Author’s response to reviews

Title: Validation of the CancerMath prognostic tool for breast cancer in Southeast Asia

Authors:

Hui Miao (ephmh@nus.edu.sg)
Mikael Hartman (mikael_hartman@nuhs.edu.sg)
Helena Verkooijen (H.M.Verkooijen@umcutrecht.nl)
Nur Aishah Taib (nuraish@gmail.com)
Hoong-Seam Wong (whseam@yahoo.com)
Shridevi Subramaniam (shridevi@crc.gov.my)
Cheng-Har Yip (chenghar.yip@gmail.com)
Ern-Yu Tan (ern_yu_tan@ttsh.com.sg)
Patrick Chan (patrick_chan@ttsh.com.sg)
Soo-Chin Lee (csilsc@nus.edu.sg)
Nirmala Bhoo-Pathy (ovenjjay@gmail.com)

Version: 1 Date: 10 May 2016

Author’s response to reviews:

We would like to thank the reviewer for the useful comments, which have all been taken into account. Below you will find a listed response to the reviewer’s queries.

Reviewer #1: The objective of this manuscript is to compare the individual subject level nodal status and survival predictions from CancerMath for breast cancer cases with 7064 observed cases of stage I to III breast cancer in Singapore Malaysia Hospital-Based Breast Cancer Registry.
For either nodal status or survival, the data in question thus involve two separate lists (or distributions), one set of subject level predictions from CancerMath and another set of subject level observed values. What is of interest is the "match" between these two values with the criterion for "match" appropriately defined (here I am using the term "match" in a rather general sense). The interest is thus in the joint performance (in terms of matching) of the predicted and observed lists, rather than on the individual lists. Many of the "match" results described in the manuscript are however based on two distributions (predicted and observed) separately, rather than their joint distributions. For example "Nodal status calculator predicted 40.6% of patients to be node positive which was lower than the observed 43.6%", in fact, reports two numbers calculated separately from the CancerMath predicted distribution and observed values and does not consider the joint distribution. While this reported information is certainly of substantial value, it is nor based on "match" at the individual subject level. Similarly, "Cancermath predicted and observed overall survival probabilities were 87.3% vs 83.4% at 5 years after diagnosis and 75.3% vs 70.4% at 10 years after diagnosis" is again based on two separate sets of calculations, one using the CancerMath predicted distribution and the other using the observed values. On the other hand, the calibration plot and the ROC curves ("The calibration plot showed underestimation for most of the groups. The AUC was 0.71 (95% CI, 0.70-0.72)") are, in fact, based on the joint distributions of the predicted and the observed values and considers the "match" at the individual subject levels. I find these to be of more value and I feel that the manuscript needs to focus more on these results and report more results based on the joint distributions.

The reviewer raises an interesting and challenging question here. Comparison of predicted and observed values by clinically meaningful prognostic subgroups are commonly demonstrated in other validation studies of prognostic models such as Adjuvant! Online and PREDICT. We presented these results so that they can be compared across different studies and can be easily understood by clinicians who might use these models. The differences between "matched" results also provide important clinical information on which subgroup this model can be applied to. We understand that model discrimination and calibration provide quantitative and valuable measures of model performance in validation study of prognostic model. Therefore we have illustrated five calibration plots and reported AUCs using different length of follow-up time. Our discussion and conclusion is also mainly focused on discrimination and calibration of the model.

In usual statistical analysis, the prediction method/model is (trained) based on the observed data and hence the question of statistical variability comes from that aspect. In this case however, the prediction model of CancerMath pre-exists and is not based on the observed data and hence there is no source of statistical variability from that aspect. The authors have been careful in their statistical analysis from this respect and I congratulate the authors.
We thank the reviewer for this comment and compliment.

In terms of prediction, I would think CancerMath predicts a probability (such as 30%) of being node positive for each subject (and similarly a probability (such as 70%) of 5 year survival and another probability of 10 year survival and so on). Thus, the cancerMath predictions of Node positive for all the subjects is list of such predicted probabilities (based on the individual CancerMath inputs for each subject). The reported 40.6% (prediction of node positive) is the median of these individual predicted probabilities. While the authors have been quite careful in describing the 40.6% as "median predicted probability" within the manuscript, I feel that the statement in the abstract of "Nodal status calculator predicted 40.6% of patients to be node positive" is NOT appropriate. The issue with the 40.6% goes actually a bit beyond. I feel that even if it is correctly described as the "median predicted probability" in the manuscript, a probable common interpretation would be "Nodal status calculator predicted 40.6% of patients to be node positive". As illustrated, even (some of) the authors interpreted it that way.

We agree with this comment and would like to thank the reviewer for pointing it out. We have changed our abstract to “the median predicted probability of positive lymph nodes is 40.6%” accordingly.

The other issue with the 40.6% vs the observed 43.6% is that these are based on the predicted and observed distributions separately and does not provide us with any idea on the sensitivity and specificity. In fact, equally important here are the positive and negative predictive values. The issue with these is that a cutoff on the predictive probabilities is needed to evaluate this. But I feel that this is the question of real interest and the authors need to report this based on a cutoff (such as default of 0.5) and also should show the ROC curve obtained by changing the cutoff.

Again, the reviewer raises an interesting question here. We think that a cut-off value may not be appropriate as CancerMath is a prognostic models rather than a diagnostic model. Clinically, diagnostic model is used for classification of current disease state which is a single binary value; the ROC curve demonstrates the trade-off between sensitivity and specificity at each possible cut-off and can be used in assigning the optimal cut-off. Whereas prognostic models are used for different clinical purposes. It incorporate the dimension of time (time-to-event data) and predict future risk of an event. It is often used for determining levels of risk, such as high, intermediate and low to aid treatment decision (not provided by CancerMath). Therefore we think discrimination and calibration are more appropriate performance measures for CancerMath.
Another general comment is that the authors report "median of the predicted probabilities" and median of the predicted survivals. I feel that from statistical viewpoint, the mean (of the predicted probabilities, survivals) is more appropriate. For example, I would think that the idea of the calibration plot is taken from the Hosmer-Lemeshow goodness of fit test and the mean is actually used in that test.

Thank you for the suggestion. We have updated our calibration plots and method section and used mean of the predicted probabilities and mean of the predicted survivals instead. The results are similar and do not change the conclusion.

Details are needed for the calibration plot in figure 2. How is the observed survival calculated. Is it the Kaplan-Meier estimate of 5 year survival for the subgroup of patients in that decile group? As before, I suggest to use average predicted survival rather than the median.

We thank the reviewer for this comment. As mentioned by the reviewer, the observed survival in each decile indeed is calculated using Kaplan-Meier estimates. We have added this in the method section. We also updated our calibration plot by using mean (average) predicted survival instead of median.

In Table 2, the n=4517. Are these who were not censored by 5 years? Are the observed and predicted number of deaths out of these subgroup only? Same question for Table 3.

Vital status was updated on 1st March 2013 for UMMC, 31st July 2013 for NUH, and 1st October 2012 for TTSH. For 5-year survival (table 2), we only included UMMC and NUH patients diagnosed in 2007 and earlier and TTSH patient diagnosed in 2006 and earlier, as patients diagnosed afterwards did not have follow up for 5 years. Similarly only UMMC and NUH cases diagnosed in 2002 and earlier were included for 10-year survival (table 3) as cases diagnosed after 2002 did not complete 10 years of follow up. Patients who were lost to follow-up, but diagnosed before the cut-off date as stated earlier were included and we use date of last contact as the endpoint. In fact there was no individual whose length of follow-up was shorter than 5 years in table 2 subset and shorter than 10 years in table 3 subset. The observed and predicted number of deaths are calculated for these two subsets.
I did not follow how the C-statistics is calculated "Outcome calculator was further evaluated using concordance statistics (C-statistics) for the entire dataset regardless of follow-up time"?

AUC can be only calculated when the outcome of interest is dichotomous and is interpreted as the probability of assigning a higher score or predicted risk to a randomly selected individual with the outcome than to another randomly selected individual without the outcome. Therefore, when outcome is time-to-event data, we have to specific the length of follow up (5-year and 10-year) to dichotomize survival outcome. However our sample size was reduced because we have to exclude recent cases who did not have 5-year follow-up as mentioned in earlier comment. An adapted method was developed by Harrel et al [ref: Evaluating the yield of medical tests. JAMA 1982;247:2543; Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med, 1996. 15(4): p. 361-87.] so that all individuals with time-to-event data can be included. The concordance statistic (C-statistic) is calculated by considering all possible pairs of patients, at least one of whom has died. If the predicted survival is larger for the patient who lived longer, the predictions for that pair are said to be concordant with the outcomes. Same as AUC, a C-statistic of 0.5 indicates no discrimination and value of 1.0 means perfect discrimination.