Author’s response to reviews

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

Authors:

Hutcha Sriplung (hutchas@gmail.com)
Seesai Yeesoonsang (seesaiy@gmail.com)
Edward McNeil (edward.m@psu.ac.th)
Surichai Bilheem (surichai.bil@gmail.com)

Version: 1 Date: 09 Apr 2016

Author’s response to reviews:

Reviewer 1: accept without revision.

Reviewer 2:

Hutcha and colleagues explored the use of a multiple imputation Method to investigate the trends in histologic types of lung cancer in a Thailand population. It is an interesting study. I have some major concerns:

1. There were some predictor variables shown in Table 3, the author should add them into Table 1, do statistical analysis, and describe in the text.

   Already added.

2. In line 12 on page 6, the author mentioned the two p values 0.235 and 0.968, I can not get them in Table 4, the author need add the p values in Table 4.

   Added in the legend of Table 4.

3. In this study, sex, age, religion, other district, and year of diagnosis were included. Could the author explain the reason? How about the other factors related to lung cancer?

A paragraph of age-period-cohort model is added at the end of the ‘Methods’ section.

Age period cohort model
Age-period-cohort (APC) regression models were used to investigate the effect of age, calendar year and birth-cohort on the incidence of a cancer. We used the classical method which fits a log-linear model with a Poisson distribution to the observed data to estimate age, period and cohort effects in a multiplicative APC model as follows:

\[ \log[R(a,p)] = f(a) + g(p) + h(c), \]

where the expected log-incidence rates \( R(a,p) \) is assumed to be equal to a linear combination of effects that adjust for age \( a \), period \( p \) and birth-cohort \( c \), where \( c=p-a \). To address the non-identifiability problem of the APC models, two-effects models (age-period and age-cohort) were first chosen and the remaining effect (cohort or period) was then identified to the respective model’s residuals using natural splines to reduce random variation. These are referred to as the AP-C and AC-P models. The analysis of APC models was performed with the Epi package for R statistical software version 3.2.2.

A sentence is added in the paragraph 3 of the ‘Results’ section.

Thus, the predictory polytomous model for multiple imputation of unknown histology should include sex and others described above as all of the variables explain the distribution of the known histology cases and should, if the missing at random (MAR) assumption was the case.

4. Figure 1 and Figure 2 shown the trend of lung cancer incidence by major histologic type. The author also concluded that "a rapid increase in the incidence of ADCA was found while the incidence of SCC in males showed no significant change and it was declining in females."

How to test the different trend change using statistical methods.

A sentence is added in the paragraph 4 of the ‘Results’ section.

With age-period-cohort model, the overall drift (slope) of ADCA in males was 5.9% per year (95%CI: 5.1%, 6.7%), that of SCC in males was 0.6% per year (95%CI: -0.3%, 1.5%), that of ADCA in females was 6.1% per year 95%CI: 5.1%, 06%, 7.2%), and SCC in females was -2.1% (95%CI: -4.4%, 0.2%).

Reviewer 3:

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.

First of all, we would like to express our great thank to your invaluable comments. We try to revise the manuscript as suggested but we are sorry that our work done is not possible to satisfy a few of your comments.

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.

A phrase is added in the ‘Background’ section.

The missing data on histological diagnosis usually happens in deep organs such as lung, liver, and brain. It occurs in situations that the patients present in late stage and the aim of treatment by physicians is just for palliation, thus, complete investigation for specific histologic type is not necessary. This situation is still true nowadays in low and middle income countries (LMIC) such as Thailand. So, the missing data on the histology of lung cancer is correlated with the stage and clinical performance of the patients but not directly with the histology of the disease itself. Data are said to be missing at random (MAR) if the probability that a value is missing may depend on observed values in the data but not additionally on the missing value itself.

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.

Some sentences are added in the ‘Background’ section.

Usually the histological type of cancers in an organ is related to different risk factors, disease course and treatment, thus, the need of the incidence and trends in the cancer of specific
histologic type that are closer to the truth than that obtained by the ignorance of the unknown histology is necessary for disease control and medical care planning. MI is a means for representing uncertainty in missing data, by producing a distribution of plausible values for a missing variable in a record, given the values of that record’s non-missing covariates [13]. The process of running a multiple imputation is explained in the Methods section below.

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.

Same as the answer for comment No. 2.

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.

Correction is made in the ‘Background’ section.

Thus, the MI technique can be used to estimate the plausible proportion of histological type distribution within a cancer registry with a proportion of unknown histology so that the distribution and incidence rates related to the histology of lung cancer can be calculated with a high confidence of accuracy that it is useful for health policy planning.

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.

I modified it like this.

The missing data on specific histological diagnosis usually happens in deep organs such as lung, liver, and brain. It occurs in situations that the patients present in late stage and the aim of treatment by physicians is just for palliation, thus, complete investigation for specific histologic type is not necessary. This situation is still true nowadays in low and middle income countries (LMIC) such as Thailand. So, the missing data on the histology is correlated with the stage and clinical performance of the patients but not directly with the histology of the disease itself. Data are said to be missing at random (MAR) if the probability that a value is missing may depend on
observed values in the data but not additionally on the missing value itself [12,13]. (Ref. 13 is the 'Multiple Imputation and Its Applications'.)

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.

Thank you for mentioning the confusion. We imputed the distribution of the number of cases that would have occurred for the two main histologic types of lung cancer and computed the age-standardized incidence rates for that two main types.

Thus, the MI technique can be used to estimate the plausible proportion of histological type distribution within a cancer registry so that the distribution and the age-standardized incidence rates of the histologic type of lung cancer can be calculated with a high confidence of accuracy and are useful for health policy planning.

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.

The paragraph is modified as following.

A polytomous regression model is given by:

$$\log(\pi_j (X)/\pi_J (X)) = \alpha_j + \beta'_j \ X,$$

where $\alpha_j$ is a constant and $\beta'_j$ is a vector of regression coefficients of $X$ explanatory variables, for $j = 1, 2, \ldots, J - 1$.

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.

Thank you. This is a black box for those who don't know the mystery of lung pathology.

Since there were cases with true unknown histology, we had to distribute those cases into one of the four categories; SCC, ADCA, large cell carcinoma, other known histology and NSCC. The fact that NSCC is a vague terminology in lung cancer pathology that can fall into one of the 4 histologic types mentioned before, the second step of MI was needed to specify which histologic type in a patient was.

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

Sorry, it is our fault for writing that phrase. This is the correction.

the effect of time on the imputed number cases and the trends in incidence rate and to ignore the change in the age structure of the population.

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?

Thank you. I agree with the comment not to use the geometric mean.

The mean (sd) age at diagnosis was 64.4 (12.4) and 63.8 (14.4) years in males and females, respectively.

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

A sentence was added in the 'Results' section.
The difference in histologic type distribution, sex, district, and period at diagnosis were different between males and females.

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

Sorry, I just wanted to describe what the table showed.

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.

Thank you. I change the words according to your suggestion.

Age, sex, and year of diagnosis were the three strongest predictors for ADCA.

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.

Thank you. I modify the sentence a bit.

The predicting model in Table 3 showed that the probability of the missing data falling in to SCC and ADCA categories varied with time as time was used as one of the explanatory variables used to run the MI.

15. Table 3: the column entitled SCC, with 1's throughout it unnecessary. It suffices to just mention that SCC is the baseline/reference.

The column of SCC is deleted from Table 3 as suggested.

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?

Muang district is the old town of Songkhla. In Thailand, Muang or Wiang district is usually the old town in almost all provinces and is the first administrative code for all provinces.

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.

I agree with the idea. However, our initial thought was that we would get 100 sets of imputation and some patients assigned to a disease could be the same person and the regression results from
100 data sets, or even combining those 100 data sets to get a big samples with duplication of cases would give more confusion to the interpretation. So we decided to use a simple comparison of the two groups, known histology and unknown histology by comparing the distribution of the MI data sets by the histologic types with the data set of those with known histology.

With this decision, we did not have the complete set of the predictor variables attached with the results of the MI. We are sorry that we could not do what you suggested. Once we run the MI again to attach the MI results to the set of predictor variables, all the results after MI would change a bit but might be enough for us to revise all the tables and graphs again.

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."

Thank you for the suggestion. I rewrite the sentence and hope it sounds better.

The actual (non-imputed) numbers and percentages of cases with squamous cell carcinoma (SCC), adenocarcinoma (ADCA), and other specified cancer are shown in the first two columns of the Table 4. After running two-step 200 rounds of MI with the polytomous regression model as described in the Method section above for prediction of the histological group in Table 3, the distribution of the estimated numbers and percentages of the three major cancer groups stratified by sex among patients with unknown histology are shown in columns 3 and 4 of the Table 4. Now the distribution of the numbers and percentages in columns 3 and 4 had probability intervals obtained by the 200 rounds of imputation.

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

Thank you very much. It is a typing mistake.

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".

Related to the comment 17, we did want to show that even with MI, the distribution of the histologic types did not changed much as the percentage of the missing histologic type in our data set was only around 30%.

A statement is added.
This results suggested a little change in the distribution of histologic type from the original distribution after MI. However, it did significantly affect the trends in the number of cases and the incidence rates in terms of the absolute values as shown in Figures 1 and 2.

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."

It is rewritten but I am not sure the accuracy in terms of statistical point of view. I, the first author, is working on the cancer registry and am learning newly developed statistical techniques to analyze the registry data. So I tend to write the statistical concepts in a simple way for those who work in cancer registries understand how to get the results rather than explaining how those statistical methods work. So I am sorry for too simple expression of statistical concepts.