Content Evaluation of a Knowledge Base

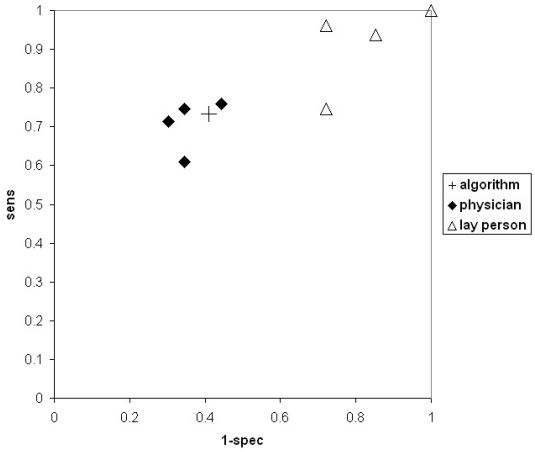
Eneida A. Mendonça, M.D., James J. Cimino, M.D.
Department of Medical Informatics, Columbia University, New York, NY, USA

Introduction. Researchers have suggested that bibliographic information should be integrated with clinical applications to facilitate access to scientific evidence. One challenge in building such a system is the construction of a medical knowledge base to support the search of online literature according to individual needs. In a previous study we described in detail the methodology used to build a knowledge base using the co-occurrence of MeSH terms in MEDLINE citations associated with the search strategies for evidence-based medicine.[1] The analysis of the relevance of the relationships between the semantic pairs generated by this process, and the clinical validity of the semantic types involved are described elsewhere.[2] The current study is focused on the content of the information extracted. This study uses the information extracted from MEDLINE citations collected and analyzed in the previous study. In that experiment, physicians identified 87 pairs of semantic types as relevant to the task of literature review.

Methods. For each of the 87 pairs, we assigned a semantic relationship based on the UMLS Semantic Network. If a direct relationship was not found, the closest level of relationship found was used. Subjects were 4 physicians, selected as experts, and 4 lay persons, selected as controls. A questionnaire containing 140 random questions was answered by each subject. A sample question is "If your patient has *Venous Thrombosis* and *Cerebrovascular Disorders*, would you be interested in articles that discuss how Venous Thrombosis occurs in Cerebrovascular Disorders?" Sensitivity and specificity were calculated for each subject using majority physician opinion as the reference standard. If the subject was a physician, his or her data were removed from the reference standard, and the criterion was adjusted. For each subject, we computed the distance between subject pairs. Bootstrapping was used to estimate the variance of these measures. This evaluation was based on a methodology used by Hripcsak and colleagues.[3]

Results. We identified 8,264 pairs of concepts based on the 87 semantic types from the previous study. Only 20 (22.99%) had a direct semantic relationship in the UMLS Semantic Net. Performance was measured by the average distance of each subject from the physicians. The average distance of the automated algorithm from the physicians was 0.089 (CI, -0.06-0.185). No physicians differed significantly from the others. The automated algorithm did not differ significantly from the physicians. All lay persons differed from physicians with highly significant p values ($p<0.01$). Sensitivity and specificity for each subject are plotted in Figure 1.

Fig 1. Sensitivity and specificity plotted on ROC axes



Conclusion. This analysis demonstrates that it is possible to extract useful medical knowledge, more specifically semantic relationships between concepts, from MEDLINE citations. The algorithm identifies relationships of a type that are of interest to clinicians. The low specificity obtained by lay persons suggests that they are interested in a broader variety of topics, perhaps because they cannot apply the same filters that a physician might use.

Acknowledgment. This publication was supported in part by National Science Foundation grant IIS-98-17434 and CNPq, Brazil, grant 20057/95-5.

References

1. Mendonça EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. Proc AMIA Symp 2000;:575-9.
2. Mendonça EA, Cimino JJ. Building a knowledge base to support a digital library. Medinfo 2001; *in press*.
3. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med. 1995; 122(9):681-8.