Reviewer’s report

Title: The Use of a Multiple Imputation Method to Investigate the Trends in Histologic Types of Lung Cancer in Songkhla Province, Thailand, 1989-2013

Version: 0 Date: 29 Jan 2016

Reviewer: Edmund Njagi

Reviewer's report:

This manuscript investigates the incidence of adenocarcinoma (ADCA) and squamous cell carcinoma (SCC) during the period 1989 - 2013, in Songkhla province of Thailand, using cancer registry data. The authors highlight unknown histology as a problem that needs to be adequately addressed in this process, and, by "equating" this to the missing data problem which has been given considerable attention in the statistical literature, they invoke multiple imputation as a possible way forward.

From a statistical perspective, it is commendable that this work clearly invests particular effort in addressing the above-mentioned problem. Indeed, this problem poses a risk of obtaining biased results and thus drawing invalid inferences in relation to the substantive question of interest, and, therefore, the importance of giving it careful attention cannot be overemphasized.

Overall, the manuscript has potential for publication, and, revision targeting accuracy of statistical content, further elucidation of specific critical statistical concepts, and clarity of presentation of the results, would benefit the manuscript greatly and make it exceedingly more understandable and insightful.

In particular:

1. The first sentence of paragraph 6 in the Background section presents an excellent opportunity to, before introducing MI, briefly introduce the various missing data mechanisms (MCAR, MAR, and MNAR), and to provide a brief reflection on both what making each of these assumptions about the missing histological diagnoses would mean to the researcher, and the implication that each assumption would have in analysing the data. This introduction and reflection would perhaps better prepare the "non-statistically" oriented reader before MI is mentioned.

2. In the same paragraph above, the statement "(MI) is a statistical method that can solve this limitation" would need rephrasing. This is to avoid creating the impression that MI actually solves the missing data problem. MI is a method used to try to achieve validity of analysis results under less restrictive (more realistic) missing data mechanisms; it would be important that this message is clearly communicated.
3. In relation to the same paragraph, the statement "by replacing unknown histologies with another one that is most likely" would also need rephrasing. The wording "another one" may create an impression that it is merely replacing a missing value with a single value, which is the definition of single imputation, as opposed to multiple imputation. It would be important to ensure that the message that MI involves imputing multiple times, properly reflecting the uncertainty of the unknown value, is communicated. Granted, one of the other sentences in the paragraph does bring out the correct message, but the statement above might best be rephrased for avoidance of confusion.

4. The statement "Thus, the technique can be used to estimate the proportion within a range of uncertainty of lung cancer cases with no information on histological type" may need some elucidation to ensure that the meaning is readily accessible to readers.

5. Overall, in relation to the paragraph mentioned in the 4 points above, it would perhaps be useful for the authors to base their definition of MI on some of the authoritative texts on the subject. A nice and comprehensive reference on MI would perhaps be the book "Multiple Imputation and Its Applications" by James Carpenter and Mike Kenward.

6. In the same paragraph above, there is the statement: "The method has been used by cancer registries in Australia to estimate …"

Going back to the article title, it reads: "The use of a multiple imputation method to investigate …"

Finally, in the last paragraph of the Background section, there is the statement: "The objective of this study is to estimate gender specific trends in incidence … using the multiple imputation method"

Looking at the three statements highlighted above, it may seem that MI is being presented as "an analysis method by itself". MI is a method for addressing missing data, a method that is usually used "alongside" the "main analysis method". In MI terminology, what I am calling the "main analysis method" here is referred to as the "substantive model".

To illustrate the point, consider the following example: suppose you have a response variable (say a continuous response variable), and a covariate (say a continuous covariate as well). Suppose you would like to analyse the relationship between the covariate and the response, using a regression model, and, unfortunately, the covariate has missing data.

If you perform your analysis, using MI to impute the missing data, you would say that you have analysed the relationship between the covariate and the response using regression analysis, while using MI to handle the missing data. If, to the contrary, you said that you have analysed the above relationship using MI, it would tend to be a little bit confusing.
In brief, the statements "The use of a multiple imputation method to investigate …", and "The objective of this study is to estimate gender specific trends in incidence … using the multiple imputation method" send a somehow confusing message regarding the place of MI in the analyses, while statements like "The use of a multiple imputation method in investigating …", or "The use of a multiple imputation method in the investigation of …", or perhaps "The investigation of the trends in histologic types …, using MI to handle missing data/missing histology" and "The objective of this study is to estimate gender specific trends in incidence … addressing missing data using the multiple imputation method" would perhaps send a clearer message as pertains the place of MI in analyses.

7. The model as spelt out on page 5 seems inadequate. On the right hand side, the "beta's" seem to be the various logit intercepts, while the parameter vector corresponding to the covariate vector "x" is missing. On the left hand side, spelling out the various probabilities as well as the link function would be useful. A useful reference in spelling out generalized linear model would perhaps be the book "Categorical Data Analysis" by Alan Agresti.

8. It would perhaps be insightful to readers if the authors could motivate their choice of the "two-step multiple imputation" approach. A reader may be interested in knowing what exactly would be the loss in imputing for both "unknown" and NSCC within one step.

9. What is meant by the statement: "… to illustrate only the effect of time on the probability of imputation"?

10. For the reader, how was reporting the geometric mean age at diagnosis better or perhaps more insightful than reporting the arithmetic mean age at diagnosis?

11. It would be good for the authors to clearly specify what they are testing using the p-value provided in Table 1.

12. Is the statement "there was a declining trend in other known histologic types throughout all periods" perhaps "too bold"?

13. "Older age, female sex and being diagnosed after the year 2000 were the three strongest predictors for ADCA". The correct interpretation/wording should be simply: "age, sex, and year of diagnosis were …" Alternatively, for instance, "males had significantly higher odds for ADCA as opposed to SCC". The book above by Agresti would perhaps be a good reference in terms of interpreting multicategory logit models.

14. "Showed that the probability of being SCC and ADCA varied with time." The book above by Agresti would perhaps be helpful in terms of the correct wording of the results.

15. Table 3: the column entitled SCC, with 1's throughout it unnecessary. It suffices to just mention that SCC is the baseline/reference.
16. The province of Songkhla contains multiple districts? In Table 3, "Other district (ref. =Muang") is provided. The other districts were grouped into one category, and compared with Muang? How was Muang chosen as the baseline category?

17. An important tool in exploring missing data is a logistic regression of the missing data indicators on the other available variables. This helps in exploring assumptions and formulating imputation models. Examples can be searched in the references mentioned above, and such an exploration would be insightful.

18. The following paragraph should perhaps be restructured and elaborated on, to help readers clearly understand the results presented in Table 4, especially how exactly the results were obtained: "With the polytomous regression model for prediction of the histological group in Table 3, the estimated number of cases with squamous cell carcinoma (SCC), adenocarcinoma (ADCA), and other specified cancer, the true numbers and percentages of the three cancers stratified by sex among patients with known histology are shown in Table 4."

19. Table 4: the number 319 under Male; is it supposed to be 219 as in Table 1?

20. The following statement perhaps needs some elucidation: "The table shows the similarity of distribution of the number of cases with known histology and those with unknown histology with Chi-square p-values of 0.235 and 0.968 for males and females, respectively".

21. The following paragraph contained in the Discussion needs rewriting to ensure accuracy, and the MI book suggested above, as well as other authoritative books written exclusively for missing data, or containing topics on missing data, e.g. "Models for Discrete Longitudinal Data" by Geert Molenberghs and Geert Verbeke, would definitely be helpful. Among the things that need to be noted is that MI has indeed been extended to MNAR situations. Additionally, it should be noted that missing data mechanism assumptions are essentially untestable; assumptions are made for a particular analysis at hand, and, the best that can be done is a sensitivity analysis:

"A common issue with multiple imputation methods is accuracy. It has been suggested that the data must be missing at random (MAR) [32]. One study demonstrated that the method works well when the percentage of missing values is between 10 and 60% [33]. However, these two studies were not based on categorical data. A simulation study on multiple imputation using a multinomial logistic regression model showed unbiased results when the values were missing at random [34]. However, missing at random is difficult to be prove."

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

Yes
Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Unable to assess

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

Yes

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests
I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal