Reviewer's report

Title: Key considerations for the experimental training and evaluation of cancer odour detection dogs: lessons learnt from a double-blind, controlled trial

Version: 1 Date: 4 December 2013

Reviewer: Marije Bomers

Reviewer's report:

Key considerations for the experimental training and evaluation of cancer odour detection dogs: lessons learnt from a double-blind, controlled trial.

In my opinion the authors have done a good job in writing about a experimental that failed and analysing & describing where and why it went wrong. Very valuable lessons to be learned here, and I think it would be worth publishing for the rest of the academic world to learn from, provided some revisions (partly major, partly minor) are made. The field is easily distracted by ‘fantastic’ results and this down- to- earth counterpart is in my opinion at least as valuable.

On several points it can do with alterations; underneath, while reading I have stated where / what I would like the authors to reconsider.

In general it’s written very comprehensible, however the phrasing can be more to the point (e.g.: ‘Ideally, dogs should never be presented samples from the same donor on more than one occasion, though in reality it may not be feasible to gain a sufficient number of training samples to do this for many types of cancer. Although pooling of samples appeared to be unsuccessful in the present study in encouraging the dogs to generalise, it would be useful to further explore the chemical evidence for whether or not pooling of urine/biological samples from different donors could assist in creating new headspace odour profiles to widen the pool of training samples.’)

The abstract should undergo major revising and needs to be more to the point and factual (see comments underneath).

The design of the study is good, although it sounds a bit misleading that 10 dogs were used, when only 3 out of 10 proved ‘trainable’. On the other hand, the efforts made to blind the testing phase are commendable. There is a relatively large number of samples used; ideally, there would be even more but as rightly stated it could be difficult to find a larger sample- selection. The selection and characteristics of donors and controls is still a bit hazy.

Abstract:

- how many samples? ‘maximum number of training and testing samples available from a large research hospital’ = rather vague and less relevant than the actual amount used.
- ‘using a study design that also provided information on canine olfactory learning’. Was the analysis of learning part of the design from the start?

- Results: what are the results? What success rate ‘is expected by chance’ and how far off are the dogs? More factual/ concrete results please rather than interpretation. I could also be worth mentioning that 7 / 10 dogs proved unfit for the job, not because their olfactory capacity was insufficient, but due to ‘character issues’.

- Conclusion: first state your conclusion; e.g. ‘… the dogs are not able to …’ and then the implications/ recommendations. I would emphasise your study illustrates that is it very easy to draw misleading conclusions about the abilities of dogs to indicate certain odours, unless extremely robust double-blind test trials are conducted.

Background:

- Clear intro.

- relatively ease = relative ease?

- Last paragraph: long sentence; without reading the rest of the article it’s a bit hazy what you actually mean to do

- What the difference between: ‘pseudoreplication of odour samples from the same donors during analyses’ and ‘samples re-used multiple times during testing, potentially making them familiar to the dog’? using donors twice vs using samples twice?

- A lot of emphasis is made on the learning aspect being a major research goal. I can imagine that this was not necessarily so on the start of the project. I can imagine that when the experiment did not succeed, the researchers investigated what went wrong, which is of course very commendable. However, in this paper it is presented like this was the idea from the start and the set-up was especially designed to do this. I wonder: what aspect in the set-up was especially designed to test whether ‘…multiple sample learning may impact upon the training of dogs.’ I would either specify this, or otherwise present your research as it is, which interesting enough whether or not you set out to analyse where it went wrong from the start.

Methods:

- Very commendable that ten different dogs were used. However, since the majority of the dogs proved to be ‘untrainable’ I wonder how they were selected? Is there an explanation that so many were not trainable?

- Impressive to have such a large number of donors

- For how long were the samples stored? More than a few months? Despite rigorous samples techniques the odour could be influenced by ‘aging’, even on -20C.

- Were the samples taken after the diagnosis had been made, so could it be the invasive procedure (e.g. taking a biopsy and therefore ‘opening’ the prostate or tumor) that leaves the smell? I take it that the healthy donors had not undergone
any invasive procedure; how many healthy donors were there (compared to BPH controls?) How sure are we that the BPH controls indeed did not have cancer; were biopsies taken in all of them? And again were the samples taken before this procedure?

- Design seems to be very adequate, although ideally more positive samples would be available, but as rightly stated it’s not that easy to get a larger number of samples. I expect wide CI on these numbers.

- Under ‘training procedure’, 2 nd paragraph: because should be became?

Results:

- Where are the patient characteristics on the CaP patients? E.g. PSA and Gleason score? Have they got severe metastasised cancer or small, relatively innocent tumours?

- I do not understand the discrepancy between the stage 3 ‘new samples’ and the formal testing results. Was the dog training with the samples (during stage 3) before the result was scored?

- The fact that only 3 / 10 dogs proved trainable is remarkable. Again, how were they selected? In my opinion this is one of the major drawbacks of the study and it should be made more clear (also in the discussion) that it’s not so much so that none of the 10 dogs were capable, but none of the 3 dogs were capable.

Discussion:

Recommendations:

- Our study illustrates that is it very easy to draw misleading conclusions about the abilities of dogs to indicate certain odours, unless extremely robust double-blind test trials are conducted. I think this should be (one of) your most important conclusion(s), and should receive more emphasis (e.g. in the abstract).

- Recommendation 4: double BLIND instead of double bind

**Level of interest:** An article of importance in its field

**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