An ontology of bacteria to help physicians to compare antibacterial spectra

Catherine Duclos PharmD, PhD, Jérome Nobécourt PhD, Gian Luigi Cartolano MD, Anis Ellini MD, Alain Venot MD, PhD
[1]Laboratoire d'Informatique Médicale et de Bioinformatique, Université Paris XIII, France

Abstract

General practitioners (GPs) may lack specialist microbiological knowledge, making it difficult for them to use documents concerning antibacterial spectra provided by French health authorities. We have developed a tool to help GPs to compare antibacterial spectra, based on an ontology of bacteria generated using OWL-DL language. This tool makes it possible to search for information concerning the antibiotic susceptibility of given bacteria, regardless of the way in which this information is expressed in the document. Applied to the whole document, the tool made 4528 spectra explicit, whereas only 3471 could be understood without microbiological reasoning. A preliminary study showed that the performance of this tool was similar to that of an expert microbiologist (94 to 98% correct responses) and better than that of unassisted GPs (84-90% correct responses).

Introduction

The French health authorities have produced a document on antibacterial spectra [1], with the aim of preserving antibiotic efficacy. This document is written by experts and, for each antibiotic, provides details of the bacteria against which it is active and the national prevalence of resistance. This document has a highly structured format, with many sections, dealing with topics such as susceptible bacteria, bacteria with intermediate susceptibility and resistant bacteria. Within these sections, bacteria are classified according to some of the characteristics used for identification (e.g. type of respiration, gram staining).

This document was designed to facilitate the distribution to GPs of up-to-date information about the national prevalence of bacterial resistance to the various antibiotics available [2,3]. Physicians are thus expected to prescribe antibiotics according to national bacterial ecology [4]. For example, in the case of suspected *Streptococcus pneumoniae* respiratory infection, a physician using this document should note that the prevalences of resistance to amoxicillin+clavulanate and of resistance to amoxicillin alone are identical in this bacterium. There is therefore no benefit to be gained from prescribing amoxicillin+clavulanate rather than amoxicillin alone.

The document is currently used for the checking of information: in empirical situations, doctors can check the susceptibility of the bacterium thought to be responsible for the infection to the antibiotics they wish to prescribe. This tool could also be used for comparisons, helping the physician to choose the most appropriate antibiotic for treating the one or several bacterial species thought to be responsible for infection.

Information about the susceptibility of bacteria to each of the antibiotics considered must be available for such comparisons to be possible. There are three possible situations. The information may be:

- Explicitly available: information is found using the exact expression for the bacterium considered (e.g. *Escherichia coli, Clostridium perfringens*),
- Implicitly available: information is not found with the exact expression considered, but can be found using a bacterial grouping containing the selected bacterium (e.g. the grouping "enterobacteria" contains *Escherichia coli*, and the grouping "gram-positive strict anaerobes contains *Clostridium perfringens*),
- Unavailable: no information is available about the bacterium considered because this information is not clinically pertinent.

The user can render implicit information explicit by determining whether the bacterium considered belongs to particular bacterial groupings. This requires microbiological knowledge about bacterial taxonomy (e.g. enterobacteria) or microbiological identification criteria (e.g. gram- negative aerobic rods).

This reasoning is straightforward for a microbiologist but GPs do not generally have sufficient knowledge to make such inferences and are therefore likely to be unable to find implicit information about the bacterium considered.

If GPs had at their disposal a tool that made the same inferences as a microbiologist, they would have no difficulties comparing antibacterial spectra.

A tool of this type should automatically search for information about the susceptibility of the bacterium considered and all the groupings to which it belongs, using a bacterial classification based on microbiological description.

The computerized microbiological classifications currently available [5,6] focus on descriptions of bacteria according to phylogenic relationships, whereas a multiperspective description of bacteria (bacterial systematics, biochemical, antigen-based, etc.) is required.

Generalist classifications, such as the MeSH, [7] could be used but are limited by certain classification features (e.g. *Clostridium perfringens* is considered to be a gram-positive endospore-forming rod but not a gram-positive strict anaerobe).

Ontologies based on description logic can be used to generate multiperspective classifications [8]. The part of SNOMED CT [9] containing bacteria is based only on bacterial taxonomy. No definition properties are associated to bacteria and some features of the classification are missing (see MeSH). The bacterial ontology contained in GALEN [10] is still being developed.

An ontology of bacteria could be used to solve this problem. Microbiological knowledge could be added to the definition of bacteria. The use of a classifier would then facilitate identification of the set of concepts (groupings of bacteria) subsuming the selected bacteria.

The aim of this study was to facilitate the comparison by GPs of antibacterial spectra for various antibiotics. We developed a tool based on an ontology of bacteria rendering explicit the most implicit knowledge about selected bacteria. We carried out a preliminary evaluation of the results obtained with this system by comparing the performance of a GP and of the tool with that of a microbiologist.

Materials and methods

Building the bacterial ontology

Information about antibacterial spectra was extracted from the "antibacterial spectra" [1] document, as previously described [11].

From this extraction, a glossary of bacterial strings used in the document was then created.

The bacterial glossary was analyzed manually, to ensure that all the lexical variants of an extracted bacterium were associated with a single corresponding concept (e.g. *E.coli* and *Escherichia coli* are the lexical variants of the *Escherichia coli* concept). For ambiguous situations (e.g. determining whether the string "streptoccocci" is associated with the "*Streptococcus* family" concept or the "*Streptococcus* genus" concept), we went back to the original document and considered the context in which the string appeared.

The final list of concepts was then reviewed to identify concepts describing a named bacterium (e.g. a species (*Chlamydia pneumoniae*) or a genus (*Chlamydia*)) and concepts describing groupings of bacteria (e.g. gram-negative aerobes).

We considered the bacterial groupings defined to have the necessary and sufficient properties for classification of the named bacteria. The atomic concepts and properties based on them could then be deduced.

A hierarchy of named bacteria was then built with "kind of" properties between species and genera. Necessary and sufficient properties were associated with genera and inherited by species. For some taxa (e.g. *Streptococcus, Staphylococcus*) other properties were added.

The ontology was built using the Protégé® ontology editor [12] and the OWL-DL language.

Building the heuristic for making knowledge explicit

An XML file called the "Explicit XML file" was created from the initial document extraction. It contained information about the antibacterial spectra ({antibiotic, bacterium} pairing associated with susceptibility and resistance prevalence). In this XML file, bacteria are expressed according the lexical variants found in the bacterial glossary. This XML file contains all the explicit knowledge.

Named bacteria were classified into bacteria groupings. The RACER® [13] inference engine was used to perform this classification.

Step 1: Searching explicit knowledge

When information is sought concerning a given named bacterium and a given antibiotic, the Explicit XML file is scanned, searching, for the given antibiotic, a string matching with one of the lexical variants of the bacterium. If a match is found then the search is stopped. If no match is found, an implicit search is then carried out.

Step 2: Searching implicit knowledge

Implicit knowledge is searched by checking for susceptibility information about at least one of the

bacterial groupings subsuming the selected named bacterium. For this purpose, the classification obtained with RACER® is browsed from the bottom to the top. At any given level, all the parents are tested before testing their ancestors. For each bacterial grouping investigated, searches are made for all its lexical variants, and string matching is performed in the Explicit XML file. If a match is obtained, the search is stopped. In other cases, the search is continued until the top of the ontology is reached, in which case, it is considered that no information is available.

Step 3: Reporting process

When a match is found for a given {antibiotic, named bacterium} pairing, information about susceptibility is extracted from the Explicit File into a new XML file called the "Processing File". The information is tagged to identify whether it was obtained through explicit or implicit processing and information is also stored for the matching bacterium (the bacterium selected for explicit processing, the bacterial grouping for implicit processing). If no match is found, no susceptibility is associated with the {antibiotic, named bacterium} pairing considered.

Evaluation

This tool is designed to help the physician to reason like a microbiologist when searching for information in the "antibacterial spectra" document. A preliminary evaluation was carried out to validate this role of the tool.

We randomly selected ten named bacteria from the whole set of named bacteria. For a limited set of 45 antibiotics (those most frequently prescribed in French general practice), we searched for antimicrobial activity against the selected bacteria in the "antibacterial spectra" document. This search was performed manually by a GP, manually by an expert microbiologist and automatically, using the computerized tool. The names of the antibiotics were hidden, so as not to influence the microbiological reasoning of the GP and the expert.

The responses of the expert microbiologists were considered to be the gold standard. The correctness of the responses provided by the GP and the computerized tool were expressed with respect to this standard. The 95% confidence intervals for the percentage of correct responses for the 450 expected spectra were calculated.

Results

Concept analysis

The concepts identified can be assigned to two groups: groupings of bacteria and named bacteria. We were able to distinguish three types of bacterial groupings: groups containing various genera and species (non specific groupings), groups containing particular species of a genus (species-specific groupings), and groups containing a fixed list of genera (taxonomic groupings). Table 1 shows the results of the concept analysis.

Concepts		N
Bacterial groupings	Non specific groupings e.g.: anaerobic bacteria, gram-negative strict aerobes rod bacteria	11
	Species-specific groupings e.g.: Coagulase-negative staphylococci, C Lancefield group streptococci	5
	Taxonomic groupings e.g.: enterobacteria, atypical mycobacteria	9
Named bacteria	Genus-named bacteria e.g.: *Yersinia, Neisseria*	62
	Species-named bacteria e.g.: *Yersinia enterolitica, Neisseria meningitidis*	92

Table 1. Bacterium-related concepts and their number of occurrences (N).

Atomic concepts and properties

Based on these groupings, we defined the following properties:
"has gram staining", "has morphology", "has respiratory mode", "has Lancefield group", "has hemolysis type" , "has coagulase type", "has genus"

These properties have as their range one of the following atomic concepts: GramStaining, RespiratoryMode, Morphology, LancefieldGroup, CoagulaseType, HemolyticType, Genus, and for domain only NamedBacteria.

Defined concepts of named bacteria and bacterial groupings

Each named bacterium is defined according to some of these properties, using both "sufficient" and "necessary and sufficient" restrictions. For example the concept "*klebsiella*" is defined by the following expression:

Klebsiella:= ∀ hasGenus : klebsiellaGenus
⊓ ∀ hasMorhology : rod
⊓ ∀ hasRespiratoryMode : Aerobes
⊓ ∀ hasGramStaining : NegativeGramStaining

The groupings have been defined with necessary restrictions, using specific properties:
- Gram staining, respiratory mode and morphology type define non specific groupings,
- Genus and Lancefield group or hemolysis or coagulase define species-specific groupings,
- Genus has been used to define taxonomic groupings.

Terminological expansion of the search of information

The Processing XML File obtained contains 20828 spectra for the 127 antibiotics against 164 named bacteria found in the antibacterial spectra document. Only 3479 spectra were found with explicit processing, 4528 were found with implicit processing and 12821 spectra remain unavailable. Of the 4528 spectra obtained from implicit information, 1673 dealt with susceptible bacteria, 84 with bacteria of intermediate susceptibility and 2771 with resistant bacteria.

Part of the classification obtained with RACER® in Protégé® is shown in figure 1. Extracts of the Explicit XML file and the Processing XML file are shown in figures 2 and 3.

Figure 1. Part of the named bacteria classification according to bacterial groupings

Category	Prevalence of resistance
Susceptible species	
Gram-positive aerobes	
Staphylococcus aureus	<10%
Staphylococcus non aureus	5%-20%

Figure 2. Extract of the Explicit XML file for the antibiotic fusidate, with bacteria expressed as lexical variants

Matching Concept: coagulase-negative staphylococci			
P+	*Staphylococcus epidermidis*	RP=	5%-20%
P+	*Staphylococcus saprophyticus*	RP=	5%-20%
Matching Concept: *Staphylococcus aureus*			
P-	*Staphylococcus aureus*	RP=	0%-10%
Matching Concept: TOP			
P+	*Bacillus anthracis*	RP=	ND
P+	*Bacillus cereus*	RP=	ND

Figure 3 Part of the Processing XML file for the antibiotic fusidate. Knowledge about *S. epidermidis* and *S. saprophyticus* is obtained from the parental "coagulase-negative staphylococci", whereas *S. aureus* was found by direct string matching and no matching was found for *B. anthracis or B. cereus* (P+=implicit processing, P-=explicit processing, RP=resistance prevalence, ND=unavailable)

Evaluation results

The 95% confidence intervals for the percentage of correct responses were [84 - 90] for the GP and [94 - 98] for the computerized tool, with a 9% difference between the corresponding means. This interval was [66%-74%] when explicit knowledge search was performed alone.

The general practitioner found it difficult to infer microbiological knowledge. A greater level of error resulted from incorrect associations: *Klebsiella* was incorrectly considered not to be an enterobacterium and *Vibrio cholerae* was wrongly considered to be an enterobacterium.

For the computerized tool, incorrect responses were related to the microbiologist using specific modes of reasoning not integrated into the heuristics of the program. When all the species of a genus were enumerated, the microbiologist asserted that the genus was equivalent to the species (e.g.: if information is available for *Pasteurella multocida*, it is deduced that it is available for *Pasteurella* in that *Pasteurella multocida* is the only pathogen of interest in the genus *Pasteurella*).

Discussion

The tool we present provides physicians with access to the reasoning of a microbiologist, making it possible to extend searches to implicit information. When applied to an antibacterial spectra document, this tool converted a large amount of implicit information into explicit information.
We previously manually constructed a bacterial classification [11]. The ontology approach is more efficient because it makes it possible to test the consistency of the classification, to apply multi-

inheritance more easily and to simplify maintenance, which may be difficult in cases of manual classification [14].

Existing ontologies were not used here because they were not suitable for the intended task. SNOMED lacks concept definition as only "is a" attributes are used to define bacteria. Pure taxonomy was not of the main goal of our work as the classifier could reengineer the classification from bacterial definitions. GALEN contains interesting definitions that are partially reused in our ontology, but also gives additional definitions not necessary for the purpose intended.

The use of subsuming reasoning based on ontology has already proved useful for the extension of information searches [15]. Our preliminary evaluation results confirm the value of this approach in another domain. However, some specific features of the reasoning of microbiologists must be taken into account. For example, information for a class can be guessed if the information provided is identical for all its subclasses.

We present here the results of a preliminary evaluation of this approach. We will now design a rigorous evaluation protocol in which antibiotics will be selected at random to balance the frequency of antibiotics corresponding to many and few bacterial groupings.

This tool was designed to improve the results of searches for information about antibacterial spectra. It should not be considered to be a computerized antibiotic decision support system like those developed in hospital settings [16].

This study is of potential interest to the group of experts in charge of writing the antibacterial spectra document, because it highlights the ambiguities of bacterial denominations that may confuse GPs attempting to use this document.

Acknowledgment

We thank the "Haute Autorité de Santé" for funding.

References

1. "Spectres Antibacteriens" 2005. Available at http://afssaps.sante.fr/pdf/5/atb.pdf, Accessed March 8, 2007
2. Moss F. McSwiggan, McNicol M, Miller D. Survey of antibiotic prescribing in a district general hospital. I. Pattern of use. Lancet. 1981; 2: 349-52
3. Yu V, Stoehr G, Starling R, Shogan J. Empiric selection by physician: evaluation of reasoning strategies. Am J Med Sci 1991; 301: 165-72
4. Manthous C, Amoateng-Adjepong Y. Empiric antibiotic use and resistant microbes: a "Catch 22" for the 21st Century. Chest. 2000; 118: 9-11
5. Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ bacterial Nomenclature). Available at: http://www.dsmz.de/bactnom/bactname.htm, Accessed March 8, 2007
6. List of Bacterial names with Standing in Nomenclature (LBSN). Available at http://www.bacterio.cict.fr, Accessed March 8, 2007
7. MeSH: Medical Subject Headings. Available at http://www.nlm.nih.gov/mesh/2006/MBrowser.html, Accessed March 8, 2007
8. Baader F, Horrocks I, Sattler U. Description Logics. In: Handbook on ontologies, Staab and Studer eds, Berlin: Springer –Verlag, 2004, p 3-28
9. SNOMED CT. Available at http://www.snomed.org/, Accessed March 8, 2007
10. OpenGalen, Open Source. Available at http://www.opengalen.org/ressources.html, Accessed March 8, 2007
11. Duclos C, Cartolano GL, Ghez M, Venot A. Structured representation of the pharmacodynamics section of the summary of product characteristics for antibiotics: Application for automated extraction and visualization of their antimicrobial activity spectra. JAMIA 2004; 11:285-293
12. Protégé. Available at http://protege.stanford.edu., Accessed March 8, 2007
13. RACER. Available at http://www.racer-systems.com/, Accessed March 8, 2007
14. Wolstencroft K, Lord P, Taberno L, Brass A, Stevens R. Protein classification using ontology classification. Bioinformatics 2006, 22:530-538
15. Henegar C, Bousquet C, Lillo-Le-Louet A, Degoulet P, Jaulent MC. Building an ontology of adverse drug reactions for automated signal generation in pharmacovigilance. Computers in Biology and Medicine 2006; 36:748-767
16. Evans RS, Pestotnik SL, Classen DC, Clemmer TP, Weaver LK, Orme JF Jr, Lloyd JF, Burke JP. A computer–assisted management program for antibiotics and others antiinfectives agents. N Engl J Med 1998; 338:232-238