Author's response to reviews

Title: Genetic analyses of smoking initiation, persistence, quantity, and age-at-onset of cigarette regular use in Brazilian families: the Baependi Heart Study

Authors:

  Andréa R. V. R. Horimoto (andreavrh@uol.com.br)
  Camila M. Oliveira (camilamacioliveira@gmail.com)
  Suely R. Giolo (giolo@ufpr.br)
  Júlia P. Soler (pavan@ime.usp.br)
  José E. Krieger (krieger@incor.usp.br)
  Alexandre C. Pereira (alexandre.pereira@incor.usp.br)

Version: 2 Date: 25 October 2011

Author's response to reviews: see over
Manuscript MS: 1031260017568451

“Genetic analyses of smoking initiation, persistence, quantity, and age-at-onset of cigarette use in Brazilian families: the Baependi Heart Study”

Dear Dr. Tim Sands
Executive Editor, BMC Medical Genetics

Please find attached the revised version of the manuscript MS: 1031260017568451. We appreciate the assistance of the reviewers, and we have made all suggested changes. Modifications incorporated in the revised version of the manuscript are in red text. The English editing was done by a fluent speaker. We hope this revised version has reached the high standards of BMC Medical Genetics and that it will be interesting for its readers.

Answers to referee 1

1. The variable "age" is never specifically defined. Is it age at onset, age at enrollment to the study, or age at a fixed time point?
   This becomes particularly clear in looking at the survival analysis for age at onset.
   a. If "age" = age at initiation, then the analysis is invalid. You cannot have the same quantity on both sides of a relationship.
   b. If "age" = age at some fixed time point, then the variable is equivalent to adding birth year to the model, and is a test for calendar time effects, e.g., that smoking started at an earlier age in prior times than it does today. If we assume that age is in years, then the HR of 1.04 suggests that older people are much more likely to start smoking at a young age.
c. If "age" = age at study enrollment, then the interpretation is a similar to b, assuming that the enrollment period is short compared to the age range. We would need this information to interpret the variables.

We reanalyzed the age-at-onset of cigarette use variable considering only the covariate sex in the model. All paragraphs related to this analysis (in Methods, Results and Discussion sections) were rewritten.

2. In the Cox model the variable "sex" was presumably deleted from the model when it was found to be "not significant". A fit that includes interactions without the main effects is, however, very sensitive to the variable coding. That is, using sex=0/1 versus sex=1/2 will give different answers. Please give details of how this was done.

We used 0 for males and 1 for females in this analysis. We also analyzed the data considering the coding 1 (males) and 2 (females) and the same results were obtained, as follow:

Model with polygenic variance component \( \sigma^2_g \)

(A) Using 0 = Male and 1 = Female

```r
> fit3 <- coxme(Surv(idini, cens1) ~ sex, random=~1|npac, varlist=kmat)
```

Null Integrated Penalized
Log-likelihood -4396.978 -4334.824 -4190.374
Penalized loglik: chisq= 413.21 on 244.98 degrees of freedom, p= 9.7e-11
Integrated loglik: chisq= 124.31 on 2 degrees of freedom, p= 0

Fixed effects: Surv(idini, cens1) ~ sex
coef exp(coef)   se(coef)     z p
sex -0.8409005  0.431322 0.09088413 -9.25 0

Random effects: ~1 | npac
Variance list: kmat
npac
Variance: 0.5766372

(B) Using 1 = Male and 2 = Female

```r
> fit3a <- coxme(Surv(idini, cens1) ~ sex1, random=~1|npac, varlist=kmat)
```

Null Integrated Penalized
Log-likelihood -4396.978 -4334.824 -4190.374
Penalized loglik: chisq= 413.21 on 244.98 degrees of freedom, p= 9.7e-11
Integrated loglik: chisq= 124.31 on 2 degrees of freedom, p= 0

Fixed effects: Surv(idini, cens1) ~ sex1
coef exp(coef)   se(coef)     z p
sex1 -0.8409005 0.431322 0.09088413 -9.25 0

Random effects: ~1 | npac
Variance list: kmat
 npac
Variance: 0.5766372

Model with polygenic $\sigma^2_g$, (npac1) and shared family $\sigma^2_f$, (npac2) variance components

(A) Using 0 = Male and 1 = Female

> fit3<-coxme(Surv(idini,cens1)~sex,random=~ 1|npac, varlist=c(kmat,fmat))

NULL Integrated Penalized
Log-likelihood -4396.978 -4331.504 -4211.972
Penalized loglik: chisq= 370.01 on 194.81 degrees of freedom, p= 5.4e-13
Integrated loglik: chisq= 130.95 on 3 degrees of freedom, p= 0

Fixed effects: Surv(idini, cens1) ~ sex
  coef exp(coef)   se(coef)     z p
sex -0.8411642 0.4312082 0.08876161 -9.48 0

Random effects: ~1 | npac
Variance list: c(kmat, fmat)
 npac1 npac2
Variance: 0.3729897 0.1750617

(B) Using 1 = Male and 2 = Female

> fit3a<-coxme(Surv(idini,cens1)~sex1,random=~ 1|npac, varlist=c(kmat,fmat))

NULL Integrated Penalized
Log-likelihood -4396.978 -4331.504 -4211.972
Penalized loglik: chisq= 370.01 on 194.81 degrees of freedom, p= 5.4e-13
Integrated loglik: chisq= 130.95 on 3 degrees of freedom, p= 0

Fixed effects: Surv(idini, cens1) ~ sex1
  coef exp(coef)   se(coef)     z p
sex1 -0.8411642 0.4312082 0.08876161 -9.48 0

Random effects: ~1 | npac
Variance list: c(kmat, fmat)
 npac1 npac2
Variance: 0.3729897 0.1750617

Model with polygenic variance component $\sigma^2_g$, stratified by sex: kmat.f = female (npac1)
and kmat.m = male (npac2)

(A) Using 0 = Male and 1 = Female

> fit3b<-coxme(Surv(idini,cens1)~sex,random=~ 1|npac, varlist=c(kmat,f, kmat.m))

NULL Integrated Penalized
Log-likelihood -4396.978 -4345.101 -4180.981
Penalized loglik: chisq= 431.99 on 289.22 degrees of freedom, p= 1e-07
Integrated loglik: chisq= 103.76 on 3 degrees of freedom, p= 0

Fixed effects: Surv(idini, cens1) ~ sex
  coef exp(coef)   se(coef)    z p
sex -0.842615 0.4312082 0.09917607 -8.5 0

Random effects: ~1 | npac
Variance list: c(kmat.f, kmat.m)
 npac1 npac2
Variance: 0.400883 0.2118333
(B) Using 1 = Male and 2 = Female

> fit3a<-coxme(Surv(idini,cens1)~sex1,random=~1|npac,varlist=c(kmat.f,kmat.m))

NULL Integrated Penalized
Log-likelihood  -4396.978  -4345.101  -4180.981
Penalized loglik: chisq=  431.99 on 289.22 degrees of freedom, p= 1e-07
Integrated loglik: chisq=  103.76 on 3 degrees of freedom, p= 0

Fixed effects: Surv(idini, cens1) ~ sex1
coef exp(coef)   se(coef)    z  p
sex1  -0.842615 0.4305831 0.09917607 -8.5 0

Random effects: ~1 | npac
Variance list: c(kmat.f, kmat.m)
   npac1 npac2
Variance: 0.400883 0.2118333

Interpretation of the results also requires knowledge of how the sex variable was coded. Does the HR of .98 mean females start out younger than males, or older?

Following our coding system:

HR female/male = 0.98

HR male/female = 1/0.98 = 1.02

Then male starts out smoking at younger age.

3. "The individual risks of age .....and 65% higher in males than the overall average risk ....". I found this sentence and subsequent ones very confusing. The std for males is sqrt(.25)= .5, exp(.5)= 1.68, exp(-.5)=.60. This says that the "average" familial effect is for their members to have a relative log-hazard that is .5 higher or lower than the population as a whole. (60% less or 68% more on the multiplicative scale).

The last paragraph of Results has been rewritten to improve its understanding.

I don't know where the last sentence comes from. Assuming additivity I get a variance for females if .31+.28, sqrt(.59)= .768, exp(.765)= 2.1 or double the risk.

The last sentence was wrong and it was rewritten.
4. I'd be surprised if the effects of age or sex satisfied proportional hazards. Did you look at this?

   We plotted Scheonfeld residuals graphics (below) and the results suggest no serious violation of the proportional hazards (slopes near to zero).

Model with the covariate sex:

![Scheonfeld residuals graphics](image)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>-0.008</td>
<td>0.0395</td>
</tr>
</tbody>
</table>

5. Minor comment: I would have found column percents more useful than row percents in table 1.

   Table 1 was modified to show column percents instead row percents.

Answers to referee 2

It is truly a missed opportunity that a heart study started in 2005-06 has such poor assessment of smoking behavior. I am willing to overlook that considering that many other studies also do not adequately assess smoking behavior, however it is up to the authors to exercise vigilance in what kind of smokers the smoking variables actually capture and how these smoking variables map onto the definitions of seemingly similar smoking variables in other studies.
We completely agree with the reviewer and have improved our assessment of smoking in the follow-up phase of our study.

1. It is difficult to believe that such a large proportion of the sample is non-smokers, especially in the 18-24 year-old group. Are these numbers consistent with the smoking prevalence in Brazil? Could there be some selection bias in this study? It is likely that individuals who self-identify as non-smokers do have some smoking experience, but likely never smoked on some sort of a regular basis to think of themselves as smokers.

The first paragraph of Background was updated to show the data of Vigitel Brazil 2010, recently published (reference 2 has also been updated). The first paragraph of Results was rewritten, and a new paragraph (the second in Results) was added to describe the sample characteristics, comparing them with Brazilian indexes. One paragraph was also added in the Discussion section to discuss the results showed in Table 1.

Therefore, the measure of “smoking initiation” likely does not reflect initiating smoking per se, but initiating some kind of regular smoking, perhaps daily smoking. There is substantial evidence that heritability is more pronounced at heavier levels of smoking, and considering the high heritability estimates for “smoking initiation” in this study, I would guess that this phenotype reflects a transition to regular or daily smoking, not initiation of smoking per se.

We rewrote one paragraph (the last in page 5) in Measures section to describe in details the questions and data collected.

Smoking initiation was analyzed as a dichotomous variable, contrasting ever versus never smokers, as also defined in other studies (Madden et al., 2004, per example).
2. Age of onset of cigarette use is actually age of onset of regular use. Again, these two definitions are very different and have different meanings. Please provide the means and range of the age of onset variable.

   Age-at-onset of cigarette use refers to regular use of tobacco. Again, the last paragraph in page 5 (Measures section) has been rewritten to detail the questions used to collect the data. In addition, the name of this variable was modified throughout the text for age-at-onset of cigarette regular use to avoid confounding.

   One sentence was added to last paragraph on page 9 to inform the mean and range of age-at-onset of cigarette use.

3. Please, remove all conclusions about lack of evidence of household effects.

   Age is the only household effect variable that was controlled for.

   We think that this analysis had not been clearly described in text in and thus we added the following paragraph on page 7 – Methods section:

   Household group analyses were also performed using the SOLAR system [13]. An additional variance parameter was added to model the effect of common environment, which is associated to any non-genetic factor shared between individuals living in the same household at the time of study. Using current residential addresses to define households, we have obtained 740 nuclear families from 95 families of the Baependi Heart Study. Household effects were investigated in both polygenic models (models I and II).

   Relatedly, please remove mention of the influence of additive genetic factors on the smoking phenotypes. Additive genetic effects cannot be estimated in family data. The authors need to make sure that it is clear that the heritability that is estimated in this study refers to broad sense heritability.
We removed the term “additive genetic effects” of the text. The sentence “All analyzed traits were significantly influenced by additive genetic factors in this population” was replaced by “All analyzed traits showed a significant familial aggregation in this population” on page 13 of the Discussion section.

4. Getting back to the point of non-specific smoking phenotypes, the results of this study need to be placed in the context of the specific smoking definitions of prior studies, not just based on how prior studies called their variables.

The last paragraph on page 13 of the Discussion section was rewritten. The difference in phenotype definitions were checked and added to the text.

5. It sounds like the cigarettes per day variable was collected as an open-ended question where no specific response categories were provided. In Table 1, this variable is shown in categories – were these categories generated by the authors and if so, what was the rationale for them? Please, provide the skew and kurtosis statistics for the cigarettes per day variable before and after normalization.

The quantity (number of cigarettes smoked per day) variable is a quantitative variable. There was no pre-defined criterion to selection of the categories showed in Table 1.

The skewness and kurtosis statistics, and qqnorm plots for quantity variable before (quantity_before) and after (quantity_after) natural log-transformation are showed below:

<table>
<thead>
<tr>
<th></th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantity_before</td>
<td>3690.204</td>
<td>194687.3</td>
</tr>
<tr>
<td>Quantity_after</td>
<td>-0.3881686</td>
<td>2.313992</td>
</tr>
</tbody>
</table>
One sentence was added in the first paragraph of the Methods - Statistical Analysis section (page 6) to provide skewness and kurtosis statistics for quantity variable after natural log-transformation.

Thank you very much for your assistance.

Best regards,

Alexandre C. Pereira, MD, PhD
Laboratory of Genetics and Molecular Cardiology
Heart Institute, University of São Paulo Medical School, São Paulo, Brazil.