Author’s response to reviews

Title: Polymorphisms in the Epidermal Growth Factor Receptor Gene and the Risk of Primary Lung Cancer: a case-control study

Authors:

Jin Eun Choi (ddochi74@hanmail.net)
Sun Ha Park (sunhapark@hanmail.net)
Kyung Mee Kim (96kmkim@hanmail.net)
Won Kee Lee (wonlee@knu.ac.kr)
Sin Kam (kamshin@knu.ac.kr)
Sung Ick Cha (sicha@knu.ac.kr)
Chang Ho Kim (kimch@knu.ac.kr)
Young Mo Kang (ymkang@knu.ac.kr)
Young-Chul Kim (kyc0923@chonnam.ac.kr)
Sung Beom Han (sbhan@dsmc.or.kr)
Tae Hoon Jung (thjung@knu.ac.kr)
Jae Yong Park (jaeyong@knu.ac.kr)

Version: 2 Date: 17 August 2007

Author’s response to reviews:

August 14, 2007
Editor-in-Chief
BMC Cancer

Dear Editor

Thank you for the review of our manuscript MS-1491149837146747, entitled “Polymorphisms in the Epidermal Growth Factor Receptor Gene and the Risk of Primary Lung Cancer: a case-control study.” We have revised our manuscript according the queries raised by the reviewers. Enclosed, please find the revised manuscript, in which all the comments of the reviewers have been addressed and the requests of the publisher have been answered.

A detailed response to each comment is attached to this letter.

Please do not hesitate to contact us if you have any questions regarding this manuscript. We again thank you and the reviewers for your consideration. We look forward to your reply.

Very sincerely yours,

Jae Yong Park, M.D., Ph.D.
Department of Internal Medicine, School of Medicine
In response to Reviewer 1

1. Provide genotype data by gender.
Reply: As requested, we have added genotype data by gender in Table 3.

2. What was the cut-off in age for younger and older individuals?
Reply: We dichotomized the subjects by median age. To clarify this point, we have edited the statistical section as follows:

In addition to the overall association analysis, we performed a stratified analysis by age (median age, < 62 years/≥62 years), gender, smoking status, cigarette exposure level (median pack-years of smoking in ever-smokers, <38 pack-years/≥38 pack-years), and tumor histology to further explore the association between EGFR genotypes/haplotypes and the risk of lung cancer in each stratum.

3. The results by smoking status and histological type are data driven since the case-control study was not designed to investigate these effects a priori. This issue has to be addressed.
Reply: As requested, we have mentioned this point in the Discussion section as follows:

Because this study was designed to evaluate the effects of EGFR polymorphisms on the risk of overall lung cancer, the stratification analyses according to age, gender, smoking status and tumor histology might have a type I error (due to multiple comparisons) and/or a type II error (due to the small number of subjects in the subgroups). Therefore, additional studies with larger sample sizes are required to confirm our findings.

4. The LD values for pairs of SNPs should be provided.
Reply: As requested, we have added the LD values for pairs of SNPs in Fig. 1. To describe the results, we have added the following sentences.

1) In the Statistical Analysis section:
The strength of LD between pairs of polymorphisms was measured by HaploView.

2) In the third paragraph of the Results Section
The five polymorphisms were not in strong LD (Fig. 1), and thus, established 29 out of the 32 (25) potential haplotypes.
5. In haplotype analysis, the authors should estimate for each haplotype frequencies in cases and controls, and then to find the P-value for the difference between estimated haplotypes in cases and controls.

Reply: As prompted by your suggestion, we estimated haplotype frequencies in cases and controls, separately, and then compared the frequency distributions between cases and controls. To clarify this point, we have edited the third paragraph of the Results section as follows:

We estimated the EGFR haplotypes of the 127378C>T, 142285G>A, 162093G>A, 181946C>T, and 187114T>C polymorphisms, and we compared their frequency distributions in the cases and controls.

We estimated the EGFR haplotypes of the 127378C>T, 142285G>A, 162093G>A, 181946C>T, and 187114T>C polymorphisms in the cases and controls, separately, and we compared their frequency distributions between the cases and controls.

6. OR can be calculated for combined genotypes.

Reply: As prompted by your request, we evaluated the combined effects of the SNPs by grouping the subjects based on the number of variant alleles each subject possessed. 1) When four SNPs (127378C>T, 142285G>A, 162093G>A and 187114T>C), which were not significantly associated with the risk of lung cancer in the logistic regression analysis for each SNP, were combined, no significant differences were observed in the frequency distributions of combined genotypes between the cases and controls, as shown below:

<table>
<thead>
<tr>
<th>Number of variant alleles</th>
<th>Cases (n=582)</th>
<th>Controls (n=582)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>77 (13.2)</td>
<td>71 (12.2)</td>
<td>0.43</td>
</tr>
<tr>
<td>1</td>
<td>88 (15.1)</td>
<td>109 (18.7)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>101 (17.4)</td>
<td>110 (18.9)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>138 (23.7)</td>
<td>130 (22.3)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>101 (17.4)</td>
<td>79 (13.6)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>42 (7.2)</td>
<td>43 (7.4)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>29 (5.0)</td>
<td>36 (6.2)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>15 (1.0)</td>
<td>4 (0.7)</td>
<td></td>
</tr>
</tbody>
</table>

2) When five SNPs, including the 181946C>T, were combined (for this analysis, we re-expressed the 181946C allele as a variant allele to facilitate comparison because the 181946T allele was protective against lung cancer in the individual logistic regression analysis), the frequency distribution of the combined genotypes in the cases was not significantly different from that in the controls, as shown below:

<table>
<thead>
<tr>
<th>Number of variant alleles</th>
<th>Cases (n=582)</th>
<th>Controls (n=582)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>34(5.8)</td>
<td>39(6.7)</td>
<td>0.43</td>
</tr>
<tr>
<td>1</td>
<td>49(8.4)</td>
<td>54(9.3)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>70(12.0)</td>
<td>80(13.8)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>84(14.4)</td>
<td>91(15.6)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>118(20.3)</td>
<td>118(20.3)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>88(15.1)</td>
<td>71(12.2)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>75(12.9)</td>
<td>65(11.2)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>32(5.5)</td>
<td>31(5.3)</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>29(5.0)</td>
<td>29(5.0)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3(0.5)</td>
<td>4(0.7)</td>
<td></td>
</tr>
</tbody>
</table>
If more than one of the polymorphisms evaluated in the study was functionally important, we would expect that the risk of lung cancer might be increased with the number of the variant risk alleles. Taking this into consideration, the results of the combined genotype analyses suggest that four other 4 SNPs might be not functionally important. We did not show the results of combined genotype analyses in the manuscript.

In response to Reviewer 2:

1. More detailed description on the case-control populations are needed, including participation rates, recruitment periods, and sites.

   Reply: In order to describe this matter, we have added the following sentences in the Study Population section as follows:

   This case-control study included 582 lung cancer patients and 582 healthy controls, and the details of the study population have been described previously (27-29). In brief, the eligible cases included all patients who were newly diagnosed with primary lung cancer between January 2001 and June 2002 at Kyungpook National University Hospital, Daegu, Republic of Korea. There were no age, gender, histological, or stage restrictions, but patients with a prior history of cancer were excluded from the study. The cases included 270 (46.4%) squamous cell carcinomas, 205 (35.2%) adenocarcinomas, 97 (16.7%) small cell carcinomas, and 10 (1.7%) large cell carcinomas.

   The control subjects were randomly selected from a pool of healthy volunteers who visited the general health check-up center at Kyungpook National University Hospital during the same period. The control subjects were frequency-matched (1:1) to the cancer cases based on gender and age (≥ 5 years).

   The control subjects were randomly selected from a pool of healthy volunteers who visited the general health check-up center at Kyungpook National University Hospital during the same period. A total of 3065 (1598 males and 1467 females) of 5578 healthy subjects agreed to participate in this study (participation rate, 54.9%). Compared with subjects that refused to participate, enrolled subjects showed similar sex (% of male, 52.5% versus 52.1%; P = 0.80) and age (52.2 ± 11.4 versus 52.1 ± 11.3; P = 0.80) distributions. From 3065 healthy volunteers, we randomly selected 582 control subjects that were frequency-matched (1:1) to the cases based on sex and age (≥ 5 years).

2. Why pack-year of smoking is treated as continuous variable in both main effect and gene-smoking interaction analysis? Have the authors perform stratified analysis among ever smokers?

   Reply: We had tested the gene-smoking interaction by stratification analysis and a logistic regression model that included the interaction term between genotype and smoking. In the regression model, smoking had been considered as both discrete and continuous variables as follows: i) pack-years of smoking, ii) square root of pack-years, iii) smoking status; never-/ever-smoker, and iv) smoking
exposure level; never-smoker/¿38 pack-years in ever-smokers/>38 pack-years in ever-smokers. Because the interaction term was not statistically significant in any of these models, we only presented the result when the continuous cumulative smoking dose (pack-years of smoking) was used in the analysis.

As requested, we have added the results of stratification analysis among ever-smokers in Table 4.

3. Why age or gender is not adjusted in Table 2? Is there any corrected p value for results in Table 2?

Reply: Since this study is age-gender 1:1 matched, there is no need for adjustment for age and gender in the investigation for overall cancer. Therefore, conditional logistic regression analysis was used to calculate odds ratios and 95% confidence intervals for overall lung cancer, with adjustment for pack-years of smoking.

To clarify this point, we have edited the statistical analysis section as follows:

Conditional logistic regression analysis was used to calculate odds ratios (ORs) and 95% confidence intervals (CIs) for overall lung cancer, with adjustment of pack-years of smoking (as a continuous variable). In addition to the overall association analysis, we performed a stratified analysis by age (median age, ¿ 62 years/>62 years), gender, smoking status, cigarette exposure level (median pack-years of smoking in ever-smokers, ¿38 pack-years/>38 pack-years), and tumor histology to further explore the association between EGFR genotypes/haplotypes and the risk of lung cancer in each stratum. The ORs and 95% CIs in the stratification analyses were calculated using unconditional logistic regression analysis, with adjustment for gender, age or pack-years of smoking, when appropriate.

4. Please present the result of the global test for the haplotype analysis.

Reply: As requested, we have presented the global P-value for the haplotype analysis in Table 5.

In response to Reviewer 3:

1. The authors first examined the frequencies of 39 candidate polymorphisms in the EGFR gene in 27 healthy Korean individuals. They indicated in the discussion that another 27 cases were also detected and those frequencies data could be added in Table 1. In addition, the detail information on the methods of examining the frequencies of 39 candidate polymorphisms needs to be provided in the Material and Methods section.

Reply: As requested, we have added the frequencies of 39 SNPs among 27 cases, and have described the methods in the Material and Methods section as follows:

Identification and selection of polymorphisms

Among the candidate polymorphisms in the EGFR gene, we initially captured 39 SNPs in the promoter region, all exons including intron-exon boundaries (10 bp of the introns on either side) and the 3¿-UTR of the gene because variants in
these regions are most likely to affect gene function (Table 1). We then examined the frequencies of the captured SNPs in a preliminary study that included 27 healthy controls and 27 lung cancer cases. Among the 39 captured SNPs, seven SNPs [127378C>T (rs2072454), 142232C>T (rs17336800), 142285G>A (rs11543848), 151904T>A (rs17290169), 162093G>A (rs10251977), 181946C>T (rs2293347), and 187114T>C (rs884225)] had minor allele frequencies greater than 10% in the 54 subjects. The 142232C>T and 151904T>A were completely or near completely linked with the 142285G>A. Thus five SNPs (127378C>T, 142285G>A, 162093G>A, 181946C>T and 187114T>C) were chosen for the association study.

2. MAF above 10% in 27 healthy Korean subjects was defined as the criteria of selecting the polymorphisms in this study. However, the polymorphism 162093G>A (rs10251977) was selected with MAF of 0.07 and the other two SNPs 142232 (rs17336800) and 151904T>A (rs17337023) with MAF above 0.10 were excluded from the subsequent analysis (Table 1). Please confirm and explain this point.

Reply: First of all, I apologize that there were some mistakes in the original manuscript. Please refer to the answer to the comment 1.

3. In discussion, the authors stressed that the SNP 162093G>A was the true locus associated with modulated the lung cancer risk, which seemed redundant and not significant, because lack of the biological evidence. The authors could not exclude the possibility of LD with other potentially functional polymorphisms (e.g. polymorphisms located in introns or flanking region, polymorphisms of insertion/deletion, and so on).

Reply: We totally agree with your comment. Thus, I have edited the paragraph as follows:

The mechanism underlying the association between the 181946C>T polymorphism and lung cancer risk remains to be elucidated. Because the 181946C>T polymorphism does not result in an amino acid change, nor does it reside within the functional domain, the observed effect of the 181946C>T polymorphism on lung cancer may be due to LD with other functional EGFR variant(s) that were not tested in this study. Therefore, additional studies are needed to detect the other functional variants in the EGFR gene and their associations with lung cancer.

4. In table 2, similar genotypes distributions were shown for 142285G>A, 181946C>T and 187114T>A, which indicated a possible LD, and the authors may try to recalculate the indicator of D prim and r square with another software (e.g. haploview).

Reply: As requested, we have recalculated LD strength using Haplovieview, and have presented the results in Fig. 1.

5. Brief description of characteristics (age, sex, smoking status and pack-years of smoking) among cases and controls was needed, though detailed information was shown in previous studies.
Reply: We have added Table 2 to present the clinical characteristics of the study population.

6. The ORs were just adjusted by pack-years of smoking in Table 2, but age was also adjusted in Table 3 and Table 4. This needs to be addressed.

Reply: Since this study is age-gender 1:1 matched, there is no need for adjustment for age and gender in the investigation for overall cancer. Therefore, conditional logistic regression analysis was used to calculate odds ratios and 95% confidence intervals for overall lung cancer, with adjustment for pack-years of smoking.

To clarify this point, we have edited the statistical analysis section as follows:
Conditional logistic regression analysis was used to calculate odds ratios (ORs) and 95% confidence intervals (CIs) for overall lung cancer, with adjustment for pack-years of smoking (as a continuous variable). In addition to the overall association analysis, we performed a stratified analysis by age (median age, ¿ 62 years/>62 years), gender, smoking status, cigarette exposure level (median pack-years of smoking in ever-smokers, ¿38 pack-years/>38 pack-years), and tumor histology to further explore the association between EGFR genotypes/haplotypes and the risk of lung cancer in each stratum. The ORs and 95% CIs in the stratification analyses were calculated using unconditional logistic regression analysis, with adjustment for gender, age or pack-years of smoking, when appropriate.

7. Other haplotypes with frequencies less than 5% were excluded in Table 4, which seemed not appropriate.

Reply: As requested, haplotypes with estimated frequencies < 5% were pooled into a single group and included in the haplotype analysis (Table 5).

8. Page 12 line 1: 181946 C-to-A should be 181946 C-to-T.

Reply: We corrected this mistake.

Thank you for your thoughtful, constructive comments.