Author's response to reviews

Title: Usability Evaluation of a Clinical Decision Support Tool for Osteoporosis Disease Management

Authors:

   Monika Kastner (monika.kastner@utoronto.ca)
   Danielle Lottridge (danielle.lottridge@utoronto.ca)
   Christine Marquez (cmarquez@utoronto.ca)
   David Newton (david.newton@utoronto.ca)
   Sharon E Straus (sharon.straus@utoronto.ca)

Version: 2 Date: 13 October 2010

Author's response to reviews: see over
Response to Reviewers
Usability Evaluation of a Clinical Decision Support Tool for Osteoporosis Disease Management

Dear Editor:

Thank you very much for the opportunity to publish our work in your journal. We have carefully considered all comments and have revised the manuscript accordingly. We addressed the two reviewers’ questions and comments below (highlighted in blue), and have also amended the manuscript accordingly (please see both marked and unmarked versions attached).

REVIEWER 1 COMMENTS:

This is a very nicely written manuscript that lays out results of a three-phase study to evaluate the usability of tools to assist patients and physicians with osteoporosis diagnosis and management. The paper that lays out the development and thus content of the tools evaluated in the current paper appears to be under review (ref #19).

This paper has now been published (see updated reference on page 23 of the manuscript).

Major Compulsory Revisions:

While the focus of the paper is clearly not on the tool development and thus content of the various tool elements (RAQ, COPE, BestPROMPT), some additional brief summary, or perhaps included in an appendix, of these tools would vastly improve the readability of the paper. I found it difficult to interpret the qualitative findings in particular without knowledge of what each of the components of the tool looked like and included. In the same vein, some clarity regarding the administration, and product, of each step would be helpful. For example, if there is sound involved in completion of the tablet RAQ, then presumably this precludes a patient completing the tool in the waiting room? Is the audio necessary or optional? Second, it was unclear to me whether or not there was something being printed after the patient completes the RAQ and, if so, how and where such printing takes place?

We added a paragraph on page 5 of the manuscript to provide more detail about the 3 components of the tool, what outputs are generated, and how this will be done.

One element that may impact the overall user-ability of these tools that is not discussed is the ability to revise the tool to reflect changes in best practices – for example, the 2010 Osteoporosis Canada guidelines for OP diagnosis and management are about to be released, and will look much different from the 2002 guidelines, in particular
incorporating a Canadian version of FRAX, a composite fracture prediction tool, and changes to BMD testing recommendations. Can the tools be readily adapted to reflect these changes? This should be included in the discussion.

We added a couple of sentences to the discussion (page 16 of the manuscript) to address how our tool can adapt to changing guidelines.

There are numerous abbreviations used (e.g. SUS, CDSS) - it would make it easier for the reader if these were restricted to perhaps just the abbreviations for the 3 tool components

We eliminated abbreviations throughout the manuscript such as: POC (point of care), FP (family physician), GIM (general internal medicine), SUS (system usability scale), and CIT (critical incident technique).

Page 15, under section 1), it is noted that all participants indicated that they would discuss the information with their physician - however, I wonder if this might be very physician specific in that those with 'responsive' physicians would be more willing to initiate the discussion -- if so, the fact that these patients were all recruited from the same FP practice might have biased this finding – this should be considered.

Although we did not address this finding specifically in terms of bias, we did indicate discuss this in our Limitations section (pg. 18 of the manuscript) - that we recruited all 19 patient participants from the patient population of one family physician, which may not be representative of other family physicians or their patients.

Minor Essential Revisions:

Page 9, paragraph 2 - please clarify the meaning of the second sentence "This process....whether the response inputs generated by the decision algorithm...".

We clarified this sentence on page 10 of the manuscript.

Tables - it might be useful to indicate the participant group for each Table – for example, Table 1 is physicians, while the other 2 tables are patients

We edited the headings for Table 1 (page 25 of the manuscript), Table 2 (pg. 26), and Tables 3 & 4 (page 27) to clarify the participant groups.

Table 2 - the footnote indicating the statistical test used for evaluation of significance is missing from the table itself - presumably the symbol is needed in the column heading on the far right?

We added the symbol (column representing the mean difference in time to RAQ completion) for the statistical test in Table 2 (page 26 of the manuscript).
Table 3 - Are the comparison groups reversed? I believe based on the data that row two should say 'Mouse/Keyboard versus Stylus' and row three should say 'Touch versus Stylus' - please clarify.

The comparisons are in the correct order - the stylus pen input device was considered the index, which was compared with either to the mouse/keyboard or touch screen.

Discretionary Revisions:

Perhaps a minor point, but on page 10 it is noted that inter-coder reliability was assessed using Kappa statistics - Kappa controls for agreement by chance, which would be fairly unlikely in the setting of qualitative data analysis, would it not? Further, the next sentence states that where there were disagreements, defined as < 90% agreement (i.e. not corrected for chance), consensus was reached. Thus, I am unclear as to the need or value of the Kappa evaluation. Please clarify.

We used a feature in NVivo to compare 2 reviewers for coding themes, but it was used mainly as a calibration exercise to ensure agreement for classifying themes. For themes that were difficult to classify (and for which the 2 reviewers did not agree), a third reviewer settled the disagreement. We clarified this on page 11 of the manuscript.

Page 13, second paragraph - it is noted that the time differences for completion with stylus pen versus touch screen was non-significant, yet it is highly probable that you had insufficient power - might simply note this?

We added a sentence to the Limitations section about this (page 18 of the manuscript).

REVIEWER 2 COMMENTS:

I heartily agree with the authors about the importance, and value, of this sort of usability testing. The paper is well researched, and I find it very hard to fault any aspect of the techniques employed. I suppose the greatest limitation is that it's a little hard to generalise from the findings, but at least it puts out there a 'data point' - e.g., these folks did pragmatic usability testing and readily found ways to improve user interaction with their system.

I have the following discretionary revision suggestions:

1. In Study 1 the authors say "the moderator simulated a patient at risk for osteoporosis and completed the RAQ on a Tablet PC" - this raises questions for me. Was it always the same simulated patient (i.e. exactly same profile)? How was this profile decided? How
might the session have gone differently with other risk levels and risk factors presenting? I can see why this approach was pragmatic, but I think it is a big limitation as compared to what could be learned from a range of patient profiles (simulated or actual).

We used input from an internal medicine specialist (with an interest in geriatrics) and a family physician to develop a case that would be the most representative of a patient at risk for osteoporosis (i.e., a 63-year old woman with at least 1 major risk factor [e.g. age or postmenopausal status] or 2 minor risk factors [e.g. smoker and excessive caffeine intake]). It was important to use a standardized case so that we could isolate usability problems independent of the patient case. We wanted to test the process of how the tool might be used in real practice (i.e., how the BestPROMPT is printed wirelessly from the patient-completed risk assessment questionnaire). We could have conducted usability sessions without the patient case by simply showing physicians the BestPROMPT sheet, but we thought it was important to first show how it can be generated and how the process might work in the context of the participant’s own practice. However, we recognize that it would have been ideal to test a range of patient profiles (with at least 5-8 participants per profile) – unfortunately this likely would have been difficult given our limited resources and low recruitment rates.

2. The authors cite Nielsen (ref 24), but I think could go a bit further in explaining the rationale for this sort of usability testing with Nielsen's argument of diminishing returns in problems found as number of testers/subjects increases.

We discussed this a bit on page 7 (first paragraph, last sentence) when we described the rationale for selecting 5-8 participants per usability study.

3. Also, Ash, Berg and Coiera's work on negative consequences of health IT is cited (ref 36) but I think should be brought out more explicitly, probably in the Intro, to explain the motivation for this sort of experiment. Further to this, I find the studies in BMJ by Eccles, Rousseau and others (2002, 2003) to be a great cautionary tale of running to an RCT without looking at basic usability. All in all, I just feel the authors should hammer the point more.

We included a couple of sentences in the Background to emphasize these points (bottom of page 5 to top of page 6 of the manuscript).

4. I was disappointed to realise that the physician interaction is entirely paper based. I suppose this fits the situation in Canada (as compared to the UK, Australia, NZ, Denmark, Netherlands and other locales where family doctors work in a computer-based fashion). This leads to a couple thoughts: (a) to at least recommend the progression to computerisation ('meaningful use' in the current US trend) to better support CDSS (including the recommended functionality to defer a prompted action); and (b) Kawamato's review of guideline success factors is cited (ref 34) but the primary recommendation - that decision support be computer-based - is essentially ignored for the physicians, which seems a basis for reflection.
We agree with the reviewer on all these issues, but have previously discussed these in some detail in our focus group study (Kastner et al, *BMC Med Inform and Dec Mak* 2010;10:40). Other barriers that we encountered during the development of our osteoporosis tool have been the lack of support from the EHR vendor involved in our study sites, specifically to integrate our software programming into their system. Currently, implementation on a wider-scale is therefore difficult not just because there are so many EHR vendors (even within one FHT), but because most vendors are independently owned and thus have their own proprietary restrictions.

5. I think the Conclusion (in the body, maybe not the abstract) should recap some of the key lessons learned for this particular software (e.g., some aspects of the refinement lessons: touchscreen less error prone than stylus). As it stands it only states the meta-level lesson that problems were found. Moreover, I think the conclusion should give some reflection of the outstanding issues of physician time, control and potential diversion from more critical tasks - these were uncovered but not convincingly resolved. The reconfirmation of these critical challenges (and with perhaps the novel twist re distraction from 'real' reason for visit) is a key message as much as the more easily dealt-with usability refinement in a narrower sense.

   We added a couple of sentences to reflect these suggestions (page 19 of the manuscript).

4. Also, just presumably a typographic error - the 11 physicians are reported "mean age ?"

   We corrected this error (page 12 of the manuscript).