C. Duclos, A. Venot

Structured Representation of Drug Indications: Lexical and Semantic Analysis and Object-Oriented Modeling

Département de Biostatistique
et Informatique Médicale,
UFR Cochin Port Royal,
Université René Descartes,
Paris 5 et AP-HP, Paris, France

Abstract: No standardized representation of drug indications is currently available that could be used in drug knowledge bases. We describe an object-oriented representation of indications that should make it possible to develop new tools for selecting drugs and checking prescriptions in computerized drug prescription systems. The model was developed using the results of a lexical and semantic analysis of drug indications, collected into a single file and processed using natural language processing software. It distinguishes both the diseases for which the drug may be given and the efficiency of the drug for a given indication. Two aspects of the model were evaluated: the differences if two independent evaluators filled the attributes independently and the loss of information induced by the use of the model. A system based on this model, making it possible for the physician to select all the drugs satisfying various criteria, is also presented.

Keywords: Drug Indications, Knowledge Base, Information Modeling, Electronic Prescribing

1. Introduction

Computerized drug prescription systems are being developed [1-3], making it possible for physicians to check their drug prescriptions. These checks are based on the use of a drug database containing structured information. Various studies have described ways in which several aspects of drug information can be correctly structured [4], for example, those concerning contraindications [5] and dose regimens [6]. The correct way to structure the "drug indication" information has not yet been determined.

Drug indication is a key piece of information, because it defines the conditions in which a particular drug may be given to a patient and with which dosing. It is obtained from clinical trial results and forms part of the Summary of Product Characteristics (SPC) established by the drug regulatory affairs authority. Existing drug databases either use free text to describe drug indications or associate very restricted coded data with free text. For instance, in the USA, FirstDatabank database, NDDF associates a code based on ICD-9-CM with each indication [7]. The French Theriaque database [8] also encodes attributes such as the type and level of efficacy of the drug.

A more complete structuring of drug indication in computer systems should increase the functions available to the physician. It may make it possible to develop new ways of selecting drugs that may be useful to the physician before she/he writes the prescription. New checks on drug prescription could be implemented by comparing medical record data with drug indications. It may also be useful for retrospective automated studies of drug use [9].

We attempted to study and model the information corresponding to drug indication. Section 1 presents the results of a lexical and semantic analysis of indication wordings that we performed to determine the type of information provided for indication and its frequency of occurrence. This analysis was then used to develop and evaluate an object-oriented model of drug indication. The third section illustrates the use of this model for the design of software that is capable of listing drugs fulfilling various drug indication criteria.

2. Lexical and Semantic Analysis of Drug Indications

2.1 Methodology of the Analysis

We used the following methodology to distinguish between the various categories of information and to evaluate their relative importance. We used the indication wordings of the French

Table 1 Frequency of adjectives often occurring in the file containing the indications of 3876 drugs.

Adjective	Frequency
Symptomatic	930
Acute	785
Chronic	484
Local	368
Severe	263
Adjuvant	101
Preventive	90

Vidal® dictionary [10]. This is an official definition based on the SPCs and checked by the national authorities. The indications (nearly 10,000) for 3,876 drugs listed in the 1998 version of the Vidal® dictionary were combined in a single file. The words were first grouped according to root or topic then assigned to grammatical categories (nouns, verbs, adjectives, and adverbs) and grouped according to notional entities (nominal complex units) the frequencies of which were calculated. This processing was performed automatically using the Nomino® software for text analysis [11]. The lexicon built with this software made it possible to perform manually a second step of semantic analysis registering the principal semantic structures contained in the indication texts, and their frequencies.

2.2 Results of Lexical and Semantic Analysis

There were 152,847 words in the file containing all the indications (2,778 different words).

The verbs "to use" and "to propose" occurred 571 and 133 times, respectively. This was because "to use" is an integral part of the expression "drug used in. . ." and "to propose" is found in the expression "drugs proposed for. . .". These expressions are reserved for drugs that are extensively prescribed (e.g., phlebotonics) but for which either the pharmacological effect has been proved but not the clinical efficacy (to propose), or for which there is no evidence of pharmacological action or clinical efficacy (to use). The frequencies of several adjectives that were often used in drug indications are given in Table 1.

Semantic analysis of the lexicon results in various groupings depending on the disease or symptom associated with the indication, the various characteristics of the treatment, and the medical procedures that necessitate the use of a drug. The main groupings and their frequencies are given in Table 2.

Table 2 Frequency of different semantic structures found in the indications related to disease, treatment and drugs used for medical procedures.

Main groupings of concepts	Frequency	Example (if illustrative)
Groupings with various instances of the concept of disease		
Disease	4650	Alzheimer's disease
Disease's location	1486	Low level urinary infection
Disease and etiology	1078	Trichophyton rubrum folliculitisis
Disease (or symptom) and severity	1046	Severe pain
Disease and chronicity	698	Chronic hepatitis
Disease and progression	60	Evolutive ulcer
Disease and level of certainty	45	Suspicion of invasive phenomenon
Disease and age	23	Recent compression
Disease and extent	26	Diffuse interstitial fibrosis
Complications of a disease	24	Secondary infection of bronchitis
Groupings with word treatment		
Treatment and its type of pharmacological action	1101	Bronchodilatator treatment
Symptomatic action of a treatment	984	Symptomatic treatment
Treatment given in association with another treatment	869	Complementary treatment
Treatment and its type of action (local or general)	418	Local treatment
Treatment chronicity	173	Basic treatment
Preventive action of a treatment	170	Preventive treatment
Relay treatment	124	Relay treatment
Reference treatment	109	Reference treatment
Curative treatment	100	Curative treatment
Treatment to be given in emergency situations	26	Emergency treatment
Expressions related to indications of drugs used for medical procedures		
Type of procedure	382	Radiological investigation
Area of the body concerned by the procedure	297	Locoregional anesthesia

3. Development of a Structured Representation of the Drug Indications

3.1 Building the Model

A structured representation was built using the results of the previous analysis and object-oriented formalism [12]. The initial version of the model was refined using a random sample of 100 indications, to obtain the final version presented here.

Our analysis showed that two parts of the description are important: the description of the disease for which the drug is indicated and information about the efficiency of the drug for that disease. Additional information is also required describing the indication of drugs used in medical procedures.

3.1.1 Information about the Aim of Drug Treatment for a Given Indication

The disease for which treatment with the drug is efficient is a fundamental part of the indication. Information about the disease should be split into three categories:

– *Generic diseases.* The broad concept of conditions targeted by the indication (e.g., dermatosis with squamous components).

– *Specific diseases.* Very often, the text of the indication also gives more specific information about the conditions that can be treated (e.g., small psoriatic lesions).

– *Excluded diseases.* These are also frequently given. They are a subclass of the generic disease group but are the diseases for which the drug is not indicated (e.g., excluding large psoriatic lesions).

The type of patient. The indication may specify the type of patient for whom the drug is indicated (e.g., infant, elderly, pregnant).

The type of action. This provides information about the expected result of the treatment (e.g., preventive, curative, symptomatic). It may also state that the drug is used purely for diagnostic procedures.

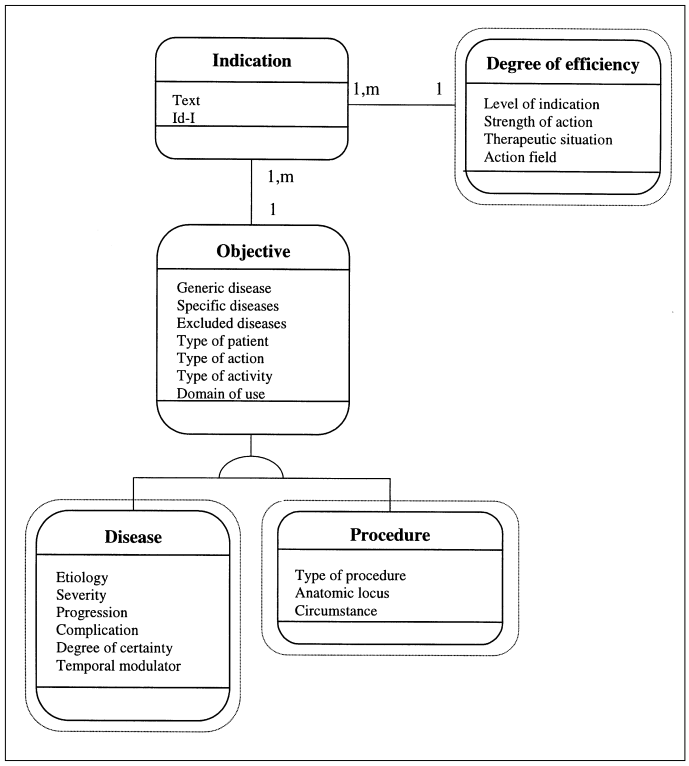

Fig. 1 Object-oriented model of drug indication (formalism defined by P. Coad and Yourdon [12]).

The type of activity. This is the pharmacological property responsible for the action of the drug (e.g., bronchodilator).
The domain of use specifies whether the drug prescription is restricted to hospital use or can be used in primary care.

Other details are often given in the indication for drugs not used for procedures.
Etiology specifies the cause of the disease corresponding to the indication (e.g., staphylococcal infection).
Severity provides information about the seriousness of the symptom or disease required for use of the drug (e.g., severe pain).
Disease progression gives details about the time course of the disease (e.g., chronic hepatitis, evolutive ulcer).
Complication gives details about of an underlying disease (e.g., secondary infection with bronchitis) or associated sign (inflammatory and painful state).
The degree of certainty specifies in the indication whether a drug can be prescribed even if the diagnosis is not sure (e.g., suspicion of an invasive phenomenon).
The temporal modulator provides more information about the disease (e.g., recent peptic ulcer).

3.1.2 Drugs Used for Medical Procedures

Three specific attributes can be distinguished for drugs used during procedures.
The type of procedure. This defines the procedure for which the drug is useful (e.g., anesthesia, radiological investigation).
The anatomic locus (or physiological system). This describes the part of the body involved in the procedure (e.g., mucous anesthesia, lung function).
The circumstances of the procedure. This provides more details about the procedure in which the drug is used (e.g., anesthesia before urological investigation).

3.1.3 Information about the Efficiency of the Drug for a Given Indication

The various attributes concerning the potential efficiency of the drug can be grouped.
The level of the indication relates to the recognised efficacy of the drug. Standard vocabulary is used by official authorities for example distinguishing three possible levels (e.g., drug indicated for, proposed for, or used for).
The strength of action of the drug indicates whether the drug can cure the disease by itself or whether it must be associated with another treatment.
The therapeutic situation. This makes it possible to define an order of preference for all the available drugs (first intention treatment, relay treatment, reference treatment).
The action field specifies whether the treatment is local or general

3.1.4 Object-Oriented Representation

All these attributes can be grouped into five class&Object. Our final object-oriented model of drug indication, using the formalism of Coad and Yourdon [12] is shown in Fig. 1.

The object "Indication" contains only an identifier and the indication which is expressed using free text. It is

Table 3
Evaluation of the differences in drug indication information representation between two evaluators working independently.
Study realized with 100 indications.

Attribute	Number of times attribute described	Number of times (percentage) where same attribute was used for same information	Number of times (percentage) where same attribute was used for different information	Number of times (percentage) where different attributes were used
Level of indication	100	100 (100)	0 (0)	0 (0)
Strength of action	17	14 (82)	0 (0)	3 (18)
Therapeutic situation	9	6 (68)	0 (0)	3 (33)
Action field	4	4 (100)	0 (0)	0 (0)
Generic disease	98	93 (95)	4 (4)	1 (1)
Specific diseases	25	22 (88)	0 (0)	3 (12)
Excluded diseases	9	8 (89)	0 (0)	1 (11)
Type of patient	14	12 (86)	0 (0)	2 (14)
Type of action	34	33 (97)	0 (0)	1 (3)
Type of activity	11	10 (90)	0 (0)	1 (9)
Domain of use	0			
Etiology	38	32 (84)	0 (0)	6 (16)
Severity	12	9 (75)	0 (0)	3 (25)
Progression	14	10 (71)	0 (0)	4 (29)
Complication	16	14 (87)	0 (0)	2 (12)
Degree of certainty	4	3 (75)	0 (0)	1 (25)
Temporal modulator	6	5 (83)	0 (0)	1 (17)
Total (all attributes)	415	378 (91)	4(1)	33 (8)

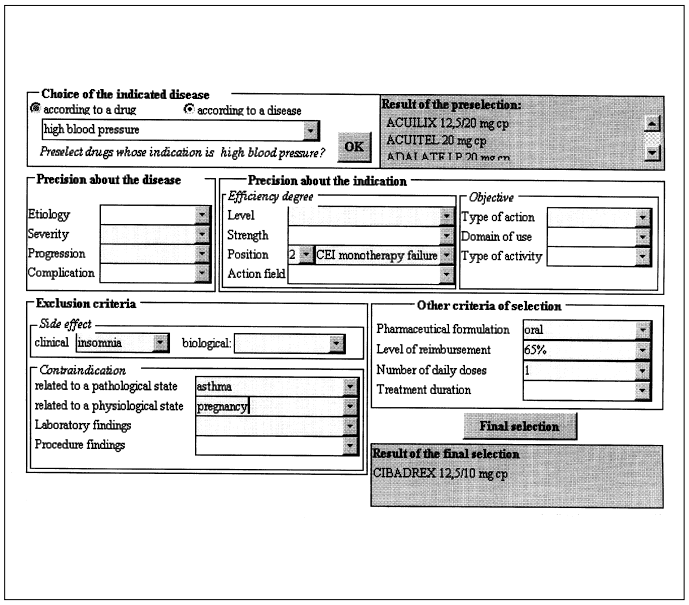

Fig. 2 Main screen of the system for constructing a query based on the indication and other criteria. This figure gives as an example a search for all the drugs indicated for high blood pressure but used only after the failure of conversion enzyme inhibitor treatment, requiring only one oral dose a day, not giving insomnia as a possible side effect and not contra-indicated by asthma and pregnancy.

connected by an instance relationship to the class "objective" and to the object "degree of efficiency". The "objective" class is specialised into "disease" and "procedure".

3.2 Evaluation of the Model

3.2.1 Methodology

We evaluated two aspects of the model. First, we investigated whether two evaluators using the model independently reached the same conclusions. Second, we investigated the extent to which the model covered all the information contained in the indication texts.

A random set of 100 indications was used to determine the interindividual variability of information representation with the model. Only drugs used for treatment were selected. Two evaluators, working independently, identified the most important information units within the indication and described the attributes of the indication. The generic wording at the beginning of the indication, the caution for use and dose regimen elements were not taken into account.

The ability of the model to cover all the information provided was evaluated by a third evaluator who did not take part in the interindividual variability study. His role was to evaluate whether the information provided by the various attributes fitted that of the indication written as free text and to identify any loss of information. A method has already been described to evaluate whether clinical classifications adequately cover medical text information [13]. 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 if there was an approximate match and complete matches were given a value of 2.

3.2.2 Results of the Evaluation

Variability Study

Three situations were observed; The two evaluators represented the same information using identical attributes, the same attributes were used for different information, or different attributes were used for the same information (Table 3). Overall the interindividual variability was low for most attributes. Nevertheless there was substantial variability for the attributes disease progression, severity and therapeutic situation. There was a correlation between the representations of the attributes disease progression and severity: both attributes were used by the two evaluators to represent "acute". The adoption of a convention would have reduced to 8% the proportion of cases in which the correspondence between the evaluators was not perfect. The same was true for the attributes "strength of action" and "therapeutic situation".

Coverage by the Model of the Information Contained in the Indications

Ninety-five percent of indications were represented without loss of information using the model. Only 5% of indications were partially represented and there were no cases of total discrepancy between the initial text and the structured representation. This model scored 1.95 over 2 using Chute's method [13].

4. The Use of Structured Drug Indications

We designed some demonstration software to illustrate the value of the model. This system makes it possible to construct a query graphically permitting the user to extract from the drug database all the drugs that satisfy various criteria.

First the physician selects criteria concerning drug indications:

- a disease, with optional specifications such as the type of patient, the etiology, severity, progression and complications.
- the degree of efficiency (level, strength, therapeutic situation and action field).
- the objective (type of action, type of activity, domain of use).

The selection may then be extended to other criteria. For example, the pharmaceutical formulation, the level of reimbursement, the number of daily doses and treatment duration can be selected.

Finally, it would also be useful to be able to exclude drugs satisfying exclusion criteria related to contraindications, interactions and side effects. We adopted the model of Liu et al. [5] for contraindications. This model distinguishes contraindications corresponding to pathological state, physiological state, and findings of investigations and procedures. Drug interactions and side effects can also be taken into account. This system was implemented using Access 97®. Figure 2 shows the user interface developed to facilitate the building of queries.

5. Discussion and Conclusion

We have designed a model to represent information concerning drug indications. Several aspects of this model will be discussed below such as the methodology used, the validity of the model and its possible generalisation, the vocabulary and coding systems to be associated with its value.

The preliminary text and semantic analysis presented in Section 2 was of great help in this study. The processing of indications collected told us a great deal about the types of information present in drug indications. The frequencies of the various categories provided quantitative information making it possible to determine the relative importance of each type of information. These tools assist in modeling by directly identifying the attributes that must be described for each object. This approach has not previously been used for information modeling in medicine and we strongly recommend its use to strengthen and improve the modeling step.

The evaluation study demonstrated that, in most cases, the model adequately represented all the information contained in the indication text. Nevertheless, caution is required when using the model in practice. Guidelines must be drawn up so that the attributes can be described unambiguously if there is significant interindividual variability.

The drug indications used to construct the model were written in French and checked by the national authorities. We need to determine whether the same model can be applied to other countries. Most attributes are not nation-specific. However the attribute "level of indication" was described as laid down by French national authorities. The wording used to describe this attribute is specific to France, but the concept is not. In the Martindale compendium [14], drug indications are formulated using bibliographic references. There is a section corresponding to drug use circumstance which could be used to describe the attribute "level of indication" by considering the style used in the wording, the present perfect "has been used" implying that drug efficacy has not been proved by clinical trials.

We did not consider the classification systems suitable for the attributes related to the disease. The choice depends on the use made of structured indications. If they are to be used for drug selection, the choice is not crucial. However if they are to be linked with the medical record of the patient to check prescriptions requiring the indication information, then complete compatibility is necessary. Only classification and coding systems with sufficient details such as SNOMED [15] or ICD 10 [16] could be adequate for coding items related to diseases. Other sources of vocabulary such as MedDRA (Medical Dictionary for Drug Regulatory Affairs) will soon be available and must be considered [17]. A study carried out of release 1.5 has shown that 37% of terms present in the indications are not described correctly [18].

We have demonstrated the use of this model to construct software to select drugs meeting various criteria. Such selections are not yet possible in commercial systems but should be easy to implement once databases with structured drug indications become available. Other potentially important uses could be made of this modeling approach. Structured indications should make it possible to implement sophisticated checks of prescriptions by matching the patient's medical data and the indications of the drugs that have been prescribed. It may also be possible to implement automated retrospective analysis of the appropriateness of prescriptions using both structured patient and drug data.

Acknowledgement

The authors thank Dr. Christophe Chailloleau for his help with the evaluation of this model.

REFERENCES

1. Linnarson R. Decision support for drug prescription integrated with computer-based patient records in primary care. Med Inform 1993; 18: 131-42.
2. De Zegher I, Venot A, Milstein C, Séné B, De Carolis B, Pizzutilo S. OPADE: Optimization of drug prescription using advanced informatics. Computer Methods Programs Biomed 1994; 45: 131-6.
3. 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: 232-8.
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. Method Inform Med 1995; 34: 318-27.
5. Liu JH, Milstein C, Séné B, Venot A. Object oriented modeling and terminologies for drug contraindications. Method Inform Med 1998; 37: 45-52.
6. Sene B, Venot A, de Zegher I, Milstein C, Errore S, de Rosis F, Strauch G. A general model of drug prescription. Method Inform Med 1995; 34: 310-7.
7. National Drug File (NDDF), www.firstdatabank.com, 1998.
8. Husson C, Mangeot A. Theriaque: Information sur le médicament et aide la prescription. In: *Informatique et Médicaments.* Venot A, Degoulet P, eds. Paris: Springer-Verlag, 1989: 208-20.
9. Coste J, Sene B, Milstein C, Bouee S, Venot A. Indicators for the automated analysis of drug prescribing quality. Method Inform Med 1998; 37: 38-44.
10. *Dictionnaire Vidal,* Paris: OVP-Editions, 1998.
11. Nomino software, HYPERLINK http://www.ling.uqam.ca/nomino.
12. Coad P, Yourdon E. *Object-Oriented Analysis.* New York: Prentice-Hall, 1991.
13. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classification. JAMIA 1996; 3: 224-33.
14. *Martindale, the extra Pharmacopoeia.* London: Pharmaceutical Press, 1998.
15. Rothwell DJ. SNOMED-based knowledge representation. Method Inform Med 1995; 34: 209-13.
16. *International Classification of diseases and related health problems* (tenth revision). Geneva: World Health Organisation, 1993.
17. Brown EG, David M. The Medical Dictionary for Regulatory Activities (MedDRA): a survey of regulatory authorities' approaches to implementation. Int J Pharmaceut Med 1998; 1: 23-7.
18. Brown EG, Clark E. Evaluation of MEDDRA in representing medicinal product data sheet information. Pharmaceut Med 1996; 10: 111-8.

Address of the authors:
Prof. Alain Venot,
Département de Biostatistique
et Informatique Médicale,
Achard 7, Hôpital Cochin,
27 rue du Faubourg Saint Jacques,
75679 Paris Cedex 14,
France.