



IDENTIFYING POTENTIAL EARLY BIOMARKERS OF ACUTE MYOCARDIAL INFARCTION IN THE BIOMEDICAL LITERATURE: A COMPARISON OF TEXT MINING AND MANUAL SIFTING TECHNIQUES

Suzy Paisley

Monday, 31 October 2016

Display Hours: 08:45 – 14:15

PRM55

B17

Paisley S, Seva J, Stevenson M, Archer R, Preston L, Chilcott J, Thornhill M
University of Sheffield, Sheffield, United Kingdom

OBJECTIVES: Text mining (TM) techniques are a novel approach to sifting large search results in systematic reviews but very little evidence exists comparing these techniques with manual sifting. This research compares TM techniques with manual iterative sifting techniques used in an existing systematic review of potential early biomarkers of acute myocardial infarction (MI).

METHODS: The existing review had retrieved 27198 references and identified 220 potentially relevant biomarkers, of which 25 were known at the outset of the review and 195 were identified by the review. The 27198 references formed the dataset for the TM evaluation. Relevance judgments were set up to identify the 195 biomarkers within the dataset. The text of a core publication, which included most of the 25 known biomarkers, was compared against each of the references using a simple TM approach. Documents were pre-processed by removing stopwords and converted into tf.idf weighted vectors. Each reference was then compared against the core publication by computing cosine similarity. The output of the TM approach was a ranked list of references.

RESULTS: TM techniques identified 189/195 biomarkers (97% sensitivity). 131 biomarkers had been identified after 1000 references had been processed (67% sensitivity at 3.7% of the abstracts) and 184 after 4000 references had been processed (94% sensitivity at 14.7% of the abstracts). The manual iterative sifting techniques used originally had sifted 17354 to identify the 195 biomarkers. 5/6 biomarkers not identified by TM techniques had been considered not relevant by the original review team.

CONCLUSIONS: Although the TM techniques applied in this study were simple, they proved to be a highly sensitive and efficient approach to identifying relevant evidence in a large dataset. Further evaluations should be carried out to apply alternative TM techniques and to increase the evidence-base on the comparative effectiveness of TM approaches in systematic reviews.