Reviewer's report

Title: Evaluation of PROforma as a language for implementing medical guidelines in a practical context.

Version: 1 Date: 14 December 2005

Reviewer: David Kaufman

Reviewer's report:

General
A study that demonstrates (or evaluates) the efficacy of a guideline execution (or modeling) language in a practical context would be a significant contribution to the field of medical informatics. The context in which this work is situated, the management of hypertensive patients through a pharmacy, represents a potentially compelling and novel context. Unfortunately, this article suffers from several significant deficiencies including inadequately stated objectives, a lack of clarity in several sections of the paper and most importantly, a complete lack of methodological description. It is impossible to discern the merit of the claims or findings presented in the results section. The results are selectively presented and stated in the form of conclusions rather than findings grounded in evidence. There is no way of knowing how the authors arrived at such conclusions on the basis of the study. In fairness to the authors, the presentation is relatively well balanced and there is no obvious bias. In fact, much of the reporting is quite critical of PROforma, highlighting significant limitations.

PROforma is described in detail in several other articles and this is clearly not the goal of this paper. In my view, this paper requires a significant re-write to be acceptable for publication. A complete description of the use of PROforma in the context of the pharmacy study may constitute a significant contribution to knowledge. This would also necessitate an adherence to standards for reporting scientific data. I don’t see how anything less would be worthy of publication in this journal. In addition, the authors would have to attend to several of the issues raised in this review.

The authors compare their work to a paper published by Peleg and colleagues (2003) who have somewhat similar aims. I would not normally raise such a comparison if the authors had not already done so. In my view, the Peleg paper reflects a carefully executed study with great attention to detail and the paper is a model of clarity in important respects.

I should clearly state that I am not an authority on guideline or knowledge representation languages. However, I would compare my level of expertise to a typical reader of this or other medical informatics journal.

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

Many of the most serious concerns are raised in the General section. The lack of detail in the methodology and the informal reporting of results are perhaps the most problematic. In addition, there should be a clear statement of the objectives and how they are operationalized in the context of this study. The general objective is to analyze the adequacy of PROforma in a specific clinical context. However, the majority of claims make only tenuous reference to specific contexts.

PROforma was evaluated in the context of 3 phases: knowledge acquisition, analysis design and system implementation. The foci of the analysis are logical adequacy, heuristic power, notational
convenience and explanation support. A well organized presentation would provide an explicit breakdown of results according to the 3 phases and on the basis of the 4 criteria. It would be extremely useful, if the authors could connect the clinical aims of the study (e.g., satisfactory outcomes) with the stated informatics aims (expressiveness of the Performa language) in some fashion. Although an extended treatment of clinical outcomes is likely beyond the scope of this paper. In general, some combination of quantitative (Tables), qualitative and anecdotal evidence could make for a compelling story. Though the selection of methods is of course entirely up to the authors and would suffice as long as they are presented with greater clarity and adhere to proper standards.

The methodological description is deficient in important respects. What are the units of data? Presumably, they differ according to the different phases. How is their adequacy judged? How were the criteria for determining (for example) logical adequacy arrived at? Is there a consensus process for making such determinations? I am particularly interested in the implementation phases and specifically would like to know about events or instance in which the system contributed to a sound decision or an unsound one or proved to be problematic in some important respect. For example, what are the factors that contributed to a false urgent referral? Can they be traced back to some limitations in the modeling language? I appreciate that some of these factors may be a result of human error or some other aspect of the system, but some diagnostic reporting could be very informative. I also realize that evaluating a knowledge representation language may not neatly fit into an experimental package the way a controlled clinical study would. But it is clear that the paper could use much greater systematicity and rigor in articulating the method and in the reporting of the results.

There are some contradictions in the paper. For example, the authors conclude that “there was no difficulty in heuristic power, but in the results discusses problems with computational completeness and the inability to express certain functions. The contradictions may be more apparent than real, but should be addressed nonetheless.

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

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

An expanded background section should include greater discussion of the motivation for guideline modeling languages.

The article should be readable by an informaticist who is not an expert in knowledge representation or clinical guidelines. It is not necessary to reduce the discussion of technical issues, but rather to provide greater exposition, examples and better graphics (perhaps, including screenshots of the implementation or the graphical modeling tool). I did not find Figures 1 and 4 to be particularly informative (or necessary). On the other hand, I appreciated Figures 2 (which is coupled to a very good textual description) and 3.

There were sections in the paper that I had difficulty following. For example, I didn’t understand how you determined which distinctions need a CIG needs to represent and in particular, the answer that you provide. It seemed circular to me, but it may just be my misread. I had a similar problem with the critique regarding the lack of support for constraints in PROforma. I don’t understand how there can be no mechanism for expressing or checking for violations of constraints in a guideline modeling language.

It would also be helpful, if the authors make clear the consequences of the limitations of PROforma (e.g., inability to group task classes into hierarchies) and the impact they had in this study.
There are a few typos and grammatical errors. For example, on page 10 it states “more convenient to visit their pharmacists rather than their GP’s surgery”.

Discretionary Revisions (which the author can choose to ignore)

It would be useful to explain the relative advantage of Task Network Model over an MLM for guideline modeling (in a sentence or so).

What are the differences (in broad strokes) between the Arrezzo and Tallis implementations?

**What next?:** Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

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

**Quality of written English:** Acceptable

**Statistical review:** No

**Declaration of competing interests:**

I declare that I have no competing interests.