## PEER REVIEW HISTORY



## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | Detection for pathway effect contributing to disease in systems epidemiology with case-control design |
| **AUTHORS** | Xue, Fuzho; Ji, Jiadong; Yuan, Zhongshang; Zhang, Xiaoshuai; Li, Fangyu; Xu, Jing; Liu, Ying; Li, Hongkai; Wang, Jia |

## VERSION 1 - REVIEW

| | |
|---|---|
| **REVIEWER** | Wei Chen<br>Tulane University, USA |
| **REVIEW RETURNED** | 06-Oct-2014 |

| | |
|---|---|
| **GENERAL COMMENTS** | This study presented the methodology of pathway effect on disease and its application to acute myeloid leukemia using a case-control study design. Omics is a state-of-the-art methodology of systems epidemiology. The method is interesting and its application to disease study is of importance. This research has merits; the application of method will add new dimensions to the literature and help move the scientific field forward. However, the presenting style of this paper needs to be improved.<br><br>Major comments<br><br>In the abstract, some key numbers should be presented in the “Results” section.<br><br>The Introduction section should be shortened. Some text can be moved to the Discussion section or Methods section.<br><br>The pathway effect was defined as product of multiple path parameters (standardized regression coefficients) in this study. In the case-control study of AML, the strength of the causal relationship was used as an exposure factor. How were the confounding factors adjusted for? In point of application to disease epidemiology study, the conventional association parameters like OR, 95%CI or p-values should be also presented in Table 2 on page 11.<br><br>The selection criteria, matching variables and disease status of 35 controls should be described. There were 98 patients in the case group. It is not a pair-matched case-control study. The frequency matching is supposed to be used in the design. Therefore, beta-D and beta-C are the mean betas in the case group and control groups, respectively, and the statistic, D (difference in betas between case and control groups), was used as an exposure factor. A significance test was performed to make decision on rejecting or accepting the null hypothesis. These points should be better addressed in “Methods”. As mentioned above, is it possible to use a traditional method like logistic regression to estimate ORs and other |

| | |
|---|---|
| | parameters? It would be easier to for readers to understand the method and its application.<br><br>The terminology “statistics” is somewhat confusing in this paper. It means path parameters in most cases in this study. It is suggested to figure out another terminology to avoid confusion, such as “parameters”, “exposure factors”, “indices” or “pathway effect measures” where appropriate.<br><br>The text related to simulation and power analyses should be shortened and save more space for application results.<br><br>How was the path “Treg→TGF-beta→Th17” chosen? Please describe whether the selection was based on physiological or metabolic pathways underlying AML or through multiple testing.<br><br>Minor comments<br><br>In the Abstract, page 2, line 26, “acute myeloid leukemia (AML)” was defined the second time in the abstract.<br><br>Page 4, line 52, instead of “and expected to”, “and are expected to” is correct.<br><br>Page 12, line 21, “acute myeloid leukemia (AML)”, AML was defined the second time.<br><br>Check the language throughout the manuscript. Pay attention to the usage of the article “a” or “the”, for example the sentences as below:<br>In the Abstract, page 2, lines 8-10, “at a (or the) human population level”<br>In the Abstract, page 2, line 13, “at a (or the) population level” and “in a great need”. |

| | |
|---|---|
| **REVIEWER** | Robin Haring<br>University Medicine Greifswald,<br>Germany |
| **REVIEW RETURNED** | 07-Oct-2014 |

| | |
|---|---|
| **GENERAL COMMENTS** | In their methodological paper “Detection for pathway effect contributing to disease in<br>system epidemiology” Jiadong J. et al. present a well-conducted study including state-of-the-art statistical analysis, a well-written manuscript & a sound discussion of their findings. Especially the real data simulation underlines the value of this paper, which I would like to recommend for publication in BMJ open.<br><br>Major comment: please include a table about the clinical characteristics of the patient-based sample: age, lab values, sex etc.; including a rationale for the choice of the applied study design (case-control) and the actual sample size (power calculations? Why N = 133?).<br><br>Minor comments: figure 2-4 need some more editing to harmonize fonts & style. |

| | |
|---|---|
| **REVIEWER** | Eiliv Lund<br>Instituteof community medicine<br>University of Tromsoe<br>Norway |
| **REVIEW RETURNED** | 22-Oct-2014 |

| | |
|---|---|
| **GENERAL COMMENTS** | This is an important methodological paper for the advancement of pathway analyses in systems epidemiology. The authors clearly argue the need for novel statistics compared to the well known methodologies of traditional survival analyses.<br><br>Some major comments:<br>1. The authors should carefully go through the manuscript and add necessary explanations so that the common epidemiologists could read the paper. As an example, page 6 second part 3) calculate z ;;; where -2zp could be difficult to understand.<br>2. Could you find a name or notation for your statistics?<br>3. Under Pathway effect p 5 you say that the distribution of the difference of pathway effect between cases and controls = D is unknown. What would be the distribution of D in the common matched analysis, the traditional nested case-control design, that is used in most genetic epidemiological research.<br>4. The paper should include an example with a theoretical larger pathway.<br>5. You write that the one possible drawback is the computational burden. At what level do you speak? PC, central university computerfacilities or bigdata problems.<br><br>There are several typing errors. |

| | |
|---|---|
| **REVIEWER** | Dr Victoria Allgar<br>University of York, England |
| **REVIEW RETURNED** | 03-Nov-2014 |

| | |
|---|---|
| **GENERAL COMMENTS** | This is a complex statistical paper, which might be more suited to a statistical journal, where the complex statistical analyses could be explored it full detail.<br><br>The practical application of the model would need a more detailed discussion, in similar terms, for a more generic reader to understand. |

## VERSION 1 – AUTHOR RESPONSE

Responses to Reviewer #1's Comments

We sincerely appreciate your effort in reviewing our manuscript. Your comments are very constructive and we have revised it accordingly. Responses to your specific comments are given below.

Major comments
1. In the abstract, some key numbers should be presented in the “Results” section.

Response: Thanks for pointing this out. We have done the modifications and added some key numbers according to your comments (see page 2, line 23-24).

2. The Introduction section should be shortened. Some text can be moved to the Discussion section or Methods section.

Response: Thanks for the suggestions. We have shortened the Introduction section and moved the text “On the other hand,…… in systems epidemiology data analysis.”(See page 4, line 10-28 in original manuscript) to the Discussion section (see page 13, line 16-34).

3. The pathway effect was defined as product of multiple path parameters (standardized regression coefficients) in this study. In the case-control study of AML, the strength of the causal relationship was used as an exposure factor. How were the confounding factors adjusted for? In point of application to disease epidemiology study, the conventional association parameters like OR, 95%CI or p-values should be also presented in Table 2 on page 11.

Response: Thanks for the valuable suggestion. The aim of our manuscript is to develop novel statistical method for detecting the pathway effect within a network between different disease status under case-control design, so the exposure unit is the pathway rather than one single factor. It is indeed important to get some association or effect parameters such as OR as you mentioned. However, unlike one single factor, it is extremely hard to define the pathway levels since one pathway usually refers to many factors with some specific topology structure. Therefore, we only put 95%CI or p-values in Table 2 of our original manuscript. Meanwhile, we have also added these comments to the Discussion section in our revision (see page 14, line 27-41).
As for the confounding factors you are concerned about, two traditional statistical procedures including stratification and regression modeling can be applied. In our method, we can use the latter to control the confounding through calculating the standardized partial regression coefficient to indicate the relationship between the factors locating on the same pathway.

4. The selection criteria, matching variables and disease status of 35 controls should be described. There were 98 patients in the case group. It is not a pair-matched case-control study. The frequency matching is supposed to be used in the design. Therefore, beta-D and beta-C are the mean betas in the case group and control groups, respectively, and the statistic, D (difference in betas between case and control groups), was used as an exposure factor. A significance test was performed to make decision on rejecting or accepting the null hypothesis. These points should be better addressed in “Methods”. As mentioned above, is it possible to use a traditional method like logistic regression to estimate ORs and other parameters? It would be easier to for readers to understand the method and its application.

Response: Thanks for your insightful comments. This unmatched case-control study is hospital-based. AML patients were diagnosed according to the French-American-British (FAB) classification system. Patients with hypertension, diabetes, cardiovascular diseases, chronic or active infection, or pregnant were excluded. Individuals with slight iron deficiency anemia, having no immunological changes, were used as controls. Because bone marrow aspiration is a quite invasive procedure, only 35 individuals were available. As you mentioned above, a significant test was performed to make decision on rejecting or accepting the null hypothesis, we have added some necessary details in methods to make it more clear (see page 7, line 8-20). Unlike traditional methods which attached little importance to the complex relationships (correlation and topological structure), we failed to obtain the OR-type parameters on the pathway level due to the impossibility to give the clear definition of pathway values (see our response to your comment 3).

5. The terminology “statistics” is somewhat confusing in this paper. It means path parameters in most cases in this study. It is suggested to figure out another terminology to avoid confusion, such as “parameters”, “exposure factors”, “indices” or “pathway effect measures” where appropriate.

Response: Thanks for your kind reminder. We preferred the term “pathway effect measures (PEM)” throughout the whole manuscript following your instructions.

6. The text related to simulation and power analyses should be shortened and save more space for application results.

Response: Thanks a lot. Simulation and power analyses are quite important to assess the performance of the proposed pathway effect measures (PEM). We have re-organized this section to make it more concise and readable in the revision.

7. How was the path “Treg→TGF-beta→Th17” chosen? Please describe whether the selection was based on physiological or metabolic pathways underlying AML or through multiple testing.

Response: The path “Treg→TGF-beta→Th17” was chosen based on physiological knowledge from clinicians. Some previous studies have focused on Treg, TGF-beta and Th17. Actually, in our original manuscript we have cited the related references [21, 22] and clarified “TGF-beta is a prerequisite for the induction of CD4+ T-cell Foxp3 expression and differentiation into Treg cells. Meanwhile, TGF-beta is also critical for human Th17 cells differentiation.” (see page 8, line 54).

Minor comments

In the Abstract, page 2, line 26, “acute myeloid leukemia (AML)” was defined the second time in the abstract.
Page 4, line 52, instead of “and expected to”, “and are expected to” is correct.
Page 12, line 21, “acute myeloid leukemia (AML)”, AML was defined the second time.
Check the language throughout the manuscript. Pay attention to the usage of the article “a” or “the”, for example the sentences as below:
In the Abstract, page 2, lines 8-10, “at a (or the) human population level”
In the Abstract, page 2, line 13, “at a (or the) population level” and “in a great need”.

Response: Thanks for your attention. All corresponding grammar and spelling errors have been corrected.

Responses to Reviewer #2's Comments

We sincerely appreciate your effort in reviewing our manuscript. Your comments are very constructive and we have revised it accordingly. Responses to your specific comments are given below.

Major comment: please include a table about the clinical characteristics of the patient-based sample: age, lab values, sex etc.; including a rationale for the choice of the applied study design (case-control) and the actual sample size (power calculations? Why N = 133?).

Response: Thanks for your valuable suggestions. Totally 133 subjects have been included in this hospital-based case-control study. AML patients were diagnosed according to the French-American-British (FAB) classification system. Patients with hypertension, diabetes, cardiovascular diseases, chronic or active infection, or pregnant were excluded. Individuals with slight iron deficiency anemia, having no immunological changes, were used as controls. Based on your advice, the clinical characteristics have been described and presented in Table 1 in our revision. On the other hand, our simulation showed that the Pathway Effect Measures could have considerate performance when the

sample size was more than 100 under the pathway length equal to 2.

Minor comments: figure 2-4 need some more editing to harmonize fonts & style.

Response: Thanks for your kind reminder. Figure 2-4 have been checked again and redrawn to harmonize fonts and style in the revision.

Responses to Reviewer #3's Comments

Thank you very much for your insightful comments. We have carefully revised the manuscript following your suggestions, which should have led to great improvement in its quality. In the following, we give our response one by one.

Some major comments:
1. The authors should carefully go through the manuscript and add necessary explanations so that the common epidemiologists could read the paper. As an example, page 6 second part 3) calculate z ;;; where -2zp could be difficult to understand.

Response: Thanks for your comments. We acknowledged that it was quite necessary to make common epidemiologists could read the paper. Since the aim of our manuscript is to develop novel statistical methods for detecting the pathway effect, it is inevitable to use some complex theoretical formulas. We have tried our best to add some detailed explanations and cited the related references [18] to make it more straight forward and easier to understand.

2. Could you find a name or notation for your statistics?

Response: Thanks. Another reviewer also questioned this. We used the term “pathway effect measures (PEM)” throughout the whole manuscript following your instructions.

3. Under Pathway effect p5 you say that the distribution of the difference of pathway effect between cases and controls = D is unknown. What would be the distribution of D in the common matched analysis, the traditional nested case-control design, that is used in most genetic epidemiological research.

Response: Thanks for pointing this out. The reason why the distribution of the difference of pathway effect between cases and controls = D is unknown, is that our proposed method is derived from the multiplying of some correlated standardized coefficients. It seems little correlated with the epidemiology design. Although our proposed Pathway Effect Measures can be extended to a matched and nested case-control design, the distribution of D is still hard to know. We have also added this comment to discussion (see page 14, line 7-18).

4. The paper should include an example with a theoretical larger pathway.

Response: Thanks a lot. We re-conducted the simulation under the pathway length 8 and 10 respectively with various sample sizes. The type I error rate was still stable, and the power trends were similar with that of smaller pathway. We have shown these results in the supplement (Figure S2).

5. You write that the one possible drawback is the computational burden. At what level do you speak? PC, central university computerfacilities or bigdata problems.

Response: Thanks a lot. We mean that the computational burden may appear when dealing with bigdata, since the bootstrap technique has been used in our proposed method.

There are several typing errors.

Response: We appreciate your effort in carefully reviewing of our manuscript. We have re-organized and carefully checked the manuscript to minimize the typing errors.

Responses to Reviewer #4's Comments

We sincerely appreciate your effort in reviewing our manuscript. Your comments are very constructive and we have revised it accordingly. Responses to your specific comments are given below.

Comment:
The practical application of the model would need a more detailed discussion, in similar terms, for a more generic reader to understand.

Response: Thanks for your comments. We acknowledged that it was quite necessary to make it more straight forward and easier to understand. Firstly, we have added some necessary details and explanations in methods to make our model much clearer (see page 7, line 8-20). Then, we have also re-organized the application section and added the corresponding clinical characteristics (see Table 1) in our revision to make it easier for generic readers to understand.

**VERSION 2 – REVIEW**

| | |
|---|---|
| **REVIEWER** | Wei Chen<br>Tulane University, USA |
| **REVIEW RETURNED** | 23-Nov-2014 |

| | |
|---|---|
| **GENERAL COMMENTS** | I am satisfied with the revision. |

| | |
|---|---|
| **REVIEWER** | Eiliv Lund<br>UiT The Arctic university of Tromsø, Norway |
| **REVIEW RETURNED** | 27-Nov-2014 |

| | |
|---|---|
| **GENERAL COMMENTS** | The Authors have adequately answered my comments, Eiliv Lund |

| | |
|---|---|
| **REVIEWER** | Victoria Allgar<br>University of York, England |
| **REVIEW RETURNED** | 16-Dec-2014 |

| | |
|---|---|
| **GENERAL COMMENTS** | The paper has been revised following the reviewers comment and is much improved. |