Reviewer's report

Title: Decision Trees: A Knowledge Representation for Automated Information Extraction of Randomized Controlled Trials

Version: 1 Date: 13 June 2008

Reviewer: Nancy Wilczynski

Reviewer's report:

General

The aim of this study was to conduct an exploratory analysis of RCT abstracts to investigate the feasibility of using decision trees as a semantic structure. The authors set out to answer 4 specific research questions outlined in the Methods section of the paper.

-------------------------------------------------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

Some aspects of the paper may be difficult for the reader to understand because of the terms/jargon used. The author should define the following terms:

Decision trees (define this early on), semantic structure, text mining, language processing, patient utilities, chance nodes, and true decision trees.

Additionally, the authors should indicate the definition they used for classifying studies as randomized controlled trials. For example, were quasi-random methods of allocation acceptable?

The 2nd sentence in the second paragraph under the section titled “Information Overload in EBM” contains an error. ACP Journal Club and Evidence-Based Medicine do not “aggregate and distill RCT outcomes into systematic reviews.” These 2 secondary journals include reviews that have been conducted by others given that they are published and pass pre-defined methods criteria.

In the Methods section of the paper, the years searched in PubMed should be stated.

Were the abstracts (the corpus) derived from CONSORT endorsing journals? Since the authors state that “In our work, we attempt to exploit the strict design principles and stringent reporting guidelines for RCTs…” one would think the CONSORT endorsing journals should be sampled as all journals indexed in Medline may not adhere to CONSORT reporting guidelines.

How did the authors determine which key decision tree elements to look for?

(page 12)
The author appear to have confused “allocation concealment” with “blinding” as the examples they provide for allocation concealment (on page 12) “…(single or double blinded, open label)…” are examples of blinding not allocation concealment.

How and when were the categories for intervention information, population information, and outcome information determined? For example, were they derived a priori or based on the corpus analysis?

Was there a reliability check for I, P, and O assignments?

The meaning of the 1st paragraph on the 14th page is unclear.

How many items were in the disputed set? (page 14 – results)

How did the authors determine that unique section headings were chosen at the discretion of the authors? (page 16)

Was the data presented on page 20 under “Study of Decision…” taken from the abstracts or the full-text of the articles?

-------------------------------------------------------------

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

First paragraph of the introduction, 3rd sentence, “While there remains…” should add “good quality” just prior to “RCTs are designed…”

First paragraph of the introduction, final sentence, “As such, RCTs have a crucial place…” should change the final 2 words from “…clinical trials.” to “…providing evidence of treatment effectiveness.”

Provide a reference for the 2nd statement (“In total, there are over 230,000 RCT citation entries in PubMed”) under the section titled “Information Overload in EBM”. The third sentence in the first paragraph of the same section should be changed to provide the actual year rather than stating “…over the last year.”

On page 6, 1st paragraph, update the sentence “By 2000, the Cochrane collaboration had produced…” by providing 2008 data.

Page 6, define the acronym “GEM”.

Page 7, move the definition of PICO up to its first occurrence.

It is not clear what is meant by the final sentence on page 7, paragraph 2, “Some of this work…”

The 4 specific questions to be answered in this study should be moved from the “Methods” to the “Introduction”.

There is no reference to table 1 in the text of the article.
The references should be reported in a consistent format. Currently there are many inconsistencies (e.g., abbreviation of the journal name is used at times while at other times the full journal title is used; volume numbers appear only some of the time).

Titles of tables should appear at the top of the table.

Should indicate in Tables 4 and 8 that data were derived from the R1 group.

Discretionary Revisions (which the author can choose to ignore)

None

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests.