010 (1)] *World Health Organization Adverse Reactions Terminology*. [Cited 2010 May 5] Available from: <http://www.unc-products.com/DynPage.aspx?id=4918>
- [WHO 2010 (2)] *World Organisation of National Colleges, Academies, and Academic Associations of General Practitioners/Family Physicians*. [Cited 2010 May 5] Available from: <http://www.globalfamilydoctor.com/wicc/icpestory.html>
- [WHO 2010 (3)] *World Health Organization. International Classification of Diseases, 10th revision*. [Cited 2010 May 5] Available from: <http://www.who.int/classifications/icd/en/index.html>
- [XML 2009] *eXtensible Markup Language*. [cited 2009 April, 23]; Available from: <http://www.w3.org/XML/>.
- [Zhang 2001] Zhang HP, Crowley J, Sox H, Olshen RA. *Tree structural statistical methods*. Encyclopedia of Biostatistics. Chichester, England: Wiley; 2001. p. 4561-73.

8. ARTICLES

This chapter presents the articles that were published in the frame of the present work.

8.1. Pubmed references

1. Can F-MTI semantic-mined drug codes be used for Adverse Drug Events detection when no CPOE is available?
Merlin B, Chazard E, Pereira S, Serrot E, Sakji S, Beuscart R, Darmoni S
Stud Health Technol Inform. 2010;160:1025-9.
2. Data-mining-based detection of adverse drug events.
Chazard E, Preda C, Merlin B, Ficheur G; PSIP consortium, Beuscart R.
Stud Health Technol Inform. 2009;150:552-6.
3. Adverse drug events prevention rules: multi-site evaluation of rules from various sources.
Chazard E, Ficheur G, Merlin B, Serrot E; PSIP Consortium, Beuscart R.
Stud Health Technol Inform. 2009;148:102-11.
4. The expert explorer: a tool for hospital data visualization and adverse drug event rules validation.
Băceanu A, Atasiei I, Chazard E, Leroy N; PSIP Consortium.
Stud Health Technol Inform. 2009;148:85-94.
5. Detection of adverse drug events detection: data aggregation and data mining.
Chazard E, Ficheur G, Merlin B, Genin M, Preda C; PSIP consortium, Beuscart R.
Stud Health Technol Inform. 2009;148:75-84.
6. Detection of adverse drug events: proposal of a data model.
Chazard E, Merlin B, Ficheur G, Sarfati JC; PSIP Consortium, Beuscart R.
Stud Health Technol Inform. 2009;148:63-74.
7. Toward automatic detection and prevention of adverse drug events.
Leroy N, Chazard E, Beuscart R, Beuscart-Zephir MC; Psip Consortium.
Stud Health Technol Inform. 2009;143:30-5.

8.2. Science Direct references

8. Détection et prévention des effets indésirables médicamenteux par fouille automatisée des dossiers patients électroniques
E. Chazard, J. Salleron, M. Génin, G. Ficheur, A. Duhamel
Revue d'Épidémiologie et de Santé Publique, Volume 58, Supplement 1, April 2010, Page S8
9. Détection et prévention des effets indésirables liés aux médicaments par data-mining
E. Chazard, C. Preda, B. Merlin, G. Ficheur, R. Beuscart
IRBM, Volume 30, Issue 4, September 2009, Pages 192-196

9. APPENDIX 1: TIME REQUIRED TO PERFORM THE DATA MINING TASK

9.1. Objective of this chapter

The main objective of this section is to answer three questions:

- How much time does it require to compute statistics on a new dataset of an already known partner?
- How much time does it require to compute statistics on a dataset of a new partner?
- How much time does it require to discover new rules?

Figure 57 displays the main steps of data extraction, data management and data analysis for one medical department of one hospital. The different steps to integrate a hospital in the analysis cycle of this work are detailed hereafter.

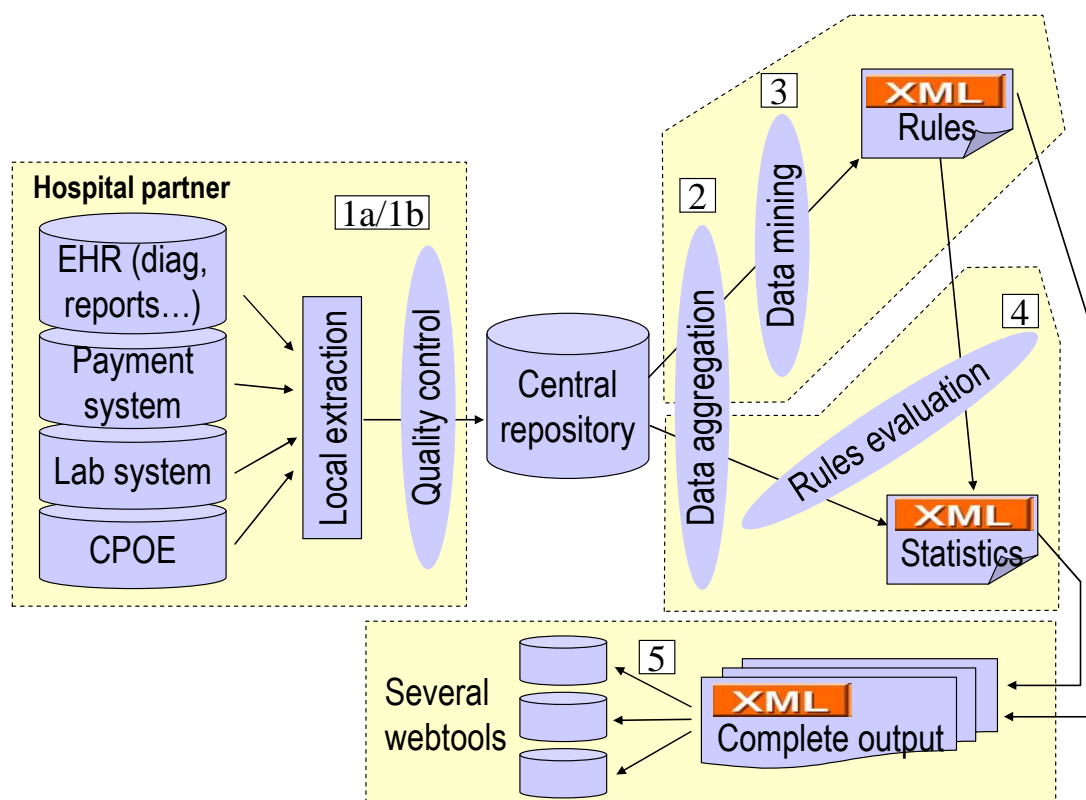


Figure 57. Global process for data extraction, data management and data analysis

9.2. Description of the tasks and needed time

9.2.1. Getting data from a hospital, first time

Steps on the diagram: 1a

Full duration: 1 man-month for the hospital,
1 man-week for the data mining team

Level of automation: Low

The hospital has to perform a data extraction from its EHRs in order to provide laboratory results, drug prescriptions, diagnoses, administrative information and free-text records. An iterative quality control is performed on the data. This quality control is described in section 2.2.5 (*Iterative quality control*) on page 57. The objective of the quality control is not to correct the dataset itself but to correct the data extraction process.

At the same time, as no French hospital is able to provide IUPAC codes for their laboratory investigations, a custom laboratory mapping has to be designed. Its aim is to transform the custom labels of laboratory parameters into precise and common laboratory parameter descriptions. The mapping process is expert-operated. For each observed laboratory parameter, it uses the labels, the declared units, the reference interval and the observed distribution of the values in the dataset.

9.2.2. Getting data from a hospital, next times

Steps on the diagram: 1b

Full duration: <6 hours

Level of automation: Full

Generally speaking, once the first extraction is performed, the extraction process is automated and next data extractions require less than 6 hours. It is then very easy to get the updates of the datasets. A fast quality control is performed on each new export.

9.2.3. Aggregating the data

Steps on the diagram: 2

Full duration: 1 hour

Level of automation: Full

Some treatments are applied to data so that they can be treated during the next steps. The design of the mapping policies (diagnoses, administrative information, drugs, and laboratory results) is now stable allowing for fully-automated transformations.

9.2.4. Discovering new rules

Steps on the diagram: 3

Full duration: ~1 month, several persons

Level of automation: Low

Discovering new rules follows a three-step procedure:

- Statistical outputs are automatically produced in order to prepare the expert sessions. Thanks to automation and scripts optimization, it requires only 2 hours. Hundreds of rules are produced but a pre-filtering session is performed and varies with respect to duration.
- Then an expert committee is in charge of filtering, validating, re-organizing and tuning the rules. This step requires that all the experts have prepared the domains that are investigated. When this is the case, 1 day with 6 experts usually allows preparing 60 rules. During the sessions, explanations of the rules are prepared.
- Finally, a long step consists of checking and testing all the results of the validation session, as well as preparing and translating the explanations of the rules in English, French and Danish.

9.2.5. Testing the rules of the central repository on an aggregated dataset

Steps on the diagram: 4

Full duration: ½ hour

Level of automation: Full

All the 236 rules of the central repository can be applied to every medical department dataset in a few minutes. The process is fully automated.

9.2.6. Publishing the datasets and the rules occurrences on the web tools

Steps on the diagram: 5

Full duration: 3 hours

Level of automation: Full

The datasets are loaded on the Expert Explorer, making it possible for physicians to review all the hospital stays. The rules occurrences are loaded on the Scorecards Tool, making it possible for the physicians of the department to know how many ADE occurred during the past month, to read the explanations, and to review the related hospital stays. The process is fully automated.

9.3. “How much time...”: use case point of view

9.3.1. Already known partner: computing statistics

Steps on the diagram: 1b, 2, 4, 5

Full duration: < 7 hours

Level of automation: Full

Each time a known partner wants to send a new dataset, the whole process requires less than 7 hours. This includes:

- the data extraction
- a fast quality control
- data aggregation
- computation of the statistics of the rules
- publication of all the files on the Expert Explorer and Scorecards

9.3.2. New partner: computing statistics

Steps on the diagram: 1a, 2, 4, 5

Full duration: 1-2 months

Level of automation: Full

If a new partner joins the project, it requires between 1 and 2 months for the whole statistics to be available on the Scorecards tool. The steps are the same as above, except that obtaining a good data extraction requires a lot of time.

9.3.3. Discovering some new rules

Steps on the diagram: 3, 4, 5

Full duration: 1 month for several persons

Level of automation: Low

All the available datasets can be used every time to discover some new rules. The duration of the process is not dependent on the amount of data. Several statistics are automatically computed but the expert-operated filtering and reorganization of the rules is time-consuming. The process also includes the description of the explanation of the rules in English, French and Danish.

10. APPENDIX 2: ODP DESCRIPTION OF THE EXPERT EXPLORER

Authors: Adrian Baceanu & Emmanuel Chazard

10.1. User requirement and Enterprise Viewpoint

The PSIP project follows two main objectives:

- To produce epidemiological knowledge on Adverse Drug Events
- To design a clinical decision support system (CDSS) implementing some ADE detection rules, those rules being deduced from data mining of the structured hospital data bases, and semantic mining of free text collections (e.g. discharge letters).

As a part of the project, an EHR visualization tool has been required. The Expert Explorer has been designed to meet the following requirements:

- **Stay display:** the tool must be able to display a given hospital stay by means of visual and comprehensive pages. This visualization tool must display medical and administrative information including diagnoses and procedures, drug prescriptions, laboratory results, and reports (e.g. discharge letters).
- **Stays review and validation:** Once a rule provided by the data-mining team is implemented in the Scorecards (next tool, presented in section 11 on page 186), the Experts must be able to review all the hospital stays that match the rule and to fill an evaluation form. The results will be used to validate the rule. The rules-review tool uses the stay-display feature.
- **Reports:** the tool also allows for reports design, execution and display. These reports will be used by physicians of the medical departments to get on the fly customized information about ADEs in their medical department.

Moreover, the tool must respect some common requirements:

- The tool must be usable for every member of the research project, some users being able to connect from another country to a hospital database.
- Anonymity and confidentiality of the datasets are mandatory [Băceanu 2009].

The functional requirements to have access to a specific scorecard are described below.

For disambiguation purposes, in the present section, the following terms will be used:

- “Rules” and “scorecards” are related to the implementation of data-mining-based rules in the Scorecards tool (presented in section 11 on page 186).
- “Queries” and “reports” don’t refer to the data-mining results. They are a feature of the Expert Explorer, which allows the users to design themselves queries and to execute them again in report pages. Those queries and reports will be called as the “reporting tool of the Expert Explorer”.

10.1.1. Use case 1

- **Goal:** To access the details of a specific stay using results of the reporting tool of the “Expert Explorer”.

- **Actors:** Medical Personnel of the hospital, Medical Experts involved in a case review, Members of the PSIP project.
- **Interface:** The “Expert Explorer”, being a web-based data visualization tool, has a detailed graphical interface. The functions of the interface can be accessed easily and are available in English, but some pages are available in multiple translations like French and Danish.
- **Preconditions:** The tool must have access to a MySQL database containing relevant information on clinical stays. The database must use the PSIP data model. Anonymity and confidentiality of the datasets are mandatory. The users must know the login credentials.

- **Basic course of events:**
 1. The user opens the application and is presented with a welcome message and the menu
 2. The user opens the “Data Set” tab, selects the working data set and clicks “Apply”
 3. The user opens the “Reports” tab, then the “Define report” option
 4. He then chooses the fields which will be displayed on the lines of the report, and the fields from the columns of the report. This way the report is defined.
 5. The user clicks the “Generate” button and the system saves the report
 6. The user clicks on the “Report Details” and the system displays the report in a table structure, with options to sort the table with a click on the columns names.
 7. Then the user can go to the “Rules” tab and select the option “Define Rule”. This will allow him designing a query.
 8. He has to select the fields, the operators and the values to construct the new rule, and he clicks the “Submit” button
 9. The system saves the query and provides the user with a link to the details of the query. The query is presented as a set of conditions and the stays that it identifies.
 10. The user reviews the details of a stay by clicking on its id, which will bring up the page with all the details for that stay (steps of the stay, medical procedures taken, ICD10 diagnoses, administrated drugs, laboratory results, laboratory charts, drug charts and free text documents linked to that stay).

- **Alternative paths:**
 - 2a. The user can use the default selected data set
 - 3a. The user can consult a previously generated report
 - 7a. The user can review a previously declared query
 - 10a. The query didn’t find any stay, the user can go back to the page where he comes from.

- **Post Conditions:** New report is added to the software. New queries are added to the software.

10.1.2. Use case 2

- **Goal:** To access the details of a specific stay and complete the pre defined questionnaire in order to determine if it was or wasn't an ADE.
- **Actors:** Medical Experts
- **Interface:** The “Expert Explorer”, being a web-based data visualization tool, has a detailed graphical interface. The functions of the interface can be accessed easily and are available in English, but some pages are available in multiple translations like French and Danish.
- **Preconditions:** The tool must have access to a MySQL database containing relevant information on clinical stays. The database must use the PSIP data model. Anonymity and confidentiality of the datasets are mandatory. The users must know the login credentials and have “Expert” access.
- **Basic course of events**
 1. The user opens the application and is presented with a welcome message and the menu
 2. The user has to login using his username and password
 3. The user opens the “Data Set” tab, selects the working data set and clicks “Apply”
 4. Then the user can go to the “Rules” tab and consult the list of previously defined rules
 5. User clicks the “See the stays” link
 6. User can select a stay from the list of stays identified by the rule
 7. User reviews the details of a stay by clicking on its id, which will bring up the page with all the details for that stay (steps of the stay, medical procedures taken, ICD10 diagnoses, drugs administrated, laboratory results, laboratory charts, drug charts and free text documents linked to that stay)
 8. User access the review page where he is asked several questions, one at a time
- **Alternative paths**
 - 3a. The user can use the default selected data set;
 - 4a. The user can directly access the details of a rule that was assigned to him for review;
 - 8a. If the expert already reviewed the stay, then the questionnaire is automatically completed with his previous answers, and he has the option to modify his answers, or cancel and go to previous page
- **Post Conditions:** The answers of the questionnaire are saved for later analysis. Rules reviewed by the user are marked as reviewed by the current user.

10.1.3. Use case 3

- **Goal:** Management of data used by the “Expert Explorer”, review of the actions of the experts

- **Actors:** Tool administrators
- **Interface:** The “Expert Explorer”, being a web-based data visualization tool, has a detailed graphical interface. The functions of the interface can be accessed easily and are available in English, but some pages are available in multiple translations like French and Danish.
- **Preconditions:** The tool must have access to a MySQL database containing relevant information on clinical stays. The database must use the PSIP data model. Anonymity and confidentiality of the datasets are mandatory. The users must know the login credentials and have “Administrator” access.
- **Basic course of events**
 1. The user opens the application and is presented with a welcome message and the menu
 2. The user has to login using his username and password
 3. The administrator opens the Admin tab
 4. He can download an excel file containing the response from questionnaires completed between a selected period of time
 5. The administrator can upload a text file containing a list of stays that need to be reviewed
 6. He can assign for review those stays for a specific user or to all users
 7. The administrator can go to the “Data set” tab and select the working data set
 8. The administrator updates the data set by uploading to the server the corresponding text files
 9. Then the user can go to the “Rules” tab and consult the list of previously defined rules
 10. The administrator can validate a specific rule from that list
 11. The administrator can delete a previously declared rule
- **Alternative paths**
 - 9a. User clicks the “See the stays” link of a rule;
 - 9b. He can assign for review the stays found by the rule;
- **Post Conditions:** Data set is updated. The experts will have new stays assigned for review. Some rules could be removed.

10.2. Information Viewpoint

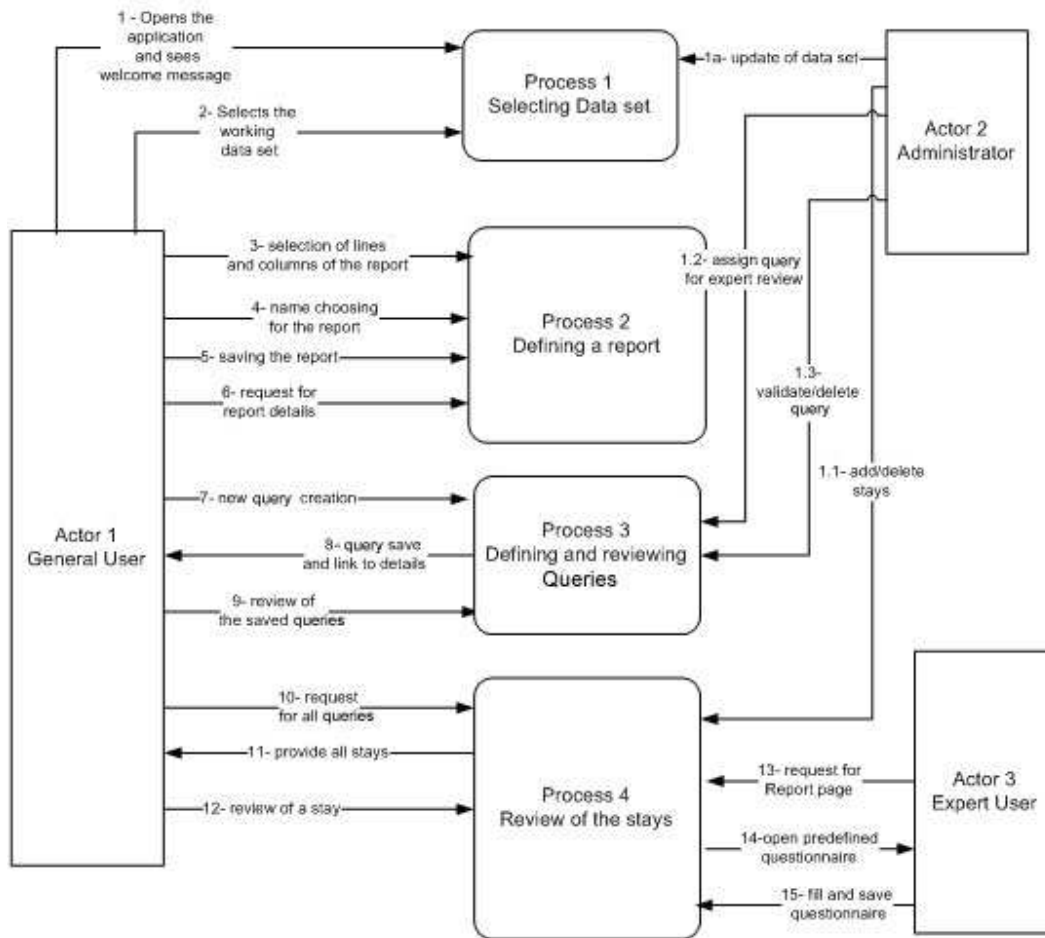


Figure 58. Data flow diagram

10.3. Computational viewpoint

The functional decomposition of the “Expert Explorer” is shown in the simplified UML diagram depicted in Figure 59.

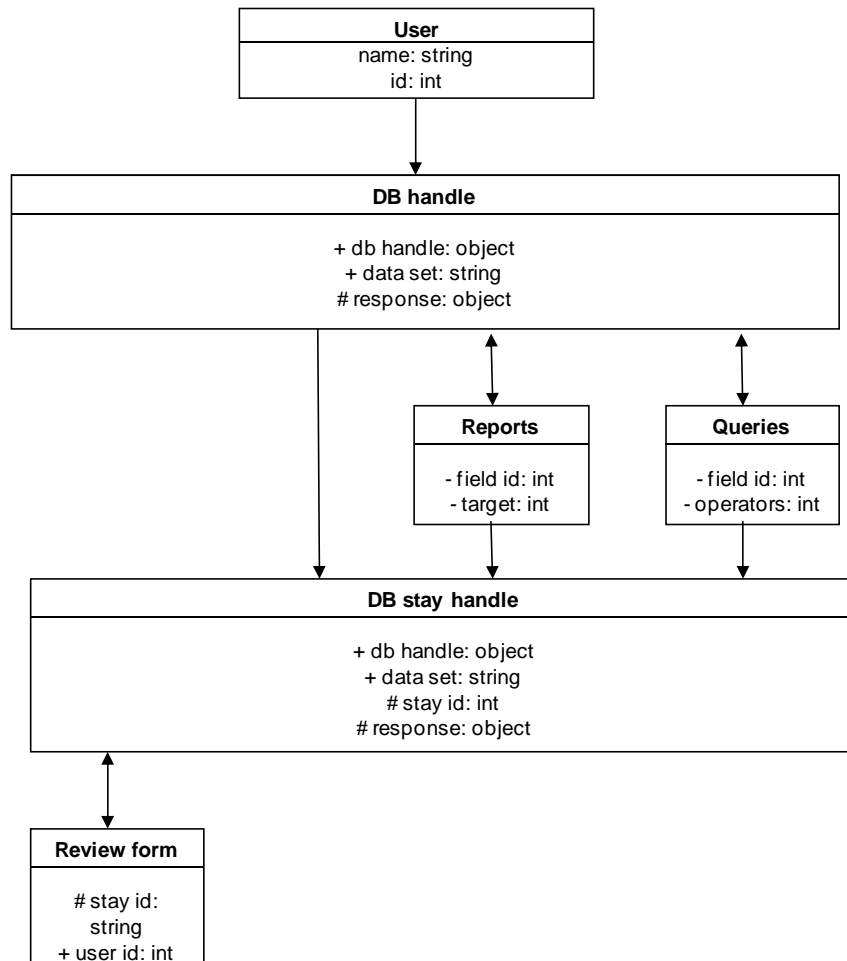


Figure 59. Simplified class diagram

10.4. Engineering viewpoint

The figure shows an overview on the distributed platform and infrastructure. The database contains the data available regarding the stays and defined reports or queries. That information is read by the tool. Then the results are processed and the data is at the moment transformed into an Extensible HyperText Markup Language XHTML file – viewable in a Web browser.

Responses from the tool are stored in the DB. Data set updates can also be performed by external applications, but in either situation, a rigorous check is performed on the data to be loaded and the results of the loading process itself.

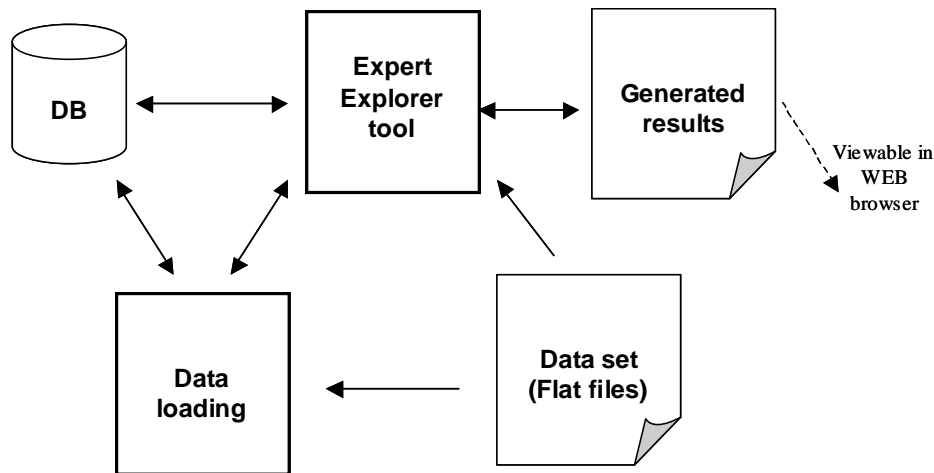


Figure 60. System context diagram showing the engineering overview

10.5. Technology viewpoint

- The application is written using HTML, CSS, PHP5, Javascript. The tool can also be installed on the local intranet of an organization, but it is dependent on a computer network and a server running the required environment
- The tool is hosted on an Apache Server and can be accessed at the web address <http://www.expert-explorer.eu/>
- The application is linked to a MySQL database composed of PSIP data model tables and additional table containing user-related information. As discussed later, an Oracle database can easily be used instead.
- Inside the application an extra table is created, called flat table, which is an aggregation of the other seven tables and has a structure specific to the data set used.

10.6. Detailed description of the interface and use flow

10.6.1. Introduction

One of the available solutions to have a data visualization tool was the Oracle Business Intelligence Suite (OBI). We needed a tool to work specially with the PSIP data model as OBI applies more easily to a Business Intelligence "star-model". It led us to develop a custom solution. This solution allows working especially with the PSIP data model, to generate laboratory results charts, drug charts, and to allow different users to fill questionnaires to evaluate the stays. This solution is a tool easy to use and focused on the goal it was designed for.

The Expert Explorer is a web-based data visualization tool. It allows representing several data from a given stay: medical and administrative information, diagnosis, medical procedures, laboratory results and drug prescriptions. The application also offers a general overview of the medical department. The medical personnel can identify particular cases in the medical unit they have in charge. This is done by some primary statistics such as death rate, distribution by patient gender or the percent of

stays that are taken care of in an ICU. The user can generate reports based on the available details of the stay. He can target specific values of the variables, and see the distribution of the stays in percent or mean values. The application is also a tool used as a support for evaluating the rules defined in the data and semantic mining process.

Using Expert Explorer one can apply primary statistics on the existing data sets: generate reports, update data sets used in the application, define queries and load queries from files, see the details of the stays corresponding to the queries.

The Expert Explorer allows for several tasks:

- report generation: reports design using basic statistics, report publication
- visualization of the hospital stays from various datasets
- queries import: the queries can be then executed as SQL queries
- validation of the stays obtained from data-mining: the expert can view the data-mining-based rules and the related stays, and validate them (or not) using a pre defined questionnaire.

10.6.2. Implementation and availability

Expert Explorer is hosted on a web server and is available to anyone that knows the login credentials.

Data is stored using a MySQL database and it uses the structure of the PSIP data model defined during the first phase of the project. As discussed later, an Oracle database can easily be used instead.

The users that access the application can fit three different profiles:

- general users, or guests that don't need a password
- medical experts that should login with a username and password and
- administrators.

10.6.3. First Page

On the first page, the user is presented with a welcome message, and the general menu.

Under the menu, on the top left, there are links to Login page, and New Account page. Those pages will be presented later. On the right, is displayed the current data set. The data set used, determines from which database the stays will be shown, and on which database the queries will be applied.

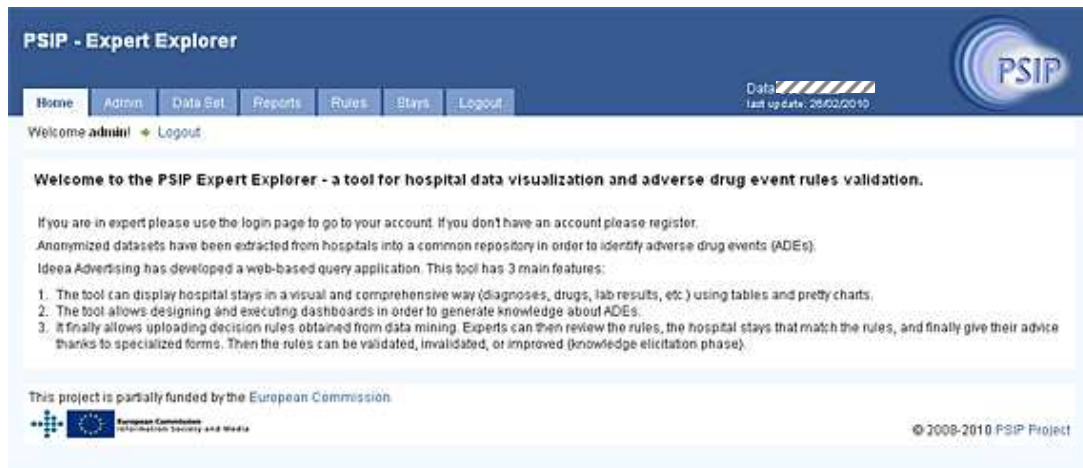


Figure 61. "Expert Explorer" welcome screen

To change the working data set, the user has to click on Data Set from the main menu. He will be presented with a new window where the available data sets are listed.

10.6.4. Data sets

Before any activity, the user should choose the data set that he wants to work on.

On the same page with data set selection, the administrators can update the data set, by uploading to the server the corresponding files or update the flat table.

The flat table is a special table, used for data mining; it represents an aggregated version of the 7 table data set, and doesn't have a fixed number of columns. That is the reason why, every time a new file is loaded, the table is recreated automatically. So, the flat table structure for different data sets could be different. Because of this when defining or loading from files the queries based on the flat the table, those queries will apply only for the current data set.

To change the working data set, simply select it from the list and click Apply button.

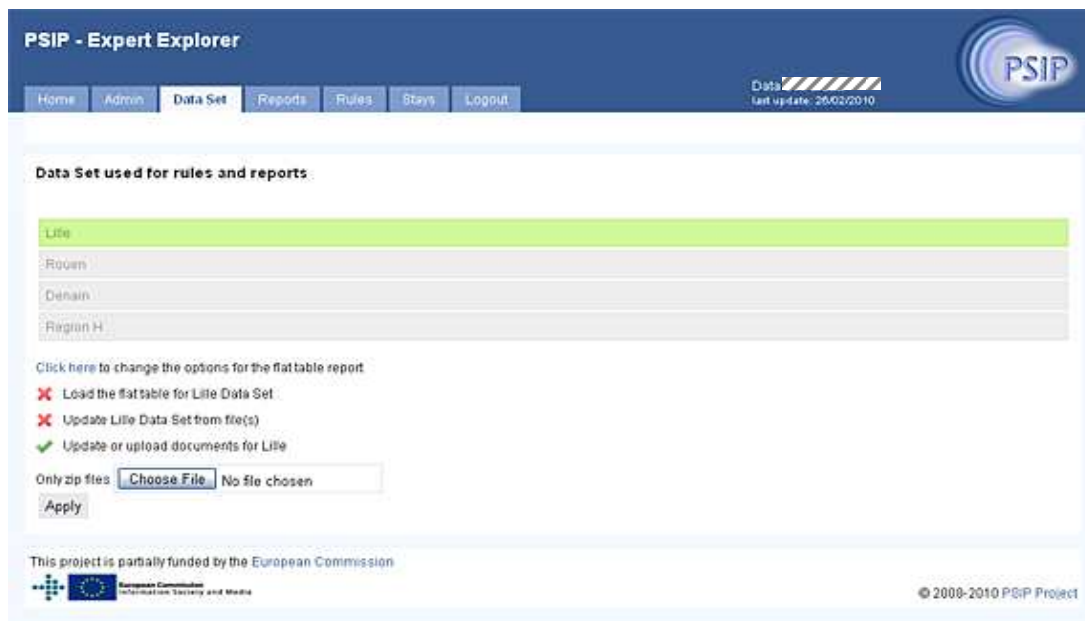


Figure 62. Data set page

The data set page offers three more options. The first one highlighted with a text and a Click here link, will redirect user to the page where he can select which columns from the flat table are displayed in a report in the rule details page. The option is only present if for the current data set (the one displayed under the menu, in the top right side) there is a flat table loaded.

The user is presented with a screen with all the outcome variables from the flat table of the current data set and he has to choose the ones which will be included in the report. X marks a variable that is not present in the report.

PSIP - Expert Explorer

Home Admin Data Set Reports Rules Stars Logout

Data last update: 26/02/2010

Return to previous page

Choose Columns Displayed in Report

<input type="checkbox"/>	mieffdeathbin	<input type="checkbox"/>	mieffcubin	<input type="checkbox"/>	mieffnb_miquant	<input type="checkbox"/>	mieffback_forthbin
<input type="checkbox"/>	mieffnb_thmdquant	<input checked="" type="checkbox"/>	miefftransferbin	<input type="checkbox"/>	dreffhemostatic	<input type="checkbox"/>	dreffhyperERca
<input type="checkbox"/>	dreffhyperERk	<input checked="" type="checkbox"/>	bieffacidose	<input checked="" type="checkbox"/>	bieffcalcalose	<input type="checkbox"/>	bieffanemia
<input type="checkbox"/>	bieffnep_cytolyse	<input type="checkbox"/>	bieffhigh_inr	<input type="checkbox"/>	bieffhyper_alb	<input type="checkbox"/>	bieffhyper_ca
<input type="checkbox"/>	bieffhyper_co2	<input type="checkbox"/>	bieffhyper_eosino	<input type="checkbox"/>	bieffhyper_glyc	<input type="checkbox"/>	bieffhyper_k
<input type="checkbox"/>	bieffhyper_na	<input type="checkbox"/>	bieffhyper_thyr	<input type="checkbox"/>	bieffhypercoag	<input type="checkbox"/>	bieffhypo_alb
<input type="checkbox"/>	bieffhypo_ca	<input type="checkbox"/>	bieffhypo_co2	<input type="checkbox"/>	bieffhypo_glyc	<input type="checkbox"/>	bieffhypo_k
<input type="checkbox"/>	bieffhypo_na	<input type="checkbox"/>	bieffhypo_thyr	<input type="checkbox"/>	bieffhypoocoag	<input type="checkbox"/>	bieffhypovemia
<input type="checkbox"/>	bieffkidney_l	<input type="checkbox"/>	biefflow_inr	<input type="checkbox"/>	bieffthrombopenia	<input type="checkbox"/>	bieffinflammation
<input type="checkbox"/>	bieffleukopenia	<input type="checkbox"/>	bieffleukocytosis	<input type="checkbox"/>	mieffduraonquant	<input type="checkbox"/>	mieffdelay_lcuquant
<input type="checkbox"/>	mieffdelay_lcuquant						

Apply

This project is partially funded by the European Commission

European Commission Information Society and Media

© 2008-2010 PSIP Project

Figure 63. Selecting flat table outcomes

The second option allows user to load the flat table for the data set selected in the list. Not the current one. The user clicks on browse, select the desired file, and then he must click Apply. If no error occurred, the file will be loaded and the total number of records loaded will be displayed.

The third option on the data set page is used to update the current data set, or the data set that is selected from the list. The user needs to load the .txt files containing data. The files must be using the structure of the 7 table data set.

10.6.5. Reports

Defining and visualizing reports was the first function implemented in Expert Explorer. This allows the user to apply some primary statistics on the available data. Can be a useful option for the medical personnel to have an overview of the medical unit they control.

The user has to choose the fields which will be displayed on the lines of the report, and the fields from the columns of the report.



Figure 64. Defining a report

For each line field the target value has to be selected. This is because the report works like this: on each line are selected only the stays that fit the selected value. The columns are used for statistic results. The user has to select the desired field and how the data will be displayed. There can always be added new lines or columns and also choose the order in which the fields are displayed in the report.

A report name can be entered, and then the Generate button will save the report. After the report is saved the user can go to Report Details to see the result.

Or, he can navigate to Saved Reports, from the main menu, and consult a list with all the reports defined so far. From that list, user can select any report and then click on View Report button to see the result of the report.



Figure 65. Saved reports list

The report is displayed in a table structure, with options to sort the table with a click on the column names.



Figure 66. Visualizing a report

If, when the report is defined, user checked the option “Last column used to list patients”, then on the last column there will be a link to a page that is listing all the stays that are identified by report’s criteria.



Figure 67. Stays identified by a report

10.6.6. Rules

This section is used to define queries, one by one, or to load queries from files. Those queries are not properly speaking the data-mining rules. It is an additional feature that can be used by any physician to get some descriptive data.

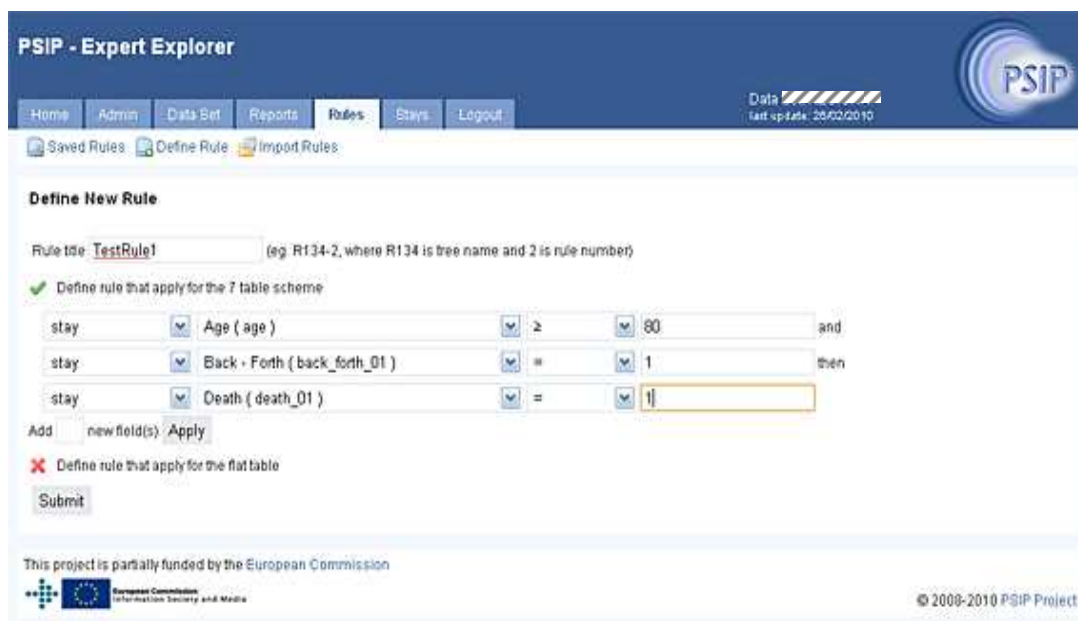


Figure 68. Defining a query

To define a query, the user has to select the fields, the operators and the values to construct an IF - AND - THEN structure which is a set of conditions leading to a result, or an outcome.

To define this structure, for each condition the user has to choose the table where the field is located, the field he targets, the operator and enter the value. The step is repeated for each condition. Values are mandatory and at least two conditions have to be entered. One special operator is REGEXP and it's assumed that the user choosing it knows how to use it (what value to enter).

There are two types of queries definition: the first defines a query using only the fields from the PSIP data model. The second one uses the fields from the flat table.

Using the "Import rules" option the user can load queries from an external file. A query is the combination of several conditions. The uploaded file should contain 1 line per condition.

- first column contains the identifier of the query
- second column contains the variable
- third column contains the comparison operator
- fourth column contains the value to compare with.

The last condition for each query will always be the outcome, the "THEN" part of the query.

Example :

```

r101-1 lab hypoalbuminemia = 1
r101-1 drug vitamin_K_antagonist = 1
r101-1 lab to_high_INR = 1
r101-2 drug proton_pump_inhibitor = 1
r101-2 admin age > 70
r101-2 lab hyponatremia = 1

```

Columns are separated with TAB.



Figure 69. Error handling while importing the queries

If there are two fields in different tables with the same name, the user will be asked to choose the correct field. (eg. In Figure 69 the field **kind** is present in two tables).

When loading queries that are based on the flat table, the application will automatically detect the field, for each condition, and will assign the query only to the current data set.

After the queries are loaded or defined, the user is provided with a link for each query to the page presenting the query details.

But first of all, he can navigate to the saved queries from the main menu.

Rule No.	Rule Definition	Stays	Valid (yes/no)	Remove
1	Age >= 80 and Sex = 1 then Diagnosis REGEXP ^(A10 A119)	See the stays	✗	
2	miEff.back_forth.bin = 1 and mi.count_diags.quantil >= 1 then miEff.death.bin = 1	See the stays	✗	
Test	Age >= 80 and Sex = 1 and Through ICU = 1 then Death = 1	See the stays	✗	
r101-1	Bio(previous hypothyroidism) > 0.5 then Drug(laxative) > 0.5	See the stays	✗	
r101-2	Bio(previous hypothyroidism) > 0.5 and Drug(laxative) < 0.5 then Drug(antibiotic; quinolone) > 0.5	See the stays	✗	
10	Bio(previous too high RR) = 1 and Med(fotage) >= 78.68 and Bio(previous hypoalbuminemia) = 1 then Bio(Low_IIR) = 1	See the stays	✗	
11	Age < 70 and ICU = 1 then Act Delay > 1	See the stays	✗	
12	Age < 70 then Duration > 10	See the stays	✗	

Figure 70. List of existing queries

On the “saved rules” page is the list of all the queries that were defined or loaded for the current data set. For each query there is a link to Query Details page. If an administrator goes to the Saved Queries page, he will have two more options for each query. The administrator can check as query as valid or delete it.

The query details page will display all the stays that are returned by (or satisfy) this query.

Below the list of the stays returned by the query, is a report based on the flat table. The columns used are defined on the Data Set page. The first line is an average of all the values from the flat table, for each column, the second line compute the averages only for the values associated to the group of stays returned by the query.

There is also a button (Download .txt file containing stays). A click on that button will provide a file in .txt file format, using tab as separator, with all the stays’ ids.

On the Query Details page, the administrator will also have an extra option, to assign for review the stays returned by the query to all experts, or only to selected expert.

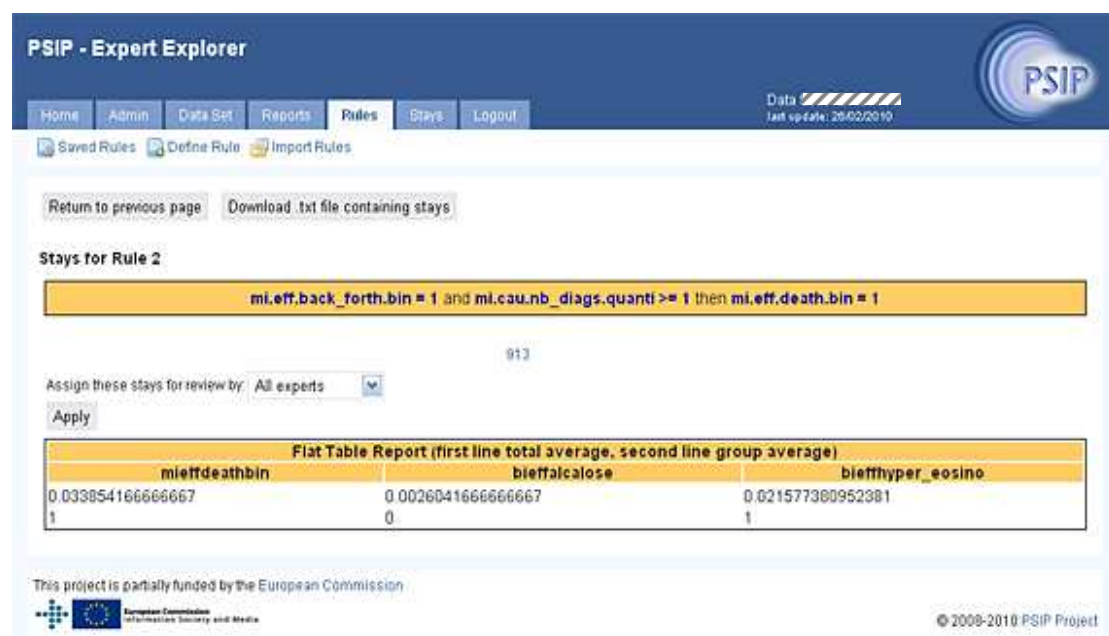


Figure 71. Query detail page

A click on a stay id will bring up the page with all the details for that stay.

10.6.7. Visualization of a stay

A hospital stay can be visualized in several contexts:

- As a standalone feature, the users are allowed to review every stays.
- In the Scorecards (the tool is presented in section 11 in page 186), every implemented rule is associated to a report which displays links to the stays that match the rule.

On the stay details page are displayed all the information available for that stay. The user only has to click on the desired tab. The following tabs are available:

- The first tab displays a description of the stay (demographics, principal diagnosis, length of stay, etc.)

- The second tab displays the different steps of the stay, each step corresponding to a medical unit visited during the stay.
- The third tab displays the medical procedures performed during the step of the stay. The procedures can be either diagnostic of therapeutic.
- The fourth tab shows the ICD10 diagnoses.
- Another tab displays the drug administrations on a tabular form, and another tab presents that information in a comprehensive way by means of a specific chart.
- The laboratory results can be displayed in a tabular form or by means of a specific chart.
- Finally, it is possible to read all the anonymized reports free text documents (e.g. discharge letter).



Figure 72. Visualization of the details for a stay

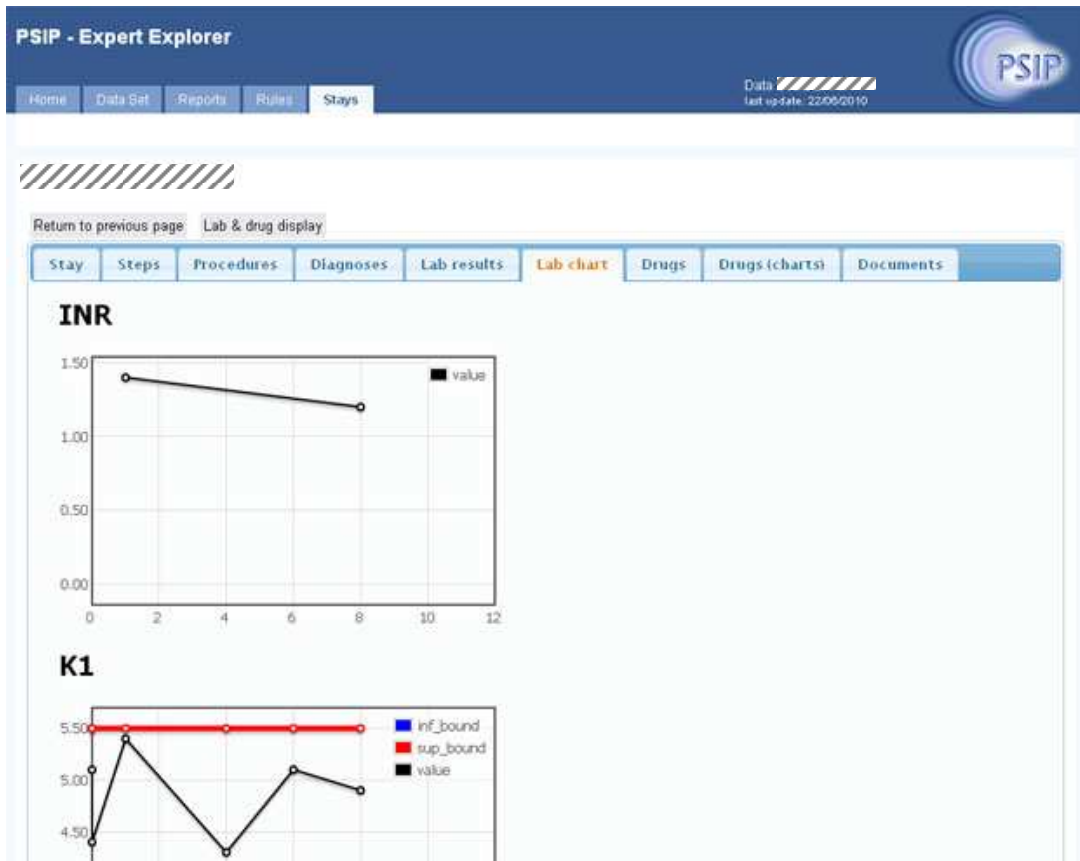


Figure 73. Laboratory charts



Figure 74. Drug charts

In addition to the existing tabs, a button labelled “Lab & drug display” will open a new window, using the full size of the screen, to ease the work of the reviewers by presenting on the same page the drug charts and the laboratory charts. In addition, two buttons on the left-top of the screen will open popups containing essential administrative information and ICD10 diagnoses.

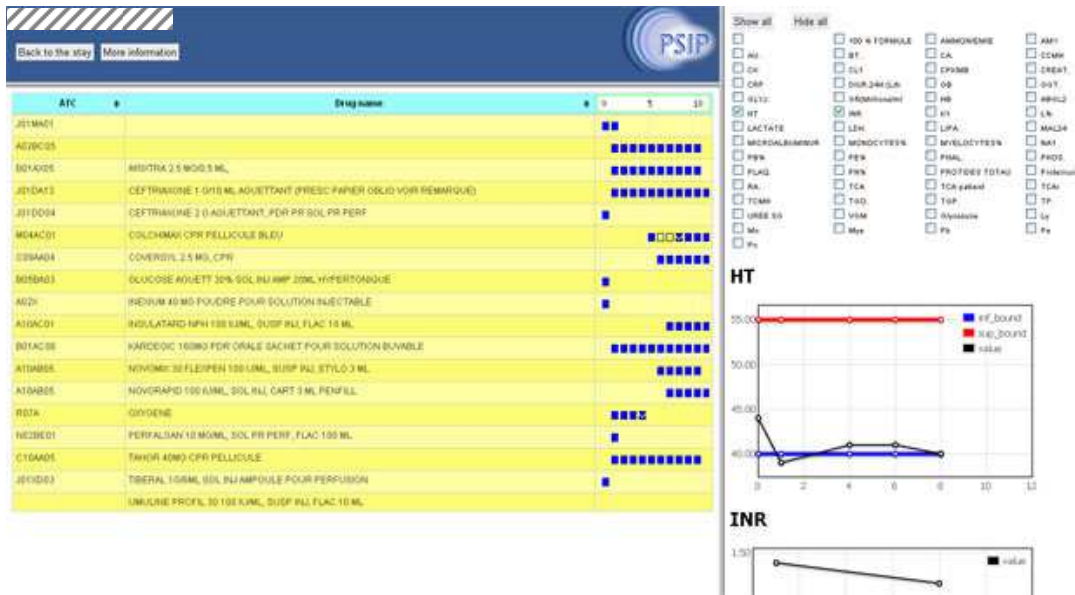


Figure 75. "Lab & drugs charts" window

10.6.8. User accounts

The program is structured on three user levels of access, depending on the role:

- Users (everybody) – can see/visualize data without having the power to modify them; can design and visualize dashboards
- Experts – can see/visualize and interpret the data; as a result to this he may register his advice and validate or invalidate the rules
- Administrator – has total access, the main role being that of implementing new rules and assign them for review; the administrator would also be in charge of updating the database's content

To register an expert, the user must click on the New Account link, and fill in all the fields in the form.

Figure 76. Registering a new user account

After the account is created, the expert should go to Login page, enter his username and password.

A correct login will redirect the user to his home page.

Id Patient	Id Stay	Age	Sex
1955	2458	80.82	male
1531	320	33.98	female
1240	520	85.31	female
504	823	87.03	female
915	1854	65.91	male
736	5	64.3	female
466	1268	72.12	female
604	913	76.42	male
744	1400	40.75	female
2029	1850	78.41	female
1327	1724	47.55	male

Figure 77. Expert home page

The expert's home page contains two tabs: one where are listed the stays that the expert need to review, and another with the stays already reviewed.

10.6.9. Experts' queries validation task

The querying functionality of the Expert Explorer can be used by the physicians to test their own ADE detection rules. In that case, a questionnaire helps them validating the queries as useful or not in ADE detection. Experts will have access to the report page. This page will be accessed by a new button, "Report", displayed only for the experts on the stay details page.

PSIP - Expert Explorer

Home Data Set Reports Rules Stays Logout

Data last update: 26/02/2010

PSIP

Save Rules Define Rule Import Rules

Return to previous page **Report**

Stay	Steps	Procedures	Diagnosis	Lab results	Lab charts	Drugs	Drugs (charts)	Documents
Age	76.4244						Medical units visited	4
Sex	male						Back and Forth between medical units	yes
Death	no						Delay next hospitalisation	-
Death Expectation	-						Transfer to another 'short hospitalisation' hospital	no
Duration	21 days							
Expected Duration	-							
Principal Diagnosis	ICD10 = R999 Mors uden specification						Through ICU	no
Number of different theoretical MDCs (Major Diagnosis Categories)	1						Through ICU Expectation	-
Number of different associated diagnosis	1						ICU Duration	0 days
Number of different acts	5						Expected ICU Duration	-
							Gravity score (SAPS)	-
							Delay before ICU	-

This project is partially funded by the European Commission

European Commission Information Society and Media

© 2008-2010 PSIP Project

Figure 78. Stay details page with the "Report" button present

The expert can review the details of each particular stay and he can go to the report page where the pre-defined questionnaire can be completed, in order to determine if it was or wasn't an Adverse Drug Event.

If the stay has a query associated, then two additional questions about the query are displayed.

If the expert already reviewed that stay, then the questionnaire is automatically completed with his previous answers, and he has the option to modify his answers, or cancel and go to previous page.

On the questionnaire page, a button in the upper part will open the stay's details in a new window, if the expert needs to consult the stay details while completing the form.

PSIP - Expert Explorer

Home Data Set Reports Files Help Logout

Welcome Adria Baccarini - Logout

Reporting Questionnaire: Stay 913 (Hospital 4)

Return to previous page Open stay details (new window)

Question 1: Adverse Event?

Was there an Adverse Event? Yes No Answer impossible Confidence score: Slight to moderate evidence

Question 2: Adverse Drug Event?

Was there an Adverse Drug Event? Yes No Confidence score: Slight to moderate evidence

Did the Adverse Drug Event occurred during the hospital stay? Yes, it occurred during the stay No, it occurred outside the stay Confidence score: Little or no evidence

Question 3: Issue of the ADE?

Did the ADE result in:

Disability: No disability Disability -> rating Answer impossible Non permanent disability

Laboratory abnormality: No Yes Answer impossible

Hematologic disorders:

Anemia Polycythemia
 Leucopenia Leucocytosis
 Thrombopenia Thrombocytosis
 Hypereosinophilia
 Other

Organ abnormality:

Renal insufficiency Hepatic cholestasis
 Hypothyroidism Hepatic cytolysis
 Hyperthyroidism

Blood coagulation disorders:

Too low INR Too high INR
 Other

Respiratory disorders:

Hypoxemia Hypercapnia
 Hypocapnia

Water electrolyte and acid base imbalance:

Hyponatremia Hypermagnesemia
 Hypokalemia Hypocalcemia
 Hypocalcemia Hypocalcemia
 Hypoalbuminemia Hyperalbuminemia
 Alkalosis Acidosis
 Other

Glucose related disorders:

Hypoglycemia Hyperglycemia

Clinical symptoms: Disability symptoms Up to one day of symptoms More days of symptoms Answer impossible

Change in therapy: No Yes Answer impossible

Prolonged hospital stay: No Yes Answer impossible

Optional comments:

Question 4: ADE preventable?

Was the ADE preventable? Yes No Confidence score:

Question 5: ADE description

Core components of the event

Anemia Fall + injury Infection
 Anticholinergic Functional decline Metabolic / endocrine
 Cardiorespiratory GI problems Neurophysiologic
 Dermatologic / allergic Gastrointestinal Renal
 Electrolyte / fluid balance Hematologic Respiratory
 EPS / TD Hepatic Syncope / dizziness
 Fall

Type of ADE

Overdosage Wrong duration Wrong time
 Underdosage Extra dose Omitted medicine or dose
 Wrong strength / concentration Drug - drug interaction
 Wrong drug Drug - food/beverage interaction
 Wrong dosage form Documented allergy
 Wrong route Drug - disease interaction
 Wrong rate (too fast or too slow) Drug - disease interaction
 Other

Drugs involved

Pariv - 10 mg/ml
 Cilastase - 100 mg/ml
 Zimelidon Novum - 10 mg/ml
 Mianserin - 50 mg
 Mianserin - 2.5 mg
 Meloxicam Barbel Viella - 5 mg/ml
 Mide SAG - 5 mg/ml
 Mycostatin - 100.000 E/ml
 Pains - 500 mg
 Promoran - 5 mg/ml
 Serenase - 5 mg/ml
 Stasolid Emulsion - 5 mg/ml
 Transaminexylofen - 100 mg/ml
 Valprocan Novum - 1 Mill. IE
 No drugs involved
 Answer impossible

Explanation:

No information about rate associated with this stay.

Validation

This project is partially funded by the European Commission

© 2008-2010 PSIP Project

Figure 79. The questionnaire

10.6.10. Administration

The administrator can review all the expert accounts. The administrator is the one that assigns rules for review to the experts. This can be done either by uploading a text file containing the stay ids or by going through the previously defined rules and assign the stay that the rules have returned.

	Id Patient	Id Stay	Age	Sex
<input type="checkbox"/>	1955	2458	80.62	male
<input type="checkbox"/>	1531	320	33.98	female
<input type="checkbox"/>	1240	520	85.31	female
<input type="checkbox"/>	504	823	87.03	female
<input type="checkbox"/>	915	1854	65.91	male
<input type="checkbox"/>	736	5	84.3	female
<input type="checkbox"/>	466	1268	72.12	female
<input type="checkbox"/>	604	913	76.42	male
<input type="checkbox"/>	744	1400	40.75	female
<input type="checkbox"/>	2029	1650	78.41	female
<input type="checkbox"/>	1327	1724	47.55	male
<input type="checkbox"/>	241	524	81.57	female
<input type="checkbox"/>	1451	1976	85.3	female
<input type="checkbox"/>	44	297	86.86	female

Figure 80. The administrator home page

To delete stays, the administrator can simply check the stays from the list and click the Remove Stays button.

To add stays, the administrator will need to load a file in the format exported from a Rule Details page, or a similar one, constructed maintaining the structure described.

The administrator can export the questionnaires filled by the experts.

This project is partially funded by the European Commission

© 2008-2010 PSIP Project

Figure 81. Export questionnaires

Also, the administrator has some additional functions available on the pages previous described.

11. APPENDIX 3: ODP DESCRIPTION OF THE SCORECARDS

Authors: Adrian Baceanu & Emmanuel Chazard

11.1. User requirements and enterprise point of view

The results of data mining can be edited in the form of “Periodic ADE Scorecards” sent to the hospital partners. The scope of the “Scorecards” application is to present the statistical results on the occurrence of ADE in order to support the discussion with the healthcare professionals.

The application is accessed only by knowing the corresponding password, based on the role of the user for a selected hospital and medical department:

- physician
- head nurse
- nurses
- pharmacist

Once the application is accessed the user can accomplish activities such as:

- select the analysis period
- consult a summary page for their department
- read individual scorecards
- review the cases.

The fact that the ADE statistics are linked with real cases has an impact on the perception of the tool by the practitioners of a medical department. It is very important for them to be able to review the cases, and to remember them as “what happened in my department last month” and not “what is theoretically known”.

The functional requirements to have access to a specific scorecard are described below.

11.1.1. Use case

- **Goal:** access a generated scorecard for a given outcome and review the cases identified by the rules
- **Actors:** medical personnel (physicians, nurses, head nurse, pharmacists, medical experts)
- **Interface:** a graphical web interface allows interactions with the application. The interface offers support for several languages (English, French, Danish) and can easily be improved to increase the number of the supported languages. The interface doesn't require advanced knowledge to be used and all the actions are performed using only the mouse
- **Preconditions:** the tool must be supplied with the latest XML files containing ADE results, formatted based on the commonly agreed structure. The user has to use the access password

Basic course of events:

1. the user logs in by selecting a hospital, department and typing a password
 2. the user is redirected to the synthesis page, where he selects the period of analysis, then he selects the desired outcomes to generate the scorecards
 3. the user is redirected to a page where he chooses a specific scorecard from the ones generated
 4. after choosing a scorecard, the scorecard page is displayed with all the implemented details when the application was designed
 5. user can review the rules leading to the outcome, characteristics of the stays, the descriptions of each rule and the stays that match the conditions and the outcome of a specific set of rules
 6. the user can go to Expert Explorer to review a selected stay, by using the list of stays from a small pop-up window next to each rule.
- **Alternative paths:**
 - 2a. the user can select one or more outcomes and review the chart with the evolution of the number of stays with adverse events, per month.
 - 6a. the user can access a “case review” page, where the stays that present the outcome are displayed, along with the rules which conditions are met by each stay
 - 6b. the user can go to Expert Explorer to review a selected stay from this page
 - 6c. a medical expert can access the review form to review the medical records.
 - **Post conditions:** the log of actions is stored for later analysis. (In case of review form access, the results are also stored).

11.2. Information view point

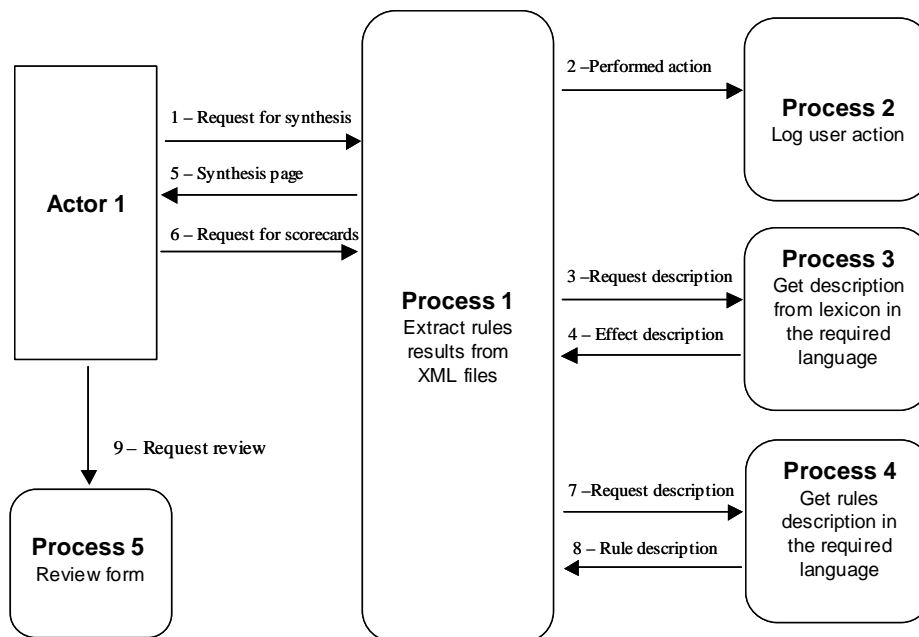


Figure 82. Data flow diagram

11.3. Computational viewpoint

The functional decomposition of the Scorecards is shown in the simplified UML diagram depicted in Figure 83. The XML extraction object is instantiated by supplying reference to the file path of the source rules. The XML extraction class provides methods for retrieval of rules results from XML files. This class creates additional class based on requirements: User Log contains methods for logging users' actions, XML lexicon with methods for retrieving description in natural language for elements like effect code and XML description with methods for extraction of rule descriptions, also in natural language. The User class can create the Review form class, with methods for review generation, validation and save for later analysis.

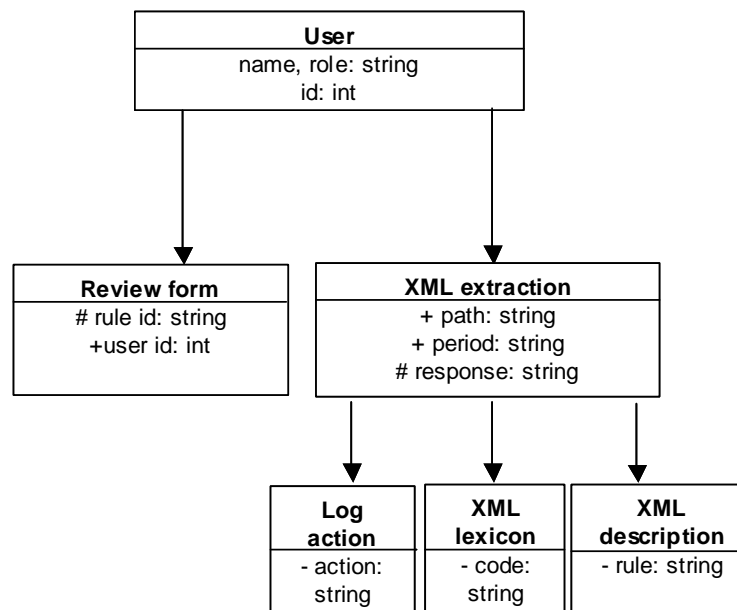


Figure 83. Simplified class diagram

11.4. Engineering viewpoint

The figure shows an overview on the distributed platform and infrastructure. The XML files containing data mining results are read by the tool. Then the results are processed and the data is transformed into an Extensible HyperText Markup Language XHTML file – viewable in a Web browser.

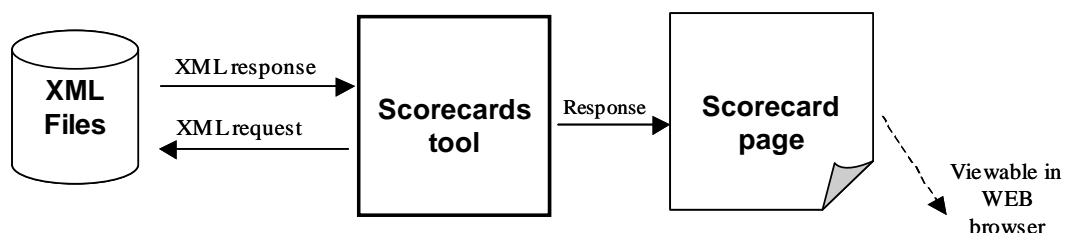


Figure 84. System context diagram showing the engineering overview

11.5. Technology viewpoint

The tool is developed using free Open Source environments like PHP and MySQL. This tool is dependent on a web server like Apache. It could be implemented in either intranet or internet networks, with restricted access based on passwords.

The data sources are the XML files containing the data mining results.

11.6. Detailed description of the interface and use flow

11.6.1. Introduction

The Scorecards tool has been elaborated to present the statistical results on the occurrence of ADE in order to support the discussion with the healthcare professionals. The Scorecards are part of the PSIP “Knowledge Management Platform”.

Periodic ADE reports or “ADE scorecards” are automatically edited and transmitted to the hospitals’ departments. This step takes place after the rules are applied on monthly exports from medical databases of participating hospitals. This screening of medical records allows to retrospectively identify hospital stays susceptible to hold an adverse drug event (ADE).

The ADE scorecards are automatically generated directly from the XML files containing ADE rules. The extraction of the datasets is cumulative, meaning that a report generated for a certain month is the result of the scan of all the stays of the previous months, starting with January, same year.

The Scorecards indicate for each type of identified adverse drug event its prevalence in the relevant department and the mean delay of appearance of the outcome after drug administration. The ADE scorecard provides additional information on the hospitalizations at risk of ADEs including the following: number of cases, patients’ characteristics (gender, age etc.), clinical contexts and categories of diseases. These reports are intended to support quality policy and management (through systematic discussions with healthcare professionals) related to medications in each medical department.

This section illustrates the design of the ADE-Scorecards displays for both the summarized data and detailed statistics and how the tool is supposed to be used.

11.6.2. Interface design and development


The website for displaying scorecards was designed using user-centered design principles [ISO 2009]. From the beginning of this work, an ergonomist was integrated in the designers-developers team to support cooperative design. Moreover, a sample of final users (4 physicians, 2 pharmacists, 3 head nurses, 6 nurses, 1 health care quality manager), participated in the design by commenting on the first mock-ups and the first version of the prototype, choosing features among parallel versions, and proposing new features and/or facilities. Results from the participatory design supported the recommendations used by the developers. A continuous evaluation procedure ensures that recommendations are taken into account [Marcilly 2010].

The tool is available using a web server. It can be accessed only by the person knowing the login credentials. It is developed as a XHTML interface, using PHP as the programming language, XML files and MySQL database for storage.

11.6.3. First page of the Scorecards

The ADE-Scorecards aim to being used by different categories of professionals (e.g. physicians, head nurses, nurses, pharmacists and maybe quality management). By logging in, users are identified (thus, the interface's language is automatically adapted and only the user's department data are displayed) [Marcilly 2010].

The first page of the interface is the login page, where user selects the hospital he wants to study (Figure 85). Depending on the hospital chosen, the user is provided with the departments of that hospital for which data exists. Once selected the department and the hospital, the user has to type a password specific both to the selected department of the hospital chosen and the category of users he is part of.



PSIP Scorecards / Tableaux de Bord

Hospital / Hopital

Department / Service

Password / Mot de passe

Connection / Connexion

Figure 85. Access page for the Scorecards

11.6.4. “Synthesis and Edition of detailed statistics” Page

After login, the user is redirected to the synthesis page. The header of each page of the Scorecards will contain information like the current hospital and department, analysis period, the user using the application, and links used for changing the language, print the page or contact the person responsible with the data.

The Scorecards are designed in such way that they could also be used in a physical environment, printed on paper. Also, the interface and data displayed are available in different languages.



Synthesis

Number of stays with adverse events

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
<input type="checkbox"/> Anemia (Hb<10g/dl)	5	4	4	4	3	5	3	5	5	2	8	5
<input type="checkbox"/> Bacterial infection (prescription of antibiotic)	2	3	2	5	2	6	4	1	3	5	7	6
<input type="checkbox"/> Diarrhoea (prescription of an anti-diarrhoeal)	1	1	0	3	1	3	2	1	1	2	4	2
<input type="checkbox"/> Diarrhoea (prescription of an antipropulsive)	1	0	0	0	0	1	1	0	0	0	0	0
<input checked="" type="checkbox"/> Fungal infection (prescription of a systemic antifungal)	3	3	2	2	3	2	2	1	3	6	1	1
<input type="checkbox"/> Fungal infection (prescription of local antifungal)	0	1	2	0	0	0	0	0	0	1	0	0
<input checked="" type="checkbox"/> Hemorrhage (prescription of hemostatic)	3	1	2	2	3	1	2	3	1	4	2	1
<input checked="" type="checkbox"/> Hemorrhage hazard (INR>4.9)	2	2	5	1	3	1	3	1	1	2	3	2
<input type="checkbox"/> Hemorrhage hazard (activated partial thromboplastin time>1.23)	0	0	0	0	0	0	0	0	0	0	1	0

Figure 86. Synthesis page

In the upper part of the synthesis page there is a table containing all ADEs and all the identified cases, classified by month (Figure 86). The user can select certain outcomes to analyze, using the checkboxes next to each outcome's name. Under the table containing the ADEs there is a graph with the evolution of the number of cases where the selected outcomes were found (Figure 87). The chart is automatically recreated when user selects different ADE outcomes in the table.

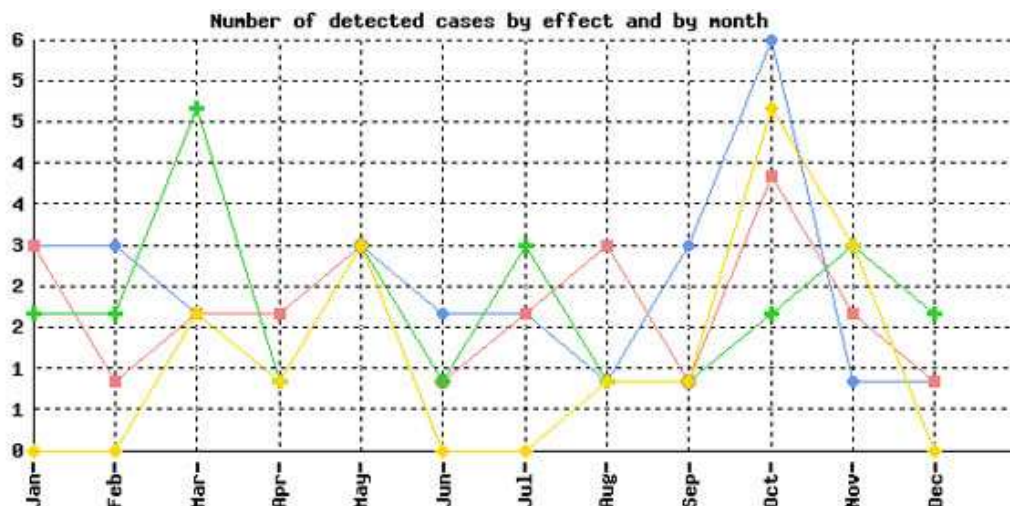


Figure 87. Number of detected cases by outcome and by month

In the lower part of the synthesis page there is a drop down menu allowing users to choose the period for which they want to get ADE statistics. A choice in this menu immediately changes the data displayed in the previous table/graph. Under the select box of the periods, there is a list with the ADE outcomes and the number of stays that

present the outcome (Figure 88). The user has to select the desired outcomes (or all outcomes) for which the scorecards will be generated.

Edit detailed statistics

Analysis period: jan-dec 2007

Detected effects

anemia (Hb<10g/dl) (53)

bacterial infection (prescription of antibiotic) (47)

diarrhoea (prescription of an anti-diarrhoeal) (21)

diarrhoea (prescription of an antipropulsive) (3)

fungal infection (prescription of a systemic antifungal) (29)

fungal infection (prescription of local antifungal) (4)

hemorrhage (prescription of hemostatic) (25)

hemorrhage hazard (INR>4.9) (26)

hemorrhage hazard (activated partial thromboplastin time>1.23) (1)

hepatic cholestasis (alkal. phosphatase>240 U/l or bilirubins>22 µmol/l) (14)

hepatic cytolysis (alanine transa.>110 or aspartate transa.>110) (9)

high a CPK rate (CPK>195 U/l) (3)

hypereosinophilia (éosinophilie>10⁹ /l) (2)

hyperkalemia (K+>5.3) (45)

hypocalcemia (calcemia<2.2 mmol/l) (25)

hyponatremia (Na+<130) (16)

increase of pancreatic enzymes (amylase>90 U/l or lipase>90 U/l) (3)

neutropenia (PNN<1500/mm3) (12)

paracetamol overdose (prescription of acetyl-cystein) (1)

prescription of vitamin K (17)

renal failure (creat.>135 micromol/L or urea>8 mmol/L) (58)

thrombocytosis (count>600,000) (12)

thrombopenia (count<75,000) (7)

Figure 88. Edition of detailed statistics

11.6.5. “Detailed statistics” Page

Once the desired outcomes are selected, and the “Generate Scorecards” button is pressed, user is redirected to the “Scorecard generation page” (Figure 89). In this page user can select an ADE result and view its generated scorecard or detailed statistics page.

Scorecards generated

- hemorrhage (prescription of hemostatic) - [View Scorecard](#)
- hyperkalemia (K+>5.3) - [View Scorecard](#)
- hypocalcemia (calcemia<2.2 mmol/l) - [View Scorecard](#)

Figure 89. Generated scorecards

For each selected adverse effect a page is generated which presents (Figure 90):

- The characteristics of identified stays, all rules together, that describe the sample of stays presenting the adverse effect, including: number of patients concerned, average age, gender proportions, proportions of diseases that might

have impact on ADEs (e.g. alcoholism, cancers, renal insufficiency) and the death rate (these deaths are not necessarily due to the adverse effect).

- Two charts: a bar plot chart representing the distribution of the number of outcomes per month during the current year and a histogram that describes the delay between the outcome and the prescription of the drug in days. The histogram uses the Struges algorithm to define the time classes on the X-axis.

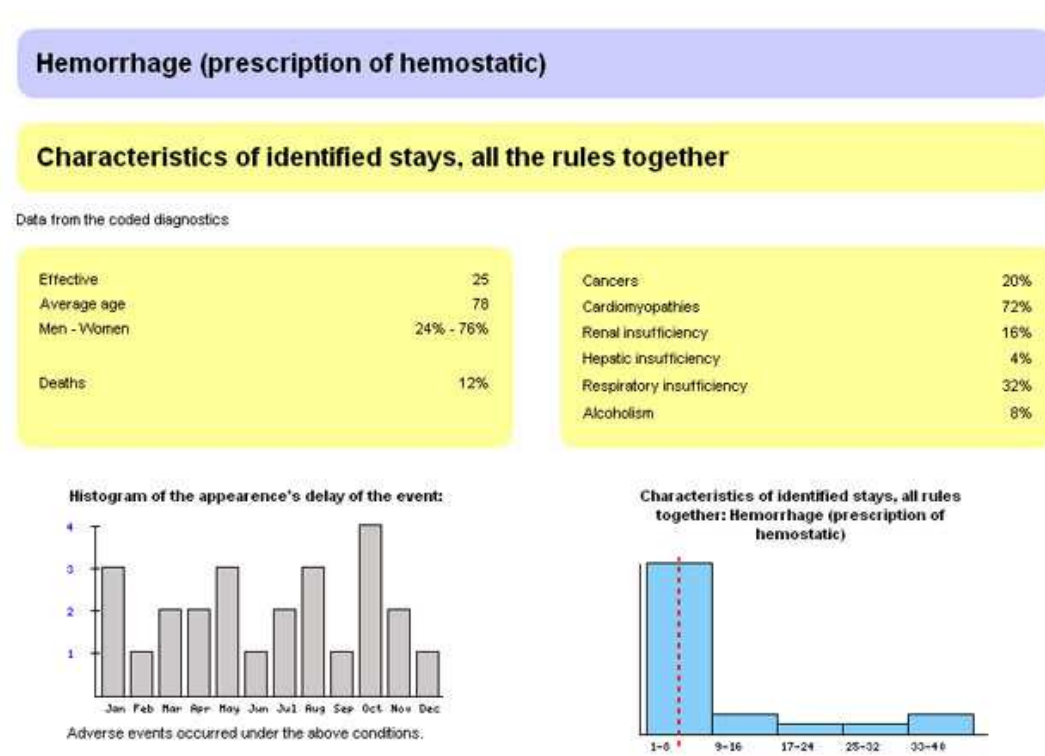


Figure 90. First part of the details page

Next, the page is completed with (Figure 91):

- The conditions (rules' premises: patients' conditions, administered drugs) potentially leading to an adverse effect with their confidence (percentage of stays for which the event occurs among the stays meeting the conditions), their median delay (from the moment when all conditions of the rule are met, the period from which over 50% of events appeared) and the number of stays they target.
- By clicking on the number of cases, a popup window (Figure 92) with the cases that match the conditions and the outcome of the rules opens, grouped by month. Further more, the details of every stay can be seen by clicking on the "View stay details", action that will give access to a synthetic view of the patients' record using an EHR visualization tool named "Expert Explorer" [Băceanu 2009].

Conditions leading to Hemorrhage (prescription of hemostatic)	Nombre de cas détectés confiance ; délai médian
(1) VKA increase the haemorrhagic risk.	18
VKA & Age < 70 & NO Respiratory insufficiency	16% ; 5j
VKA & Age ≥ 70 & NO Respiratory insufficiency	14% ; 3.5j
VKA & Age < 70 & Respiratory insufficiency	13% ; 6j
VKA & Age ≥ 70 & Respiratory insufficiency	12% ; 6j
GLOBAL	25

Confidence (a%): percentage of stays for which the effect occurs among the stays meeting the conditions.
Median delay: from the moment when all conditions of the rule are met, period from which over 50% of effects will be appeared.

Figure 91. Middle part of the details page

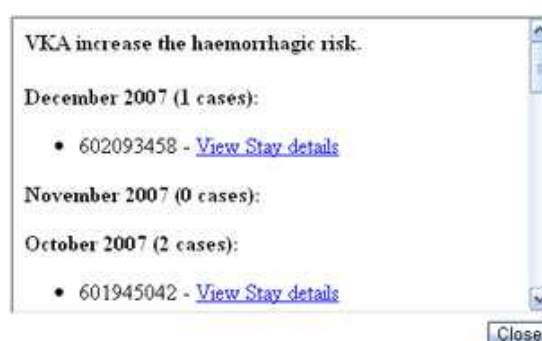


Figure 92. Pop-up window showing the cases

In the lower part of the page there are footnotes for each rule (Figure 93). They are composed of the textual representation of the rule, as a set of conditions and description of the conditions, which may contain a longer description of the rules, scientific explanations and references, and advice. An icon with an arrow pointing to the upper part of the document will move the focus to the start of the page if the users click on it.

Details of rules

[1] VKA increase the haemorrhagic risk.

VKA & Age < 70 & NO Respiratory insufficiency → Hemorrhage (prescription of hemostatic)
VKA & Age < 70 & Respiratory insufficiency → Hemorrhage (prescription of hemostatic)
VKA & Age ≥ 70 & NO Respiratory insufficiency → Hemorrhage (prescription of hemostatic)
VKA & Age ≥ 70 & Respiratory insufficiency → Hemorrhage (prescription of hemostatic)

Vitamin K antagonists increase the haemorrhagic risk.

In case of a vitamin K antagonist treatment, the dosage has to be adapted and the clinical and biological monitoring have to be increased.


 TOP

Figure 93. Details of rules in the bottom of the page

11.6.6. “Review cases” Page

In the header of a Scorecard page, a link called “Review Cases” will redirect the user to another page. While the scorecards help getting drug-related knowledge, this page would be used to list the cases to review, for a given outcome. The interest is that on a scorecard, for each rule we can get all the related stays, but the same stay can match several rules and then appear several times. The aim is here to present each stay only once, and to show for a given stay all the rules that fired.

The “stay-wise” approach can be justified as follows. The Knowledge Elicitation task of the PSIP Project has two objectives: to provide information on ADEs and to provide ADE prevention methods. The Scorecards answer the first objective: they aim of bringing new comprehensive knowledge to the physicians. But the same rules must be validated with respect to the second objective, for a CDSS use. In a CDSS, alerts will be displayed for a unique stay. Several rules can fire at the same time. This corresponds to a “stay-wise” approach to the rules. That is the reason why the validation of the rules is performed “stay-wise” and not “rule-wise”. In addition, for concrete cases, the rules might combine together.

On this page all the stays that are selected at least once in the scorecard for the current outcome are displayed (Figure 94). Below each stay, the rules where the stay was detected are displayed (as set of conditions) and the confidence of the rule follows.

If the user using the tool has certain access rights (like the medical experts), then a specific button is added beside each stay. This button allows for reviewing the stay by means of a specific questionnaire explained in next section.

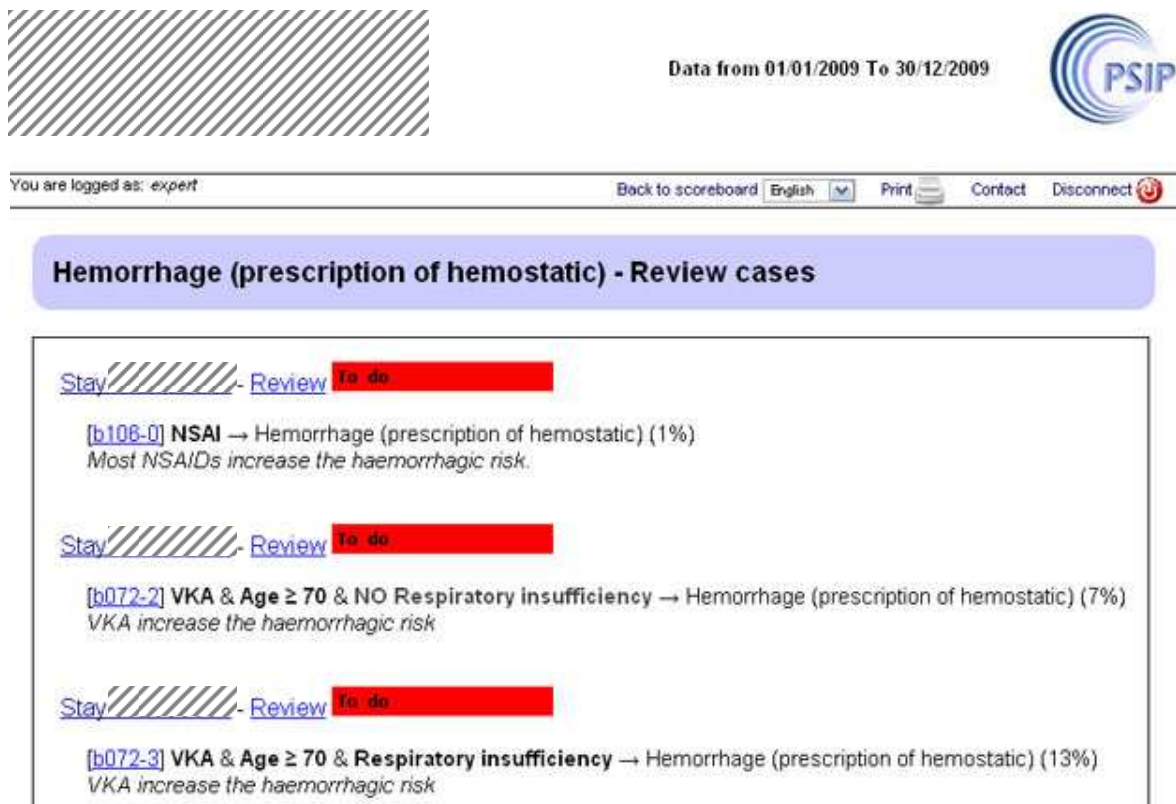


Figure 94. Review cases page

On the bottom of the page (Figure 95), the rules explanations are displayed, similar to the bottom of the scorecard page.

[b077] HMWH increase the haemorrhagic risk

High weight heparin → Hemorrhage (prescription of hemostatic)

High molecular weight heparins increase the haemorrhagic risk

In case of a high molecular weight heparin treatment, the dosage has to be adapted and the clinical and biological monitoring have to be increased.

[b072] VKA increase the haemorrhagic risk

VKA & Age < 70 & NO Respiratory insufficiency → Hemorrhage (prescription of hemostatic)

VKA & Age ≥ 70 & NO Respiratory insufficiency → Hemorrhage (prescription of hemostatic)

VKA & Age ≥ 70 & Respiratory insufficiency → Hemorrhage (prescription of hemostatic)

Vitamin K antagonists increase the haemorrhagic risk

In case of a vitamin K antagonist treatment, the dosage has to be adapted and the clinical and biological monitoring have to be increased.

[b106] Most NSAIDs increase the haemorrhagic risk.

NSAI → Hemorrhage (prescription of hemostatic)

Most non-steroidal anti-inflammatory drugs can increase the haemorrhagic risk by acting on the platelets aggregation.

Monitor the minor bleedings during treatment with NSAIDs.

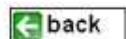




Figure 95. Bottom of the "Review cases" page

11.6.7. "Review cases" questionnaire

For each outcome, and for each stay where at least one rule led to the specified outcome, the user has the possibility to review the stay. A small picture on the "review cases" page indicates the user if the review is undone, in progress or done. Each form has the same structure: an introductive part with useful information, a link to visualize the manual of the form, and the form itself.

The first form (Figure 96) is an introductive form about the specified outcome. It consists of one unique question: is the outcome really present, missing or characterized by an aberrant value?

Medical records revue form		Effect - introduction
Stay	1	
Effect	hemorrhage hazard (INR?4.9)	
User	2	
Be careful Entry already existing in the database. Any submission will modify the previous entry.		



* : mandatory answer * : optional answer ?? Help with data capture ??

Q1 : Effect verification * (only one possible answer)		
Lack of the effect	Aberrant value	Present effect
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
<input type="button" value="confirm"/>		

Figure 96. First form of the questionnaire: existence of the outcome

The second form (Figure 97) is specific to each rule leading to the specified outcome with the current stay. If there are three rules leading to the outcome, the form will be displayed three times, once for each rule. It consists of three questions:

- What are the motives for the non applicability of the rule for the stay?
- According to the user, how important is the contribution of the cause(s) of the rule to the outcome?
- How important is the certainty of the user's answer?

Medical records review form		Rule 1 / 2
Stay	1	
Rule	b001-0	
VKA & 5-HT agonist & NO respiratory obstruction → hemorrhage hazard (INR?4.9)		
User	2	
Be careful Entry already existing in the database. Any submission will modify the previous entry.		

* : mandatory answer * : optional answer ?? Help with data capture ??

Q1 : Motives for the non applicability of the rule on the stay * (several possible answers)							
Lack of a cause	Route of administration	Dose	Too short delay	Too long delay	Chronology	Other	None
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
If the answer is 'Other', point out the motive *							

Q2 : According to you, the cause(s) of the rule * (only one possible answer)						
Do not contribute to the effect	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Fully contribute to the effect
	0	1	2	3	4	
Explanation *						



Q3 : Certainty of your answer * (only one possible answer)						
I am not certain of my answer	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	I am fully certain of my answer
	0	1	2	3	4	
What information are you missing? *						
<input type="button" value="confirm"/>						

Figure 97. Second form of the questionnaire: one form per rule

The third form (Figure 98) is a conclusion about the Adverse Drug Event, defined as a set of causes and an outcome. It aims at evaluation of how much such alerts could be useful in a CPOE context. It consists of two questions:

- For all rules giving rise to this outcome during the stay, does the user think that at least one of the causes contained in the rules has led/participated in the outcome?
- For all rules giving rise to this outcome during the stay, does the user think that the information contained in the set of rules are relevant for understanding the outcome?

For both questions, the user is asked about his certainty of his responses.

Medical records revue form		Effect - conclusion
Stay	1	
Effect	hemorrhage hazard (INR?4.9)	
User	2	
Be careful Entry already existing in the database. Any submission will modify the previous entry.		

* : mandatory answer * : optional answer ?? Help with data capture ??

Q1 : For all rules giving rise to this effect during the stay, do you think that at least one of the causes contained in the rules has led/participated in the effect? * (only one possible answer)						
Not at all	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Quite
	0	1	2	3	4	
Are you sure of your answer? * (only one possible answer)						
I am not certain of my answer	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	I am fully certain of my answer
	0	1	2	3	4	
Q2 : For all rules giving rise to this effect during the stay, do you think that the information contained in the set of rules are relevant for understanding the effect? * (only one possible answer)						
Not at all	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Quite
	0	1	2	3	4	
Are you sure of your answer? * (only one possible answer)						
I am not certain of my answer	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	I am fully certain of my answer
	0	1	2	3	4	
<input type="button" value="confirm"/>						

Figure 98. Third form of the questionnaire: cause-to-effect relationship

12. APPENDIX 4: DESCRIPTION OF THE OUTPUT OF THIS WORK (USE OF THE XML FILES)

This appendix describes the main output of the present work. This output is a set of XML files that are regularly updated and automatically loaded by our partners for several purposes:

- automatic loading into the Scorecards and the Expert Explorer
- automatic loading into a CDSS, so that the same rules can be used by several software programs through a Connectivity Platform.

12.1. Mapping XML files

12.1.1. Overview

Those files allow linking binary variables used for rules execution to native fields of the data model, *e.g.* “too high INR” means “the value of INR is greater than 5”.

There is no difference between the documentation of condition variables and the outcome variables for the mapping step. The use as “condition” or “outcome” is embedded in the rules.

In June 2010 the mapping policies allow to generate 589 binary variables suitable as “condition variables” (48 from diagnoses, 500 from drugs, 36 from laboratory results, 5 from administrative information) and among them 67 allow to generate “outcome variables”.

12.1.2. Diagnosis mapping: mapping_diag.xml

12.1.2.1. Preview

```
<?xml version="1.0" encoding="UTF-8"?>
<variables>
  <variable name="di1.tumor">
    <code value="C000"><![CDATA[Malignant tumor ...]]></code>
    <code value="C001"><![CDATA[Malignant tumor ...]]></code>
    ...
  </variable>
  <variable name="di1.hemato_hemostasis">
    <code value="D65"><![CDATA[Intravascular disseminated coagulation]]></code>
    <code value="D66"><![CDATA[Hereditary deficiency in factor VIII]]></code>
    ...
  </variable>
  ...
</variables>
```

12.1.2.2. Example

Merge together all the ICD10 diagnoses available for the current stay:

- diagnoses of the “diag_step_stay” table

- principal diagnoses of the “step_stay” table
- principal diagnoses of the “stay” table

Please notice that ICD10 codes are used without dot (“.”). If there is at least one of the codes in the list then the binary variable “di1.tumor” is set to 1 for any length of stay. If no code is found, the variable “di1.tumor” is set to 0 for any length of stay.

12.1.2.3. Notes

Also refer to the corresponding section according to the use you intend to have:

- *12.3 - How to implement the rules for a prospective use (transactional use of the CDSS)?* on page 212
- *12.4 How to implement the rules for a retrospective use (retrospective use of the CDSS, dashboards, confidence computation)?* on page 214

For acute diseases (like “di₂.xxx”) use the diagnoses of the current stay. In a prospective use, if available, use the admission ground.

For chronic diseases (like “di₁.xxx”) use all the previously available diagnoses of the current patient. An arbitrary temporal limit can be fixed: only the previous stays that are less than 2 years old can be used.

Many countries use customized versions of ICD10: some codes are forbidden and replaced with more precise codes. Here is how to proceed:

- The visible extensions must not be exported:
 - in Denmark supplementary letters might exist (e.g. A000A) then drop them (e.g. A000)
 - in France supplementary digits following a “+” character might exist (e.g. A000+00) then drop them (e.g. A000)
- However some codes might remain as customized codes. If a code does not match any category, then try to shorten it digit per digit and see if any father code could match a category.

Most of the codes do not match any category.

Nota bene: As diagnosis codes are not stamped with any date, when an acute disease is described using an ICD10 code it is not possible to know if it is the admission ground or if it happened during the stay. This difficulty has been taken into account while designing the mapping policies. The implementation must remain basic.

12.1.3. Drugs mapping: mapping_drug.xml

12.1.3.1. Preview

```
<?xml version="1.0" encoding="UTF-8"?>
<variables>
  <variable name="dr1.addictiveDisorderDrug">
    <code value="N07BA01"><![CDATA[Nicotine]]></code>
    <code value="N07BA02"><![CDATA[Bupropion]]></code>
    <code value="N07BB01"><![CDATA[Disulfirame]]></code>
    <code value="N07BB03"><![CDATA[Acamprosate]]></code>
    <code value="N07BB04"><![CDATA[Naltrexone]]></code>
    <code value="V03AB16"><![CDATA[Ethanol]]></code>
  </variable>
</variables>
```

```

</variable>
<variable name="dr1.adrenalHorm_glucocorticoids">
    <code value="H02AB09"><![CDATA[Hydrocortisone]]></code>
</variable>
...
</variables>

```

12.1.3.2. Example – dr1.* or dr2.*

For a given stay, use the ATC codes of the “drug” table. If there is at least one of the codes in the list then the binary variable is set to 1, else 0.

Example with the *dr1.addictiveDisorderDrug* binary variable: this variable refers to the ATC codes N07BA01, N07BA02, N07BB01, N07BB03, N07BB04 and V03AB16.

If the current stay contains at least one of those codes, then “dr1.addictiveDisorderDrug”=1, else “dr1.addictiveDisorderDrug”=0.

12.1.3.3. Example – dr1.suppr.* or dr2.suppr.*

The drug mapping policy also allows building the **drug discontinuation** variables. The name of the variable is constructed by replacing the first dot “.” by the “**suppr.**” string. For instance, the “dr1.**suppr.**addictiveDisorderDrug” variable is always built:

- if dr1.addictiveDisorderDrug==0 then dr1.**suppr.**addictiveDisorderDrug=0 (no drug => no discontinuation of the drug)
- if dr1.addictiveDisorderDrug==1 then
 - if the drug is still administered during the last available day, then dr1.**suppr.**addictiveDisorderDrug=0
 - if the drug has been discontinued (delay>1day) then dr1.**suppr.**addictiveDisorderDrug=1

12.1.3.4. Notes

Also refer to the corresponding section according to the use you intend to have:

- *12.3 How to implement the rules for a prospective use (transactional use of the CDSS)?* on page 212
- *12.4 How to implement the rules for a retrospective use (retrospective use of the CDSS, dashboards, confidence computation)?* on page 214

12.1.4. Lab results mapping: mapping_lab.xml

12.1.4.1. Important notes

Please notice that:

- If the hospital is able to provide IUPAC codes, then use “ALL” as the hospital name, in order to use the generic mapping mechanism. If the hospital is not able to provide IUPAC codes, a customized mapping has to be designed by the data mining team. In that case, look for the name of the hospital in the mapping file.

- Several different laboratory results anomalies could lead to value 1 of a given binary variable. The results for each binary variable should be given by an “OR” operator or a “MAX” function: variable equals 1 if at least one abnormal value is observed.
- When temporal considerations are required, apply the LOCF (Last Observation Carried Forward) interpolation, i.e. for a given day, in the absence of any new measurement of a laboratory parameter, use the latest available measurement from the previous ones.

The normality bounds are defined with absolute references.

Also refer to the corresponding section according to the use you intend to have:

- *12.3 How to implement the rules for a prospective use (transactional use of the CDSS)?* on page 212
- *12.4 How to implement the rules for a retrospective use (retrospective use of the CDSS, dashboards, confidence computation)?* on page 214

12.1.4.2. Preview

```
<?xml version="1.0" encoding="UTF-8"?>
<variables>
  <hospital name="ALL">
    <variable name="bi.hep_cytolyse" setting="NPU19651" operator="sup" abs_ref="70" />
    <variable name="bi.hep_cytolyse" setting="NPU19654" operator="sup" abs_ref="70" />
    ...
  </hospital>
  <hospital name="rouen">
    <variable name="bi.hep_cytolyse" setting=" ALAT (TGP)" operator="sup" abs_ref="70" />
    <variable name="bi.hep_cytolyse" setting=" ASAT (TGO)" operator="sup" abs_ref="70" />
    ...
  </hospital>
  ...
</variables>
```

12.1.4.3. Example

Let’s take the example of an hospital that produces IUPAC codes. The section of interest is hospital[@name==”ALL”], and we’ll use the “IUPAC” and “value” fields of the data model in the “bio” table.

If you find one measure of NPU019651 where “value”>70 (provided by the condition as the “absolute reference” XML property @abs_ref) then “bi.hep_cytolyse” is set to 1 as long as there is no re-testing or as long as there is still an abnormal laboratory result. The binary variable “bi.hep_cytolyse” is set to 0 elsewhere.

Since the bounds are absolute references, records of NPU019651 with or without upper bounds embedded in the “bio” table can be used.

12.1.4.4. General principles

Operator and reference value:

- if @operator=”inf” then use the “<” comparison operator and use the provided absolute reference @abs_ref.

- if @operator="sup" then use the ">" comparison operator and use the provided absolute reference @abs_ref.

12.1.4.5. Additional laboratory mapping computations

The following computation has to be performed in addition:

*IF bi.thrombopenia==1 & bi.leukopenia==1 & bi.anemia==1
THEN bi.pancytopenia=1 ELSE bi.pancytopenia=0*

Use the maximum of the start dates as the start date.

Use the maximum of the stop dates as the stop date (elsewhere it could lead to wrong associations).

12.2. Rules XML files

12.2.1. Lexicon: lexique.xml

12.2.1.1. Principle

Variables can be replaced with more user-friendly names. The name to use depends on the position of the variable in the rule: condition (use "condition" marquees) or outcome (use "effect" marquees). English, French and Danish labels are provided.

12.2.1.2. Preview

```
<?xml version="1.0" encoding="UTF-8"?>
<lexique>
  <effect id="bi.hypo_na">
    <label language="fr"><![CDATA[apparition d'une hyponatrémie (Na+<130) ]]></label>
    <label language="en"><![CDATA[appearance of hyponatremia (Na+<130) ]]></label>
    <label language="dk"><![CDATA[bi.hypo_na]]></label>
  </effect>
  <effect id="bi.thrombopenia">
    <label language="fr"><![CDATA[Apparition d'une thrombopenia (nb<75 000)]]></label>
    <label language="en"><![CDATA[appearance of thrombopenia (count<75,000) ]]></label>
    <label language="dk"><![CDATA[bi.thrombopenia]]></label>
  </effect>
  ...
  <condition id="bi.hypo_na">
    <label language="fr"><![CDATA[hyponatrémie]]></label>
    <label language="en"><![CDATA[hyponatremia]]></label>
    <label language="dk"><![CDATA[bi.hypo_na]]></label>
  </condition>
  <condition id="bi.thrombopenia">
    <label language="fr"><![CDATA[thrombopénie]]></label>
    <label language="en"><![CDATA[thrombopenia]]></label>
    <label language="dk"><![CDATA[bi.thrombopenia]]></label>
  </condition>
  ...
</lexique>
```

12.2.1.3. Example

The fictive rule:

bi.hypo_na => bi.thrombopenia

Can be written in English as:

hyponatremia => appearance of thrombopenia (count<75,000)

The fictive rule:

bi.thrombopenia => bi.hypo_na

Can be written in English as:

thrombopenia => appearance of hyponatremia (Na+<130)

12.2.1.4. Details

Variable initial names follow a defined convention:

Kind of variable	Initial name
Diagnosis: acute disease	diag2.xxx
Diagnosis: chronic disease	diag1.xxx
Drug	dr1.xxx or dr2.xxx
Drug discontinuation	dr1.suppr.xxx or dr2. suppr.xxx
Lab result	bi.xxx
Medical information	mi1.xxx or mi2.xxx

The strings are included in CDATA containers and encoded in UTF-8. There is no HTML code, so that the string can be used either in HTML pages or graphic user interfaces of software.

12.2.2. Rules repository: rules_yyyy-mm-dd.xml

12.2.2.1. Important notes

The rules are using one to several conditions that are linked together with the “AND” operator. The conditions use several binary variables. The construction of those binary variables is documented in the section 12.1 (*Mapping XML files*) above.

12.2.2.2. Preview [XML structure modified]

In September 2010, the repository contains 236 validated rules. Only the rules where @validated=1 must be used.

```
<?xml version="1.0" encoding="UTF-8"?>
<rules>
  <rule validated="1" effect="bi.high_inr" root="b003" number="b003-0">
    <condition kind="subgroup">
      <field>dr1.antithrombotic_vitKantagonist</field>
      <operator><![CDATA[=]]</operator>
      <ref>1</ref>
    </condition>
```

```

        <condition kind="cause">
            <field>dr1.thymoanaleptic_SSRI</field>
            <operator><![CDATA[=]]></operator>
            <ref>1</ref>
        </condition>
        <condition kind="segmentation">
            <field>di1.resp_obstruc</field>
            <ref>0.5</ref>
            <operator><![CDATA[<]]></operator>
        </condition>
    </rule>
    ...
</rules>

```

12.2.2.3. Example

The preview section above shows a rule. A rule always contains the following attributes:

- **@validated:** only use rules with the “1” value.
- **@number:** a unique numeric identifier. This unique identifier is a string.
- **@root:** the first part of the @number attribute. The free-text comments can be found by means of that attribute.
- **@effect:** the outcome of the rule

Then the rule is a various number of conditions linked together with the AND operator.

Each condition has a unique attribute:

- **@kind:** a string like “subgroup” or “cause” or “segmentation”. Those kinds have no consequence about the implementation of the rules. Briefly speaking:
 - Subgroup conditions help defining a subgroup from which some statistics are computed, such as the relative risk.
 - Cause conditions are the ones that are explained in the comments of the rule
 - Segmentation conditions do not explain why an outcome occurs, but have an impact on the statistics that are computed.

Each condition is composed by 3 elements:

- **field:** the variable name
- **operator:** an operator (<, >, <=, >=, =) contained in a CDATA container. It might have sometimes to be trimmed
- **ref:** a reference value

12.2.3. Pre-rules repository: rules_root_yyyy-mm-dd.xml

This file is not to be used for rule implementation. This file is generated, and then it is automatically transformed into the rules_yyyy-mm-dd.xml file. This file does not contain one record per rules, but one record per root of rules. Then the roots of rules are automatically split into one or several rules.

There are only a few differences with the the rules_YYYY-mm-dd.xml file:

- the @number attribute does not exist, as the @root attribute is a primary key
- for conditions where @kind=segmentation, the @operator attribute is not defined.

A transformation tool is used and automatically creates all the combinations of segmentation conditions, using alternatively “<” and “≥” operators. As a consequence, a root-rule containing k segmentation conditions ($k \geq 0$, generally $k \in \{0,1,2,3\}$) will automatically be split into 2^k rules.

12.2.4. Rules contextualization: rules_result_YYYY-mm-dd.xml

12.2.4.1. Important notes

Rules execution is enforced in all the available medical departments. The support and the confidence of each rule are provided for each department and are not applicable somewhere else. That is contextualization. The occurrences of a given rule can be found using **test_rule@number** as unique identifier.

A different file is computed for each hospital * year. e.g. Denain_2007M12 for the complete year 2007, Denain_2009M11 from January 2009 to November 2009, etc.

Confidence = probability of having the outcome knowing that the conditions are met

Support = probability of having the outcome and matching the conditions at the same time

Confidence = $P(\text{outcome} \mid \text{condition}_1 \cap \dots \cap \text{condition}_k) = \text{numerator} / \text{denominator}$

Support = $P(\text{outcome} \cap \text{condition}_1 \cap \dots \cap \text{condition}_k) = \text{numerator} / \text{total number of cases}$

12.2.4.2. Status of a rule in a specific place

While enforcing each rule to run in a medical department, several cases could occur. The status of the rule is summed up using 0,1,2,3 or 4:

is the outcome variable available?

- NO => 0

- YES => *are the cause variables available?*

- NO => 1

- YES => *do some stays match the causes (denominator>0)?*

- NO => 2

- YES => *do some stays match the causes*

and have the outcome (numerator>0)?

- NO => 3

- YES => 4

12.2.4.3. Preview

```
<?xml version="1.0" encoding="UTF-8"?>
<results>
  <test_partner id="denain_chir" total="5420">
    <test_effect id="bi.kidney_i">
```



```

<test_rule number="b031-0">
  <statut>4</statut>
  <num>2</num>
  <denom>8</denom>
  <total>5420</total>
  <ratio>0.25</ratio>
  <p_fisher>0.00155140994901065</p_fisher>
  <rr>32.2875</rr>
  <id_stay_neg id_hosp="4">601554883; 601633720</id_stay_neg>
  <id_stay id_hosp="4">601554883; 601633720</id_stay>

  <delay><mean>5</mean><sd>2</sd><q0>0</q0><q1>1</q1><q2>3</q2>
  <q3>4</q3><q4>6</q4><q5>12</q5><data>1;1;2;2;3;3;3;3;4;6;8;12</data></delay>
  <trace variable="mi1.age.quanti" mean="77.47" sd="15.1" />
  <trace variable="mi2.death.bin" mean="0" sd="0" />
  <trace variable="mi2.icu.bin" mean="0" sd="0" />
  <trace variable="mi1.gender.bin" mean="0" sd="0" />
  <trace variable="mi2.high_duration.bin" mean="0" sd="0" />
  <trace variable="di1.cancer" mean="0.3333" sd="0.5774" />
  <trace variable="di1.cardiovasc_myocardial" mean="0.6667" sd="0.5774" />

  <trace variable="di1.renal_insuf" mean="0.3333" sd="0.5774" />
  <trace variable="di1.resp_insuf" mean="0" sd="0" />
  <trace variable="di1.hepatic_insuf" mean="0.3333" sd="0.5774" />
  <trace variable="di1.alcool" mean="0.3333" sd="0.5774" />
  <period year="2009" month="01" >
    <id_stay id_hosp="1"></id_stay>
    <num>0</num>
  </period>
  <period year="2009" month="02" >
    <id_stay id_hosp="1">602918577;602942571</id_stay>
    <num>2</num>
  </period>
</test_rule>
<GLOBAL>
  <num>85</num>

  <delay><mean>5</mean><sd>2</sd><q0>0</q0><q1>1</q1><q2>3</q2>
  <q3>4</q3><q4>6</q4><q5>12</q5></delay>
  <trace variable="mi1.age.quanti" mean="67.92" sd="17.7" />
  <trace variable="mi2.death.bin" mean="0" sd="0" />
  <trace variable="mi2.icu.bin" mean="0" sd="0" />
  <trace variable="mi1.gender.bin" mean="0.5" sd="0.5222" />
  <trace variable="mi2.high_duration.bin" mean="0" sd="0" />
  <trace variable="di1.cancer" mean="0.25" sd="0.4523" />
  <trace variable="di1.cardiovasc_myocardial" mean="0.6667" sd="0.4924" />

  <trace variable="di1.renal_insuf" mean="0.08333" sd="0.2887" />

```

```

<trace variable="di1.resp_insuf" mean="0.3333" sd="0.4924" />
<trace variable="di1.hepatic_insuf" mean="0.25" sd="0.4523" />
<trace variable="di1.alcool" mean="0.1667" sd="0.3892" />
</GLOBAL>
...
</ results >

```

12.2.4.4. Example

The fictive rule above can be linked to the [Rules repository: rules_yyyy-mm-dd.xml] file using its *@number* attribute. The values displayed above are valuable only in the “denain_chir” medical department.

- **status=4** means that all the variables were available, some stays matched the conditions and some of them had the outcome (see explanations above).
- **denom=8, num=2, ratio=0.25** mean that 8 stays matched the conditions and 2 of them had the outcome (25 %)
- **total=5420** means that a total of 5420 stays were scanned. This information is available as a marquee nested in the test_rule marquee but also as an attribute in the test_partner marquee. For historical reasons, it is redundant.
- **rr=32, p_fisher=0.0016** mean that the risk was increased by 32 and a Fisher test returned a 0.16 % p value.
- **id_stay_neg** provides the stays that met the conditions at the same time but not the outcome
- **id_stay** provides the stays that met the conditions and the outcomes at the same time (stays for review)
- **delay** provides more information about the delay between (1) the first day where all the causes are present at the same time (2) the day where the outcome finally occurs. The median (refer to the *q2* attribute) of this delay could be useful to predict when the outcome could occur. The information is detailed as follows:
 - **mean** mean
 - **sd** standard deviation
 - **q0** minimum
 - **q1** first quartile (25th percentile)
 - **q2** second quartile (50th percentile, median)
 - **q3** third quartile (75th percentile)
 - **q4** maximum
 - **data** all the delays, sorted and coma-separated (useful for histograms)
- **trace** elements allow to have the average and the standard deviation of some interesting variables
- **period** provides for a specified combination of the year (*@year* attribute, e.g. 2009) and month (*@month* attribute, e.g. 01, 02...12) the stays that met the conditions and the outcomes at the same time. The information is detailed as follows:
 - **id_stay** the stays, separated by a semi-colon
 - **num** the frequency of the matching stays

Those information are not complete when the status is lower than 4.

Once per outcome, a marquee named GLOBAL is added after the rules. Its structure is similar to the structure of the test_rule marquee. It allows for describing some basic statistics related to an outcome. It describes at the same time all the unique stays that have fired at least one validated rule that was able to predict the current outcome. There is no doubleton. The available marquees are:

- **num** number of stays that match at least one of the rules that can lead to the current outcome
- **delay** provides more information about the delay between (1) the first day where all the causes are present at the same time (2) the day where the outcome finally occurs. When a stay fires different rules, the causes might be matched at different times. Only the maximal delay is kept. The information is detailed as follows:
 - **mean** mean
 - **sd** standard deviation
 - **q0** minimum
 - **q1** first quartile (25th percentile)
 - **q2** second quartile (50th percentile, median)
 - **q3** third quartile (75th percentile)
 - **q4** maximum
 - **data** all the delays, sorted and coma-separated (useful for histograms)
- **trace** elements allow to have the average and the standard deviation of some interesting variables

Note that the computation of the Relative Risk and the P Value of the Fisher's Exact Test has changed. When a rule includes some conditions with @kind=subgroup, those conditions are first applied in order to determine the "basal" population. The other stays are dropped. Then, the other conditions are applied and tested.

Example: VKA & PPI => INR>5

Where VKA is a @kind=subgroup condition and PPI is a @kind=cause condition.

First apply the subgroup condition: Basal population: VKA=1

Then apply the other conditions.

Cases: $VKA=1 \cap PPI=1$

Cases with outcome: $INR>5 \cap VKA=1 \cap PPI=1$

Controls: $VKA=1 \cap PPI=0$...and not $VKA=0 \cup PPI=0$

Controls with outcome: $INR>5 \cap VKA=1 \cap PPI=0$

Relative risk: $P(INR>5 | VKA=1 \cap PPI=1) / P(INR>5 | VKA=1 \cap PPI=0)$

and not $P(INR>5 | VKA=1 \cap PPI=1) / P(INR>5 | VKA=0 \cup PPI=0)$

The Fisher's exact test evaluates the independency between PPI=1 and INR>5 only inside the subgroup VKA=1. The test does not evaluate independency between $(VKA=1 \cap PPI=1)$ and INR>5.

12.2.5. Rules explanations: rules_explanations_YYYY-MM-DD.xml

The use of this file is not mandatory. It has been conceived to add some clear explanations to the rules. It meets 3 different needs:

- 1- to help dashboards generation
- 2- to generate some clear documents for pharmacologists or, more generally speaking, for rules sharing.
- 3- to prepare user-friendly outputs for patients in WP10

For that purpose, the document presents a set of HTML labels for the rules. Those labels will be available in different languages (Fr, En, Dk) and for different users (physicians, patients, nurse). The labels are:

- 1- a short label
- 2- a long explanation
- 3- a set of recommendations

The labels are not linked with the rules but with their roots. As a consequence, some rules that share the same root identifier also share the same explanations, as the segmentation conditions, which are the only differences between them, do not add any sense but only add precision to the confidence computation. In the rule XML file, use the **rule@root** and not the rule@number attribute: it will match a **rule@id** marquee of the explanation file.

12.3. How to implement the rules for a prospective use (transactional use of the CDSS)?

12.3.1. General considerations

This “how to” section describes how to use the data and the mapping definitions to execute the rules in a prospective way, i.e. in a real clinical situation when a physician enters a drug prescription and is able to receive alerts or messages. This section allows defining if the rule fires (yes/no). But a concrete implementation of the rules should use the meta-rules as defined in section 5.3.2 (*Meta-rules for the implementation into a CDSS*) on page 142.

12.3.2. “Cause” conditions of the rule

12.3.2.1. Definition of the administrative--related conditions (mi1.*)

No specific remark.

12.3.2.2. Definition of the drug-related conditions (dr1.*)

Use the list of the different drugs that are:

- administered not more than 5 days earlier (i.e. from day-5 to day+0)
- or proposed as a new prescription during day+0

12.3.2.3. Definition of the drug-discontinuation-related conditions (dr1.suppr*)

Use the list of the different drugs that are discontinued not more than 5 days earlier, which means prescribed from day-6 to day-1 and not prescribed anymore or not planned to be prescribed in day+0.

12.3.2.4. Definition of the diagnoses-related conditions (di1.* and di2.*)

Proceed as follows:

- for the diag1.* causes, use all the available diagnoses for the current patient including the past 2 years (the mapping will only keep the chronic diseases, don't care about the acute/chronic distinction)
- for the diag2.* causes, if available, use the admission motive

12.3.2.5. Definition of the lab-result-related causes (bi.*)

Use the **latest available measure** for each parameter, whatever the date. This comes down to perform a last observation carried forward (LOCF) interpolation.

12.3.3. Outcome of the rule

If the rules were applied as is, it wouldn't be necessary to trace the outcome, as we hope it hasn't yet occurred. But several meta-rules as proposed in section 5.3.2 (*Meta-rules for the implementation into a CDSS*) on page 142 could bring interesting information and require the outcome to be traced. The mapping of the outcome variables follows the same process as the mapping of the condition variables.

12.3.4. How to manage several rules that predict the same outcome?

The rules provided by data mining may be redundant, i.e. several rules that predict the same outcome could fire at the same time for different reasons. The redundancy has been reduced by means of automatic split of the roots into rules but still remains. We propose a very simple way to manage that:

- group all the rules by identical outcome
- for a given outcome, execute all the related rules, memorize the output but do not point it out.
- if several rules from the same group have fired, keep only the rule that provides the highest confidence for the current place.

For the confidence calculation, the stays are restricted as follows:

- stays shorter than 2 days are dropped (length of stay=discharge date – admission data; equals zero when the patient is admitted and discharged the same day)
- stays without any drug prescription are dropped (the confidence is not changed as all the rules involve at least one drug)

12.4. How to implement the rules for a retrospective use (retrospective use of the CDSS, dashboards, confidence computation)?

12.4.1. General considerations

⚠ Warning: This section, unlike the prospective part described above, is not involved in triggering alerts. We explain here the techniques that are used by the data mining team to compute the confidences of the rules. This is complex, but has been thought so that the prospective implementation is simple and reliable.

This computing might have two other utilities:

- to recalculate the confidences of the rules with the CDSS
- to generate retrospective dashboards in the Expert Explorer

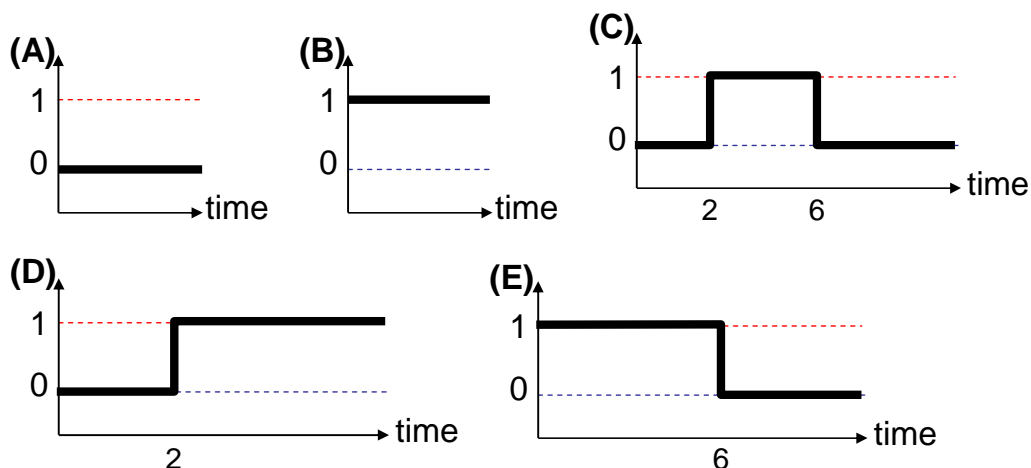
12.4.2. The “event” concept

The data-mining step and the rules-confidence computation use the “event” concept. All the data are aggregated in order to get events. An event is a triplet of 3 values:

- the status of the event (0 / 1)
- the start date of the event (integer / NA)
- the stop date of the event (integer / NA)

This triplet allows describing 5 different shapes (with additional temporal information, day 0 standing for the admission date):

Case	Status	Start	Stop
A	0	-	-
B	1	-inf	+inf
C	1	2	6
D	1	2	+inf
E	1	-inf	6



For the implementation, depending on the language that is used, the [-inf] value can be replaced by a [-1] numeric value, and the [+inf] value can be replaced by a [999] numeric value. The date comparisons will remain valid.

12.4.3. From data to events

12.4.3.1. Variables like mi1.* (administrative information, conditions)

Unlike the other kinds of variables, the status, start and stop dates are defined in the code. Those variables are very special and they are enforced as events in order to guarantee the abstraction of the process although the concept of event is not appropriate.

Variable	Status	Start date	Stop date
Mi1.age.quanti	Stay.age (nb: not binary but float)	-inf	+inf
Mi1.geo_dpt.bin	Stay.geo_dpt_01	-inf	+inf
Mi1.geo_reg.bin	Stay.geo_reg_01	-inf	+inf
Mi1.icu.bin	Stay.through_icu_01	Stay.delay_icu	+inf
Mi1.gender.bin	Stay.sex	-inf	+inf

12.4.3.2. Variables like mi2.* (administrative information, outcomes)

Unlike the other kinds of variables, the status, start and stop dates are defined in the code. Those variables are very special and they are enforced as events in order to guarantee the abstraction of the process although the concept of event is not appropriate. **The implementation is useless** as we finally do not provide rules that lead to those outcomes.

Variable	Status	Start date	Stop date
Mi2.back_forth.bin	Stay.back_forth_01	Stay.duration	+inf
Mi2.death.bin	Stay.death_01	Stay.duration	+inf
Mi2.icu.bin	Stay.through_icu_01	Stay.delay_icu	+inf
Mi2.tardive_icu.bin	If(Stay.through_icu_01==1 and Stay.delai_icu > 3 ; 1 ; 0)	Stay.delay_icu	+inf
Mi2.transfert.bin	Stay.transfer_01	Stay.duration	+inf
Mi2.nb_mu_quanti	Stay.nb_mu (NB: not binary but integer)	Stay.duration	+inf
Mi2.thmdc.bin	If(stay.nb_th_mdc > 1;1;0)	Stay.duration	+inf
Mi2.high_duration.bin	If(stay.duration > stay.duration_exp + 1.96 * stay.duration_sd ;1;0)	stay.duration_exp + 1.96 * stay.duration_sd	+inf
Mi2.high_icu_duration.bin	If(stay.duration_icu > stay.duration_icu_exp + 1.96 * stay.duration_icu_sd ;1;0)	stay.duration_icu_exp + 1.96 * stay.duration_icu_sd	+inf

Mi2.early_rehosp.bin	If(stay.delay_next_hosp > 0 and stay.delay_next_hosp < 10 ;1;0)	Stay.duration	+inf
----------------------	--	---------------	------

12.4.3.3. Variables like *dr1.** and *dr2.** (drugs, conditions or outcomes)

The transformation is generic and simple. Let's assume that a binary *dr*.** variable corresponds to a list (a,b,c,d) of drugs:

- IF at least a drug from the list (a | b | c | d) is administered
- THEN:
 - status = 1
 - start=min(administration date of drugs a,b,c or d)
 - stop=max(administration date of drugs a,b,c or d)
- ELSE
 - status=0
 - start=NA (you can also use start=+inf to simplify some comparisons)
 - stop=NA (you can also use stop=-inf to simplify some comparisons)

We are aware that, proceeding this way, pauses in drug prescription are ignored.

12.4.3.4. Variables like *dr1.suppr.** (drug discontinuations, conditions)

Let *dr1.suppr.mydrug* be the variable of interest. Use the event named *dr1.mydrug* defined beforehand. For the current stay:

- IF status of *dr1.mydrug* ==1
- THEN apply the default values:
 - status = 1
 - start= stop of *dr1.mydrug*
 - stop=+inf
- ELSE
 - status=0
 - start=NA (you can also use start=+inf to simplify some comparisons)
 - stop=NA (you can also use stop=-inf to simplify some comparisons)

12.4.3.5. Variables like *bi.** (lab results, conditions or outcomes)

The data are available in a dataframe *df1* {kind, date, value}, with 1 line per laboratory measurement. Let *df_final* be the dataframe with 1 line par stay where you intend to store the events.

The first step is to transform the dataframe *df1* into a modified dataframe *df2* {variable_name, date, status(0/1)}. In this new dataframe *df2*, there are as many lines as in *df1*, but there are 2 differences:

- the native “kinds” are replaced with the new “variable name” and doing that, there is an aggregation of the lab kinds that help to describe the same lab-related anomaly

- the values are replaced with the status where 0 stands for normal values and 1 stands for abnormal values according to the mapping policy.

Split the dataframe *df2* into 2 subsets:

- *df3* contains only normal values (status=0)
- *df4* contains only abnormal values (status=1)

Use *df4* to define in *df_final* the status, start and stop of the events when the status is set to 1:

- status=1
- start=min(dates for the current variable_name and the current stay in *df4*)
- stop=max(dates for the current variable_name and the current stay in *df4*)

For the stays that are not present in *df4*, write defaults values in *df_final*:

- status=0
- start=NA (you can also use start=+inf to simplify some comparisons)
- stop=NA (you can also use stop=-inf to simplify some comparisons)

Finally use *df3* to realize a LOCF interpolation in *df_final*:

- for each stay and variable_name, restrict *df3* to measures that are posterior to the latest measure that can be found in *df4* using the same criteria
- formally add in *df3* a normal value for every stays and variable_name, using +inf as date
- each time the status is set to 1 in *df_final*, extend the stop date in *df_final* up to the minimal date that remains in the *df3* dataframe minus 1 (first normal measure after the latest abnormal measure, or +inf).

12.4.3.6. Variables like di1.* or di2.* (diagnoses, conditions)

Simply use the aggregation as defined before.

- IF status==1
- THEN formally use widest dates:
 - start= -inf
 - stop=+inf
- ELSE
 - start=NA (you can also use start=+inf to simplify some comparisons)
 - stop=NA (you can also use stop=-inf to simplify some comparisons)

12.4.4. Using events to compute the confidence of a rule

To obtain the confidence of a rule, you have to compute the numerator and the denominator, and then to divide the numerator by the denominator.

12.4.4.1. Computing the numerator of a rule

The numerator (support in absolute number) is the number of stays that match the conditions and the outcome, conditions and outcomes being chronologically compatible. Follow this procedure for each outcome in a temporary dataset:

- First restrict the events that are used as outcomes: an event that occurs before day2 cannot be considered as an outcome
 - IF status of the outcome ==1 & start of the outcome < 2
THEN status of the outcome =0.
- Then examine every conditions of the rule and set to 0 all the conditions that are not chronologically compatible with the outcome:
 - IF status of the condition==1 & start of the condition > start of the outcome THEN status of the condition=0.
 - IF status of the condition==1 & (stop of the condition)+5 < start of the outcome THEN status of the condition=0.

The count of the stays matching at the same time the causes and the outcomes despite the changes above provides the numerator for the current rule.

12.4.4.2. Computing the denominator of a rule

The denominator is the number of stays that match the conditions, but not necessarily the outcome. It is the sum of 2 numbers:

- the number of stays that match the conditions and the outcome, it is the numerator as defined above
- the number X of stays that match the conditions and not the outcome. Here is how to compute X:
 - examine all the stays that match the causes and not the outcome
 - for each stay:
IF min(start dates of the causes)>(max(start dates of the causes)+5)
THEN the causes are not present together, the stay is excluded
ELSE the causes are present together, the stay is included
 - X = number of included stays

13. APPENDIX 5: VALIDATION OF THE USE OF SEMANTIC MINING FOR ADE DETECTION

13.1. Introduction

13.1.1. Objective

In the present work of data-mining-based ADE detection, semantic mining is used to provide drug codes extracted from free-text reports to the data mining task when no CPOE is available. The objective of this appendix is to validate the use of a semantic mining tool, F-MTI, for ADE detection.

13.1.2. Rationale in Semantic Mining evaluation

Semantic Mining is mainly oriented towards automatic indexing. For the evaluation of automatic indexing, different criteria can be measured, according to the literature [Makhoul 1999, Van Rijsbergen 1979, Nakache 2005]. The quality of the automatic indexing is evaluated by comparing the results of this automatic indexation (the candidate set) and the results of a gold standard (the gold standard set) on an evaluation dataset. The gold standard is the manual indexing performed by a human expert. An example is provided in Figure 99.

Discharge letter	Semantic mining	Expert encoding
<i>Dear colleague, Your patient Mrs XX has been admitted in our department in relation with a <u>carpal tunnel syndrome</u> (...) She is known by our department because of her recent history of <u>femur neck fracture</u> (...) Her <u>levothyroxine sodium</u> treatment has been followed up (...)</i>	G56	G56
	S72	(history)
	(not explicit)	E03

*E03: hypothyroidism
 G56: carpal tunnel syndrome
 S72: femur neck fracture*

Precision: semantic mining has found G56&S72 but only G56 is true => P=0.5

Recall: semantic mining should have found G56&E03 but only found G56 => R=0.5

Figure 99. Example of semantic mining applied on a discharge letter; precision and recall computation

For that purpose, different measures are commonly recognized as pertinent:

- **Precision (P)** is the number of indexing terms present in both candidate and gold standard sets divided by the total number of indexing terms in the candidate set. It measures the ratio of signal.
- **Recall (R)** is the number of indexing terms present in both candidate and gold standard sets divided by the total number of indexing terms in the gold standard set. It measures how well gold standard indexing terms are retrieved.

- **F-measure (F)** is the weighted harmonic mean of precision and recall. The traditional F-measure or balanced F-score is: $F = 2 * P * R / (P + R)$ where F is the F-measure, P is the precision and R is the recall.

Supplementary parameters were introduced to add a supplementary weight to precision or recall depending on the task that are to be evaluated:

- **Silence** corresponds to the proportion of terms not extracted (silence=1-Recall; false negatives).
- **Noise** corresponds to the proportion of false terms extracted by the system (Noise=1-Precision; false positives).
- **Purity** evaluates the proportion of indexation mistakes (extraction of a false term) avoided by the system.

In the present work, three main metrics are calculated to show the performance of F-MTI indexing compared to the gold standard manual indexing: Precision, Recall, F-measure. These metrics are often used to evaluate the performances of automatic indexing tools [Pereira 2008, Névéol 2007, Névéol 2006].

13.2. Material and Methods

13.2.1. Evaluation, Step 1: extraction of ATC codes from free-text documents: agreement between F-MTI and experts

The aim of this first phase is to measure the accuracy of the extraction of the drug names included in the various free-text documents by means of the F-MTI Semantic Mining Analyzer. Several de-identified discharge letters are obtained:

- 4,000 from the Rouen University Hospital (F), from which 50 are used for the validation task
- 10,000 from the Denain General hospital (F) , from which 32 are used for the validation task

The drug names extracted by automatic semantic mining (F-MTI) are compared with the ones obtained from human medical expertise (Figure 100).

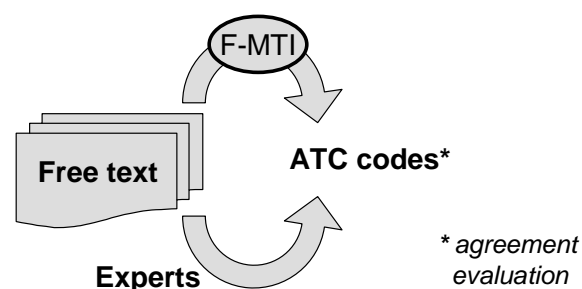


Figure 100. First validation step

In the discharge letters, the drug names appear as brand or commercial names in 90% of cases, or as international names (INN). The list of brand names and INN names available in France are provided by the Vidal Company.

F-MTI indexing tool is used to extract the drug names and index them into ATC Codes: the results are gathered in the candidate set. The gold standard set is the result of the manual indexing performed by a human expert: the gold standard set. Human

experts are a pharmacist and a medical archivist in Rouen; and two physicians in Denain.

In each free-text document, the Experts list:

- the drug names recorded in the document (this is the “gold standard”),
- the drug names extracted by the F-MTI semantic tool.

Those lists are used to compute the precision and the recall.

13.2.2. Evaluation, Step 2- extraction of ATC & ICD10 codes from free text: agreements between F-MTI and EHR

In the Denain General Hospital, both the CPOE and the free-text documents are available. In this phase, the results of the semantic mining of the free-text documents (for the identification of the drugs prescribed or administered to the patient) are compared with the ones registered in the CPOE (Figure 101). This phase allows for computing the concordance between semantic mining analysis results and CPOE extraction for the identification of the drugs potentially linked with ADEs. This phase is only feasible in a hospital equipped both with a Hospital Information System containing the free-text documents and a CPOE System.

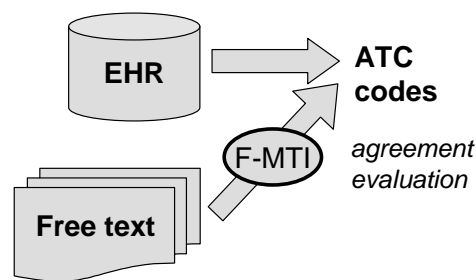


Figure 101. Second validation step

Thirty seven anonymized patients’ complete electronic health records (EHRs) from the Denain General Hospital are used. Those records include:

- data from the EHR and the CPOE: ATC codes for drugs,
- the free-text documents and the results of the automatic indexing of these letters by Semantic Mining (F-MTI): ATC codes.

The Method consists of the careful comparison of the codes obtained from semantic mining of the free-text documents with the codes contained in the CPOE. The recall R and the precision P are computed.

13.2.3. Evaluation, Step 3- validation of the use of the semantic mining results for data-mining-based ADE detection

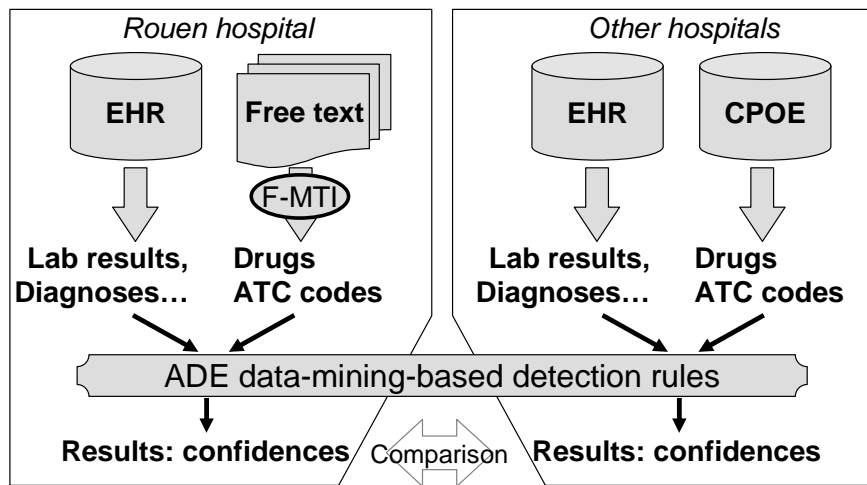


Figure 102. Third validation step

This third validation phase consists of exploring the results of data-mining-based ADE detection rules when drugs are obtained from Semantic Mining of the various free-text documents, in case of absence of CPOE. This is done by studying the frequency of potential ADEs in the Rouen university hospital and comparing this frequency with the ones observed in hospitals where a CPOE is implemented (Copenhagen and Denain) (Figure 102).

The Material is represented by the data-mining-based detection rules obtained from the present work. Those rules are described in the results, in section 0 (

Decision rules integrated in the central rule repository) on page 102. Those rules are a set of conditions that can lead to a traceable ADE. For each rule, the confidence is computed in Denain, Copenhagen and in the Rouen University Hospital where Drugs are obtained from Semantic Mining Analysis. Each rule is characterized by its confidence (1: proportion of outcome knowing that all the conditions are met) and its support (2: proportion of records matching both conditions and outcome).

$$\text{Confidence} = P(O | C_1 \cap \dots \cap C_k) \quad (1)$$

$$\text{Support} = P(O \cap C_1 \cap \dots \cap C_k) \quad (2)$$

The Method consists of the comparison of the confidences of the rules in the different places:

- the Rouen hospital where ATC codes are extracted from summaries,
- the other hospitals where ATC codes are extracted from CPOEs. The datasets from Denain and Copenhagen are pooled together to have only 2 datasets to compare. Moreover, pooling all the other datasets allows to get a better estimate of the confidences of the rules.

For each rule, all the stays that match the conditions of the rule are considered. The aim is then to test the independency between two binary variables using a Fisher's exact test:

- the occurrence of the outcome (0 = "No" / 1 = "Yes"),

- the drug extraction method (CPOE/semantic mining)

For a given rule two results can be obtained:

- if p value < 0.05 then there is a significant difference between the confidence of the rule in Rouen and in other hospitals (the variables are not independent)
- if p value > 0.05 then no significant difference is observed between the confidence of the rule in Rouen and in other hospitals.

None of those results is interesting rule by rule. If significant p value is obtained for one rule, it is not surprising the confidences of the rules depend on the context in which they are used in (the patients, the practices and the knowledge are different). But if most of the rules look like having similar confidences in Rouen than in other places, it is an argument to say that the results of rules evaluation are consistent in Rouen compared with other hospitals.

13.3. Results

13.3.1. Evaluation Step 1- extraction of ATC codes from free-text documents: agreement between F-MTI and experts

The main results in the Rouen university hospital are:

- the overall Precision is **P = 0.84**
- the overall Recall is **R = 0.93**
- the F-measure is **F = 0.88**

The main results in the Denain General Hospital are:

- the overall Recall is **R = 0.88**
- the overall Precision is **P = 0.88**
- the F-measure is **F = 0.88**

These results are coherent although the hospitals use different Hospital Information Systems, employ different physicians and take in care different populations of patients.

They appear as so successful as compare to the literature [Evans 1996, Sirohi 2005, Gold 2008] particularly in the context of the French language where some particular difficulties have to be overcome such as negations, or some verbal passive forms.

13.3.2. Evaluation Step 2- extraction of ATC codes from free text: agreements between F-MTI and EHR

The ATC codes from the semantic mining are considered as “candidates” while the ATC codes from the CPOE are given as “the “gold standard”. The results are:

- the overall Recall is **R = 0.37**,
- the overall Precision is **P = 0.73**,
- the F-measure is **F = 0.49**

13.3.3. Evaluation Step 3- validation of the use of the Semantic Mining results for data-mining-based ADE detection

The comparison between Rouen and other hospitals datasets is performed on each rule separately. An example of rule is provided below:

Rule:

Vitamin K antagonist (VKA) & Antiepileptic & History of too low INR
→ VKA overdose (detected by INR>4.9)

Confidence in Rouen:

6 stays match the conditions, 2 of them present the outcome
Confidence = 33%

Confidence in other hospitals:

206 stays match the conditions, 43 of them present the outcome
Confidence = 21%

Fisher's exact test: p=0.61 (not significant)

In that example, no significant difference is observed between Rouen and other hospitals pooled together. The same method is applied on the 236 validated rules. A significant difference between the pooled confidence and the Rouen confidence can be observed in 48 rules (20.3% of the rules).

If the results in Rouen had been really similar to the results in the other hospitals, the proportion of rules with non significant Fisher's test would have been around 5%.

13.4. Discussion

13.4.1. Ability of F-MTI to extract codes from free-text reports

Predicting ATC codes from the discharge summaries looks successful although that task was performed on unstructured free text: the F-measure is 88%, which is a good score. However, the F-MTI tool still has to be improved. It encounters difficulties to recognize brand names in the discharge summaries due to identified problems that are currently being corrected.

Some additional problems are linked with incorrect spelling of the names in the discharge summaries. Some brand names are written improperly with dash ("-") or underscore ("_") or with an incorrect space " " (e.g. *di-antalvic*, *diffu k*, *di hydan*, *cacit D*, *calcidose vit D*, *co renitec*). On the contrary, some brand names are written without dash ("-") or underscore ("_") or space (" "), as normally they should have to (e.g. *chibroproscar* instead of *chibro-proscar*; *bi preterax* instead of *bipreterax*). Some other misspellings or mistyping are quite frequent (e.g. *triapridal* instead of *tiapridal*, *genopevaryl* instead of *gynopevaril*, *dextropropoxifene* instead of *dextropropoxyfene*, *piperacetam* instead of *piracetam*, *ketoderme* instead of *ketoderm*).

Some mistakes are redundant, e.g. the brand name is *cacit D3*. It is not automatically indexed and *cacit* is indexed instead of it. The same is occurring with *di-antalvic* & *antalvic* and *calcidose Vit D* & *calcidose*.

Some mistakes are more difficult to correct, as they refer to ambiguous terms. For instance in the lab results section of a discharge summary, *Albumin* refers to a lab result, while *Albumin* is also the brand name of a drug. This ambiguity will have to be handled.

13.4.2. Ability of F-MTI to provide ATC codes instead of a CPOE

The comparison between ATC codes extracted from the discharge summaries and ATC codes extracted from the CPOE shows a low agreement: the F-measure is 49%, and this is mainly due to a low recall: 37% (Figure 103). Those bad scores probably attest in fact a poor agreement between the discharge summaries and the information of the CPOE.

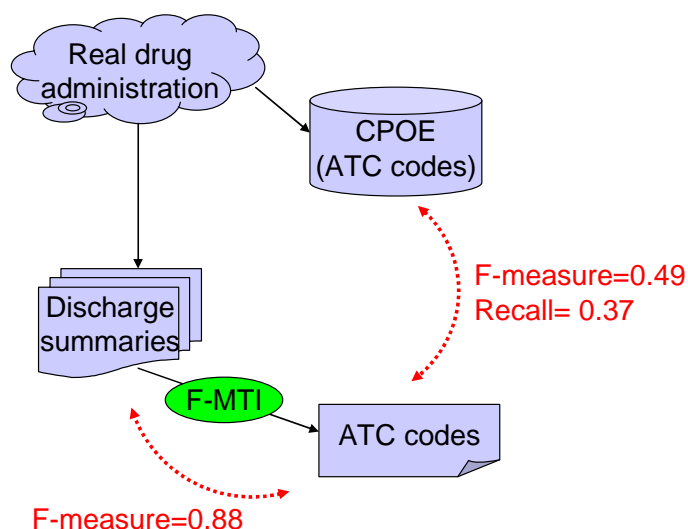


Figure 103. Agreements measured in the 1st and 2nd evaluation steps

Several situations lower the recall in the 2nd validation step:

- Most often, only the main treatments are listed in the discharge summaries. Particularly, temporary treatments are nearly always omitted, such as oxygen, rehydration solutions, pain killers, laxatives, anti-diarrheal drugs, etc.
- Finally, when the patient dies, there is no mention of any treatment in the discharge summary.

Several situations lower the precision in the 2nd validation step:

- The treatment the patient usually takes before the admission is very often mentioned. Though, part of this treatment is often discontinued.
- The treatment the patient is prescribed at discharge is always described, but sometimes those drugs were not prescribed during the stay.
- Finally, in some cases, a patient who suffers from a chronic illness such as diabetes is asked to continue the treatment and to provide himself his drugs during the hospitalisation. In that case, the drugs are not registered in the CPOE but are mentioned in the discharge summary. Such a situation is illegal, and in that case the ATC codes extracted from the discharge summaries are more reliable than those extracted from the CPOE.

13.4.3. Ability of F-MTI to be used for ADE detection

The 3rd validation step shows uncertain results. First the method only compares the confidences of the rules, and 20% of the rules get different confidences in Rouen and in other hospitals, although a proportion around 5% would be expected. As the proportion of ADEs is very low, using several hospitals pooled together for the

control group allowed getting more reliable estimates of the confidences. But in order to conclude, it would be more interesting to test both methods (F-MTI vs. CPOE) in the same hospital, using the same records. This could be performed in a further work.

Using semantic mining on free-text records cannot work in several circumstances:

- When codes are extracted from free text, there is no information about temporal constraints. As a consequence, for a given rule, it is not possible to be sure that the conditions were matched before the outcome occurred.
- As a consequence of the previous point, it is not possible to take drug discontinuations into account as conditions, as in the following rule:
VKA & discontinuation of a laxative → hemorrhage hazard (INR>4.9)
That rule usually works with a 18% confidence.
- It is not possible to use drugs as outcome signals, as it is not possible to be sure that the drug was administered far from the admission, as in the following rule:
Ticarcillin → Fungal infection (detected by administration of antifungal drug)
That rule usually works with a 16% confidence.
- In the discharge summaries there is an insufficient presence of “comfort” drugs which are often involved in pharmacokinetic interactions, as in the following rule:
VKA & acetaminophen → hemorrhage hazard (INR>4.9)
That rule usually works with a 10% confidence.

At the opposite, the method seems to work well in most of the rules, that involve the “main drugs” of the patient and that lead to a laboratory-related outcome that is easy to trace, such as the following rule:

VKA & hypoalbuminemia → hemorrhage hazard (INR>4.9)
That rule usually works with a 20% confidence.

13.5. Conclusion

A semantic mining tool is probably not able to automatically discover ADE prevention rules from past hospital stays. It is not able to prevent ADEs as the discharge summaries and letters are always written after the end of the stay. Nevertheless, semantic mining of those documents can help to retrieve administered drugs in absence of CPOE in order to compute the confidence of ADE detection rules. Doing that, semantic mining of the free-text documents can allow for ADE detection in past hospital stays, but would require an expert-operated review to confirm all the potential ADE cases.

14. APPENDIX 6: VALIDATED RULES

The following tables present the validated rules. The rules are grouped by module, a module being defined as a common outcome. Statistics are limited to the confidence, the delay of appearance and the Fisher's test p value. The statistics are only displayed for each hospital (in alphabetical order: Denain, Frederiksberg, Lille, Nordsjaelland and Rouen). Cells in pink are used to display significant results.

14.1. Anemia (Hb<10g/dl)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b102-0	NO cancer & NSAID	74/1061=7% delay=3j p=0.2119	150/5033=3% delay=3j p=1	6/194=3.1% delay=2.5j p=0.0022	5/2635=0.2% delay=3j p=0	13/851=1.5% delay=3j p=0.0036
b189-0	proton pump inhibitor	35/356=9.8% delay=3j p=0.188	163/3148=5.2% delay=3j p=0	13/308=4.2% delay=3j p=0.0013	7/1633=0.4% delay=3j p=0.0461	11/237=4.6% delay=2j p=0.0837

14.2. Hepatic cholestasis (alkalin phosphatase>240 UI/l or bilirubins>22 µmol/l)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b075-0	NO every traumatism & high weight heparin	3/78=3.8% delay=2j p=0.2476	0/2=0% delay= p=1	1/21=4.8% delay=2j p=0.7205	0/686=0% delay= p=0	0/6=0% delay= p=1
b076-0	NO every traumatism & low weight heparin	10/1014=1% delay=2.5j p=0.0022	14/2031=0.7% delay=4j p=0.2532	5/54=9.3% delay=2j p=0.825	0/1010=0% delay= p=0	2/405=0.5% delay=9.5j p=0.277

b105-0	NSAI	11/1177= 0.9% delay=3j p=0.0005	50/5361= 0.9% delay=5j p=0.9236	20/209= 9.6% delay=2j p=0.5362	0/2788= 0% delay= p=0	5/927= 0.5% delay=8j p=0.037
--------	------	--	--	---	------------------------------------	---

14.3. Hepatic cytolysis (alanine transaminase>110 UI/l or aspartate transaminase>110 UI/l)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b073-0	high weight heparin	0/86= 0% delay= p=1	0/2= 0% delay= p=1	2/24= 8.3% delay=3.5j p=0.2588	32/639= 5% delay=2j p=0.102	0/6= 0% delay= p=1
b074-0	low weight heparin & age < 70	2/600= 0.3% delay=2j p=0.1208	4/943= 0.4% delay=5j p=1	0/26= 0% delay= p=0.6213	23/688= 3.3% delay=2j p=0.6006	1/294= 0.3% delay=17j p=1
b074-1	low weight heparin & age ≥ 70	1/595= 0.2% delay=4j p=0.0272	3/1122= 0.3% delay=2j p=0.4799	0/33= 0% delay= p=0.3948	11/286= 3.8% delay=2j p=0.8741	2/143= 1.4% delay=3.5j p=0.2126
b195-0	proton pump inhibitor	2/364= 0.5% delay=1.5j p=0.5822	19/3192= 0.6% delay=3j p=0.1174	4/318= 1.3% delay=3j p=0.0011	39/1566= 2.5% delay=2j p=0.0028	4/243= 1.6% delay=6j p=0.048

14.4. High a CPK rate (CPK>195 UI/l)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b094-0	statin & age < 70	1/241= 0.4% delay=2j p=1	1/585= 0.2% delay=9j p=1	1/48= 2.1% delay=2j p=0.4404	97/1431= 6.8% delay=2j p=0.301	0/102= 0% delay= p=0
b094-1	statin & age ≥ 70	6/481= 1.2% delay=3j	5/1001= 0.5% delay=3j	0/185= 0% delay=	75/923= 8.1% delay=2j	0/135= 0% delay=

		p=0.0666	p=0.027	p=0.1387	p=0.0107	p=0
--	--	----------	---------	----------	----------	-----

14.5. Hemorrhage hazard (INR>4.9)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b002-0	VKA & thymoanaleptic & NO selective serotonin reuptake inhibitor	10/61= 16.4% delay=6j p=0	0/17= 0% delay= p=1	0/1= 0% delay= p=1	0/3= 0% delay= p=1	0/2= 0% delay= p=1
b003-0	VKA & selective serotonin reuptake inhibitor & NO respiratory obstruction	2/21= 9.5% delay=5j p=0.0376	4/39= 10.3% delay=7.5j p=0	0/1= 0% delay= p=1	0/5= 0% delay= p=1	0/8= 0% delay= p=1
b003-1	VKA & selective serotonin reuptake inhibitor & respiratory obstruction	3/11= 27.3% delay=6j p=0.0005	0/5= 0% delay= p=1	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>
b004-0	VKA & proton pump inhibitor & NO benzamide neuroleptic	9/33= 27.3% delay=4j p=0	0/101= 0% delay= p=1	0/4= 0% delay= p=1	1/35= 2.9% delay=4j p=0.2812	0/4= 0% delay= p=1
b004-1	VKA & proton pump inhibitor & benzamide neuroleptic	0/4= 0% delay= p=1	2/10= 20% delay=2j p=0.0004	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>
b007-0	VKA & quinolone & age < 70	2/11= 18.2% delay=1.5j p=0.0108	0/8= 0% delay= p=1	<i>No stay</i>	1/7= 14.3% delay=3j p=0.0638	<i>No stay</i>
b007-1	VKA & quinolone & age ≥ 70	2/34= 5.9% delay=5.5j p=0.0889	1/20= 5% delay=3j p=0.0603	0/1= 0% delay= p=1	0/3= 0% delay= p=1	<i>No stay</i>
b008-0	VKA & macrolide	5/45= 11.1% delay=5j	1/8= 12.5% delay=2j	<i>No stay</i>	0/1= 0% delay=	0/1= 0% delay=

		p=0.0005	p=0.0246		p=1	p=1
b009-0	VKA & cycline	0/1= 0% delay= p=1	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>
b010-0	VKA & azole antibiotic	3/16= 18.8% delay=5j p=0.0015	0/9= 0% delay= p=1	0/2= 0% delay= p=1	0/2= 0% delay= p=1	0/2= 0% delay= p=1
b011-0	VKA & cephalosporin & age < 70	1/3= 33.3% delay=3j p=0.0435	1/22= 4.5% delay=5j p=0.0661	<i>No stay</i>	1/5= 20% delay=3j p=0.046	<i>No stay</i>
b011-1	VKA & cephalosporin & age ≥ 70	0/5= 0% delay= p=1	5/66= 7.6% delay=4j p=0	0/2= 0% delay= p=1	0/5= 0% delay= p=1	<i>No stay</i>
b012-0	VKA & amoxicilline and clav.ac. & age < 70	6/26= 23.1% delay=4j p=0	0/2= 0% delay= p=1	<i>No stay</i>	1/9= 11.1% delay=3j p=0.0813	<i>No stay</i>
b012-1	VKA & amoxicilline and clav.ac. & age ≥ 70	15/73= 20.6% delay=6j p=0	1/10= 10% delay=6j p=0.0306	0/1= 0% delay= p=1	0/8= 0% delay= p=1	<i>No stay</i>
b013-0	VKA & other beta lactam	2/3= 66.7% delay=10j p=0.0006	0/4= 0% delay= p=1	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>
b014-0	VKA & penicillin & age < 70 & NO diuretic	3/15= 20% delay=4j p=0.0012	0/4= 0% delay= p=1	<i>No stay</i>	1/9= 11.1% delay=3j p=0.0813	0/1= 0% delay= p=1
b014-1	VKA & penicillin & age < 70	3/20= 15% delay=4j p=0.0029	0/5= 0% delay= p=1	<i>No stay</i>	0/12= 0% delay= p=1	<i>No stay</i>

	& diuretic					
b014-2	VKA & penicillin & age ≥ 70 & NO diuretic	9/35= 25.7% delay=4j p=0	2/21= 9.5% delay=4j p=0.0019	No stay	0/2= 0% delay= p=1	0/1= 0% delay= p=1
b014-3	VKA & penicillin & age ≥ 70 & diuretic	9/66= 13.6% delay=7j p=0	1/21= 4.8% delay=1j p=0.0632	0/1= 0% delay= p=1	0/7= 0% delay= p=1	No stay
b015-0	VKA & aminoglycoside	1/4= 25% delay=8j p=0.0576	0/10= 0% delay= p=1	No stay	0/6= 0% delay= p=1	No stay
b016-0	VKA & glycopeptide	1/4= 25% delay=1j p=0.0576	0/3= 0% delay= p=1	No stay	0/2= 0% delay= p=1	No stay
b017-0	VKA & sulfamide	0/5= 0% delay= p=1	0/5= 0% delay= p=1	No stay	0/5= 0% delay= p=1	0/3= 0% delay= p=1
b018-0	VKA & hypoalbuminemia & NO low INR	7/22= 31.8% delay=2j p=0	0/20= 0% delay= p=1	No stay	No stay	No stay
b018-1	VKA & hypoalbuminemia & low INR	4/32= 12.5% delay=7j p=0.0012	0/22= 0% delay= p=1	0/1= 0% delay= p=1	No stay	No stay
b020-0	VKA & systemic antifungal & NO griseofulvin	2/14= 14.3% delay=7.5j p=0.0174	1/12= 8.3% delay=3j p=0.0366	No stay	0/2= 0% delay= p=1	0/2= 0% delay= p=1
b021-0	VKA & type 3 antiarrhythmic & NO diuretic	1/7= 14.3% delay=3j p=0.0987	0/4= 0% delay= p=1	No stay	0/5= 0% delay= p=1	No stay

	& age < 70					
b021-1	VKA & type 3 antiarrhythmic & NO diuretic & age ≥ 70	10/33= 30.3% delay=5j p=0	1/4= 25% delay=5j p=0.0124	<i>No stay</i>	0/1= 0% delay= p=1	<i>No stay</i>
b021-2	VKA & type 3 antiarrhythmic & diuretic & age < 70	1/23= 4.3% delay=4j p=0.2896	0/7= 0% delay= p=1	<i>No stay</i>	0/13= 0% delay= p=1	0/3= 0% delay= p=1
b021-3	VKA & type 3 antiarrhythmic & diuretic & age ≥ 70	8/54= 14.8% delay=8j p=0	2/14= 14.3% delay=3j p=0.0008	0/1= 0% delay= p=1	1/9= 11.1% delay=4j p=0.0813	<i>No stay</i>
b022-0	VKA & antiobesity	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>
b023-0	VKA & thyroid hormone & NO antiepileptic	3/39= 7.7% delay=3j p=0.0193	0/11= 0% delay= p=1	<i>No stay</i>	0/15= 0% delay= p=1	<i>No stay</i>
b023-1	VKA & thyroid hormone & antiepileptic	9/41= 22% delay=5j p=0	0/12= 0% delay= p=1	<i>No stay</i>	2/10= 20% delay=4j p=0.0037	<i>No stay</i>
b025-0	VKA & antiepileptic	38/216= 17.6% delay=5j p=0	7/138= 5.1% delay=6j p=0	0/4= 0% delay= p=1	3/36= 8.3% delay=4j p=0.0045	0/13= 0% delay= p=1
b027-0	VKA & fibrate	1/18= 5.6% delay=2j p=0.2347	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>
b028-0	VKA & anti-gout	2/24= 8.3% delay=3.5j	0/13= 0% delay=	0/1= 0% delay=	0/5= 0% delay=	0/2= 0% delay=

	& NO antiepileptic	p=0.048	p=1	p=1	p=1	p=1
b028-1	VKA & anti-gout & antiepileptic	4/21= 19.1% delay=4.5j p=0.0002	1/10= 10% delay=19j p=0.0306	0/1= 0% delay= p=1	0/2= 0% delay= p=1	<i>No stay</i>
b029-0	VKA & hypothalamo hypophyseal hormone	0/2= 0% delay= p=1	<i>No stay</i>	<i>No stay</i>	0/1= 0% delay= p=1	<i>No stay</i>
b030-0	VKA & antispasmodic	5/20= 25% delay=6j p=0	0/2= 0% delay= p=1	<i>No stay</i>	0/3= 0% delay= p=1	<i>No stay</i>
b031-0	VKA & alcoholism	5/41= 12.2% delay=5j p=0.0003	0/1= 0% delay= p=1	0/2= 0% delay= p=1	0/1= 0% delay= p=1	<i>No stay</i>
b032-0	VKA & tocopherol (vit E)	0/2= 0% delay= p=1	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>
b033-0	VKA & antineoplastic	1/7= 14.3% delay=6j p=0.0987	1/5= 20% delay=2j p=0.0154	<i>No stay</i>	0/2= 0% delay= p=1	<i>No stay</i>
b034-0	VKA & peripheral vasodilatator & NO sympathomimetic drug	4/51= 7.8% delay=3.5j p=0.0065	0/1= 0% delay= p=1	0/2= 0% delay= p=1	0/7= 0% delay= p=1	<i>No stay</i>
b034-1	VKA & peripheral vasodilatator & sympathomimetic drug	3/20= 15% delay=3j p=0.0029	<i>No stay</i>	0/1= 0% delay= p=1	0/1= 0% delay= p=1	<i>No stay</i>
b035-0	VKA & type 1 antiarrhythmic	4/33= 12.1% delay=5j p=0.0013	0/10= 0% delay= p=1	<i>No stay</i>	0/6= 0% delay= p=1	0/2= 0% delay= p=1
b036-0	VKA & hepatic insufficiency	4/24= 16.7% delay=4j	0/6= 0% delay=	0/1= 0% delay=	0/1= 0% delay=	<i>No stay</i>

		p=0.0004	p=1	p=1	p=1	
b037-0	VKA & systemic steroidal anti inflammatory & NO anxiolytic	4/71=5.6% delay=5.5j p=0.0202	0/22=0% delay= p=1	No stay	0/12=0% delay= p=1	0/3=0% delay= p=1
b037-1	VKA & systemic steroidal anti inflammatory & anxiolytic	9/44=20.5% delay=11j p=0	1/16=6.3% delay=5j p=0.0485	No stay	0/6=0% delay= p=1	No stay
b038-0	VKA & anti-diarrheal	9/41=22% delay=3j p=0	2/9=22.2% delay=2j p=0.0003	0/2=0% delay= p=1	0/8=0% delay= p=1	0/1=0% delay= p=1
b040-0	VKA & suspension of anti-diarrheal	4/25=16% delay=3j p=0.0004	1/3=33.3% delay=1j p=0.0093	0/2=0% delay= p=1	0/8=0% delay= p=1	0/1=0% delay= p=1
b043-0	VKA & hypocalcemia & NO low INR	2/12=16.7% delay=4.5j p=0.0129	1/6=16.7% delay=6j p=0.0185	0/4=0% delay= p=1	No stay	0/1=0% delay= p=1
b043-1	VKA & hypocalcemia & low INR	1/18=5.6% delay=7j p=0.2347	0/5=0% delay= p=1	0/2=0% delay= p=1	No stay	No stay
b045-0	VKA & opioid	15/106=14.2% delay=4j p=0	4/125=3.2% delay=6j p=0.0006	No stay	0/18=0% delay= p=1	0/15=0% delay= p=1
b047-0	VKA & acetaminophen/paracetamol & age < 70	3/24=12.5% delay=3j p=0.005	1/49=2% delay=11j p=0.1415	0/1=0% delay= p=1	1/11=9.1% delay=3j p=0.0984	0/12=0% delay= p=1
b047-1	VKA & acetaminophen/paracetamol & age ≥ 70	14/95=14.7% delay=3.5j p=0	6/98=6.1% delay=5j p=0	0/7=0% delay= p=1	0/9=0% delay= p=1	0/15=0% delay= p=1
b048-0	VKA	7/96=7.3%	4/145=2.8%	No stay	1/42=2.4%	0/16=0%

	& NSAID	delay=3j p=0.0005	delay=5.5j p=0.001		delay=4j p=0.3273	delay= p=1
b049-0	VKA & osmotical laxative	18/129= 14% delay=6j p=0	3/58= 5.2% delay=5j p=0.0008	0/3= 0% delay= p=1	0/10= 0% delay= p=1	0/14= 0% delay= p=1
b052-0	VKA & immunomodulation factor	3/12= 25% delay=5j p=0.0006	0/3= 0% delay= p=1	<i>No stay</i>	0/1= 0% delay= p=1	<i>No stay</i>

14.6. Lithium overdose (to high a lithium level)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b128-0	Lithium	0/38= 0% delay= p=0	1/63= 1.6% delay=1j p=0.0105	0/7= 0% delay= p=0	0/55= 0% delay= p=0	0/5= 0% delay= p=0

14.7. Heparin overdose (activated partial thromboplastin time>1.23)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b053-0	high weight heparin & hepatic insufficiency	0/5= 0% delay= p=1	<i>No stay</i>	<i>No stay</i>	0/22= 0% delay= p=0	<i>No stay</i>
b054-0	high weight heparin & chronic renal insufficiency	1/16= 6.3% delay=8j p=0.5112	0/1= 0% delay= p=1	1/5= 20% delay=3j p=0.3246	0/198= 0% delay= p=0	<i>No stay</i>
b055-0	high weight heparin & NSAID	1/21= 4.8% delay=8j p=0.6093	<i>No stay</i>	1/5= 20% delay=3j p=0.3246	0/423= 0% delay= p=0	0/1= 0% delay= p=0

b056-0	high weight heparin & systemic steroidal anti inflammatory	0/16=0% delay= p=1	No stay	2/4=50% delay=2.5j p=0.0304	0/93=0% delay= p=0	No stay
b057-0	high weight heparin & plasma substitutes	1/2=50% delay=3j p=0.0855	No stay	No stay	0/10=0% delay= p=0	No stay

14.8. Hypereosinophilia (éosinophilia>10⁹/l)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b069-0	high weight heparin	1/86=1.2% delay=12j p=0.3186	0/2=0% delay= p=1	0/26=0% delay= p=1	0/699=0% delay= p=1	0/6=0% delay= p=1
b070-0	low weight heparin & age < 70	1/600=0.2% delay=12j p=0.5119	2/941=0.2% delay=4.5j p=0.2418	0/29=0% delay= p=1	0/745=0% delay= p=1	0/294=0% delay= p=1
b070-1	low weight heparin & age ≥ 70	3/595=0.5% delay=3j p=0.7427	5/1117=0.4% delay=5j p=0.8311	0/33=0% delay= p=1	0/300=0% delay= p=1	0/143=0% delay= p=1
b096-0	quinolone	3/414=0.7% delay=4j p=0.4252	14/888=1.6% delay=3j p=0.0002	0/111=0% delay= p=0.1551	0/486=0% delay= p=1	1/38=2.6% delay=8j p=0.0812

14.9. Hyperkalemia (K⁺>5.3)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b059-0	NO renal failure	2/22=9.1%	0/1=0%	0/5=0%	7/182=3.8%	0/5=0%

b060-0	& high weight heparin NO renal failure & low weight heparin & age < 70	delay=1.5j p=0.1274 7/523= 1.3% delay=4j p=0.0271	delay= p=1 2/835= 0.2% delay=5.5j p=0.4436	delay= p=1 0/20= 0% delay= p=0.0961	delay=3j p=0.1698 4/470= 0.9% delay=3j p=0	delay= p=1 1/289= 0.3% delay=5j p=0.3379
b060-1	NO renal failure & low weight heparin & age ≥ 70	11/331= 3.3% delay=6j p=0.6069	0/745= 0% delay= p=0.0322	0/14= 0% delay= p=0.2384	7/76= 9.2% delay=2j p=0.3434	1/117= 0.9% delay=1j p=1
b061-0	NO renal failure & high weight heparin & diabetes	1/3= 33.3% delay=1j p=0.0826	0/1= 0% delay= p=1	No stay	1/15= 6.7% delay=2j p=1	No stay
b062-0	NO renal failure & low weight heparin & diabetes & NO laxative	2/93= 2.2% delay=4.5j p=1	0/39= 0% delay= p=1	No stay	5/93= 5.4% delay=2j p=0.833	0/11= 0% delay= p=1
b062-1	NO renal failure & low weight heparin & diabetes & laxative	3/17= 17.7% delay=13j p=0.0113	0/8= 0% delay= p=1	0/2= 0% delay= p=1	0/4= 0% delay= p=1	0/2= 0% delay= p=1
b063-0	NO renal failure & high weight heparin & angiotensin conversion enzyme inhibitor	1/7= 14.3% delay=1j p=0.1822	No stay	0/1= 0% delay= p=1	2/73= 2.7% delay=2.5j p=0.2376	No stay
b064-0	NO renal failure & low weight heparin & angiotensin conversion enzyme inhibitor	11/197= 5.6% delay=7j p=0.0267	0/176= 0% delay= p=1	0/5= 0% delay= p=1	3/147= 2% delay=2j p=0.026	0/22= 0% delay= p=1
b065-0	NO renal failure & high weight heparin & NSAID	1/6= 16.7% delay=1j p=0.1584	No stay	0/2= 0% delay= p=1	1/101= 1% delay=2j p=0.0142	No stay
b066-0	NO renal failure	10/201= 5%	2/1032= 0.2%	0/5= 0%	6/243= 2.5%	1/289= 0.3%

b083-0	& low weight heparin & NSAID renal failure & high weight heparin	delay=5j p=0.0783 4/59= 6.8% delay=8j p=0.0851	delay=5j p=0.1702 0/1= 0% delay= p=1	delay= p=1 3/17= 17.7% delay=2j p=0.7197	delay=2j p=0.0075 17/410= 4.1% delay=3j p=0.0498	delay=1j p=0.3379 0/1= 0% delay= p=1
b084-0	renal failure & low weight heparin & age < 70	1/73= 1.4% delay=3j p=0.7243	1/101= 1% delay=13j p=0.4056	0/5= 0% delay= p=1	3/208= 1.4% delay=3j p=0.0009	2/5= 40% delay=4j p=0.0009
b084-1	renal failure & low weight heparin & age ≥ 70	15/247= 6.1% delay=3j p=0.0048	9/362= 2.5% delay=5j p=0.0001	1/16= 6.3% delay=4j p=0.7127	6/184= 3.3% delay=2.5j p=0.0699	2/24= 8.3% delay=7.5j p=0.0225
b085-0	renal failure & high weight heparin & diabetes	0/19= 0% delay= p=1	<i>No stay</i>	1/5= 20% delay=4j p=0.5274	4/102= 3.9% delay=3j p=0.4156	<i>No stay</i>
b086-0	renal failure & low weight heparin & diabetes & NO laxative	1/59= 1.7% delay=12j p=1	1/39= 2.6% delay=5j p=0.1816	0/2= 0% delay= p=1	3/104= 2.9% delay=3j p=0.1604	0/1= 0% delay= p=1
b086-1	renal failure & low weight heparin & diabetes & laxative	2/14= 14.3% delay=6.5j p=0.058	0/13= 0% delay= p=1	0/2= 0% delay= p=1	0/18= 0% delay= p=0.6265	<i>No stay</i>
b087-0	renal failure & high weight heparin & angiotensin conversion enzyme inhibitor	0/23= 0% delay= p=1	<i>No stay</i>	2/6= 33.3% delay=3j p=0.1975	9/236= 3.8% delay=3j p=0.1059	0/1= 0% delay= p=1
b088-0	renal failure & low weight heparin & angiotensin conversion enzyme inhibitor	7/125= 5.6% delay=3j p=0.0901	4/122= 3.3% delay=4.5j p=0.0035	1/7= 14.3% delay=4j p=1	4/180= 2.2% delay=3j p=0.0136	2/9= 22.2% delay=5j p=0.0032
b089-0	renal failure	0/14= 0%	<i>No stay</i>	0/3= 0%	12/252= 4.8%	0/1= 0%

b090-0	& high weight heparin & NSAID renal failure & low weight heparin & NSAID	delay= p=1 5/114= 4.4% delay=3j p=0.2576	2/268= 0.7% delay=7.5j p=0.3999	delay= p=1 0/6= 0% delay= p=1	delay=3j p=0.299 7/257= 2.7% delay=3j p=0.0095	delay= p=1 1/14= 7.1% delay=3j p=0.1306
b138-0	NO renal failure & suspension of osmotical laxative	7/134= 5.2% delay=3j p=0.1055	1/487= 0.2% delay=1j p=0.5199	28/147= 19.1% delay=5j p=0.0541	8/259= 3.1% delay=3j p=0.0204	0/131= 0% delay= p=0.6351
b139-0	NO renal failure & suspension of other laxative	9/176= 5.1% delay=6j p=0.0986	0/372= 0% delay= p=0.2663	9/50= 18% delay=5j p=0.4003	3/15= 20% delay=6j p=0.0694	0/177= 0% delay= p=0.2467
b140-0	NO renal failure & suspension of propulsive laxative	0/1= 0% delay= p=1	0/1= 0% delay= p=1	<i>No stay</i>	0/5= 0% delay= p=1	0/2= 0% delay= p=1
b141-0	NO renal failure & peripheral sympatholytic	9/164= 5.5% delay=5j p=0.051	1/214= 0.5% delay=1j p=1	5/54= 9.3% delay=3j p=0.4187	6/235= 2.6% delay=3j p=0.0099	0/31= 0% delay= p=1
b143-0	NO renal failure & beta blocker & NO calcium blocker	7/439= 1.6% delay=4j p=0.1329	2/981= 0.2% delay=29.5j p=0.2376	3/66= 4.5% delay=3j p=0.0252	18/785= 2.3% delay=2.5j p=0	2/118= 1.7% delay=3.5j p=0.3285
b143-1	NO renal failure & beta blocker & calcium blocker	9/113= 8% delay=3j p=0.0045	1/233= 0.4% delay=6j p=1	3/22= 13.6% delay=2j p=1	2/188= 1.1% delay=2j p=0.0005	0/23= 0% delay= p=1
b146-0	NO renal failure & angiotensin conversion enzyme inhibitor & age < 70	5/251= 2% delay=12j p=0.5587	0/486= 0% delay= p=0.1807	2/40= 5% delay=2j p=0.1063	7/876= 0.8% delay=3j p=0	1/70= 1.4% delay=19j p=0.5088
b146-1	NO renal failure & angiotensin conversion enzyme inhibitor & age ≥ 70	17/528= 3.2% delay=6j p=0.5818	5/723= 0.7% delay=3j p=0.4186	9/81= 11.1% delay=3j p=0.6146	7/225= 3.1% delay=2j p=0.0381	3/84= 3.6% delay=12j p=0.0471

b147-0	NO renal failure & potassium sparing diuretic	3/181= 1.7% delay=3j p=0.4919	4/274= 1.5% delay=4j p=0.0514	3/24= 12.5% delay=3j p=1	13/242= 5.4% delay=3j p=0.5952	0/23= 0% delay= p=1
b148-0	NO renal failure & suspension of potassium lowering diuretic	3/150= 2% delay=1j p=0.8012	2/480= 0.4% delay=3j p=1	22/107= 20.6% delay=5.5j p=0.0393	41/574= 7.1% delay=5j p=0.5375	0/55= 0% delay= p=1
b149-0	NO renal failure & potassium	17/726= 2.3% delay=4j p=0.4743	7/1573= 0.4% delay=11j p=0.8499	1/9= 11.1% delay=2j p=1	3/223= 1.3% delay=2j p=0.0004	3/125= 2.4% delay=7j p=0.1219
b152-0	NO renal failure & amoxicilline and clav.ac. & age < 70	3/369= 0.8% delay=3j p=0.0136	0/117= 0% delay= p=1	0/21= 0% delay= p=0.1007	10/388= 2.6% delay=3.5j p=0.0006	0/5= 0% delay= p=1
b152-1	NO renal failure & amoxicilline and clav.ac. & age ≥ 70	11/250= 4.4% delay=2j p=0.1209	0/153= 0% delay= p=1	2/38= 5.3% delay=4j p=0.151	1/41= 2.4% delay=3j p=0.5194	0/1= 0% delay= p=1
b155-0	NO renal failure & suspension of sulfamid or sulfonamid	1/31= 3.2% delay=16j p=0.5904	0/75= 0% delay= p=1	2/35= 5.7% delay=7j p=0.2132	10/163= 6.1% delay=4j p=1	0/10= 0% delay= p=1
b156-0	NO renal failure & NSAID & age < 70	7/340= 2.1% delay=4j p=0.4994	2/2232= 0.1% delay=5.5j p=0.0007	0/24= 0% delay= p=0.0645	10/931= 1.1% delay=2.5j p=0	1/586= 0.2% delay=32j p=0.0111
b156-1	NO renal failure & NSAID & age ≥ 70	13/410= 3.2% delay=6j p=0.6429	3/1757= 0.2% delay=7j p=0.0282	4/64= 6.3% delay=3j p=0.0901	8/233= 3.4% delay=2j p=0.0576	5/258= 1.9% delay=6j p=0.1622
b158-0	NO renal failure & suspension of systemic steroidal anti inflammatory	6/131= 4.6% delay=5j p=0.2757	3/330= 0.9% delay=5j p=0.2383	3/94= 3.2% delay=6j p=0.0008	28/608= 4.6% delay=6.5j p=0.0485	0/192= 0% delay= p=0.2463
b159-0	NO renal failure & digitalis glycoside	6/137= 4.4% delay=8j p=0.2868	3/572= 0.5% delay=7j p=0.7682	2/18= 11.1% delay=2j p=1	2/97= 2.1% delay=2.5j p=0.093	2/52= 3.8% delay=8.5j p=0.0919

b160-0	NO renal failure & immunomodulation factor	1/52= 1.9% delay=2j p=1	1/22= 4.5% delay=11j p=0.1068	1/23= 4.3% delay=4j p=0.2347	6/199= 3% delay=2.5j p=0.0408	0/4= 0% delay= p=1
b164-0	renal failure & suspension of osmotical laxative & age < 70	3/17= 17.7% delay=3j p=0.0113	1/54= 1.9% delay=10j p=0.2424	3/14= 21.4% delay=14j p=0.4278	21/118= 17.8% delay=8j p=0	0/11= 0% delay= p=1
b164-1	renal failure & suspension of osmotical laxative & age ≥ 70	4/78= 5.1% delay=2.5j p=0.2839	2/239= 0.8% delay=1j p=0.3466	31/188= 16.5% delay=5j p=0.2457	21/195= 10.8% delay=11j p=0.0256	0/26= 0% delay= p=1
b165-0	renal failure & suspension of other laxative & NO hepatic cholestasis	6/102= 5.9% delay=4j p=0.0686	2/178= 1.1% delay=1j p=0.2311	8/56= 14.3% delay=2.5j p=0.8445	0/20= 0% delay= p=0.6372	0/32= 0% delay= p=1
b165-1	renal failure & suspension of other laxative & hepatic cholestasis	7/16= 43.8% delay=5j p=0	0/14= 0% delay= p=1	9/26= 34.6% delay=6j p=0.0058	<i>No stay</i>	0/8= 0% delay= p=1
b166-0	renal failure & propulsive laxative	0/3= 0% delay= p=1	<i>No stay</i>	0/1= 0% delay= p=1	<i>No stay</i>	<i>No stay</i>
b167-0	renal failure & peripheral sympatholytic	6/133= 4.5% delay=5j p=0.2791	0/90= 0% delay= p=1	1/64= 1.6% delay=2j p=0.0011	8/377= 2.1% delay=3j p=0.0001	1/5= 20% delay=1j p=0.0486
b169-0	renal failure & beta blocker & NO thrombin inhibitor	24/326= 7.4% delay=5j p=0	14/742= 1.9% delay=3j p=0	6/107= 5.6% delay=2j p=0.0073	29/1208= 2.4% delay=2j p=0	3/56= 5.4% delay=5j p=0.0163
b169-1	renal failure & beta blocker & thrombin inhibitor	12/76= 15.8% delay=3j p=0	<i>No stay</i>	<i>No stay</i>	1/108= 0.9% delay=2j p=0.0098	<i>No stay</i>
b172-0	renal failure & angiotensin conversion enzyme inhibitor	24/404= 5.9% delay=4j p=0.0005	5/516= 1% delay=7j p=0.1902	16/166= 9.6% delay=2j p=0.087	26/1374= 1.9% delay=2.5j p=0	2/27= 7.4% delay=2j p=0.0281

	& NO opioid					
b172-1	renal failure & angiotensin conversion enzyme inhibitor & opioid	7/106= 6.6% delay=6j p=0.0301	16/276= 5.8% delay=4.5j p=0	4/32= 12.5% delay=1.5j p=1	6/198= 3% delay=2.5j p=0.0407	1/23= 4.3% delay=18j p=0.2059
b173-0	renal failure & potassium sparing diuretic	8/115= 7% delay=6j p=0.0159	14/291= 4.8% delay=6.5j p=0	2/37= 5.4% delay=2.5j p=0.151	8/420= 1.9% delay=2j p=0	2/18= 11.1% delay=5.5j p=0.0129
b174-0	renal failure & suspension of potassium lowering diuretic & NO thrombin inhibitor	10/148= 6.8% delay=5j p=0.009	3/453= 0.7% delay=2j p=0.504	52/256= 20.3% delay=4j p=0.0011	152/1409= 10.8% delay=8j p=0	0/31= 0% delay= p=1
b174-1	renal failure & suspension of potassium lowering diuretic & thrombin inhibitor	5/32= 15.6% delay=3j p=0.0019	<i>No stay</i>	<i>No stay</i>	2/101= 2% delay=2.5j p=0.0659	<i>No stay</i>
b175-0	renal failure & potassium	18/368= 4.9% delay=3j p=0.0217	31/1242= 2.5% delay=8j p=0	2/4= 50% delay=1.5j p=0.0951	8/487= 1.6% delay=2j p=0	2/54= 3.7% delay=13j p=0.098
b176-0	renal failure & thrombin inhibitor	24/206= 11.7% delay=5j p=0	<i>No stay</i>	0/1= 0% delay= p=1	2/156= 1.3% delay=2.5j p=0.0044	<i>No stay</i>
b178-0	renal failure & amoxicilline and clav.ac.	15/287= 5.2% delay=5j p=0.0259	1/159= 0.6% delay=4j p=0.56	4/61= 6.6% delay=3.5j p=0.1235	6/273= 2.2% delay=2.5j p=0.0016	1/3= 33.3% delay=2j p=0.0295
b179-0	renal failure & suspension of aminoglycoside	5/26= 19.2% delay=3j p=0.0007	6/252= 2.4% delay=9.5j p=0.0018	5/9= 55.6% delay=8j p=0.0038	49/188= 26.1% delay=9j p=0	0/7= 0% delay= p=1
b180-0	renal failure & suspension of vitamin K antagonist	3/197= 1.5% delay=2j p=0.379	0/91= 0% delay= p=1	1/11= 9.1% delay=6j p=1	9/69= 13% delay=12j p=0.0432	0/9= 0% delay= p=1
b181-0	renal failure & suspension of sulfamid or sulfonamid	1/26= 3.8% delay=1j	1/43= 2.3% delay=3j	2/33= 6.1% delay=8j	19/252= 7.5% delay=5j	0/2= 0% delay=

		p=0.5269	p=0.1982	p=0.3018	p=0.5146	p=1
b182-0	renal failure & NSAID & NO potassium sparing diuretic	20/381= 5.2% delay=3j p=0.0062	14/1192= 1.2% delay=6.5j p=0.0023	8/101= 7.9% delay=2j p=0.0697	30/1168= 2.6% delay=3j p=0	3/76= 3.9% delay=3j p=0.0365
b182-1	renal failure & NSAID & potassium sparing diuretic	3/34= 8.8% delay=5j p=0.0704	8/168= 4.8% delay=7j p=0	0/10= 0% delay= p=0.3735	3/253= 1.2% delay=3j p=0.0001	0/8= 0% delay= p=1
b184-0	renal failure & suspension of systemic steroidal anti inflammatory	5/87= 5.7% delay=4j p=0.0996	4/211= 1.9% delay=4.5j p=0.0227	12/80= 15% delay=3.5j p=0.7372	51/378= 13.5% delay=11j p=0	0/29= 0% delay= p=1
b185-0	renal failure & digitalis glycoside	10/94= 10.6% delay=3j p=0.0003	12/504= 2.4% delay=7j p=0	3/49= 6.1% delay=2j p=0.1369	4/181= 2.2% delay=2.5j p=0.0137	0/17= 0% delay= p=1
b186-0	renal failure & immunomodulation factor	1/19= 5.3% delay=3j p=0.4211	0/21= 0% delay= p=1	0/5= 0% delay= p=1	2/133= 1.5% delay=1.5j p=0.012	0/3= 0% delay= p=1

14.10. Hypocalcemia (calcemia<2.2 mmol/l)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b191-0	NO hypoalbuminemia & proton pump inhibitor	19/332= 5.7% delay=3j p=0.202	39/2559= 1.5% delay=3j p=0.5436	43/257= 16.7% delay=2j p=0.107	76/1583= 4.8% delay=2j p=0.0038	1/190= 0.5% delay=2j p=0.2823

14.11. Hypokalemia (K⁺<3.0)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b150-0	NO renal failure	12/234= 5.1%	No stay	No stay	2/146= 1.4%	0/1= 0%

	& thrombin inhibitor	delay=3j p=0.0003			delay=2j p=0.3306	delay= p=1
--	----------------------	----------------------	--	--	----------------------	---------------

14.12. Hyponatremia (Na⁺<130)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b188-0	proton pump inhibitor & age < 70	8/164= 4.9% delay=4.5j p=0.1429	21/1228= 1.7% delay=2j p=0.8233	4/77= 5.2% delay=3.5j p=0.1248	28/952= 2.9% delay=3j p=0.0317	1/123= 0.8% delay=5j p=1
b188-1	proton pump inhibitor & age ≥ 70	9/194= 4.6% delay=7j p=0.1223	61/1951= 3.1% delay=5j p=0	9/219= 4.1% delay=3j p=0.0002	19/626= 3% delay=3j p=0.1215	1/109= 0.9% delay=1j p=1

14.13. Renal failure (creat.>135 μmol/l or urea>8.0 mmol/l)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b091-0	quinolone & age < 70	6/174= 3.4% delay=2j p=0.0404	12/309= 3.9% delay=5j p=0.5183	2/26= 7.7% delay=2.5j p=0.2913	21/256= 8.2% delay=3j p=0.254	0/18= 0% delay= p=1
b091-1	quinolone & age ≥ 70	24/204= 11.8% delay=3j p=0.0306	25/541= 4.6% delay=3j p=0.0843	4/71= 5.6% delay=3j p=0.0051	15/198= 7.6% delay=2j p=0.1966	2/15= 13.3% delay=5j p=0.1087
b097-0	NSAI & NO aspirin & NO potassium lowering diuretic & age < 70	0/110= 0% delay= p=0.0003	7/1550= 0.5% delay=4j p=0	0/6= 0% delay= p=0.5973	7/85= 8.2% delay=3j p=0.5957	1/504= 0.2% delay=4j p=0
b097-1	NSAI & NO aspirin	4/22= 18.2% delay=2j	30/530= 5.7% delay=4j	1/8= 12.5% delay=2j	2/6= 33.3% delay=2.5j	1/117= 0.9% delay=1j

b097-2	& NO potassium lowering diuretic & age ≥ 70 NSAI & NO aspirin & potassium lowering diuretic & age < 70	p=0.0812 2/8= 25% delay=1.5j p=0.1192	p=0.0039 6/79= 7.6% delay=2.5j p=0.0463	p=1 <i>No stay</i>	p=0.1242 0/9= 0% delay= p=0.6107	p=0.1263 0/5= 0% delay= p=1
b097-3	NSAI & NO aspirin & potassium lowering diuretic & age ≥ 70	2/12= 16.7% delay=6.5j p=0.2308	11/133= 8.3% delay=2j p=0.0046	0/1= 0% delay= p=1	0/5= 0% delay= p=1	2/9= 22.2% delay=1.5j p=0.0431
b197-0	angiotensin conversion enzyme inhibitor & age < 70	21/321= 6.5% delay=2j p=0.5172	27/659= 4.1% delay=2j p=0.2606	7/62= 11.3% delay=3j p=0.2281	136/1725= 7.9% delay=2j p=0	4/85= 4.7% delay=4j p=0.5613
b197-1	angiotensin conversion enzyme inhibitor & age ≥ 70	138/965= 14.3% delay=3j p=0	96/1345= 7.1% delay=4j p=0	29/271= 10.7% delay=3j p=0.0007	142/1045= 13.6% delay=2j p=0.0007	13/119= 10.9% delay=6j p=0.0004

14.14. VKA underdose (INR<1.6)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b019-0	VKA & griseofulvin	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>
b026-0	VKA & antiepileptic	30/200= 15% delay=4.5j p=0	12/117= 10.3% delay=3j p=0.0013	0/4= 0% delay= p=1	8/31= 25.8% delay=2j p=0.0113	1/10= 10% delay=21j p=0.3334
b039-0	VKA & anti-diarrheal	9/37= 24.3% delay=4j p=0	1/7= 14.3% delay=4j p=0.2329	0/2= 0% delay= p=1	3/7= 42.9% delay=3j p=0.0278	<i>No stay</i>
b041-0	VKA	1/17= 5.9% delay=2j	0/2= 0% delay=	0/2= 0% delay=	3/7= 42.9% delay=3j	<i>No stay</i>

	& suspension of anti-diarrheal	p=0.2886	p=1	p=1	p=0.0278	
b042-0	VKA & digitalis glycoside & NO chronic renal insufficiency & age < 70	3/12= 25% delay=5j p=0.0015	4/34= 11.8% delay=4j p=0.0362	<i>No stay</i>	2/10= 20% delay=3j p=0.2757	2/3= 66.7% delay=16j p=0.0045
b042-1	VKA & digitalis glycoside & NO chronic renal insufficiency & age ≥ 70	17/85= 20% delay=5j p=0	6/88= 6.8% delay=4.5j p=0.1441	0/1= 0% delay= p=1	2/10= 20% delay=2.5j p=0.2757	2/6= 33.3% delay=4j p=0.021
b042-2	VKA & digitalis glycoside & chronic renal insufficiency & age < 70	<i>No stay</i>	0/1= 0% delay= p=1	<i>No stay</i>	0/1= 0% delay= p=1	<i>No stay</i>
b042-3	VKA & digitalis glycoside & chronic renal insufficiency & age ≥ 70	0/3= 0% delay= p=1	1/8= 12.5% delay=3j p=0.2614	<i>No stay</i>	2/3= 66.7% delay=2j p=0.0296	<i>No stay</i>
b051-0	VKA & immunomodulation factor	3/9= 33.3% delay=2j p=0.0006	2/3= 66.7% delay=5.5j p=0.004	<i>No stay</i>	0/1= 0% delay= p=1	<i>No stay</i>
b078-0	VKA & high INR & NO hypoalbuminemia	19/73= 26% delay=2j p=0	0/8= 0% delay= p=1	0/3= 0% delay= p=1	1/5= 20% delay=1j p=0.4196	<i>No stay</i>
b078-1	VKA & high INR & hypoalbuminemia	5/14= 35.7% delay=2j p=0	2/8= 25% delay=31j p=0.0332	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>
b079-0	VKA & high INR & NO hypocalcemia	23/86= 26.7% delay=2j p=0	2/16= 12.5% delay=31j p=0.1174	0/1= 0% delay= p=1	1/5= 20% delay=1j p=0.4196	<i>No stay</i>

b079-1	VKA & high INR & hypocalcemia	1/3= 33.3% delay=4j p=0.0583	<i>No stay</i>	0/1= 0% delay= p=1	<i>No stay</i>	<i>No stay</i>
b080-0	VKA & beta lactam & age < 70	6/30= 20% delay=1.5j p=0	6/30= 20% delay=4.5j p=0.0007	<i>No stay</i>	4/19= 21.1% delay=2.5j p=0.1247	1/3= 33.3% delay=7j p=0.1143
b080-1	VKA & beta lactam & age ≥ 70	15/95= 15.8% delay=6j p=0	8/79= 10.1% delay=4.5j p=0.0089	0/2= 0% delay= p=1	4/14= 28.6% delay=2.5j p=0.0484	0/1= 0% delay= p=1
b081-0	VKA & prokinetic & NO high INR	2/12= 16.7% delay=5.5j p=0.0225	2/12= 16.7% delay=1.5j p=0.0711	<i>No stay</i>	0/1= 0% delay= p=1	<i>No stay</i>
b081-1	VKA & prokinetic & high INR	2/5= 40% delay=4.5j p=0.0037	0/3= 0% delay= p=1	<i>No stay</i>	<i>No stay</i>	<i>No stay</i>
b082-0	VKA & antiacid	5/22= 22.7% delay=5j p=0.0001	<i>No stay</i>	<i>No stay</i>	1/1= 100% delay=3j p=0.1031	<i>No stay</i>

14.15. Neutropenia (PNN<1500/mm3)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b100-0	NO cancer & NSAID	7/1084= 0.6% delay=7j p=0.0864	9/5059= 0.2% delay=2j p=0.0033	<i>Missing variables</i>	0/2655= 0% delay= p=0	0/853= 0% delay= p=0.2523
b190-0	proton pump inhibitor	6/364= 1.6% delay=6.5j p=0.3227	12/3196= 0.4% delay=4.5j p=1	<i>Missing variables</i>	0/1655= 0% delay= p=0	2/244= 0.8% delay=38.5j p=0.0492

14.16. Increase of pancreatic enzymes (amylase>90 UI/l or lipase>90 UI/l)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b122-0	azole antibiotic	6/165= 3.6% delay=2.5j p=0.0001	7/466= 1.5% delay=1j p=0.1316	0/50= 0% delay= p=1	4/146= 2.7% delay=3j p=0.3173	1/201= 0.5% delay=2j p=0.4441
b123-0	cephalosporin & NO hepatic insufficiency	0/40= 0% delay= p=1	36/2952= 1.2% delay=3.5j p=0.0302	0/66= 0% delay= p=1	7/664= 1.1% delay=2j p=0.2096	1/238= 0.4% delay=2j p=0.505
b123-1	cephalosporin & hepatic insufficiency	0/5= 0% delay= p=1	8/74= 10.8% delay=2j p=0	0/2= 0% delay= p=1	1/25= 4% delay=2j p=0.3498	0/2= 0% delay= p=1
b124-0	Statin	1/723= 0.1% delay=2j p=0.1693	15/1585= 0.9% delay=3j p=0.7723	0/229= 0% delay= p=0.3595	14/2400= 0.6% delay=2j p=0	1/237= 0.4% delay=6j p=0.5034
b125-0	systemic steroidal anti inflammatory	1/737= 0.1% delay=1j p=0.1687	19/1446= 1.3% delay=4j p=0.0701	0/173= 0% delay= p=0.6094	14/958= 1.5% delay=2.5j p=0.6883	1/290= 0.3% delay=2j p=0.5815
b126-0	proton pump inhibitor & NO hepatic insufficiency	1/318= 0.3% delay=3j p=1	33/3062= 1.1% delay=6j p=0.1789	0/304= 0% delay= p=0.111	21/1571= 1.3% delay=2j p=0.2301	1/239= 0.4% delay=1j p=0.5066
b126-1	proton pump inhibitor & hepatic insufficiency	0/48= 0% delay= p=1	9/126= 7.1% delay=2j p=0	0/20= 0% delay= p=1	1/48= 2.1% delay=2j p=0.563	0/5= 0% delay= p=1

14.17. Pancytopenia

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b101-0	NO cancer	0/1084= 0%	1/5060= 0%	0/210= 0%	0/2655= 0%	0/853= 0%

	& NSAID	delay= p=0.0385	delay=1j p=0.0518	delay= p=0.5869	delay= p=0	delay= p=0.0327
--	---------	--------------------	----------------------	--------------------	---------------	--------------------

14.18. Thrombocytosis (count>600,000)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b118-0	azole antibiotic	3/166= 1.8% delay=6j p=0.1943	8/468= 1.7% delay=5j p=0.002	0/47= 0% delay= p=0.4051	0/158= 0% delay= p=0	1/201= 0.5% delay=7j p=1
b119-0	low weight heparin	18/1197= 1.5% delay=5j p=0.0262	13/2063= 0.6% delay=4j p=0.3053	1/59= 1.7% delay=4j p=0.7171	0/1045= 0% delay= p=0	4/437= 0.9% delay=5j p=0.4978
b120-0	systemic antifungal	3/128= 2.3% delay=9j p=0.112	6/396= 1.5% delay=5.5j p=0.0132	0/15= 0% delay= p=1	0/290= 0% delay= p=0	1/60= 1.7% delay=7j p=0.3326
b121-0	quinolone & age < 70	5/183= 2.7% delay=4j p=0.0254	11/312= 3.5% delay=5j p=0	2/27= 7.4% delay=4j p=0.2454	0/278= 0% delay= p=0	1/20= 5% delay=1j p=0.1248
b121-1	quinolone & age ≥ 70	2/234= 0.9% delay=6.5j p=1	3/572= 0.5% delay=9j p=0.7608	0/78= 0% delay= p=0.1044	0/208= 0% delay= p=0	0/18= 0% delay= p=1

14.19. Thrombopenia (count<75,000)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b067-0	NO cancer & high weight heparin & age < 70	1/17= 5.9% delay=4j p=0.1919	<i>No stay</i>	0/4= 0% delay= p=1	0/367= 0% delay= p=0	0/4= 0% delay= p=1

b067-1	NO cancer & high weight heparin & age ≥ 70	0/59= 0% delay= p=1	0/1= 0% delay= p=1	0/19= 0% delay= p=1	0/284= 0% delay= p=0	0/2= 0% delay= p=1
b068-0	NO cancer & low weight heparin & NO hemostasis disorder (diag)	7/1030= 0.7% delay=7j p=0.0883	4/1960= 0.2% delay=2j p=0.8103	0/49= 0% delay= p=1	0/949= 0% delay= p=0	0/403= 0% delay= p=0.2208
b068-1	NO cancer & low weight heparin & hemostasis disorder (diag)	0/13= 0% delay= p=1	0/1= 0% delay= p=1	0/1= 0% delay= p=1	0/19= 0% delay= p=0	<i>No stay</i>
b099-0	NO cancer & NSAID & NO hepatic insufficiency	2/1029= 0.2% delay=11.5j p=0.0002	3/5012= 0.1% delay=2j p=0.0001	1/205= 0.5% delay=4j p=0.328	0/2613= 0% delay= p=0	0/851= 0% delay= p=0.0044
b099-1	NO cancer & NSAID & hepatic insufficiency	5/55= 9.1% delay=3j p=0.0006	0/46= 0% delay= p=1	0/4= 0% delay= p=1	0/42= 0% delay= p=0	0/2= 0% delay= p=1
b107-0	NO cancer & antiH2	0/54= 0% delay= p=1	0/41= 0% delay= p=1	0/2= 0% delay= p=1	0/36= 0% delay= p=0	<i>No stay</i>
b108-0	NO cancer & platelet aggregation inhibitor & NO NSAID	4/165= 2.4% delay=2j p=0.1489	1/119= 0.8% delay=4j p=0.2749	0/95= 0% delay= p=0.3853	0/220= 0% delay= p=0	0/5= 0% delay= p=1
b109-0	NO cancer & potassium lowering diuretic	16/1105= 1.4% delay=3j p=0.457	10/2810= 0.4% delay=4.5j p=0.3004	2/326= 0.6% delay=3j p=0.1644	0/1979= 0% delay= p=0	1/163= 0.6% delay=3j p=0.5719
b110-0	NO cancer & proton pump inhibitor & NO hepatic insufficiency	2/276= 0.7% delay=3.5j p=0.5842	8/2790= 0.3% delay=2j p=0.8349	4/272= 1.5% delay=3j p=1	0/1503= 0% delay= p=0	0/145= 0% delay= p=1
b110-1	NO cancer & proton pump inhibitor	3/44= 6.8% delay=2j	3/123= 2.4% delay=9j	1/16= 6.3% delay=3j	0/44= 0% delay= p=0	0/3= 0% delay= p=0

	& hepatic insufficiency	p=0.017	p=0.0043	p=0.2121	p=0	p=1
b111-0	NO cancer & acetaminophen/paracetamol	8/1876=0.4% delay=3j p=0	7/4663=0.2% delay=5j p=0.0471	0/332=0% delay= p=0.0039	0/1340=0% delay= p=0	0/1159=0% delay= p=0.0001
b112-0	NO cancer & beta blocker & NO hepatic insufficiency	4/773=0.5% delay=4.5j p=0.0546	2/1776=0.1% delay=3j p=0.2172	4/191=2.1% delay=3.5j p=0.5001	0/2293=0% delay= p=0	0/162=0% delay= p=1
b112-1	NO cancer & beta blocker & hepatic insufficiency	8/95=8.4% delay=3j p=0	0/23=0% delay= p=1	1/13=7.7% delay=3j p=0.1758	0/54=0% delay= p=0	No stay
b113-0	NO cancer & beta lactam & NO hepatic insufficiency	7/1046=0.7% delay=3j p=0.0665	13/3681=0.4% delay=2j p=0.2515	0/209=0% delay= p=0.051	0/1695=0% delay= p=0	0/334=0% delay= p=0.3797
b113-1	NO cancer & beta lactam & hepatic insufficiency	4/82=4.9% delay=3.5j p=0.0186	3/90=3.3% delay=2j p=0.0018	0/11=0% delay= p=1	0/57=0% delay= p=0	0/1=0% delay= p=1
b114-0	NO cancer & quinolone & NO hepatic insufficiency	2/310=0.6% delay=3j p=0.5926	6/728=0.8% delay=2j p=0.0118	2/93=2.2% delay=3.5j p=0.6393	0/409=0% delay= p=0	0/31=0% delay= p=1
b114-1	NO cancer & quinolone & hepatic insufficiency	1/38=2.6% delay=2j p=0.3794	1/28=3.6% delay=1j p=0.0726	0/4=0% delay= p=1	0/15=0% delay= p=0	No stay
b115-0	NO cancer & type 3 antiarrhythmic & NO hepatic insufficiency	5/257=1.9% delay=2j p=0.2521	0/109=0% delay= p=1	0/65=0% delay= p=0.6169	0/701=0% delay= p=0	0/13=0% delay= p=1
b115-1	NO cancer & type 3 antiarrhythmic & hepatic insufficiency	0/6=0% delay= p=1	No stay	0/1=0% delay= p=1	0/10=0% delay= p=0	1/1=100% delay=2j p=0.005
b116-0	NO cancer	0/118=0%	4/218=1.8%	0/58=0%	0/243=0%	0/126=0%

b116-1	& antiparasitic & NO hepatic insufficiency NO cancer & antiparasitic & hepatic insufficiency	delay= p=0.406 0/11= 0% delay= p=1	delay=1.5j p=0.0026 0/15= 0% delay= p=1	delay= p=1 0/3= 0% delay= p=1	delay= p=0 0/14= 0% delay= p=0	delay= p=1 <i>No stay</i>
b117-0	NO cancer & selective serotonin reuptake inhibitor & NO hepatic insufficiency	1/255= 0.4% delay=4j p=0.3773	2/999= 0.2% delay=5.5j p=1	0/123= 0% delay= p=0.2392	0/240= 0% delay= p=0	0/106= 0% delay= p=1
b117-1	NO cancer & selective serotonin reuptake inhibitor & hepatic insufficiency	1/13= 7.7% delay=1j p=0.1503	0/22= 0% delay= p=1	0/5= 0% delay= p=1	0/6= 0% delay= p=0	1/1= 100% delay=2j p=0.005

14.20. Diarrhea (prescription of an anti-diarrheal)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b192-0	proton pump inhibitor	11/362= 3% delay=6j p=0.7441	45/3190= 1.4% delay=6j p=0.0001	0/326= 0% delay= p=0	0/1655= 0% delay= p=0	4/242= 1.7% delay=9.5j p=0.2543

14.21. Diarrhea (prescription of an antipropulsive)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b196-0	proton pump inhibitor	1/367= 0.3% delay=4j p=0.352	14/3196= 0.4% delay=5j p=0.086	0/326= 0% delay= p=0	0/1655= 0% delay= p=0	3/243= 1.2% delay=6j p=0.4007

14.22. Bacterial infection (detected by the prescription of antibiotic)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b104-0	NO beta lactam & NO antineoplastic & NSAID & age < 70 & NO diabetes	24/257=9.3% delay=3j p=0.6037	71/1780=4% delay=3j p=0	0/17=0% delay= p=0	0/1010=0% delay= p=0	9/467=1.9% delay=3j p=0.0001
b104-1	NO beta lactam & NO antineoplastic & NSAID & age < 70 & diabetes	9/82=11% delay=2j p=0.8553	9/140=6.4% delay=1j p=1	0/4=0% delay= p=0	0/347=0% delay= p=0	1/17=5.9% delay=7j p=0.5885
b104-2	NO beta lactam & NO antineoplastic & NSAID & age ≥ 70 & NO diabetes	63/460=13.7% delay=3j p=0.0213	199/1598=12.5% delay=3j p=0	0/113=0% delay= p=0	0/593=0% delay= p=0	23/217=10.6% delay=3j p=0.0004
b104-3	NO beta lactam & NO antineoplastic & NSAID & age ≥ 70 & diabetes	27/161=16.8% delay=3j p=0.0124	35/177=19.8% delay=3j p=0	0/27=0% delay= p=0	0/218=0% delay= p=0	3/23=13% delay=5j p=0.1072

14.23. paracetamol overdose (detected by the prescription of acetyl-cystein)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b129-0	acetaminophen/paracetamol	0/2007=0% delay= p=0	0/5129=0% delay= p=0	0/373=0% delay= p=0	0/1447=0% delay= p=0	0/1312=0% delay= p=0

14.24. Fungal infection (detected by the prescription of local antifungal)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b130-0	systemic steroidal anti inflammatory	14/325=4.3%	1/487=0.2%	0/64=0%	0/799=0%	0/220=0%

b130-1	& age < 70	delay=5.5j p=0.0004	delay=6j p=0.3795	delay= p=0	delay= p=0	delay= p=0.3984
	systemic steroidal anti inflammatory & age ≥ 70	26/406= 6.4% delay=4j p=0	9/957= 0.9% delay=3j p=0.29	0/110= 0% delay= p=0	0/187= 0% delay= p=0	1/69= 1.4% delay=3j p=0.4198

14.25. Fungal infection (detected by the prescription of a systemic antifungal)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b132-0	azole antibiotic & NSAID < 1	3/105= 2.9% delay=4j p=0.2463	18/215= 8.4% delay=3j p=0	0/40= 0% delay= p=0	0/118= 0% delay= p=0	1/48= 2.1% delay=4j p=1
b132-1	azole antibiotic & NSAID ≥ 1	4/20= 20% delay=7.5j p=0.0003	7/125= 5.6% delay=3j p=0.0121	0/11= 0% delay= p=0	0/40= 0% delay= p=0	0/116= 0% delay= p=0.1074
b133-0	antiparasitic & NSAID < 1	3/122= 2.5% delay=4j p=0.4532	16/164= 9.8% delay=3j p=0	0/53= 0% delay= p=0	0/345= 0% delay= p=0	1/41= 2.4% delay=4j p=0.612
b133-1	antiparasitic & NSAID ≥ 1	4/29= 13.8% delay=10.5j p=0.0012	6/95= 6.3% delay=3j p=0.0114	0/17= 0% delay= p=0	0/73= 0% delay= p=0	0/113= 0% delay= p=0.1765
b131-0	amoxicilline and clav.ac.	28/923= 3% delay=4j p=0.001	4/427= 0.9% delay=1.5j p=0.1531	0/144= 0% delay= p=0	0/790= 0% delay= p=0	1/11= 9.1% delay=1j p=0.2227
b134-0	other beta lactam	5/23= 21.7% delay=3j p=0	18/126= 14.3% delay=4j p=0	0/3= 0% delay= p=0	0/101= 0% delay= p=0	1/3= 33.3% delay=8j p=0.0662
b194-0	proton pump inhibitor & age < 70	2/165= 1.2% delay=6.5j	25/1225= 2% delay=2j	0/82= 0% delay=	0/993= 0% delay=	8/128= 6.3% delay=4.5j

		p=1	p=0.8288	p=0	p=0	p=0.0064
b194-1	proton pump inhibitor & age ≥ 70	9/200= 4.5% delay=4j p=0.0052	69/1942= 3.6% delay=6j p=0	0/244= 0% delay= p=0	0/662= 0% delay= p=0	7/109= 6.4% delay=3j p=0.0095

14.26. Hemorrhage (detected by the prescription of hemostatic)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b077-0	high weight heparin	2/85= 2.4% delay=7.5j p=0.1386	0/2= 0% delay= p=1	0/26= 0% delay= p=0	0/699= 0% delay= p=0	0/6= 0% delay= p=1
b071-0	low weight heparin	5/1190= 0.4% delay=2j p=0.1411	24/2051= 1.2% delay=3j p=1	0/66= 0% delay= p=0	0/1045= 0% delay= p=0	1/436= 0.2% delay=1j p=0.6956
b072-0	VKA & age < 70 & NO respiratory insufficiency	1/68= 1.5% delay=3j p=0.4102	2/114= 1.8% delay=23.5j p=0.3914	0/4= 0% delay= p=0	0/51= 0% delay= p=0	0/16= 0% delay= p=1
b072-1	VKA & age < 70 & respiratory insufficiency	2/46= 4.3% delay=6.5j p=0.0484	0/14= 0% delay= p=1	0/1= 0% delay= p=0	0/16= 0% delay= p=0	<i>No stay</i>
b072-2	VKA & age ≥ 70 & NO respiratory insufficiency	10/214= 4.7% delay=7.5j p=0	7/209= 3.3% delay=5j p=0.0121	0/12= 0% delay= p=0	0/35= 0% delay= p=0	0/21= 0% delay= p=1
b072-3	VKA & age ≥ 70 & respiratory insufficiency	15/190= 7.9% delay=7j p=0	0/31= 0% delay= p=1	0/2= 0% delay= p=0	0/12= 0% delay= p=0	0/1= 0% delay= p=1
b106-0	NSAI	6/1182= 0.5% delay=6j p=0.3524	44/5352= 0.8% delay=3j p=0.0011	0/227= 0% delay= p=0	0/2788= 0% delay= p=0	2/925= 0.2% delay=1.5j p=0.1025

14.27. VKA overdose (detected by the prescription of vitamin K)

ID	Conditions	Hospital_1	Hospital_2	Hospital_3	Hospital_4	Hospital_5
b127-0	VKA & NO respiratory insufficiency	11/283=3.9% delay=6j p=0	8/324=2.5% delay=5.5j p=0.0033	0/16=0% delay= p=0	0/86=0% delay= p=0	0/38=0% delay= p=1
b127-1	VKA & respiratory insufficiency	17/237=7.2% delay=7j p=0	0/45=0% delay= p=1	0/3=0% delay= p=0	0/28=0% delay= p=0	0/1=0% delay= p=1

15. TABLE OF FIGURES

Figure 1. Decision rules induction in Procedure A	19
Figure 2. Approach for decision rules in Procedure B.....	20
Figure 3. The probability of a rule to be applicable when some variables are missing strongly decreases when the number of predictors is high.	22
Figure 4. Partitioning with Agglomerative Hierarchical Clustering.....	27
Figure. 5 Example of dendrogramm.....	27
Figure 6. Distribution of the correlation coefficient between cause variables (about 10^5 values).....	33
Figure 7. Distribution of <i>Log10(p value of the correlation coefficient's nullity test)</i> (about 10^5 values).....	33
Figure 8. Scheme of an artificial neuron	34
Figure 9. Example of decision tree.....	36
Figure 10. Description of the architecture of the PSIP project	45
Figure 11. Ariane's thread.....	47
Figure 12. Ariane's thread – Available data and SPCs (right part: details about the data repository)	49
Figure 13. Compromise to reach in building a common data mode	50
Figure 14. Simplified representation of the data scheme	52
Figure 15. Rule-based detection of ADEs with a CPOE (left) or without CPOE (right)	59
Figure 16. Ariane's thread – Rules induction, storage into a repository, and machine evaluation.....	67
Figure 17. List of outcomes extracted from the SPCs.....	70
Figure 18. The 4 first steps of the procedure to “fish” ADEs	75
Figure 19. Aggregation engines and mapping policies	76
Figure 20. Different possible shapes of the events	77
Figure 21. Example of administered drug aggregation: anticoagulant drugs .	78
Figure 22. Different positions of the Acetyl-Salicylic acid (Aspirin) in the ATC classification	79
Figure 23. Places of Rifampicin and Isoniazid in the ATC classification	79
Figure 24. Transversal and hierarchical redundancy of the classification	80
Figure 25. Example of laboratory result aggregation: the Digoxin and Potassium values of a stay	81
Figure 26. Automatic XML output of R scripts.....	90

Figure 27. XML data scheme	91
Figure 28. Rules inclusion in the repository	92
Figure 29. Automated machine evaluation of the rules in every department ..	92
Figure 30. Breadcrumb trail – Results of data mining, rules of the repository and related statistics, web tools	95
Figure 31. First rule gives $p(\text{too low INR during stay})=86\%$ instead of 1% ..	97
Figure 32. Second and third rules give $p(\text{too low INR during stay})= 67\%$ and 80%	97
Figure 33. Decision tree: circumstances that lead to a leukopenia (leukocyte count $< 3 \times 10^9/l$)	100
Figure 34. Flowchart of the different outcomes explored in this work	101
Figure 35. Length of stay in control group (left) and potential ADE group (right)	113
Figure 36. Length of stay in control group (left) and potential ADE group (right)	115
Figure 37. Length of stay in control group (left) and potential ADE group (right)	117
Figure 38. Login page of the Scorecards	119
Figure 39. Synthesis page of the Scorecards	120
Figure 40. List of generated scorecards.....	121
Figure 41. Scorecard of hyperkalemia ($K^+ > 5.3$)	122
Figure 42. When required, a popup displays the IDs of the potential ADE cases.	123
Figure 43. Main screen of the Expert Explorer.....	124
Figure 44. The 3 zones of the Expert Explorer main page.....	125
Figure 45. Use of the lab panel.....	126
Figure 46. When required, a popup displays information about the rules that fire on the current stay	127
Figure 47. Analysis of the administered drugs	128
Figure 48. The "more information" popup	129
Figure 49. List of the potential ADE cases to review.....	130
Figure 50. Two different ways to compute the positive predictive value of a set of rules	136
Figure 51. Example of ADE case matching the rule <i>VKA & hypoalbuminemia</i> \rightarrow <i>too high INR</i> . Left: with binary transformation. Right: without binary transformation.....	140
Figure 52. Meta-rule n°1: static filter	142
Figure 53. Meta-rule n°2: temporal filter.....	143
Figure 54. Meta-rule n°3: rules with laboratory-related outcomes.....	143

Figure 55. Retrospective use of the set of XML files.....	145
Figure 56. Prospective use of the set of XML files.....	145
Figure 57. Global process for data extraction, data management and data analysis.....	159
Figure 58. Data flow diagram.....	167
Figure 59. Simplified class diagram.....	168
Figure 60. System context diagram showing the engineering overview.....	169
Figure 61. "Expert Explorer" welcome screen.....	171
Figure 62. Data set page.....	172
Figure 63. Selecting flat table outcomes.....	173
Figure 64. Defining a report.....	174
Figure 65. Saved reports list.....	174
Figure 66. Visualizing a report.....	175
Figure 67. Stays identified by a report.....	175
Figure 68. Defining a query.....	176
Figure 69. Error handling while importing the queries.....	177
Figure 70. List of existing queries.....	177
Figure 71. Query detail page.....	178
Figure 72. Visualization of the details for a stay.....	179
Figure 73. Laboratory charts.....	180
Figure 74. Drug charts.....	180
Figure 75. "Lab & drugs charts" window.....	181
Figure 76. Registering a new user account.....	181
Figure 77. Expert home page.....	182
Figure 78. Stay details page with the "Report" button present.....	183
Figure 79. The questionnaire.....	184
Figure 80. The administrator home page.....	185
Figure 81. Export questionnaires.....	185
Figure 82. Data flow diagram.....	187
Figure 83. Simplified class diagram.....	188
Figure 84. System context diagram showing the engineering overview.....	188
Figure 85. Access page for the Scorecards.....	190
Figure 86. Synthesis page.....	191
Figure 87. Number of detected cases by outcome and by month.....	191
Figure 88. Edition of detailed statistics.....	192
Figure 89. Generated scorecards.....	192
Figure 90. First part of the details page.....	193

Figure 91. Middle part of the details page.....	194
Figure 92. Pop-up window showing the cases.....	194
Figure 93. Details of rules in the bottom of the page.....	194
Figure 94. Review cases page.....	195
Figure 95. Bottom of the "Review cases" page.....	196
Figure 96. First form of the questionnaire: existence of the outcome.....	197
Figure 97. Second form of the questionnaire: one form per rule.....	198
Figure 98. Third form of the questionnaire: cause-to-effect relationship	199
Figure 99. Example of semantic mining applied on a discharge letter; precision and recall computation	219
Figure 100. First validation step.....	220
Figure 101. Second validation step.....	221
Figure 102. Third validation step.....	222
Figure 103. Agreements measured in the 1st and 2nd evaluation steps	225

16. TABLE OF TABLES

Table 1. Definition of the requirements in respect with the procedure.	20
Table 2. The hospital stay table	53
Table 3. The steps of the hospital stays table.....	55
Table 4. The diagnoses table.....	55
Table 5. The medical procedures table.....	55
Table 6. The drug table.....	56
Table 7. The lab results table.....	56
Table 8. The reports table.....	56
Table 9. The semantic mining table	56
Table 10. Example of quality control and abnormal values for a numeric variable such as the age	57
Table 11. Example of quality control and abnormal values for a string variable such as the principal diagnosis, encoded in ICD10	58
Table 12. Example of ATC-to-ATC interaction table. For each pair of ATC codes, the identifier of the interaction is provided if an interaction is described	63
Table 13. List of outcomes that are traced in the data	71
Table 14. First example of ADE case and inferred variable	73
Table 15. Second example of ADE case and inferred variable.....	73
Table 16. Third example of ADE case and inferred variable.....	74
Table 17. Examples of information classification: potential condition / potential outcome.....	82
Table 18. General considerations about various sources of rules.....	87
Table 19 Traceable outcomes that enabled to discover ADE detection rules	98
Table 20 Traceable outcomes for which no case or too few cases can be observed in the dataset	98
Table 21 Traceable outcomes that don't allow for ADE detection rule discovery.....	99
Table 22. Modules and corresponding number of rules	108
Table 23 Anticoagulation niche	109
Table 24 Proton pump inhibitor niche	109
Table 25 Rules out of the niches.....	110
Table 26 Count of rules per origin.....	111
Table 27. Categories of drug alerts [Shedlbauer 2009].....	132

Table 28. Summary of the grand challenges of clinical decision support system [Sittig 2008]	138
---	-----