Methodology for the analysis and representation of the medical information about drugs in the Summary of Product Characteristics (SPC)

Alain Venot MD, PhD, Catherine Duclos PharmD

Department of Medical Informatics, Cochin University Hospital, 27 rue du Fg St Jacques, 75014 Paris, France

ABSTRACT

We present a methodology for the representation of the medical knowledge in the drug SPCs. It includes four steps, the two first of which are automated. All instances of a particular SPC text are gathered into a single file. Lexical analysis of the content of this file is performed and a lexicon with the occurrence of words and groups of words is built. Semantic analysis is carried out considering the concepts underlying each word of the lexicon and the most important concepts are kept. This semantic analysis results in a list of attributes which are then included in an object-oriented model. We have used this method to structure drug indications. This application clearly illustrates the advantages of this method over purely manual analysis. This method could be generalized for all categories of medical information about drugs.

INTRODUCTION

In all industrial countries, drug agencies such as the FDA are involved in the registration of new drugs the properties of which are described in a Summary of Product Characteristics. SPCs consist of several sections including the composition, indications and contraindications of the drug. They are used to construct and maintain databases about drugs. It is now important to develop a structured representation of this knowledge so that these databases can be integrated into computerized decision support systems that generate alerts and reminders for the physician. Such tools have already been shown to be effective at improving the quality of care and decreasing costs in hospital settings [1].

The information we have about drugs is of two types: pharmaceutical and medical. Pharmaceutical information includes the composition, type of pharmaceutical form and presentation of the drug. It does not include any medical data and is derived from fabrication and registration characteristics. Such information is easy to structure. In contrast, medical information about drugs summarizes the results obtained with patients during various clinical trials. The corresponding sections of the SPCs are expressed in several sentences of natural language and are far more difficult to represent in a structured way without the loss or modification of information.

The multiple levels of drug description have already been dealt with [2,3] but very few studies have been published that have dealt with either the structures of drug databases [4-6] or more specifically with the problem of the nature and representation of the medical information contained in these SPCs. The published studies have not covered all aspects of this information and have dealt with drug contraindications [7]. However this issue must be debated and discussion should be initiated concerning the options avaliable to database owners for solving these problems.

We propose herein a methodology well suited to the representation of the medical information in SPCs but not restricted to this domain. We demonstrate the great value of the preliminary lexical and semantic analysis of the SPC texts for developing an object-oriented model bringing together the information in a structured way. We first present the general methodology that we recommend. We then illustrate its use by presenting results obtained by analyzing the information concerning drug indications and discuss the validity and application of this approach.

METHODOLOGY OF THE ANALYSIS

Domain of application

We considered only the problem of representing the information contained in several sentences of natural language in which a reference terminology has been used and for which numerous instances may exist. This corresponds well for instance to the indications, contraindications and precautions sections of the drug SPCs because the expressions used in these texts are checked and standardized by the national authorities and thousands of drugs exist.

The four steps of the analysis

Rather than trying to build directly a model for the representation of this information, we thought that it would be useful to carry out a preliminary systematic study of the content of the SPCs.

Our proposed methodology involves four steps: the gathering of all the instances of a particular text into a single file, the automated lexical analysis of the content of this file, the semantic analysis of the resulting lexicon and the object-oriented modeling of this information. The two first steps are automated, the others manual. This methodology is illustrated in Figure 1.

The first step involves selecting a given section (e.g. drug indications) and gathering into a single file the corresponding texts which are usually stored in a database.

During the second step of lexical analysis, a lexicon is built and the occurrences of the various words or expressions are calculated. All the words contained in the single compiled file are classified into grammatical categories (nouns, verbs, adjectives, and adverbs) after a preliminary transformation into their canonical forms. Software is available that performs this analysis automatically. The frequencies of all the words and expressions (nominal complex units) are calculated. Some of these frequencies are useful for preliminary determination of the major concepts that must not be omitted from a structured representation.

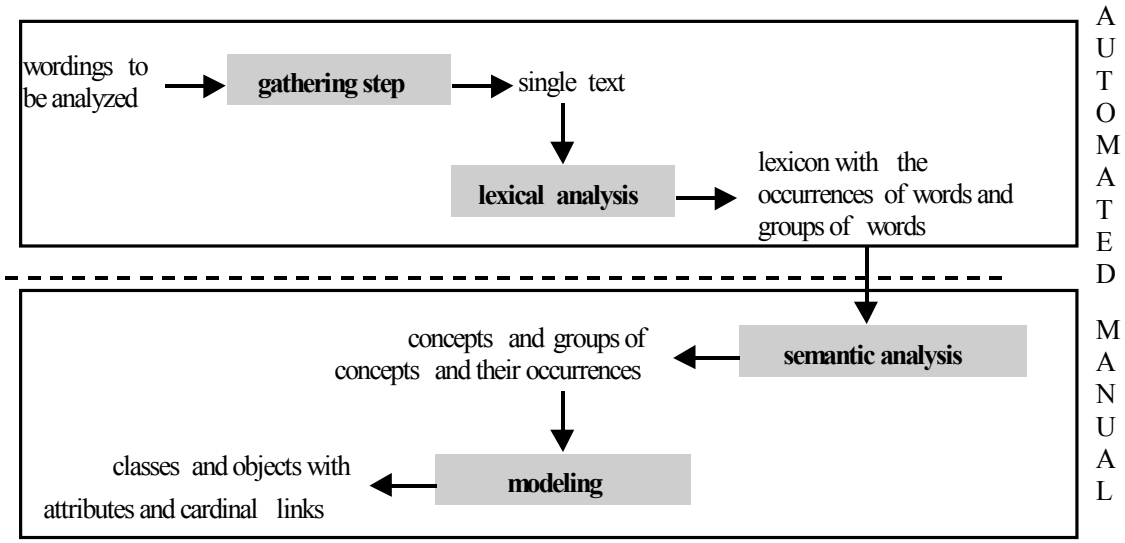

Figure 1. The four steps for the representation of medical information drugs.

The third step is the semantic analysis of the lexicon. This lexicon consists of numerous instances of various concepts. These concepts must be determined. There is currently no way to perform this semantic analysis automatically and it must therefore be performed manually. Each word or expression must be linked to a more general concept (for instance "chronic" or "progressive" disease relates to the concept of disease progression; the adjectives chronic and progressive become possible instances for this concept). The semantic analysis results in a list of these general concepts and the frequency with which instances of each concept are found in the section of information analyzed. Not all concepts are of equal importance. The aim is to achieve a final

structure that is not too complex. It is therefore necessary to decide which concepts must be kept, bearing in mind the use that can be made of them in a computerized drug prescription system. This step produces a list of concepts that can be regarded as attributes describing various aspects of the medical information about the drugs.

During the last step the results of the semantic analysis are taken into account, the list of attributes, beeing used to build an object-oriented model. The attributes identified during the semantic analysis are grouped into various classes and objects. Concepts giving additional insight into given aspects of the drug indication (e.g. the degree of efficacy of a drug for a given indication) are grouped together. Relationships between classes and objects and cardinal links are then determined.

Finally the model must be evaluated. No model can be produced that stores in a structured way all the pieces of information present in the original text. Information is therefore lost. We evaluated the extent of this loss by applying a method already used to investigate whether clinical classifications, adequately cover medical text information [8].

APPLICATION TO THE REPRESENTATION OF INFORMATION CONCERNING DRUG INDICATIONS

The value of structured drug indications

Drug indication is a key piece of information because it defines the conditions in which a particular drug may be given to a patient. Structuring drug indication in computer systems should increase the use that the physician can make of this information.

It should make it possible to develop new ways of selecting drugs that may be useful to the physician before he writes the prescription. New queries should make it possible to produce lists of drugs satisfying more sophisticated indication-based criteria. These queries could also include the association of criteria related to the indications and to other aspects of drugs such as contraindications. For instance, a more refined representation of drug indication information (and of other medical information sections of the SPCs) should make it possible to list all the antibiotics commercially available, active for the oral treatment of a given type of infection with no more than two intakes per day, not contraindicated in various physiological conditions and diseases.

New checks on drug prescription could also be implemented based on structured indications, comparing medical record data with drug indications.

Structured indications are also necessary for retrospective automated studies of drug use [9].

Automated lexical analysis of drug indications

We analyzed the wordings of the indications of 3876 drugs listed in the 1998 version of the French Vidal dictionary [10] which directly uses the SPCs checked by the national authorities. These indications were combined into a single file which was then automatically analyzed using Nomino® software [11]. A lexicon was built based on the resulting analysis. The occurrences of many words or groups of words were interesting. For instance, the adjectives acute and severe were found 785 and 263 times respectively among the almost 10 000 indications gathered. Similarly the nominal complex units "symptomatic treatment" and "auxiliary treatment" were found 815 and 375 times respectively, illustrating the importance of the concepts behind these expressions.

Semantic analysis based on the indication lexicon

Based on the lexicon we identified more general concepts, as illustrated by Table 1. For instance, "bronchodilator treatment" can be regarded as an instance of the general concept "type of activity" which was found 1101 times in the SPCs. Similarly, the various instances, "symptomatic, preventive... treatment" relate to the concept of "type of action" of the treatment, this concept being found 1254 times in the SPCs. Other examples are given in Table 1.

The semantic analysis resulted in a list of 21 concepts that we considered as the attributes of the objects to be determined.

Object-oriented modeling

The 21 attributes were grouped into five classes or objects using the formalism described by Coad and Yourdon [12]. The main object gathers the attributes describing the objectives of the indication. The disease concerned by the indication is a fundamental attribute of this object but the type of action, and type of activity are also attributes necessary to represent the information present in the original text of the

indication. Another object gathers attributes describing the degree of efficacy of the drug for the given indication. This object includes attributes such as the strength of action and the level of indication, which can be very useful if one wishes to create a list of all the drugs active against a given disease, with a similar level of efficacy. Additional attributes were gathered into an object specifically devoted to drugs used during procedures (e.g. anesthesia, radiological investigations) and not used to treat a disease or symptom.

For a random sample of 100 drugs we studied the coverage by the model of the information contained in the indications. We used an evaluation method described elsewhere [8]. For each indication in this sample, a value of zero was given if there was no reasonable match between the wording of the free text and the structured representation. A value of one was given for an approximate match and a value of 2 for a perfect match. We found that 95% of the indications were represented by the model with no loss of information. There were no cases of total discrepancy between the initial text and the structured representation. Overall the model obtained the score of 1.95, the maxximum possible score being 2.

Examples of words or groups of words in the lexicon	**Main grouping concepts (attributes)**	**Occurrence of the concept**	**Class and Object**
Bronchodilator treatment	Type of activity	1101	Objective of the indication
Symptomatic treatment, Preventive...	Type of action	1254	
To propose, to use...	Level of indication	890	Degree of efficacy
Complementary treatment, associated treatment...	Strength of action	869	
Reference treatment, relay treatment...	Therapeutic situation	259	
Local treatment, general treatment...	Action field	418	

Table 1 : Examples of semantic analysis based on the indication lexicon: groups of words associated with the concept of treatment

DISCUSSION AND CONCLUSION

The complementary advantages of lexical and conceptual approaches for managing medical information have already been stressed [13]. We propose here a methodology based on these approaches, which is ideal for structuring the medical information within SPCs.

This methodology has many advantages over the classical approach, which involves manual analysis of the information within SPCs. These advantages should be emphasized and particular aspects discussed.

It is very easy to gather all the SPCs sections under consideration into a single file. In the example presented here, we have gathered together indications for drugs at the commercial product level. The same drug with two different commercial names and identical registered indications therefore appears twice in our analysis. It would also have been possible to perform the same operation by gathering together the indications at a generic level taking into account the International Non proprietary Name (INN), pharmaceutical form and strength. But such an approach is problematic. It may happen in several countries that two identical drugs marketed by two different firms have not exactly the same registered indications.

The preliminary automated building of a lexicon of words and groups of words permits to give a flat view of the various categories of information present in the thousands of SPCs of all the drugs. Such a view is impossible to derive from simply reading the SPCs. The semantic grouping and calculation of occurrences makes it possible to distinguish between words and groups of words that are particularly

important because they are frequent and those that are less important because they are restricted to a small number of drugs. For those words and groups of words restricted to very few drugs, the frequency of prescription of the drug has also to be taken into account to retain only the important categories of information in the final model. Lexical analysis also makes it possible to evaluate the homogeneity of the vocabulary used in the various sentences and may be used to improve and standardize the writing of SPSs.

Semantic analysis produces a list of attributes considerably simplifying the modeling step. It is necessary to simply group the remaining attributes into classes and objects. The proposed approach is effective, as shown by the evaluation performed for the drug indication model because information was total for more than 95% of drugs.

One disadvantage of this method is that it is not possible to carry out automated analysis at the moment. The development or improvement of semantic networks for drugs would be useful for the development of automated procedures.

The final step presented here is the building of an object-oriented model. Other information modeling techniques producing relational models or conceptual graphs could also be considered.

Terminology problems must not be underestimated. For each attribute of the model, a set of possible values has to be adopted. If possible, sources with standardized terminology should be preferentially used. The most difficult terminology problem is the choice of a classification system for diseases and symptoms. This choice is crucial if the structured indications are to be linked with the medical record to check prescriptions. However this problem is beyond the scope of this paper.

It would be of great value to apply this methodology to the various sections of medical information in the SPCs. We are currently studying the precautions for use of drugs. New types of reminders could result from the systematic use of structured drug information in drug prescription systems.

ACKNOWLEDGMENTS

We would like to thank Christophe Chailloleau for help with the evaluation of this model and Mrs Bonjean from OVP-Vidal who provided the CD-ROM of all the drug indications.

REFERENCES

1. Evans RS, Pestotnik SL, Classen DC, Clemmer TP, Weaver LK, Orme JF, Lloyd JF, Burke JP. A computer-assisted management program for antibiotics and other antiinfective agents. New Engl J Med 1998; 338: 232 – 238.
2. CEN/TC 251/PT014 ENV 12610, 1997 Medical informatics - Medicinal product identification.
3. Sperzel WD, Broverman CA, Kapusnik-Uner JE, Schlesinger JM. The need for a concept-based vocabulary as an enabling infrastructure in Health Informatics. Proc AMIA Annu Fall Symp 1998, 865-869.
4. Milstein C, de Zegher I, Venot A, Sene B, Pietri P, Dahlberg B Modeling drug information for a prescription-oriented knowledge base on drugs. Methods Inf Med 1995; 34: 318-327.
5. Francois M, Joubert M, Fieschi D, Fieschi M Modeling and implementing a database on drugs into a hospital intranet. Comput Biol Med 1998 Sep;28(5):553-65
6. Keller F, Frankewitsch T, Zellner D, Simon S, Czock D, Giehl M Standardized structure and modular design of a pharmacokinetic database. Comput Methods Programs Biomed 1998 Feb;55(2):107-15
7. Liu JH, Milstein C, Séné B, Venot A. Object-oriented modeling and terminologies for drug contraindications. Meth Inform Med 1998; 37: 45 - 52.
8. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classification. JAMIA 1996;3:224-233.
9. Coste J, Séné B, Milstein C, Bouee S, Venot A. Indicators for the automated analysis of drug prescribing quality. Meth Inf Med 1998; 37: 38-44
10. Dictionnaire Vidal, Paris, OVP-Editions, 1998.
11. Nomino software, www.ling.uqam.ca/nomino, 1998
12. Coad P, Yourdon E. Object-oriented analysis. New York: Prentice-Hall, 1991.
13. Rassinoux AM, Miller RA, Baud RH, Scherer JR Modeling concepts in medicine for medical language understanding. Meth Inform Med 1998; 37: 361-372.