International Journal of Medical Informatics (2005) **74**, 563—571

ELSEVIER

International Journal of Medical Informatics

www.intl.elsevierhealth.com/journals/ijmi

Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach

Cédric Bousquet[a,b], Corneliu Henegar[a], Agnès Lillo-Le Louët[b], Patrice Degoulet[a], Marie-Christine Jaulent[a,*]

[a] *INSERM U729, Faculté de médecine Broussais Hôtel Dieu, 15 rue de l'Ecole de Médecine, 75006 Paris, France*
[b] *Centre Régional de Pharmacovigilance, Hôpital Européen Georges Pompidou, Paris*

Received 29 October 2004; received in revised form 17 March 2005; accepted 6 April 2005

KEYWORDS
Adverse drug reaction reporting systems;
Terminology;
Automatic data processing;
Knowledge representation (computer);
Description logic;
Ontological modeling

Summary Automated signal generation is a growing field in pharmacovigilance that relies on data mining of huge spontaneous reporting systems for detecting unknown adverse drug reactions (ADR). Previous implementations of quantitative techniques did not take into account issues related to the medical dictionary for regulatory activities (MedDRA) terminology used for coding ADRs. MedDRA is a first generation terminology lacking formal definitions; grouping of similar medical conditions is not accurate due to taxonomic limitations.

Our objective was to build a data-mining tool that improves signal detection algorithms by performing terminological reasoning on MedDRA codes described with the DAML + OIL description logic. We propose the PharmaMiner tool that implements quantitative techniques based on underlying statistical and bayesian models. It is a JAVA application displaying results in tabular format and performing terminological reasoning with the Racer inference engine.

The mean frequency of drug-adverse effect associations in the French database was 2.66. Subsumption reasoning based on MedDRA taxonomical hierarchy produced a mean number of occurrence of 2.92 versus 3.63 ($p < 0.001$) obtained with a combined technique using subsumption and approximate matching reasoning based on the ontological structure. Semantic integration of terminological systems with data mining methods is a promising technique for improving machine learning in medical databases.

1. Introduction

The World Health Organization (WHO) defines a signal in pharmacovigilance as ''any reported information on a possible causal relationship between

* Corresponding author. Tel.: +33 1 42 34 69 86.
E-mail addresses: cedric.bousquet@spim.jussieu.fr (C. Bousquet), marie-christine.jaulent@spim.jussieu.fr (M.-C. Jaulent).

1386-5056/$ — see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2005.04.006

an adverse event and a drug, the relationship being unknown or incompletely documented previously'' [1]. Manual screening of whole pharmacovigilance databases for signal detection becomes no more feasible due to large collections of pharmacovigilance cases. The WHO Uppsala Monitoring Center collects case reports from 67 participating countries and owns a huge database of three million case reports. The Food and Drug Administration (FDA) in the United States manages the adverse event reporting system (AERS) that contains more than two million case reports. This database is growing fast, 250,000 new cases are entered every year.

The WHO [2] and the FDA [3] are currently using automated detection algorithms based upon Bayesian analysis in order to achieve signal generation. Evans et al. propose proportional reporting ratios for signal detection in the adverse drug reactions on-line information tracking (ADROIT) database in the United Kingdom [4]. Van Puijenbroek is investigating the use of reporting odds ratios and logistic regression for detecting drug-drug interactions in the Lareb pharmacovigilance center in The Netherlands [5]. The pharmaceutical industry is investigating the interest of automated methods.

The objective is to support pharmacovigilance experts for signal detection. A recent evaluation of the WHO Bayesian approach showed good overall sensitivity but rather low specificity [6]. One limitation is the small number of occurrences of each drug-event association in the database [7]. Moreover, performances of these quantitative methods are limited by the fact that they do not take into account the semantic information existing in the controlled vocabularies used to code adverse events in case reports.

The medical dictionary for regulatory activities (MedDRA) is a new terminology currently used for recording and reporting adverse drug event data in most countries [8]. It has already been adopted for the coding of adverse effects in pharmacovigilance databases of many governmental organizations and pharmacovigilance units in the pharmaceutical industry. This terminology proposes many terms and a fine granularity, which allows to code in a precise way descriptions of adverse effects such as they are described by physicians.

The part of the medical informatics community involved in developing terminologies is moving towards the supply of third generation systems according to Rossi-Mori's classification [9]. Desiderata were proposed by Cimino for the development of new medical terminologies [10]. We showed in a previous study that MedDRA does not commit to several desiderata and is a first generation system [11]. It organizes terms in several system organ classes (SOCs) but is not really multi-axial. Terms are not described in a formal way by means of a description logic. Grouping of terms by the means of high level categories is not accurate. MedDRA would benefit from formal definitions in an adverse drug reaction terminology.

In this paper, we present an original approach meant to improve signal detection. We built a tool (''PharmaMiner'') that performs terminological reasoning among drug-event pairs to group semantically linked adverse events and applies Bayesian (information component) and statistical analysis methods on these groups to detect potential signals. The terminological reasoning relies on a formal ontology of adverse events built from MedDRA. We report comparative results obtained on a subset of the French national Pharmacovigilance database while applying first grouping of cases with the MedDRA terminology and second terminological reasoning within an ontological structure.

2. Background

2.1. MedDRA

The MedDRA hierarchy is organized in five levels as follows [8]: system organ class, high level group terms (HLGT), high level terms (HLT), preferred terms (PT) and low level terms (LLT). The PT and HLT levels are recommended for signal generation and the LLT level for coding adverse events. The September 2002 version (5.1) of MedDRA comprises 26 SOCs, 332 HLGTs, 1683 HLTs, 16102 PTs and 56981 LLTs.

The finer granularity of MedDRA terms to describe medical data compared to other terminologies (WHO-ART, COSTART) implies less occurrences of a given drug-event association within the database, and lowers the performance of automated signal detection algorithms. Some studies have argued that an automatic grouping of cases expressing similar medical conditions would increase the power of the detection algorithms [12,13].

Moreover, there is no formal representation of MedDRA terms. The multiaxiality of MedDRA is limited to the fact that a PT can belong to several SOCs and cannot be compared to the mutiaxial properties of other terminological systems such as SNOMED, which allows multiple views on data according to the axis (morphological, topographic, etc.) [14]. Special search categories (collections of PTs assembled from various SOCs) have been introduced in MedDRA to group terms with close meaning