Author’s response to reviews

Title: Population risk factor estimates for abdominal aortic aneurysm from electronic medical records: a case control study

Authors:

Diane T Smelser (dtsmelser@geisinger.edu)
Gerard Tromp (gctromp@geisinger.edu)
James R Elmore (jelmore@geisinger.edu)
Helena Kuivaniemi (shkuivaniemi@geisinger.edu)
David P Franklin (dfranklin@geisinger.edu)
H Lester Kirchner (hlkirchner@geisinger.edu)
David J Carey (djc Carey@geisinger.edu)

Version: 3
Date: 14 October 2014

Author’s response to reviews: see over
Dear Dr. Shipley,

We would like to submit our revised manuscript entitled: “Population risk factor estimates for abdominal aortic aneurysm from electronic medical records: a case control study” by myself, G Tromp, JR Elmore, H Kuivaniemi, DP Franklin, HL Kirchner and DJ Carey, for publication in BMC Cardiovascular Disorders.

We are grateful for the two expert Reviewers for their insightful comments. Below we provide detailed responses to the specific comments.

**Reviewer 1:**

1. “The response from the authors is very good, and the revised manuscript is constructively altered.”

   **Response:** The reviewer commented on our response to the previous review, but did not suggest any further revisions.

**Reviewer 2:**

1. Abstract (& elsewhere): ‘protectively associated’ could easily be interpreted causally, because ‘protective’ is causal language. I still prefer positive and negative associations as terminology.

   **Response:** We agree that protectively associated can imply causality, so we have revised the manuscript and replaced ‘protectively’ with ‘negatively’.

2. Methods: It wasn’t clear to me about when the variables were measured. If we think of date of diagnosis as an index date for cases, was it required that you only use data for that subject on or before the index date? What is the corresponding index date for controls?

   **Response:** The index date for the controls was the date that the data pull occurred. Since AAA is a late onset disease, most patients were in their 70s or 80s so most comorbidities have already occurred. We clarified this in the Data Source section of the paper: “All diagnoses, laboratory measures and clinical
values from primary care and specialty clinic visits (as of the date of the data extraction) were extracted. Age was defined as the age at AAA diagnosis for cases, and age at data extraction for the controls.”

3. Methods: I am a little concerned about the meta-analysis step. My understanding is that variable selection is done separately for each bootstrap sample. That would suggest that for two different samples the selected model might have different variables. If you then want to use a meta-analysis technique to estimate a coefficient, you will be combining estimates from different models. In one model you might be conditioning on 5 variables, and in another you might be conditioning on 8, etc. The coefficients have different interpretations in each model (unless you assume all coefficients of variables not selected are exactly equal to 0). Some clarification here would be great.

**Response:** The Reviewer raises an important point. We revised the methodology for the bootstrap/bagging analysis. In order to correct the issue of the coefficients having different interpretations, we added an additional bootstrap with 14 fixed variables in each model, without stepwise elimination. For the most part the estimates remain almost unchanged as compared to the original method. The largest changes tend to be for variables retained less often, as the reviewer mentioned. We compared the new estimates to the old ones and also looked at it as a proportion of the estimate (similar to a Coefficient of Variation). In the 14 fixed variable result, 6 were biased in a positive direction, 5 were biased in a negative direction and 3 were neutral. This demonstrates that a) biases exist, b) they are relatively small, and c) overall the biases seem to counteract so that any given estimate might be slightly over- or under-estimated. None of the biases severely affect the estimates.

We revised the Statistical Analyses section of the paper “A final model was generated using variables that were consistently retained in most bootstrap iterations. A second set of 2,500 bootstrap data sets were generated and analyzed using logistic regression with the final model of a fixed number of variables (table 2), i.e., each bootstrap set was analyzed with the same model. Regression estimates were recorded for each iteration and the estimates aggregated using meta-analytic techniques (using random-effects weighting). Variables were ranked by how often they were retained in the model, and by the $P$ value, which was based on the mean z score weighted by the number of iterations the corresponding variable was included in the model. The 14 highest ranked variables were then fixed in a second bootstrap analysis (no stepwise elimination). We considered variables statistically significant at $P < 0.05$, two-sided.”

We also revised Figure 2 to reflect the estimates using the revised method.
We appreciate the thoughtful comments and suggestions that improved our manuscript, and look forward to hearing from you soon.

Yours Sincerely,

Diane T. Smelser, PhD
Research Scientist