Peer Review File

Article information: https://dx.doi.org/10.21037/tcr-23-1825

Reviewer A

Comment: *This study aimed to identify clinicopathologic prognostic variables and construct a nomogram for the prediction of overall survival in young breast cancer patients with regional lymph node metastases. However, there have been similar studies on young women using the same SEER database in the past. In subgroup analyses with lymph node metastasis, which is a clear risk factor, there is little new clinical impact.*

Reply: Thank you for your feedback and for pointing out the existence of prior studies focusing on young breast cancer patients in the SEER database. While we acknowledge the presence of similar research in this specific subgroup, we believe that our study makes a valuable contribution to the field of breast cancer prognostic research. Our study aims to construct a nomogram for the prediction of overall survival, which, to our knowledge, has not been specifically developed for young breast cancer patients with regional lymph node metastases using the SEER database. By doing so, we seek to provide a tool that can assist clinicians in individualized prognostic assessment and treatment decision-making for this particular patient population. Furthermore, we recognize the importance of considering the clinical impact of our findings, especially in light of existing research. We will ensure that our study emphasizes any novel insights or additional understanding that it brings to the field, thereby contributing to the broader knowledge base on prognostic variables in young breast cancer patients with regional lymph node metastases.

We appreciate your perspective and will carefully address the need to highlight the distinctiveness and clinical relevance of our study within the context of existing literature on this topic.

Reviewer B

Comment: *The article is clear, the study design is well-defined, and the findings address the main research question. Furthermore, the topic is of great relevance and is well-supported in the article.*

Reply: It's great to hear that the article you're referring to has clear study design, well-defined findings that address the main research question, and is well-supported, particularly considering the relevance of the topic. It's important to recognize high-quality research and the significance of its findings in contributing to the respective field.

If you have any specific questions or if there's anything else you'd like to discuss about this article or any related topic, feel free to share, and I'd be happy to provide further assistance or discuss related matters.

Reviewer C

Comment 1: *There are 1000+ papers proposing nomograms in breast cancer.*

Needs a mini-review of what has already been presented, and discuss what is the novelty of the present submission.

Reply 1: We sincerely appreciate the valuable comment. Nomograms in breast cancer are statistical tools used for prognosis, prediction, and decision-making. They utilize multiple factors, such as patient characteristics, tumor characteristics, and treatment variables, to provide personalized estimates for outcomes such as survival, recurrence, and response to therapy. These nomograms have gained attention due to their ability to provide individualized risk assessments and assist clinicians in making informed treatment decisions. Some common themes addressed in previous papers include the development and validation of nomograms for various breast cancer subtypes, the incorporation of novel biomarkers and imaging techniques into nomograms, and the evaluation of nomogram performance in different patient populations. To the best of our knowledge, our study is the first population-based comprehensive retrospective study focus on YBC patients with regional lymph node metastasis.

Changes in the text: (see Page 4, line 30 and Page 5, line 1-9).

Comment 2: *Abstract: should specify how "young" was defined. In the text, mentions <=40 years, but paper lacks justification of the 40 years cutoff.*

Reply 2: Thank you a lot for reminded us of this important point. We added the definition of "young" in the abstract and provided detailed explanation of young breast cancer in our revised manuscript.

Changes in the text: (see Page 1, line 23 and Page 4, line 1-6).

Comment 3: *Training and validation set: what decided the choice of the size of split? How different partitions would have affected the results? Random split: what was used to assess robustness?*

Reply 3: Thank you very much indeed for your comments. The choice of the size of the training and validation split is often determined by various factors, such as the amount of available data, the complexity of the model, and the desired level of accuracy. A common practice is to use a 7:3 or 8:2 split for training and validation, respectively. However, the specific split size can vary depending on the study design and data characteristics. In our study, using the *caret* package in R, the eligible participants were randomly split into a training set and a validation set, conforming to a frequently-used 7:3 ratio. Different partitions of the data can affect the results of the study, as they may lead to different model performance estimates and generalizations. To minimize the potential impact of partitioning on the results, we can perform multiple splits, with different random seeds, and report the average or median performance across the splits to obtain more robust estimates. To assess the robustness of the model to random splits, we can perform repeated random splits using different random seeds and report the variance or standard deviation of the performance metrics across the splits.

Comment 4: *Methods use the prediction of 2, 3 and 5 years OS. This is cumbersome*

and moreover makes the assumption that the 2, 3 and 5 years should be considered as distinct separate outcomes. 1) If the proportional hazards assumption holds, then a single OS outcome would be sufficiently representative, e.g. median OS. 2) But if the proportional hazards do not hold, then the validity of the Cox models (from which the nomograms were constructed) is questionable.Use of the restricted mean survival time would be preferable than the above pointwise OS.

Reply 4: Thank you very much indeed for your constructive and insightful comments. You have raised some important points about the use of pointwise overall survival (OS) predictions at 2, 3, and 5 years versus alternative approaches. If the proportional hazards assumption holds, using a single OS outcome, such as median OS, can indeed be sufficient to represent the overall survival. This assumption implies that the hazard ratio remains constant over time. In this case, the median OS can provide a concise summary of the survival distribution. However, it is essential to assess the proportional hazards assumption critically before making this choice.

If the proportional hazards assumption does not hold, the validity of the Cox models used to construct the nomograms may be questionable. In such cases, using pointwise OS predictions at different time points might not fully capture the dynamic nature of the survival process. Alternative approaches should be considered to account for time-dependent effects properly.

One preferable alternative you mentioned is the use of the restricted mean survival time (RMST). RMST provides an overarching measure of survival over a specified time horizon, considering the area under the survival curve up to that time point. Unlike pointwise predictions, RMST captures the entire survival experience and is less sensitive to the specific time points chosen. When using RMST, one can summarize the survival experience in a single value for the desired time horizon. It offers a more comprehensive and interpretable measure of the overall survival rather than distinct separate outcomes at specific time points.

However, after careful consideration, we decided to keep using the pointwise OS for following reason. We aimed to construct a novel nomogram to estimate individualized risk based on patient and disease characteristics. To use the nomogram, each level of variables was assigned a specific point on the scale. By summing the points from each variable, a total point was obtained for the individual patients. We can then predict 2, 3 and 5 year OS probability by projecting the total points to the total score scale of the nomogram. While using the RMST is a preferable alternative for longer follow-up periods, it might not be as essential for shorter durations. In our case, given the relatively short follow-up, pointwise OS predictions could still provide valuable information about survival at those specific time points.

Comment 5: *The follow-up is quite very short, 41 months, less than 4 years!*

Reply 5: Thank you for this great comment! Apologies for any confusion caused. If the average follow-up period is only 41 months (less than 4 years), it is indeed

a relatively short duration. Our study is clearly a retrospective cohort study that is inevitably biased by patients' selection in the SEER database. Since the SEER database has collected distant metastatic sites and immunohistochemical data since 2010, which is the basis for dividing breast cancer into different molecular subtypes, young breast cancer patients with regional lymph node metastasis from January 2010 and December 2015 were searched according to the AJCC 7th edition TNM staging system in the SEER database. In such cases, using pointwise overall survival (OS) predictions at 2, 3, and 5 years may still be valid and practical. These time points can provide meaningful information about survival at specific landmarks within the follow-up period. Given the relatively short average follow-up, We decided to add this inevitable limitation in our discussion section.
Changes in the text: (see Page 13, line 12-14).

Comment 6: *SEER release version full citation should be specified as recommended by SEER.*
Reply 6: We apologize for any previous oversight. When referencing data or information from the Surveillance, Epidemiology, and End Results (SEER) program, it is indeed important to provide the full citation in accordance with SEER's recommendations. To ensure accuracy, we referred to the SEER program's official website for the most up-to-date citation guidelines. Thank you a lot for reminded us of this important point.
Changes in the text: (see Page 5, line 22-23).

Comment 7: *Should clarify if T, N were clinical or pathological classification. T-stage and N-stage are very coarse categorizations, paper lacks consideration of number of nodes, lymph node ratio or tumor size.*
Reply 7: Thank you for bringing that to our attention. The distinction between clinical and pathological classification is crucial, as it impacts the accuracy and interpretation of the findings. In the SEER database, derived AJCC T stage and N stage typically refers to the pathological staging, which is determined based on the pathological analysis results of the surgically resected tissue. We admitted that T-stage and N-stage are very coarse categorizations in our study, whereas, they are most commonly used clinically. Your insightful point about the importance of considering the number of nodes, lymph node ratio, and tumor size is well taken. Previous study found that high nodal tumor burden (>2 positive lymph nodes) were more likely to occur in young women diagnosed with breast cancer. We will take this into consideration for a more comprehensive analysis in our future research. Thank you for your valuable input.
Changes in the text: (see Page 7, line 2-4).

Comment 8: *Age reaches significance in multivariate model of Table 2. But: 1) no clarification how the variable was coded. 2) if age was categorized according to a cutoff (median), should present the analysis justifying that the categorized cutoff is*

better than age as a continuous variable. 3) if continuous, should verify the linearity (for example https://pubmed.ncbi.nlm.nih.gov/16212670/), and if needed consider fractional polynomials, for example https://pubmed.ncbi.nlm.nih.gov/21953493/ or https://pubmed.ncbi.nlm.nih.gov/35382744/ .

Reply 8: Thank you for pointing out the importance of clarifying the coding of the age variable in our univariate and multivariate model. According to Date Description for SEER Research and Research Plus (https://seer.cancer.gov/data-software/documentation/seerstat/nov2021/#ss-variables), this data item represents the age of the patient at diagnosis for this cancer. The code is three digits and represents the patient' s actual age in years up to age 84. Age 85 and over are grouped as 85+. Age 100 and over are grouped as 100+. Thus, age was utilized as a continuous variable in our analysis and expressed as medians with quartiles. After univariate and multivariate cox regression analyses, we found that age was not ultimately regarded as independent predicted variables of OS in our target population (P=0.052). We agree on the significance of assessing linearity and considering alternative modeling approaches. We will conduct the suggested linearity assessment and explore methodologies such as fractional polynomials in our future further study.

We appreciate your valuable insights and will incorporate these critical methodological considerations into our future study to ensure the robustness and clarity of our findings. Thank you for bringing these important points to our attention.

Changes in the text: (see Page 6, line 19).