Terminology for the description of the diagnostic studies in the field of EBM

Natalia Grabar, Ludovic Trinquart, Isabelle Colombet

Terminology for the description of the diagnostic studies in the field of EBM

Natalia GRABAR[a], Ludovic TRINQUART[b], Isabelle COLOMBET[c]
[a] *CNRS STL UMR 8163, Université Lille 3, rue Barreau, 59653 Villeneuve d'Ascq, France*
[b] *French Cochrane Center, France, AP-HP, Paris France*
[c] *Université Paris Descartes, Paris, F-75006 France; HEGP AP-HP, 20 rue Leblanc, Paris, F-75015 France*

Abstract. *Diagnostic systematic reviews is a relatively new area within the Evidence-Based Medicine (EBM). Their indexing in Pubmed is not precise, which complicates their detection when a systematic review is to be realized. In order to provide an assistance in the selection of relevant studies, we propose to develop a terminology describing this area and the organization of its terms. The terminology is built with a bottom-up approach. It contains 255 terms organized into five hierarchical levels. Only a small proportion of these terms (13%) are already registered in MeSH. This terminology will be exploited in a dedicated web service as a main tool for the detection of relevant diagnostic studies.*

Keywords. Evidence-Based Medicine; Review, Systematic; Language; Natural Language Processing; Terminology

1. Introduction

The aim of systematic reviews (SR) is to provide a synthesis of multiple primary research studies concerned with a given clinical question. Such syntheses are a part of the Cochrane Collaboration effort and published in the Cochrane library. The library is thereby a knowledge base which can be used by health professionals for supporting decisions within the frame of the Evidence-Based Medicine (EBM). The vast majority of SRs addresses the efficacy of interventions to treat or prevent diseases. Other SRs focus on diagnostic or prognostic studies. These reviews can be methodologically challenging. In particular, an essential step is to identify all relevant studies to be included in the review. Identifying diagnostic test accuracy studies is more difficult than searching for randomized trials. First, an exhaustive search strategy should involve several electronic bibliographical databases. Second, the indexing of diagnostic studies is imperfect as there is not a unique keyword for an accuracy study comparable with the term “randomized controlled trial” [1]. Third, methodological electronic search filters for diagnostic studies (which aim to restrict the search to articles that are most likely to be diagnostic studies) are not recommended

because they can lead to the omission of a substantial number of relevant studies [2,3]. Fourth, supervised machine learning methods used for the automatic selection of relevant studies for therapeutic SRs [4-7] are not efficient because of the small amount of existing diagnostic reviews. Consequently, reviewers often have to screen for eligibility very large number of references, most of them being irrelevant to the clinical question of interest. The whole process is performed manually which is a real burden to reviewers. We propose to help the process of selection of relevant articles with a semantic information retrieval system through a terminological resource. To our knowledge, no such resource have been yet designed and published.

Two kinds of approaches are distinguished when creating terminologies, namely the top-down (main high-level concepts are defined and then populated) and bottom-up (terms are observed within the exploited material and then organized into classes, sub-classes etc). Corpora of textual documents and Natural Language Processing (NLP) methods are often used in bottom-up approaches [8-9]. Transformation-based approaches have also been proposed, they exploit HTML and XML metadata [10] or databases [11-12]. In our work, we use corpora and NLP methods, because textual material is easily accessible and contains data actually and naturally used in the area of interest. Other related works should be mentioned. For instance, an ontology of EBM has been proposed [13]. It attempts a modelization of this area and it targets particularly relations which may exist between patient records and meta-analysis results. Another work proposes an ontology related to SRs and meta-analyses [14]. It contains 128 elements exploited for manual tagging of five Randomized Controlled Trials studies in neurosurgery. Intra and inter-annotator comparison shows that such ontologies allow to obtain a high annotation agreement (kappa rating from 0.53 to 0.82) and an improvement in the quality of reporting. We aim at creating a terminology dedicated to diagnostic studies.

2. Material and Methods

Material. We exploit a set of corpora and the MeSH terminology [15].The main subset of corpora is composed of scientific literature and reports related to diagnostic studies. It contains: 6 reference articles dedicated to description of the STARD initiative and its main concepts, and 20 diagnostic studies, among which 15 are full-text articles and 5 are abstracts. References and full text of these articles are available upon request. These are supposed to be instantiations of the STARD initiative and to describe studies performed within the EBM framework. This *diagnostic* corpus contains 105,000 occurrences (or words). Additional corpora are used to ensure the specificity of terms, they cover other types of SRs: *prognostic* (6 citations, 36,000 occ.), *therapeutic* (7 citations, 36,779 occ.) and *observational* (6 citations, 39,800 occ.). MeSH terminology [15] is typically used for indexing the scientific literature in Pubmed database, among which for indexing the SRs. We expect that MeSH provides several terms relevant to diagnostic accuracy studies reviews. If new terms are found in the corpora, and according to the expert validation, they may be considered as additional relevant terms for MeSH.

Method. Our method carries out extraction of terms and their alignment with MeSH.

Another step is dedicated to the evaluation and structuring of the extracted data.

Automatic acquisition and alignment of terms. Corpora are first pre-processed through the Ogmios platform [16]. This platform performs the segmentation into words and sentences, POS tagging (assignment of part-of-speech categories: *cancers/Noun*, *cancerous/Adjective*) and lemmatization (definition of the normalized form of words: *cancers* => *cancer*) with TreeTagger [17]. The step of term extraction is carried out with the syntactic rule-based parser Y_ATE_A [18]. Once the terms are extracted from corpora, they are aligned with the MeSH terminology. For all the extracted terms, their frequencies are computed in each processed corpus. This information is assumed to help the selection and validation step: frequencies of terms may be indicative of their specificity to the diagnostic area. Indeed, if terms occur only or more often in diagnostic corpus they show a high specificity, otherwise their specificity to the diagnostic area is lower.

Evaluation and structuring. An independent evaluation was performed manually by two experts (a physician and a biostatistician with experience in SR). In cases of disagreements, consensus was established further to discussions. Each extracted term was examined, together with its distributions and frequencies across the corpora. Global inter-expert agreement was assessed with chance-corrected kappa statistics and with simple raw specific agreement indexes, which are the conditional probability, given one expert gives a result, that the other expert gives the same result [19]. Structuring was performed through a bottom-up approach: selected terms were categorized into categories and then sub-categories, according to their semantics.

3. Results and Discussion

Processing of diagnostic corpus led to the extraction of 7,448 terms, among which 1,218 (16.3%) are already registered in MeSH, while 6,230 are new terms. The acquisition on other corpora produced the following results: observational corpus provides 1,640 terms among which 722 (44%) in MeSH; prognostic corpus provides 2,383 terms among which 531 (22.3%) in MeSH; therapeutic corpus provides 1,602 terms among which 590 (36.8%) in MeSH. Table 1 contains an example of the extracted terms together with their frequencies in various corpora. If an extracted term is also recorded in MeSH, we indicate in the first column its MeSH hierarchical tree (*i.e.*, E, G or N), otherwise it is provided by Y_ATE_A. We then indicate frequencies of the extracted terms (frequency in diagnostic corpus F_{tot}, and separately in methodological documents F_{met} and studies F_{stu}). We also indicate the number of diagnostic corpus documents in which these terms occurred (total number N_{tot}, and separately number of methodological documents N_{met} and of studies N_{stu}). The last three columns indicate the frequencies of these terms in the three other corpora.

Table 1: Excerpt of the extracted data.

	Terms	*Diagnostic*						*Prog*	*Obs*	*Ther*
		F_{tot}	F_{met}	F_{stu}	N_{tot}	N_{met}	N_{stu}	F_{tot}	F_{tot}	F_{tot}
E01	diagnosis	194	77	117	19	6	13	13	27	6
E05	roc curve	14	4	10	8	2	6	2	0	0
N06	prevalence	10	6	4	2	1	1	3	11	0
Y_ATE_A	diagnostic accuracy	150	122	28	13	6	7	10	0	0

Y_ATE_A	diagnostic performance	30	12	18	3	2	1	0	0	0
N06	confidence intervals	20	5	15	14	4	10	7	3	8
Y_ATE_A	characteristics curve	2	1	1	2	1	1	0	0	0
Y_ATE_A	clinical trials	12	6	6	8	4	4	4	8	38

Further to the expert evaluation, a set of 219 terms is selected. Among these, 26 (13%) are already registered in MeSH (E (n=11), G (n=2) and N (n=11) MeSH trees), while 193 are provided only by Y_ATE_A. The inter-expert agreement is NN. An additional set of 36 terms have been added by experts, which gives a total of 255 terms. The additional terms are often variations of the extracted terms (*i.e.* abbreviations: *npv*, *ppv*) or terms suggested by the extracted data (*dor* and *cut point* never occurred individually but within larger terms and have been added as individual entry). Within the initial set of 7,448 extracted terms, only 3% of these have been selected. The rejection rate is very important. Some of the rejected terms are indicated in lower part of table 1. Among the rejected terms we observe: (1) common errors usually observed with automatic term extraction methods due to tagging errors; (2) sequences non relevant to a terminology (journals, authors, ...); (3) too general terms (*public health*, *confidence intervals*, *characteristics curve*); (4) terms non specific to diagnostic studies (*clinical trials*). Specificity of the material needed for the task and current shortcomings of the automatic term extraction may explain such rejection rate. With this kind of data, where rate of selection is both globally low and heterogeneous between experts, inter-expert agreement kappa is low (0,106), although average positive (selection) and negative (rejection) agreements are respectively 0.14 and 0.84. Exploitation of such methods allows to construct a terminology where no existing semantic resources are available and to insure that this terminology will be relevant to the processing of real data. A low number of MeSH terms within the validated data indicates that diagnostic area is poorly covered by MeSH. If MeSH were to be enriched with such terms, the indexing of diagnostic studies would be more precise and would help realization of SRs.

Next and final step of the work is dedicated to the structuring of the selected terms. Five levels of terms have been defined. Figure 1 shows the four higher levels corresponding to categories of terms. These four broad categories represent main aspects for diagnostic studies. Notice that nearly all the MeSH terms are positioned under the *Test characteristics* tree, which indicates again the necessity of such a resource.

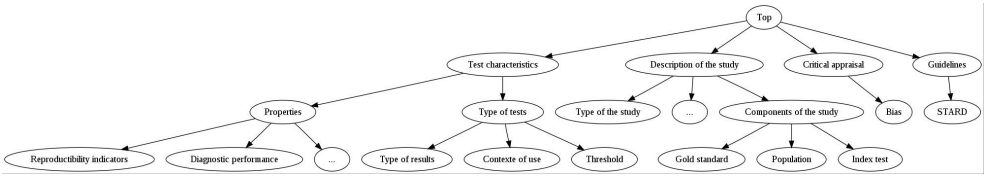

Figure 1: Hierarchical tree of the terminology.

4. Conclusion and Perspectives

We presented an experience in building a terminology of diagnostic studies within the

EBM area. We exploited automatic methods for term extraction and for their alignment with an existing terminology (MeSH). Only small part of the acquired and validated terms is already recorded in MeSH. This indicates that MeSH may be enriched with some of the terms from the constructed terminology in order to provide assistance in indexing the diagnostic studies. The validated terms have also been structured, and the resulting semantic resource contains five hierarchical levels. We plan to exploit and evaluate this resource within the webservice dedicated to the automatic selection of literature [20].

Acknowledgment. This work is part of the ReSyTAL project, supported by a research grant from French PHRC, designed to facilitate the selection of relevant scientific literature as well as realization of diagnostic SRs.

References

[1] Haynes RB and Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from medline: analytical survey. *BMJ* 2005;330(7501):1162–3.
[2] Leeflang M, Scholten R, Rutjes A, Reitsma J, and Bossuyt P. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *Clin Epidemiol* 2006;59(3):234–40.
[3] Meade M and Richardson W. Selecting and appraising studies for a systematic review. *Ann Intern Med* 1997;127(7):531–7.
[4] Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, and Aliferis C. Text categorization models for high-quality article retrieval in internal medicine. *J Am Med Inform.* 2005;12(2):207–16.
[5] Cohen A, Hersh W, Peterson K, and Yen P. Reducing workload in systematic review preparation using automated citation classification. *JAMIA* 2006;13(2):206–19.
[6] Demner-Fushman D, Few B, Hauser S, and Thoma G. Automatically identifying health outcome information in medline records. *JAMIA* 2006;13(1):52–60.
[7] Kilicoglu H, Demner-Fushman D, Rindflesch T, Wilczynski N, and Haynes R. Towards automatic recognition of scientifically rigorous clinical research evidence. *J Am Med Inform Assoc* 2009;16(1):25–31.
[8] Condamines A and Rebeyrolle J. CTKB : A corpus-based approach to a terminological knowledge base. In: Proceedings of Computerm'98, Coling-ACL'98. 1998:29–35.
[9] Maedche A and Staab S. Mining ontologies from text. In: Dieng R and Corby O, eds, EKAW 2000.
[10] Giraldo G and Reynaud C. Construction semi-automatique d'ontologies à partir de DTDs relatives à un même domaine. In: Actes Ingénierie des Connaissances (IC). 28-30 mai 2002.
[11] Krivine S, et al. Construction automatique d'ontologies à partir d'une base de données relationnelles : application au médicament dans le domaine de la pharmacovigilance. In: IC 2009.
[12] Kamel M and Aussenac-Gilles N. Construction automatique d'ontologies à partir de spécifications de bases de données. In: IC 2009, 2009.
[13] Pisanelli D, Zaccagnini D, Capurso L, and Koch M. An ontological approach to evidence-based medicine and meta-analysis. In: MIE 2003, 2003:543–8.
[14] Zaveri A, Cofiel L, Shah J, et al. Achieving high research reporting quality through the use of computational ontologies. *Neuroinformatics* 2010;8(4):261–71.
[15] National Library of Medicine, Bethesda, Maryland. Medical Subject Headings, 2001. *www.nlm.nih.gov/mesh/meshhome.html*.
[16] Hamon T, Nazarenko A, Poibeau T, Aubin S, and Derivière J. A robust linguistic platform for efficient and domain specific web content analysis. In: RIAO 2007, Pittsburgh, USA. 2007.
[17] Schmid H. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK. 1994:44–9.
[18] Aubin S and Hamon T. Improving term extraction with terminological resources. In: FinTAL 2006, number 4139 in LNAI. Springer, August 2006:380–7.
[19] Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43:551-558
[20] Trinquart L, Fanet A, Grabar N, and Colombet I. A unique web service to facilitate the study selection process in systematic reviews. In: Joint Colloquium of the Cochrane & Campbell Collaborations, 2010.