Author's response to reviews

Title: Randomised trial investigating the relationship of response rate for blood sample donation to site of biospecimen collection, fasting status and reminder letter: the 45 and Up Study

Authors:

Emily Banks (emily.banks@anu.edu.au)
Nicol Herbert (nicol.herbert@saxinstitute.org.au)
Kris Rogers (kris.rogers@saxinstitute.org.au)
Tanya Mather (tanya.mather@anu.edu.au)
Louisa Jorm (louisa.jorm@saxinstitute.org.au)

Version: 2 Date: 25 July 2012

Author's response to reviews: see over
Dear Dr Boutron,

Re: Randomised trial investigating the relationship of response rate for blood sample donation to site of biospecimen collection, fasting status and reminder letter: The 45 and Up Study

Thank you very much for your email and for the helpful and constructive comments by your reviewers. We provide a point-by-point response to each of these below and have uploaded the amended paper. The reviewers’ comments are given in italics.

We very much hope you will now find our paper suitable for publication in *BMC Medical Research Methodology*. Please do not hesitate to contact me at emily.banks@anu.edu.au or +61 2 6125 0328 if you require anything further.

With best wishes,

Yours sincerely,

Professor Emily Banks, on behalf of the authors
1. MAJOR COMPULSORY REVISIONS

1.1 Page 7 Intervention – Commercial pathology centres: The last sentence on page 7 states that “collection of blood samples was undertaken by existing pathology centre staff who were specifically trained by Link-Up Study staff to measure height, weight, waist circumference, blood pressure and heart rate”. Did the participating commercial pathology centres have a large number of individual staff all of whom collected samples from study participants? If so, how was it possible to train them to take measurements and collect specimens in accordance with study protocols?

There were only two commercial pathology centre sites in this trial, and those sites did not have a large number of staff. Therefore, it was possible to provide some basic training in the collection of the physical measurements (height, weight, waist circumference, blood pressure and heart rate). Commercial pathology centre staff were not given specific training to collect specimens in accordance with the study protocol. The business practice of the commercial pathology centres is to use Pathology Request Forms, which specify the collection instructions. As noted in the discussion (page 26), the physical measurements training provided to the commercial pathology centre staff would be difficult to replicate on a larger scale.

1.2 Page 11 Randomisation: I think it would be OK to indicate more specifically the reason for making the decision to include a fourth arm in Parramatta after the trial has started, and why it was preliminary results from Wagga Wagga that prompted this. If it was poor response rates at commercial pathology centres when a reminder letter was not issued then there would be no harm in making such a statement.

We have now added a statement in the randomisation section to indicate that the reason for making the decision to include a fourth arm in Parramatta was because of poor response rates in the preliminary results from Wagga Wagga.

1.3 Page 13 Statistical methods:

(i) Paragraph two states that “Response rate was analysed in two models and expressed as a participation rate ratio (RR)” but it would be better to say that “Response rate was analysed in two models and ASSOCIATIONS WITH EXPOSURE VARIABLES expressed as a participation rate ratio (RR)”;

We have amended the manuscript in accordance with the reviewer’s suggestion; this revision is part of removing reference to characteristics of participants versus non-participants, suggested by the second reviewer.

(ii) The sentence on the “log binomial model” (presumably this is just logistic regression?) appears to be largely repeated for models 1 and 2, so perhaps just state this once and then point out in a separate paragraph the differences between model 1 and model 2;

We have now mentioned one model only, in keeping with the reviewer’s suggestion and the revisions mentioned above.

(iii) The statement that “model fit was verified by deviance and chi-squared” is vague and I would leave this out;

We have now removed this statement.

(iv) Stating that “the overall effect of a variable was assessed with a likelihood ratio test” is also a bit vague, so perhaps say that “nested models (e.g. those with and without a given exposure variable) were compared using the likelihood ratio test LRT)”;

We have now removed this statement.
This section has been removed, due to the revisions suggested by the second reviewer.

(v) Similarly, be more direct with the assessment of heterogeneity by saying that “heterogeneity of effect between areas was assessed by including the relevant interaction term and comparing this model to the model without the interaction term using the LRT”;

This section has been removed, due to the revisions suggested by the second reviewer.

(vi) While the paired t-test is not an unacceptable way to compare self-reported and measured height, this only addresses the null hypothesis that the mean difference is zero, which is not an especially relevant hypothesis. What one is really interested in here is comparing these two methods of measurements, and determining the extent of the agreement between them. The Bland-Altman procedure (Lancet 1986) plots the paired difference in measurements against the within-pair average measurement to generate a prediction interval for the mean difference, and to assess (visually) whether this mean difference depends on the level of measurement. I suggest the authors try this out and attempt to incorporate some basic results from this procedure.

Concern is often expressed that dedicated clinics and dedicated study staff are necessary to make physical measurements for cohort studies and that use of health services staff may lead to reductions in data quality. The main reason for comparing self-reported and measured height and weight for the two different types of site was to find empirical evidence of issues with data quality. However, the paper was already quite long and we wanted to explore whether there were large or systematic differences between measurements from the dedicated clinics and from the pathology services. We are acquainted with Bland-Altman plots and have used them in a validation study based on the dataset used for the current manuscript (Ng et al 2011, attached). We considered the inclusion of Bland-Altman plots potentially excessive for the current paper.

In keeping with the reviewers’ suggestion, we have created Bland-Altman plots for comparison of the measurements at the two types of site, but have not shown the figures for the current paper, for the reasons outlined above. We have instead reported on their results as follows:

“Bland-Altman plots suggested no large or systematic variation in the differences between self-reported and measured height and weight according to the type of data collection site (data not shown).”

We have also added reference to Bland-Altman plots in the statistical methods section. We are happy to include the plots if the reviewers and editor consider they would add to the paper.

1.4 Page 15 Results: Many of the estimated rate ratios (RR) are close to 1.00 with large p-values, but the important thing is that the 95% confidence intervals do not include values of the RR that indicate a strong departure from the null hypothesis, that is, in these cases a relevant difference in the response rates can effectively be ruled out since the RR is estimated with high precision. A statement highlighting that there is little evidence for large differences in response rates would be useful.

We have now added a statement highlighting that there is little evidence for large differences in response rates, as follows:

“For the three randomised elements of this study above, relating to the three research aims, a meaningful difference in the RR can effectively be ruled out, since they are estimated with high precision. Hence, there is little evidence for large differences in RR.”
1.5 Page 17 Physical measures etc: The mean differences in measured heights and weights are worth reporting, but the correlations are pointless since they will always be very high for two methods measuring the same underlying quantity even if one is systematically in error. Better to ditch the correlations and report a summary from the Bland-Altman procedure (see above).

Please see response to comment 1.3 (vi).

1.6 Page 20 Discussion; The statement that “the randomised elements of this trial were analysed by intention-to-treat, to avoid biases resulting from the use of observational data” is rather clumsy since intention-to-treat analysis has not been previously mentioned, and it seems to imply that bias is a feature of observational data; what I think the authors meant is that “per protocol” analyses that compare groups defined by the intervention actually received (rather than that randomly allocated) can introduce bias in estimation since the comparison groups are no longer exchangeable. I would just leave this sentence out.

We have deleted this sentence, in keeping with the reviewer’s suggestion.

1.7 Pages 21-24 Discussion: The discussion is quite long, and I think it digresses on page 21 with the paragraph “the proportion of people willing to participate etc.” The material on pages 22-23 on dedicated clinics and existing commercial pathology centres is rather long-winded and could be consolidated into a single paragraph or two shorter paragraphs.

We have now shortened the discussion and made it more focussed.

MINOR ESSENTIAL REVISIONS

Background
Page 4 In the first paragraph I feel it would be better to refer to “biomarkers” rather than “bio-analytic data” since the latter seems to be, well, a made-up term! And in the second paragraph one might equally replace “biodata” with “biomarker”. “Biodata” also appears on page 12 and in the first sentence of the discussion.

We used the term “bio-analytic data” to refer to both the blood samples and the physical measurements. We apologise for being overly creative here and have followed the reviewer’s suggestion, using the term “biomarker”.

Methods
Page 5 Data collection centres: Wagga Wagga is hardly a metropolis, but it is a city of about 50,000 people and is a major regional centre, so describing it as “rural” is inaccurate. I feel it would be better to replace “rural” with “regional”.

We have replaced “regional” with rural, as suggested.

Randomisation
Page 10 “151.00S” should be “151.00E”.

Thank you very much for picking this up. We have amended this as suggested.

REVIEWER TWO: Bruno Giraudreau
2.1 General comment
The present manuscript reports the results of a pragmatic trial aimed at increasing response rate for blood sample donation. The research question is pragmatic and of real relevance because very few is known on the subject and because collecting data and biospecimens always raises difficulties in large cohort studies. However, I think there are too many objectives in the present study, and too many
results which are reported. This makes the paper hard to read and we come to forget the main objective of the study. Thus:

- We have actually two trials: one in an urban area, the other one in a rural area
- We have actually 3 interventions (site of biospecimen collection, fasting, reminder), and Figure 1 is absolutely necessary to understand the randomization design
- Individual factors associated to the response rate are indeed of interest, but this comes down to report both the results of a randomized trial and of an observational study

2.2 I do not well understand the relevance of results reported in Table 4:
- Why comparing height, weight, blood pressure measurements between dedicated clinics and pathology services? This is in no way “data quality”.
- What is the relevance of assessing the difference between self-reported and measured height and weight?

Concern is often expressed that dedicated clinics and highly trained study staff are necessary to make physical measurements for cohort studies and that staff from health services may not be capable of taking measurements in the same way. The main reason for comparing self-reported and measured height and weight for the two different types of site was to find empirical evidence of systematic differences that might indicate issues with data quality. We agree that it is not an orthodox measure of data quality, but these measures were pragmatic and defined prior to analysis. We have clarified the reason for comparing height, weight and blood pressure measurements, as follows:

“Concern is often expressed that dedicated clinics and highly trained study staff are necessary to make physical measurements for cohort studies and that staff from health services may not be capable of taking measurements in the same way. Among individuals who went on to participate in the Link-Up Study, the biospecimen yield and features of the biomarker data (height, weight, heart rate and systolic and diastolic blood pressure) were compared for those attending the dedicated clinic and those attending the pathology centre, in order to investigate whether there was any empirical evidence of systematic differences that might indicate issues with data quality.”

- Worse: what is the relevance of reporting correlation between self-reported and measured height and weight?

We have now removed reference to the correlation between self-reported and measured height and weight, and have focussed on absolute differences and interpretation of the Bland-Altman plots.

2.3 In the end, doing so makes the paper long (discussion section is 6 page long!) and hard to read. I would strongly suggest to focus on the results of their randomized trials (and indeed, actually, there are two randomized trials, cf supra) and to report them in agreement with the CONSORT Statement and its extensions.

We have now removed reference to the observational comparison of participants and non-participants and report the remaining trial in keeping with the CONSORT Statement.

Specific comments

2.4 The conclusion of the abstract does not provide enough information.

We have amended the conclusion section of the abstract to include the main finding of the study. We would be happy to add more, but are currently constrained by the word count.
2.5 The end of the “Background” section explicitly reports the hierarchy of the research questions investigated in this study:

- First question: influence of the site of data collection
- Second question: influence of the fasting status
- Third question: influence of a reminder letter

Therefore, I think the statistical analysis plan should follow this hierarchy, which is not presently the case.

We apologise for any lack of clarity in this element of the paper. If the reviewer/editor requires that the background, statistical methods and results sections follow the same order, then the order of research questions needs to be:

1. Influence of fasting status
2. Influence of reminder letter
3. Influence of site of data collection

This is because of the need to have the results shown in this order. Having demonstrated the lack of meaningful effect of fasting status and reminder letter on response rates, we were then able to present these results to give the final assessment of the impact of site of data collection, without having to have the reader account for these other differences in their interpretation. If we reported on response rates according to the type of data collection site first, the reader would be wondering how we factored in the other randomisation arms.

Hence, in order to meet the reviewer’s request, we have changed the order of the research questions throughout so they are consistent.

2.6 Secondary aims listed in the end of the “Background” section should better be dropped from the present paper, for greater clarity. Moreover, I do not well understand how is assessed the “feasibility of collection of biospecimens and data on physical measures”. Is this assessed through the primary outcome?

We have removed the secondary aims, in keeping with the reviewer’s suggestion.

2.7 I would strongly suggest reporting results as if two independent randomized trials had been conducted: one in an urban area, and the other in a rural area. I would add that sample size calculation had not been performed considering data of these two areas would be pooled.

It is routine for randomised trials to include multiple study sites and to test for variation in results according to site of recruitment. In this case, we had planned a priori to combine results across urban and rural sites to analyse data on fasting status and data collection type, as well as showing results separately for the urban and rural sites. While we agree that the study could potentially be seen as two independent randomised trials, we are concerned that portraying it as such would detract from the clarity of the paper, and would not reflect accurately the original protocol. We have now clarified this, along with the power calculations (see point 2.12, below). We also note that this point was not seen to be of concern for Reviewer One.

2.8 Participant information: I supposed participant consented for the collection, but were not informed of the trial hypothesis. If such, this should be mentioned. Indeed, not fully informing participants of the study hypothesis is fundamental in the present study, since it is the only way of having some form of blinding in the present study. So a complete reporting of this point is necessary. Moreover, we should have a section talking about blinding in this study.

Thank you for this suggestion. We have added the following section to the paper:

“Blinding
Although invitees were told in the study information leaflet that they were part of a pilot project and that their participation would assist “in deciding the best way of implementing
the Link-Up Project”, they were not aware of the trial hypotheses or randomisation to differing data collection site types, fasting status and reminder letter receipt. In analysing and interpreting the data, the authors were aware of the randomisation status of study participants.”

2.9 For greater clarity, I would suggest having a “Design” section explaining the randomization strategy, in agreement with the hierarchical objectives of the study. Of note, this section is advised in the CONSORT guidelines.

We followed the CONSORT guidelines and have the required section entitled “Randomisation”, starting on page 10. We are a little unsure what the reviewer means in addition to this and see the term “Design” as a more general heading, that would seem out of place as the heading for this section.

2.10 Page 9-10: there is a complete description of how blood specimen were collected and processed for storage but, in the end, quality of blood specimens is only assessed through the number of aliquots. So I wonder whether such a precision is necessary.

We have now shortened this section. We have tried to retain enough background information on specimen collection methods necessary to interpret the data on biospecimen yield. Practical considerations for different methods of collection can have a bearing on yields and the ability of collections sites of differing sophistication to manage them.

2.11 Data were collected during a one month period. Why having chosen such a short delay?

The short data collection period reflects the real-world scenario of gathering biospecimens from the finite number of 45 and Up Study participants in each region. The duration of data collection was based on the number of eligible participants, estimated response rates and the likely throughput at each centre.

2.12 I do not well understand the sample size calculation. The hypothesis is a RR of 1.3, and a 30% response rate is expected. If 30% is the response rate in the control group, the expected response rate in the experimental group is 39%, and the sample size required (for a chi-square test, although I recognize that the main analysis was to fit a log binomial model) is then 437 per group. If 30% is the overall mean response rate, it comes down that the response rate in the control group is 26% and the expected response rate in the experimental group is 34%. The sample size required is the 514 per group. So precisions are expected. Otherwise, authors never took into account multiplicity in statistical analysis. This point must be discussed, and authors have to explain us how they dealt with it.

Power analysis
We agree with Reviewer Two’s analysis that it would take 514 persons per group (if each site were treated as a separate study) to achieve 0.8 power with difference of 9% (and reference proportion of 30%). With the sample sized specified it would be possible to estimate a hypothesised RR of 1.35 (difference of 10.5%). Personnel have changed during the project and we have been unable to confirm the details further provided in the manuscript – we apologise for this oversight and are grateful to the reviewer for their comments.

Based on the original ethics application and study design, the response rate was initially expected to be 50%, and it was expected to pool between study sites to test for the effect of fasting status and data collection type (in original ethics application). Assuming the 50% response rate and target power of 0.8, the number required per group (across sites) to detect a 7.6% absolute difference in response is 675. Across the two sites, 334 were recruited for a total of 668 (slightly less than 337.5 per area).
We propose a rewritten version of the details of the power analysis:

“We estimated we would need 668 participants in each treatment group across the area, based on an initially estimated response rate of 50% and difference in response rate of 7.6% between treatment groups for power of 0.796. We allocated 334 persons to each area and planned to pool across areas to compare treatment effects of data collection type (pathology vs. dedicated clinic).”

**Multiple outcome measurement**

Hypotheses about our main outcome (difference in response rate between data collection types) were specified prior to data collection and subsequent hypotheses about effect of collection mode on secondary outcomes (measurements and aliquot yield) were developed after the study had commenced but before data was available when these outcomes were identified as important components of the study. We have not completed any formal adjustment of p-values for multiple outcomes in our study (e.g. Bonferroni corrections) as our hypotheses were predetermined and because of the effect of Type II error. It would be preferable for readers to interpret any p-values with the number of tests made in mind for each type of outcome, so we propose adding this statement to the limitations of the study:

“We have not made a formally corrected (e.g. Bonferroni correction) for multiple testing (multiple treatment comparison and multiple outcomes) in our study. We have calculated five p-values for the main outcome (response rate), seven p-values for measurement differences, and 16 p-values for blood aliquot yield. The chance of a Type I error increases with the number of tests performed, so our results should be interpreted with the number of tests for each outcome in mind.”

2.13 **The result section should be reported in agreement with the objective hierarchy previously specified.**

We have now made the order of the objectives consistent throughout the paper (see response to comment 2.5).

In the end, I do think this paper is of interest, because of its originality and because of its relevance. However, it should be re-written to be more concise.