Pharmacology and pharmacokinetics provide information that makes it possible to understand drug-body interactions. Such information is useful in the reasoning process (1) for choosing a drug to prescribe; (2) for comparing drugs; (3) for giving advice on drug administration; (4) for identifying contraindications, interactions, or therapeutic strategies in physiological conditions or particular diseases; and (5) for listing drugs according to some of their properties.

Pharmacokinetics concerns the descriptive and quantitative study of what happens to a drug in the body to which it is administered. Mathematical modeling is used for the study of the dynamic processes involved.

In compartmental modeling, the body is represented as a whole, made up of virtual compartments between which exchanges occur. The compartmental approach uses mathematical models comprising differential equation sets to predict the serum level of a drug according to the dose, the intake interval, the number of administrations, and the mode of administration. These models have been integrated into software to assist clinicians in determining the optimal dosage schedule.
Thus, modeling choices available do not allow direct handling of knowledge data by the clinician and need to be integrated into a software system for particular tasks such as dosage schedule optimization.

Most pharmacokinetics data are found in textual form in drug compendia and drug databases. Their sources are either summarized data produced by the drug regulatory agencies or data from studies published during drug development. In these databases, the information about pharmacokinetics is always expressed in natural language. Sometimes information is organized in sections with labels representing generic concepts such as “absorption,” “distribution,” “metabolism,” and “elimination” or narrower concepts such as bioavailability, bioavailability by oral route, bioavailability by oral route for tablets, bioavailability by intramuscular route, and effect of food on absorption. Finally, the contents of the pharmacokinetic description are similar, whether the source is European (e.g., Vidal) or American (e.g., AHFS DI).

Pharmacokinetics data are written in natural language, which can be stored in electronic databases or drug compendia for consultation by practitioners but is not well suited to the development of automated prescription assistance. The conceptual and formal representation of the domain ontology is essential to handle the knowledge effectively in computerized systems. The OpenGALEN Common Reference Model contains pharmacokinetics descriptions such as drug absorbing, excreting, metabolizing, and protein binding process, but no current drug ontology uses it. The “Alcohol and Other Drugs Thesaurus” describes an ontology of pharmacokinetics aimed at expressing pharmacokinetics processes quantitatively. We are not aware of a more finely detailed model of pharmacokinetics.

The objective of this work was to build a finely structured representation of pharmacokinetics that could be used in a computerized system to facilitate the use of pharmaco-therapeutic information currently contained only in electronic drug compendia. Prescribers could then use a computerized system to get answers to such questions as: (1) which anti-hypertensive drugs are mainly eliminated in urine, (2) which anti-infective drugs diffuse in bone, and (3) for which anti-asthmatic drugs are there pharmacokinetic data for children. It would also be easy to build comparative lists to give an overview of differential properties of drugs (for example: compare the metabolism of cardiac glycoside drugs).

Our aim was to elaborate an object-oriented model that can represent in a structured and exhaustive way the information contained in pharmacokinetics sections of SPCs. Here we describe our methodology for knowledge acquisition based on domain analysis and natural language processing, and the method used to evaluate the content coverage and the accuracy of the developed model. We present both the object-oriented model for representation of pharmacokinetics information in UML formalism and an evaluation of the model.

Methods

We used a two-step approach to identify the information elements contained in the pharmacokinetics section of the SPCs. First, we analyzed the compartmental and physiology-based pharmacokinetics modeling by identifying the main concepts on which they rely. Second, we used natural language processing tools to analyze the contents of pharmacokinetics texts extracted from SPCs. We then built an object-oriented model of pharmacokinetics information, which was evaluated for its ability to represent the initial text information and maintain its sense.

Knowledge Acquisition Method

Establishing the Foundations of the Model

First, we studied established pharmacokinetics knowledge using educational pharmacokinetics sources. We manually identified the fundamental concepts of pharmacokinetics and defined the experimental schedule required to produce real data for use in compartmental or physiology-based modeling.

This first stage of analysis led to a preliminary model of the domain. This model describes the categories of knowledge found in pharmacokinetics. These categories are considered the main concepts that should be found in pharmacokinetics texts and are used to group terms in the following analysis.

Terminological Analysis of Pharmacokinetics Section of SPCs

Data sources. The pharmacokinetics texts were taken from the Vidal drug database that contains the information specified by the French drug agency in the SPCs. 5,293 different drugs had SPCs with a pharmacokinetics section, and 1,950 unique texts were extracted (two or more different drugs can have the same pharmacokinetics text). We combined them in a single file.

Natural language processing tool. We used the robust syntactic analysis tool Lexter to analyze the text file. It performs a morpho-syntactic analysis of the sentences of a French corpus, and yields a dependency network of candidate terms (CTs). CTs are presented in uninflected form (lemma form). They are either simple (single noun, adjective, verb, adverb) or complex (words arrangements in the phrase such as nominal syntagma or adjec-
tival syntax (adv/st + adj/adjective)). Complex CTs are divided into a head and an expansion (e.g., the nominal syntax “substantial protein binding” is separated into a head “protein binding” and an expansion “substantial”). Through the terminological network provided by Lexter, each CT is linked to all of the CTs for which it is the head or the expansion. Each CT is also linked to the textual units where it appears (e.g., the CT “plasmatic protein,” a nominal syntagma, appears in the textual unit “the level of plasmatic protein binding is low [14 to 21%]”).

**Terminological exploration.** The documentation analyzed was substantial and therefore the expected number of CTs was high. The most common words occur frequently and constitute the majority of the word tokens in the text. We ordered CTs according to their syntactic category and their frequency of occurrence and then examined the CTs that occurred more than 3 times.

- **Selection of CTs.** Pharmacokinetics-specific CTs were manually assigned to a main concept defined in the preceding analysis. CTs with important meaning not directly related to pharmacokinetics were assigned to a “nonspecific” concept. This first analysis led to the identification of the main semantic groupings.

- **Relation between CTs.** The lexical environment of words with similar meaning (such as binding, fixation) was explored. The terminological network of these CTs and the textual units in which they appeared helped us to find the co-occurring CTs. These CTs were listed according to their frequency of co-occurrence. We repeated this analysis for all the pharmacokinetics-specific CTs.

### Object-oriented Model Building

For CTs with similar meanings we compared the lists of co-occurring CTs and deduced common underlying concepts (e.g., absorption is a reaction, metabolism is a reaction, absorption occurs in the stomach, metabolism is hepatic, so that we can deduce that a reaction occurs in a location). These concepts were then organized either as classes, attributes, or class relationships. They were then added to the basic model (class addition, class partition, class generalization, or class specialization). The representations of a sample of 10 randomly selected texts were assessed, and the findings were used to refine the model.

### Object-oriented Representation

To represent the object-oriented model of pharmacokinetics, UML formalism was chosen. The relationships are either:

- **Inheritance relationship:** Class A inherits from Class B or Class B subsumes Class A (Class A _ ► Class B)
- **Aggregation relationship:** Class A is part of Class B or Class B is composed of Class A (Class A _ ◊ Class B)
- **Association relationship:** Class A is associated with Class B (Class A _ Class B)

The cardinality constraints are noted as follows: 0...1 = 0 to 1; 1...* = 1 to many, 0...* = 0 to many, 1 = 1, * = many.

### Evaluation Method

#### Material

To perform the evaluation, we used 100 pharmacokinetic texts randomly extracted from the 1,950 French SPC texts available. The selected texts had between 1 and 26 sentences and 14 to 717 words. The sentences were manually converted by one of us (CD) in an Access database that reproduced the structure of the model. This took almost 3 hours per text although the time varied according to the length of text. A total of 5,225 class instances and 9,347 attribute instances were created. All of the attributes specified in the model were filled at least once. To facilitate assessment, we used colored forms that presented (1) the identification of the drug (name, INN, dosage, route), (2) the whole pharmacokinetics text, and (3) each sentence of the text in both natural language and structured format. We used a set of colors to distinguish class attributes, attribute values, and relationships (Fig. 1).

### Evaluation Criteria

The following criteria were used to evaluate the model:

- **Completeness:** measures the ability of the model to represent all information units contained in the initial text; it also shows if there are concepts missing from the pharmacokinetics model. This criterion has four values: completely, almost completely, little or not represented.

- **Semantic accuracy:** measures the ability of the model to store the meaning of the initial whole text. This criterion shows the descriptive competence of the pharmacokinetics model and the potential ambiguity of some concepts. It has four values (entirely, substantially, little, or not distorted).

### Method

For each structured text, eight experts working independently were asked to assess these two criteria. They compared the structured representation to the initial text for each sentence and assigned a value to each criterion. When they found a defect of representation or of meaning they attributed a cause: either a defect of the model or a defect of the textual source. For example: the sentence “in case of renal impairment it is not necessary to modify the dosing schedule” would not be represented by the model because it is not pharmacokinetics but dosing schedule.

In this first analysis, experts assigned a value to the whole text according to the model point of view.

The eight evaluators first worked on a preliminary sample of five texts to familiarize themselves with the model and the evaluation principle. For the evaluation, 25 structured texts (selected from the sample of 100 texts)
were assigned randomly to each evaluator. Each of the 100 texts was evaluated by two experts, each blind to the findings of the other. Any disagreement between evaluators was resolved using the Delphi method: for each evaluator involved in the disagreement, results of the evaluation of the other evaluator were presented and then a new evaluation was asked.

Statistics
We determined the 95% confidence interval of the frequency of each criterion value. We performed ANOVA non-parametric tests to determine:

- If there was a significant difference between the corpus of texts assigned to each evaluator (number of sentences)
- If there was a significant difference between the evaluators' responses.

Results
Knowledge Acquisition Results
Generic Model of Pharmacokinetics
We constructed a preliminary global model of pharmacokinetics from the results of the domain study using UML formalism (Fig. 2). The model distinguishes three entities: the administration protocol, the real pharmacokinetic process description and the mathematical model building.

From an administration protocol in which the route and dosage regimen are specified (dose, administration frequency, number of intakes): the real pharmacokinetics process can be described as such:

- The parent compound or its metabolites undergo a series of reactions including absorption, distribution, metabolism and elimination processes.
• It is studied by assays for the drug in samples (of, for example, blood or urine) collected at determined times.

A mathematical pharmacokinetics model can be elaborated using the experimental data and its parameters (for example half-life, clearance, distribution volume).

Lexico-semantic Analysis

The pharmacokinetics text file contained about 300,000 words and was composed of texts of 5 to 1,043 words.

The Lexeter lexico-semantic extraction based on the 1,950 pharmacokinetics texts provided a lexicon of 206,500 entries (17,520 different CTs) and 20,761 textual units. Of the 17,520 different CTs, 3,132 occurred more than 3 times and only 222 occurred more than 100 times. Nouns and nominal syntagma were the most frequent syntactic units.

We selected CTs representing direct instances of major pharmacokinetics concepts described in the generic model, such as CTs related to an administration protocol, a process, an observational datum and a parameter. There were 592 such specific pharmacokinetics CTs. We also selected CTs representing useful concepts such as quantification, variation, type of disease, or a population. These nonspecific terms represented 535 CTs. Table 1 shows an example of how the more frequent CTs were grouped according to their similar meanings in pharmacokinetics-specific or nonspecific concepts.

Study of the lexical environment of CTs with similar meanings yielded the set of common CTs that frequently co-occur. For example, fixation was most frequently described with CTs related to a binding site, a numerical value, a substance name, a precision, an ordinal value, and a study type. These CTs also co-occurred with “binding.” We could, therefore, deduce that the concepts underlying “binding” and “fixation” are common (Table 2).

Model Building

We compared the co-occurring concepts for CTs with near meanings to identify class attributes, class relationships, and new classes. For example, instances of a reaction co-occurred most frequently with other CTs that described where, when, why, and how the reaction occurred, on what the reaction acted, and how the reaction was assessed. To describe a reaction class, its characteristics need to be identified. The reaction class should be linked to classes describing (1) properties, (2) mechanisms, and (3) substances; a measurement class is needed to illustrate it.

The new classes, relationships, and attributes were then used to refine the generic model. The initial reaction class was then linked to a property class and a mechanism class. Parent compound and metabolite were subsumed by a substance class that was linked to an origin class and a state class. Process classes and reaction classes were subsumed by a more abstract class called “explanatory data.” Observational data was replaced by a measurement class, which had links with temporal-coordinate class, value class and evolution class. The administration protocol was divided into four classes (administered product, administration mode, dosing schedule, and treatment schedule). The administration protocol class was aggregated with two new classes—“measurement protocol” and “population features”—to form a new abstract class called “experimental protocol.”

Description of the Final Model of Pharmacokinetics

The pharmacokinetics model describes the reality of the processes obtained under experimental conditions, the compartmental pharmacokinetics model and its parameters, and the variations obtained under particular con-
• Information about the real pharmacokinetics process knowledge

Information about the real pharmacokinetics process can be separated into that which explains the behavior of the drug in the body through time, and illustrative information, which includes the results of the measurements made on samples collected at given times.

Explanatory data describe the drug-body interaction. The changes of a drug in the body can be described as a succession of four processes: absorption, distribution, metabolism, and elimination. Each can be broken up into one or more reactions with one or more mechanisms. The reactions involve a substance, which changes states because of alterations in its state of binding, activity, localization, and biotransformation. The whole process—reactions, substances, and mechanisms can have certain properties (e.g., intestinal absorption is slow) and be ordered (e.g., urinary elimination occurs mainly by glomerular filtration).

The real pharmacokinetic process can be investigated by measurement with results expressed quantitatively or qualitatively and associated with a temporal co-ordinate. These measurements can change in time with one or more phases.

• Experimental protocol

The experimental protocol describes the conditions for obtaining data on the real process. It consists of an administration protocol, a measurement protocol and a population that has certain features.

The administration protocol is made up of information about the administered drug (e.g., tablet of drug X), the mode of administration (e.g., administration by the oral route on an empty stomach), the dosing schedule (e.g., repeated administration every 8 hours of 300 mg of X), and the treatment schedule (e.g., an 8-day course).

The measurement protocol describes the conditions of measurements (e.g., by gas chromatography).

The population features concern the population sample used for experimentation. These features consist of general features (e.g., man or dog) and particular features that give details about the physiologic type (e.g., a child less than 12 years old), the genetic type (e.g., slow acetylator), or the pathologic type (e.g., renal insufficiency with a creatinine clearance lower than 30 ml/min).
• **Mathematical model**

The information in the mathematical model describes the selected compartmental kinetic model applicable to the real processes observed. This information concerns either the model structure (e.g., monocompartmental model) or the pharmacokinetics model parameters (e.g., the half life is 3 hours).

• **Information about the influence of factors causing variation**

This describes the changes in measurements, parameters, or processes related to the changes in experimental conditions (e.g., the passage through the hematoencephalic barrier is increased when there is meningeal inflammation; the drug half-life is higher for children; the unbound fraction is higher in patients with hypoalbuminemia).

**Model Evaluation Results**

We did not observe any significant difference between the corpus of texts (number of sentences) evaluated by each evaluator (p > 0.657) or between evaluator responses (p > 0.591 for completeness and p > 0.456 for semantic accuracy).

Evaluation indicated that the model gives a good representation of pharmacokinetics information: it was able to represent pharmacokinetics information completely in 89% of the cases (CI 95% [83%–95%]), and in 11% of the cases in an “almost complete” way (CI95% [5%–17%]). The following sentence was judged “almost not distorted”: “Metabolic transformation by hepatic microsome enzymes (inducible)” was translated into the model as: “Metabolism is a reaction, has hepatic area, has the generic site microsome, has the specific site enzyme, has the property inducible.” The deformation of meaning is on the inducibility that is linked to metabolism instead of enzyme. There were no cases of significant or entire distortion.

The evaluators agreed on 93 texts and disagreed on only seven texts. The use of Delphi method to solve these cases was effective. A consensus was quickly obtained for all seven texts after a new evaluation. Each evaluator saw 25 texts and judged 21–25 times that the information was completely represented, 0–4 times that the information was almost completely represented, 23–25 times that the information was not distorted, and 0–2 times that the information was almost not distorted.

**Discussion and Conclusion**

We aimed to develop a pharmacokinetics model capable of representing all information contained in the pharmacokinetics section of SPCs. We used both general knowledge about pharmacokinetics and specific descriptions of drug pharmacokinetics available in SPCs. General knowledge helped us to define the generic structure of the pharmacokinetic model. This top-down approach, although performed manually, gave a quick organization for the main concepts, because the domain is delimited and is already the object of mathematical modeling. To refine this generic structure, we assessed the knowledge present in pharmacokinetics
sections of SPCs. For this bottom-up approach, we combined natural language processing results using manual semantic analysis of the relevant pharmacokinetics candidate terms specifying generic concepts from top down analysis. The bottom-up and top-down approaches are standard methods in knowledge engineering and are used in the medical field, but they are viewed as two separate ways of constructing a model. Our approach can be viewed as a middle-out approach, the most important concepts lead to generalization and special-
Generic pharmacokinetics concepts were used to select domain-specific candidate terms (CTs) that were then used as reference points to navigate in the initial text and to discover other CTs that often co-occur with them. The study, focused on specific CTs with similar and near meanings, allowed exploration of the common lexical environment of pharmacokinetic-specific CTs. By grouping all the co-occurring CTs, it was possible to indicate the links between them and to generate concepts that are new and more precise that would be introduced into the initial model to extend it.

We initially selected very few CTs from the wide-ranging lexicon produced by the natural language processing tool. The selective exploration of the contexts of occurrence related to these selected CTs optimized the semantic analysis: the co-occurring CTs were mainly describers of the CTs selected.

The results of the lexico-semantic analysis were then easily transposed into an object-oriented representation: co-occurring CTs were either attributes, new classes or class relationships. A large number of sentences was analyzed, and consequently the model was very rich but kept its initial generic structure.

The evaluation method used to validate the model had its strengths and weaknesses. As a detailed description of pharmacokinetics, the model has to be able to describe all information contained in SPCs. We therefore chose to represent the entirety of the textual information and not select elements from these texts as in some other evaluation studies. According to the criteria of Friedman et al., the model was frozen before the evaluation, the reference standard was established (sample of 100 randomized texts), but, unfortunately, the developer of the model participated in the evaluation. We chose one of the model developers to convert the texts into the model format because the task was very time-consuming, even for someone already familiar with the structured representation of pharmacokinetics (the task took almost 3 hours per document). We believe that it did not alter the objectivity of the evaluation because each structured representation was twice evaluated by independent evaluators. The model is not designed to evolve with time; therefore we did not need a methodology that could be reused for further evaluations, as used by Rocha et al. or Zweigenbaum et al. However, we compared the evaluators to judges who assess how well the structured representation corresponds to the initial text.

The evaluation of the model suggested that it adequately represents any information contained in the pharmacokinetics section in SPC. The completeness criteria indicate that some concepts are missing (such as causal relation), but the defects of representation related to the pharmacokinetics model were minor. The major defects in representing texts were related to the content of the initial text, which sometimes was vague and inconsistent (missing data or comparison between values which have not been quantified) or did not deal with pharmacokinetics (pharmacodynamics data, drug interactions). The “semantic accuracy” criterion showed that the information was mostly not distorted and the few distortions were judged to be minor. Lastly, the pharmacokinetics model satisfies the criteria of conciseness: the model does not store useless definitions (none of the classes or attributes were never filled). We conclude that the model developed satisfies the desiderata of completeness and coherence of coverage described by the Canon Group.

Our model is not unique but is of value because it has been evaluated and is apparently applicable to all SPCs. To avoid redundancy of attributes, we chose to create abstract classes that offer a high level of generalization, increasing the complexity of the model.

The model stores descriptive knowledge about pharmacokinetics, whereas drug databases focus on numerical information. The fine structure of the model covers all the information contained in SPCs but could also store information found in bibliographic reviews of drug pharmacokinetics. The production of structured data is unfortunately time consuming, but this could be improved by developing a knowledge editor that could recognize and extract automatically the frequent patterns of information.

However, simplified versions of the model could be created according to projected applications. This requires specifying classes and their attributes according to their frequency of use and their functional interest. For example each instance of a reaction could be a new class (absorption class or protein-binding class) and their specific properties added as attributes. The property class would then disappear. Such simplified versions could be used to structure certain elements of pharmacokinetics, but cannot represent all of the information contained in all pharmacokinetics texts.

As apparent during evaluation of the model, the quality of the textual source of information (SPCs) is often insufficient. Using the model to enter legal information about pharmacokinetics into drug databases may lead to an information gain and help clarify the content. This information could then be used by a computerized system to select drugs for a given indication according to pharmacokinetics selection criteria or to compare drug pharmacokinetic behavior in a given therapeutic class.

References

25. Bourigault D. LEXTER, a terminology extraction software for knowledge acquisition from texts. In 9th Knowledge Acquisition for Knowledge Based System Workshop (KAW’95), Banff, Canada, 1995.
Building and Evaluation of a Structured Representation of Pharmacokinetics Information Presented in SPCs: From Existing Conceptual Views of Pharmacokinetics Associated with Natural Language Processing to Object-oriented Design

Catherine Duclos-Cartolano and Alain Venot

J Am Med Inform Assoc 2003 10: 271-280
doi: 10.1197/jamia.M1193

Updated information and services can be found at:
http://jamia.bmj.com/content/10/3/271.full.html

These include:

References
This article cites 20 articles, 4 of which can be accessed free at:
http://jamia.bmj.com/content/10/3/271.full.html#ref-list-1

Article cited in:
http://jamia.bmj.com/content/10/3/271.full.html#related-urls

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/