Reviewer’s report

Title: Can routine data be used to support cancer clinical trials? A historical baseline on which to build: retrospective linkage of data from the TACT (CRUK 01/001) breast cancer trial and the National Cancer Data Repository

Version: 0 Date: 21 Mar 2017

Reviewer: Jody Ciolino

Reviewer’s report:

The authors present an interesting idea in their manuscript that, if developed further, may increase cancer clinical trial efficiency in some instances; it also has the potential to increase efficiency and ease of data capture for trials in fields other than cancer in which ideally participants would be followed for longer periods of time (i.e., longer than five years as the authors note). Registries and other sorts of 'routine' data collection routes may in general be rich sources of data, but there are inherent biases if one chooses to use this as a primary source of data collection. The authors should explain this issue in greater detail. There are multiple other major (and some minor) issues that should be addressed prior to consideration of this paper publication in Trials. They are outlined here:

1. The term 'characterises' in the background section of the abstract should be replaced with something less vague. It is unclear what this really means.

2. Several statements require more evidence/specific references to provide greater strength. For example, there should be a reference or greater evidence that there is a tendency to curtail follow-up after five years. Is this anecdotal or is there a specific reference or references that allow the authors to make this claim?

3. The authors should define 'routine' sources of data capture.

4. The details surrounding methodology is lacking; this is probably my largest concern for this paper. The authors should explain 'data cleaning, resolution of incomplete and incorrectly formatted information..." (page 6, line 41)...what was the process of data cleaning and how were incomplete and incorrectly formatted data 'resolved'? What are examples of inconsistencies? Were the decisions and ultimate final dataset(s) agreed upon in some objective way or via study team consensus?

5. I have a similar comment regarding decision of a suitable 'proxy' event to identify recurrences. What were candidate proxy events and how were they determined/agreed upon?

6. There is reference to an exploratory multiple logistic regression model, but these results are not reported. Again, detail surrounding these methods are missing. Which software was used, what levels of significance were used, etc. Further, it would seems that rather than simply
presenting % agreement, more descriptive measures such as Kappa statistics (with confidence limits) and correlation coefficients/Bland-Altman results, etc. would be more informative.

7. It should be clearer as to what the authors mean by 'one row per patient' - One would assume this means duplicates were removed. How did this process occur? What makes one record the 'correct' one?

8. The discussion makes reference to 'good' quality in the first sentence. This term seems subjective and/or a clear definition of 'good' should be added.

9. The conclusion makes reference to 'the goal' - whose goal is this? The authors should explain and elaborate on these points. In addition, the last sentence regarding 'working with NCRAS...to validate routine data...' should lay out more concrete steps using less vague terminology. How do the authors plan to 'work with NCRAS' and what is the method of validation?

10. The authors should consider discussion of strengths and limitations of this project along with wider applicability to other fields or datasets.

**Level of interest**
Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

**Quality of written English**
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

**Quality of figures**
All images and figures within the manuscript should be genuine i.e. without evidence of manipulation. No specific feature within an image may be enhanced, obscured, moved, removed, or introduced. If you have concerns about the veracity of the figures you should choose the first option below.

**Statistical review**
Is it essential that this manuscript is seen by an expert statistician? If so, please give your reasons in your report.

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:
1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal