MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Knowledge Discovery Based on the Identification of Relationships Among Medical Concepts by Clustering Candidate Descriptors

Shizuka Yamazaki[a], Mariko Iwasawa[b], Shin-ichi Nakayama[b], Natsuo Onodera[b], and Yusuke Tanigawara[a]

[a] *Department of Pharmacy, School of Medicine, Keio University, Japan*
[b] *Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan*

Abstract

The discovery of hidden connections within the medical literature could give new and plausible knowledge. This study proposes and examines a new methodology for hypothesis generation and knowledge discovery using cluster analysis. We obtained a knowledge overview from a network of medical concepts by clustering candidate descriptors.

Keywords:

cluster analysis, information storage and retrieval, medical subject headings

Introduction

It is important to obtain knowledge on relationships between various phenomena observed in clinical practice. All relationships are not always direct connections, however. The identification of relationships among medical concepts from the literature database could show such indirect connections.

Methods

Don R. Swanson, in his the ABC model, showed that unknown relationships between a chemical substance (A) and a disease (C) could be found by identifying a concept (B) that connects A and C in the literature (Figure 1) [1]. Closed and one-directional open discovery processes were reported by Marc Weeber et al. [2]. Most prior works used ranking based on term frequency. Such methodology may bring on only a part of the concepts. For effective discovery of useful relationships, we propose grouping candidate descriptors by means of hierarchical clustering. Three types of discovery processes were examined experimentally. In the closed process (a), the B terms were MeSH terms that co-occurred in the document sets A and C: these B terms were grouped into several clusters. In the one-directional open process (b), the B terms were MeSH terms in the set C: these B terms were grouped and the B terms in each group were used to extract set A from MEDLINE: the A terms were also grouped. In the two-directional open process (c), the A and C terms were MeSH terms in the set B: those terms were grouped. In each process, the stop list of MeSH terms used was similar to that applied in Swanson's method [3]. These processes were applied in an attempt to identify relationships between fish oil and Raynaud's disease from MEDLINE documents as an experimental test of Swanson's hypothesis. Additionally, we searched for particular terms with the relevant subheading: C-term(s)/th [therapy] - B-terms(s)/de [drug effect] - A-term(s)/tu [therapeutic use] OR A-term(s)/pd [pharmacology].

Results

In our experiment (a), 56 documents on "Raynaud Disease/th" and 40 on "Fish Oils/tu OR Fish Oils/pd" from 1983 to 1985 in MEDLINE had 29 co-occurring MeSH terms, included 7 B-terms/de. For example, "Platelet Aggregation/de" and "Blood Viscosity/de" were found among the MeSH terms (Figure 2). When we did not use subheadings in searching for documents, the document sets increased in size to 511 hits for "Raynaud Disease" and 262 for "Fish Oils", but the number of co-occurring terms was the same and the resulting knowledge overview was similar. In our experiment (b), MeSH terms appearing in 56 "Raynaud Disease/th" documents were clustered, which generated 5 groups of B terms (Figure 3). One of the groups of B terms facilitated the discovery of documents on Fish Oil. In addition, we got the relationships between Raynaud's disease and the serotonin reuptake inhibitor Fluoxetine, and also the calcium channel blocker Felodipine, Nisoldipine, Isradipine. In our experiment (c), the relationships between fish oil and variant angina pectoris were obtained (Figure 4).

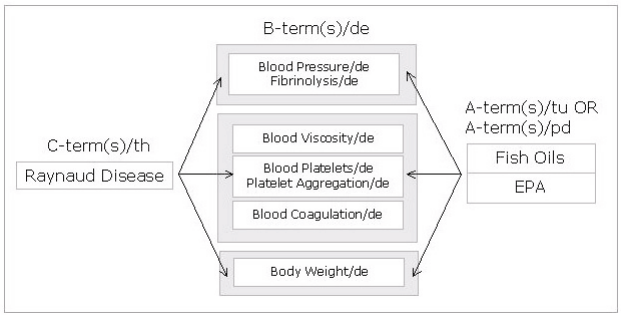

Figure 2 - Closed discovery process

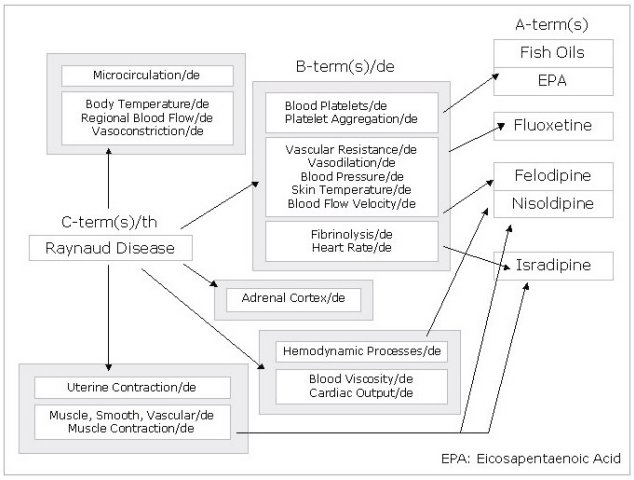

Figure 3 - One-directional open discovery process

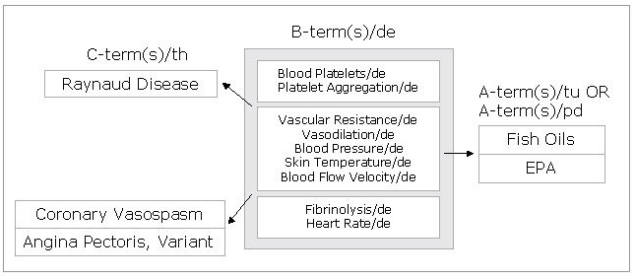

Figure 4 - Two-directional open discovery process

Conclusion

By grouping candidate descriptors using cluster analysis, we succeeded in detecting invisible connections between medical concepts. It is important to obtain knowledge on relationships between various phenomena observed in clinical practice. All relationships are not always direct connections, however. Our new methodology was effective in identifying hidden connections between medical concepts. The resulting knowledge overview could be helpful in clinical practice.

References

[1] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine. 1986. 30(1), 7-18.

[2] Weeber M, Klein H, de Jong-van den Berg LTW, and Vos R. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. Journal of the American Society for Information Science and Technology. 2001: 52(7), 548-557.

[3] Swanson DR. Smalheiser NR. Torvik VI, Ranking indirect connections in literature-based discovery: The role of medical subject headings. Journal of the American Society for Information Science and Technology. 2006: 57(11), 1427-1439.

Address for correspondence

Shizuka Yamazaki
Department of Pharmacy, School of Medicine,
Keio University, Japan
E-mail: yshizu@sc.itc.keio.ac.jp