Reviewer's report

Title: Comparison of measurement methods with a mixed effects procedure accounting for replicated evaluations (COM3PARE): Method comparison algorithm implementation for head and neck IGRT positional verification

Version: 2 Date: 11 August 2014

Reviewer: Gareth Webster

Reviewer's report:

Major compulsory revisions: none

Minor essential revisions: none

Discretionary revisions:

1. The authors define three tests for the ‘interchangeability’ of two methods, but there is little demonstration or discussion of how sensitive these tests are to outliers. For example, a comparison of, say, ultrasound-based prostate setup vs kVCBCT may be in good agreement for the most part, but vary substantially in the presence of bilateral hip prosthesis or a very large patient. Would the authors expect their inter-subject variability test to highlight these problems? What if the same dataset contained a few large patients (that made the ultrasound worse) and a couple of bilateral hip replacements (making the CBCT worse)? Would this test ascribe the same variability to both techniques or would it spot that the variability came from different patients? Some explicit discussion of this would be of interest.

2. Could the authors please comment on the criteria set for interchangeability: it could be argued that P<0.05 requires that a fairly marked difference is seen between techniques. If comparison gave P=0.10 for each test, this is still pretty likely that there is a difference but the tests given here would seem to imply that the techniques compared were interchangeable. Although there is a test to highlight a systematic variation in mean values >1mm, there doesn’t appear to be anything comparable for inter or intra-subject variability. I can’t see any requirement of minimum sample sizes that might mitigate this effect, would the authors recommend limits on sample size?

3. In the clinical example used in the paper the test uses paired data that are acquired approximately simultaneously. Do the authors anticipate this being the only scenario in which the COMPARE strategy holds, or would they also advocate its use in two imaging strategies used on separate but similar cohorts of patients? In either scenario, would the authors place any limits on how much data (or what ratio of data between the 2 methods) are required for a reliable result?

4. While the manuscript is very well-written it may be worth the authors
considering that parts of it are a challenging read for the not-particularly-mathematical mind, so more frequent interpretation throughout in terms of the clinical scenario being explored would be helpful.

P12, line 26: …comparison….?

**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