ABS ANNALS OF BREAST SURGERY
AN OPEN ACCESS JOURNAL TO BRIDGE BREAST SURGEONS ACROSS THE WORLD

Peer Review File

Article information: https://dx.doi.org/10.21037/abs-22-31

Reviewer A

Thank you for submitting this interesting article on the use of artificial intelligence in the management of breast cancer.
I would like to make some considerations
Comment 1. Most of the readers are clinicians so they are not familiar with much of the terminology used such as machine learning (ML), a brief description in the introduction would be welcome. The same happens with the various techniques that can be used in the ML expressed in the results or table number 2.
Reply 1. The following was added to the Introduction, Changes "Machine Learning (ML) is a type of Artificial Intelligence that uses computer algorithms and statistical models that transform and analyze datasets for the purpose of discovering new relationships. Some early examples of ML include Linear and Logistic regression, and support vector machines that uncover the maximum separation between groups." (lines 60-63). Also, the paragraph was changed and reference listed for further explanation of ML, "A variety of ML decision trees are often used in medical differential diagnoses (12)." (line 116).

Comment 2. I think the use of low/high adherence can be confusing. In the work presented here, it can be deduced that the delay in diagnosis and treatment is due more to the greater difficulty of access to medical or health resources than to non-compliance with scheduled appointments. Most patients with high ADI rates and low "adherence" have no private insurance, are black, have more advanced tumors, and consequently more mastectomies. Therefore, we could conclude that more than patients with low adherence, they would be patients with a delay in diagnosis and treatment due to their socioeconomic conditions. On the other hand, I totally agree with you on the need to include new variables in the studies, such as the ADI, as well as gender, age, ethnicity, etc.
Reply 2. In this study we sought to determine factors that lead to delay in definitive surgery and chose to use the terms adherence versus non-adherence to differentiate the groups as we were also exploring time horizons. Reviewer A has correctly interpreted our findings, in that all those factors (as we report in the results section and further discuss in the 1st paragraph of the Discussion (See lines 180-188). We also agree with the reviewer that our use of low/high adherence was confusing. We have changed all references from high/low adherence to adherent or non-adherent respectively, including column headers of table 1.

Comment 3. The World Health Organization (WHO) defines adherence to treatment as the correct follow-up of the prescribed treatment and the degree of compliance with scheduled appointments.
Reply 3. This is exactly why we chose to use the terminology, as there are many appointments between diagnosis and making it to the operating room. The purpose of

our research was to be innovative in that the past literature on adherence was covered in our text, whereas our findings were specific to the social determinants of health that decrease adherence. While we considered using data points such as arrived appointments, we ultimately decided that would not substantially contribute to the knowledge base. Therefore, we don't think it necessary to change our paper with regard to this comment (See lines 81-85). We will however likely used arrived and canceled appointments as measures in subsequent intervention trials. See also response to Reviewer A comment 2 above.

Reviewer B

This is quite interesting paper but I'd like to make some recommendation for the authors:
Comment 1. The goal and purpose of this paper are not clearly stated. Please make a correction. Reply 1. The following was reworded to clarify goal and purpose, "In our own practice, we noticed delays in care for some of our patients that seemed more related to non-medical factors. Thus, we are uniquely positioned to explore whether distinctive traits of our breast cancer patients such as their genetics, physical characteristics, and health habits. The goal of the present study was to investigate discrepancies that exist in social determinants of health (SDOH) at the urban core that might provide insight into how locally derived intervention may benefit our patients and other regional medical systems. The overall al purpose was to do the exploratory analysis of a ML model developed to predict adherence for underserved women newly diagnosed with breast cancer." (lines 76-83).

Comment 2. Discussion is focused on ML application in medicine that was already proven once and again. In my opinion is not necessary and seems irrelevant to this study, but it may be that this Journal's viewership needs this explanation. This should be reviewed by Editor.
Reply 2. We have followed this reviewer's advice and the text and references 28 and 29 were deleted from the Discussion section. Should the editor decide it provides clarity or is necessary we can easily add it back to the manuscript.

Comment 3. Conclusions and strengths are not clearly stated. Please rephrase.
Reply 3. The Conclusion was rephrased as, "allowing intervention at an earlier period to mitigate the delays. When applied widely with interventions, (perhaps by a nurse navigator), it has the potential to reduce significant health disparities in one of the country's most health disparate counties." (lines 283-285).

Comment 4. (line54) "insurance status differences" - such statement is strictly USA related, the article will be presented to global audience so it is advised to generalize and there should be explicitly explained.
Reply 4. Changed "the presence or absence of health insurance coverage" (line 54-55)

Comment 5. (line58) citation 6 is from 2012, more recent reference should be used.
Reply 5. Dankwa-Mullan et al., 2021 citation added (line 59)

Comment 6. (line59&65) abrv. ML is first used in line 59 but is expanded in line 65. It is common practice to show full text of abbreviation at first use.
Reply 6. The order was reversed (lines 60-62)

Comment 7. (line69') "In Florida,[...]" is should be: the state of Florida in USA.
Reply 7. Changed (line 73)

Comment 8. (line69") "higher for Whites than Black" I'd recommend rephrase, e.g. "related to ethnicity"
Reply 8. Changed phase to “related to race.” Hispanic vs non-Hispanic is ethnicity (lines 73, 74)

Comment 9. (line 72) "unique characteristics" this statement is not clear, please expand/explain or rephrase.
Reply 9. Phase deleted and traits expanded to, “Thus, we are uniquely positioned to explore whether distinctive traits of our breast cancer patients such as their genetics, physical characteristics, and health habits.” (lines 78, 79)

Comment 10. (line102&103) 6 is not a median in range 1-10. I'd change to "above average". Reply 10. Changed to “above average” (line 112-114)

Comment 11. (line106) why only gradient boosted trees method is explained?
Reply 11. This was deleted. The paragraph was changed and reference listed for further explanation of ML, “A variety of ML decision trees are often used in medical differential diagnoses (12).” (line 116).

Comment 12. (line112) used libraries needs citations.
Reply 12. References were added for libraries (line 123).

Comment 13. (line128') "neural networks" are mentioned in text but there are none in results – Table 2.
Reply 13. Entire paragraph was deleted from text.

Comment 14. (line128") Only one kernel for SVM was tested. Why this one? The other could possibly give better results. Please explain.
Reply 14. While a radial or Gaussian kernel could have also been chosen, we decided to use a linear kernel.

Comment 15. (line150) "before optimization" optimization is mentioned in text but not further explained. How was it performed, how much it improved the results?
Reply 15. The sentence, “Several rounds of model selection and optimization were conducted amongst the different models, for the different time horizons for surgery, and to optimize the parameters/hyperparameters of each model.” was added (lines 163-165).

Comment 16. (line151) Why adaboost was selected the best? Based on which evaluation metrics? Please explain. F1 scores were the metrics?

Reply 16. The section was rephrased, “The AdaBoost model using a time horizon of 110 days had an acceptable sensitivity and specificity with a before and after optimization AUC and F1 of 0.799 and 0.855 and 0.820 and 0.856 respectively, and was retained as the classification model. A comparison of the various ML classification techniques are presented in Table 2 in order of the metrics AUC, F1, and degree of accuracy and precision.” (lines 165-169).

Comment 17. (line152) "tuned-up AdaBoost" how was it tuned-up? please explain.

Reply 17. The writer of this section used "tune up" and "optimization" interchangeably, so for clarity only the term “optimization” was used (line 166). There are several rounds of model selection/optimization or tuning-up: 1) picking from amongst the different models, 2) picking from amongst the different time horizons for delayed surgery (i.e., how we arrived at 110 days), and 3) tuning the parameters/ hyperparameters of each model (i.e., optimization), (See No.15).

Comment 18. (line154) why are you giving F1score before optimization, please change to after or give both.

Reply 18. We have revised this paragraph and now optimization is before and after AUC and F1score. The sentence was rewritten as, “The AdaBoost model using a time horizon of 110 days had an acceptable sensitivity and specificity with a before and after optimization AUC and F1 of 0.799 and 0.855 and 0.820 and 0.856 respectively, and was retained as the classification model.” (lines 165-167).

Comment 19. The optimal time-horizon of 110 days should be explained in more detail.

Reply 19. The following phrase was added in the Procedures to the sentence, “to determine the optimal time horizon to maximize accuracy and precision between the groups in the ML model.” (lines 130- 132) and added to the Discussion section was, “In our analysis, the optimal time horizon for surgery was 110 days. This is 20% longer than the 90 day cut off described by Ho et al. (17) who found that outcome was no worse in patients who had delayed treatment of greater than 90 days post-diagnosis based on tumor stage.” (lines 189-191) and “As illustrated in Figure 2, the delay in adherence increased through the first 60 days after diagnosis and after that the gap never closes..” (lines 201-202).

Comment 20. (line198-199) Reliable conclusion that should be given more focus.

Reply 20. The Conclusion was rewritten in its entirety, see other Reviewer comments. (lines 279-287).

Comment 21. Please improve the clarity of Table1. Headers without data should be emphasized either with bold or larger margin or aligning.

Reply 21. Margins of Table 1 were realigned to improve clarity.

Comment 22. There are missing data in Table2 in AUC in lines 417&418. Please

explain or correct.

Reply 22. For a Support Vector Machine or in a Ridge Regression Classifier the output are binary and it is not possible to draw a ROC curve unless you vary the learning thresholds and hyperparameters for these two models. The software libraries that support these two programming methods automatically optimize for the learning threshold/hyperparameters to give you the optimal classifier that fits the best model, so they cannot be made into a continuous graph which would allow calculation of the area under the curve (AUC). Scores other than the AUC/ROC are used for comparison or a transformation that includes the other point in the optimization process that would be included. Graphs can be made by transforming the data into a logistic classifier and varying the hyperparameters, but that just adds extra work without improving upon the original answer.

Comment 23. Please improve the clarity of Table 2. It is unclear is it sorted by any column. Best value in column should be marked (preferably with bold).

Reply 23. The AdaBoost technique results were bolded in Table 2.

Reviewer C

This study uses machine learning models to predict adherence to timely treatment for patients with breast cancer at the UF Health hospital in Jacksonville, Florida. The study design and methods are very interesting. It demonstrates the importance of non-medical factors (eg, ADI) in a cancer treatment outcome compared to medical factors.

Major concerns:

Comment 1. Stage is an important factor for patients receiving timely treatment. I don't think you can impute 37% missing data of stage without corrupting the results. Instead, missing stage itself is of interest. Exploring the reasons behind the missingness is also necessary. Are patients with missing stage more likely to receive treatment or delay their treatment? Are patients from high ADI areas more likely to have missing stage? These are important questions behind data quality and disparity. I recommend including and excluding missing stage in sensitivity analyses.

Reply 1. After careful review, it was discovered that the 37% missing data was a carry-over from a previous draft. The Strengths and Limitations was entirely rewritten to state, “For this study, we gathered extensive EHR and tumor registry data from our diverse, predominantly underserved patient population, which creates a rich foundation for exploration of potential systematic barriers faced by women. However, missing data for specific variables in the EHR and tumor registry data are present, which thwarts further investigation into this area (29, 30). By combining tumor registry with EHR data, we can improve the reliability and validity of the data for analysis. It is difficult to collect all data without missing some information, as was the case in our study. The problem of linking databases has resulted in low overall match rates due, in part, to different variable definitions and missing data within each database (31). Ways to statistically account for missing data, such as omitting all cases with missing information or analyzing them as a separate group, may produce biased results (32). At the same time, the convention of imputation of all cases with unknown stage proportionally to the known stages increase the probability of

mistakenly assuming that the stage distribution of the unknown and observed stages are equivalent." (lines 259-271)." We do agree with this reviewer that missing stage may be another factor and will incorporate that into future analyses.

Comment 2. It is arguable whether all women with stage 0 breast cancer should receive surgery. I recommend including a statement explaining the inclusion stage 0 patients or performing a sensitivity analysis by excluding stage 0 patients. Standard of care for stage 0 Breast cancer is surgery.
Reply 2. The sentence, "Women with Stage 0 breast cancer were included in the present study as standard of care for Stage 0 breast cancer remains surgery." was added (lines 94-96).

Minor issues:
Comment 3. Line 141. "The median age of the participants was 59 years". Was the median age at the time of diagnosis?
Reply 3. Yes, age at diagnosis. Text added, also see Table 1 (line 90).

Comment 4. What is the geographic scale of the area deprivation index? I believe it is block group level if you used the data from the University of Wisconsin-Neighborhood Atlas.
Reply 4. The geographic scale was the census tract block group level (lines 113-115).

Comment 5. Line 148. "Figure 2 depicts the adherence over non-adherence ratio…" Maybe I am wrong but is the adherence over non-adherence ratio a measure of odds? The value of odds could be greater than 1, but in Figure 2, the maximum value of the Y axis is 1. This is misleading in that it could be interpreted as the percentage of patients receiving treatment over time (cumulative). I suggest describing this measure clearer and including that description in the figure title or note.
Reply 5. "Adherence (ratio)" does not mean adherence over non-adherence, but adherence over both populations (adherence + non-adherence), which is why the upper limit of Y-axis is 1. My guess is that you didn't mean to talk anything about adherence/non-adherence ratio at all in your manuscript. The phrase, "adherence over both populations (adherence + non-adherence)," was added to the text. (line 161).

Comment 6. Line 162. Patients received lumpectomy, mastectomy, or no treatment. Thus, mastectomy is a highly confounding variable with lumpectomy. I don't think you can say "while lumpectomy remained in the model, mastectomy was not a significant factor" because if you take lumpectomy out of the model, then mastectomy may become an important variable.
Reply 6. The last sentence of that paragraph, "Moreover, while lumpectomy remained in the model, mastectomy was not a significant factor." was deleted. We suspect type of surgery (lumpectomy, not mastectomy) remained in the model because it is an outpatient procedure and rarely involves coordinating with plastic/reconstructive surgery as might be needed in a patient undergoing mastectomy.

Comment 7. Line 174. “Ho et al. (18) found that outcome was no worse in patients who had delayed treatment of greater than 90 days post-diagnosis based on tumor stage.” Line 177. “Delayed first treatment of greater than 90 days post diagnosis was associated with worse outcome in patients with invasive non-metastatic and metastatic breast cancer.” The above two descriptions are contradictory to each other.
Reply 7. Added “On the other hand,” to clarify difference between 30 day and 90 day delays in the text. (line 193).

Comment 8. Line 215. “Data-driven prediction models, however, often mistakenly draw causal effects without the necessary parameters or their predictions.” The models themselves don’t draw causal effects. It is human interpretation that mistakenly draws causal effects from these models.
Reply 8. Changed to “mistakenly misinterpreted as having causal effects” (lines 234, 235)

Reviewer D

Overall, this is an important issue that has clinical, public health and psychosocial implications. The paper is relatively transparent in its method but would benefit from consideration of the following:

Comment 1. Lines 87-101: the paper focuses on BR outcomes but the preamble is too long perhaps and could be compressed into a single paragraph.
Reply 1. Deleted from the Introduction, “Outcome studies that utilized ML have demonstrated exceptional performance in the prediction for a wide range of conditions and situations from the probability of survival in an intensive care setting, assistance with clinical decisions, to large population studies for the prevention of suicide (7-11).”

Comment 2. From line 103: the paper suggests that little is known about BR outcomes and geography which tallies with the available evidence. However, some contextual anchoring of the research using what is known about BC, BR and access and outcomes more broadly would be useful. Linking the research described briefly to the broader research base on geography, urban-rural divide in cancer treatment and management more generally would be useful.
Reply 2. The urban-rural divide was addressed and the reference Zipkin et al. 2022 added. (lines 213-217).

Comment 3. Lines 110-112: not sure these are needed - they seem to repeat earlier content.
Reply 3. This sentence was edited to, “and later refined to produce national percentiles and state-based deciles.” (lines 108, 109).

Comment 4. Line 116 Study Design: were BRACA patients included? Also, there is a tendency here and throughout to refer to variables 'included' when for clarity stating which ones were used would be preferable (e.g. Patient Characteristics line 138-148 is

an example of this).

Reply 4. BRCA positive patients were not included in the study. The following was added to the Participant section, “The CPT codes 77066 and 77065 or ICD-9-CM 174.* and ICD-10-CM C50.* were used to identify participants who met inclusion criteria.” (lines 93, 94). The variables, age at diagnosis, cancer stage at diagnosis, race, insurance type at diagnosis, type of surgery, and Area Deprivation Index were written out in the text. (lines 103, 104).”

Comment 5. Line 154: were readmissions included as a separate complication?

Reply 5. No, readmission was not a separate complication.

Comment 6. Line 164-166: the use f 1.5 year cut-off for inclusion could have removed patients with lengthy BR complications. How many patients were excluded using this heuristic? Characteristics compared to included in ways that are important clinically?

Reply 6. We limited our inclusion dates to the 1.5 year cut off because we didn’t want the data to be skewed by a few patient outliers who might not have gotten their surgery until long after 18 months. Additionally, patients with BR (assuming this is breast reconstruction) complications would have been included because their index surgery would have been prior to the 1.5 years. The following sentence was added, “Patient surgeries that were delayed beyond 18 months were not in the analysis to avoid skewing the data with a few outliers.” (lines 132, 133).

Comment 7. Lines 169-171: other variables known to be linked with clinical outcomes such as SES, education, marital status, social support, other medical or even psychiatric diagnoses were not included. Were these available? Also, there is no indication as to whether those included and their analyzed data were first diagnoses or second, recurrences and so forth.

Reply 7. The ADI scale score clearly was defined as a composite measure of 17 census variables designed to describe socioeconomic disadvantage based on income, education, household characteristics, and housing and others. Therefore, we did not deem it necessary to repeat the same information separately as this would confound the findings more than add to understanding. (lines 106-108).

Comment 8. Lines 243-248: there are other data that might be available that could have been useful to support the claims being made (or alluded to) about distance to travel. The issue here is one of causality or probably causal direction and process. For example, if distance per se is a barrier to positive BR outcomes and distance limits access because of 'travel' (as implied) then were there data on number of appointments missed or rearranged by patients because of geographic distance to travel? This is linked to a broader point about the messaging in the discussion (e.g. lines 296-300 suggests that alerting patients to risk of not attending follow-up appointments will help remedy the disparity between far and not far groups BR outcomes and by implication that the differences between the groups lie in factors within the patients (e.g. even aspects of individual differences such as Type D behaviour or health beliefs). This whole issue of the underlying processes that could

link variables is not considered. It is alluded to as 'complex' but the conclusions imply otherwise. A little consideration (in the form of caveats for example) might help nuancing the conclusions and mitigate any simplifications.

Reply 8. Agreed. The following was added to the Limitations section, “There are likely other unaccounted for underlying processes that may contribute to non-adherence. Anecdotally, we can describe missed appointments and patient reported transportation barriers, however, the current study was unable to account for these patient issues. Collecting data on such factors may help elucidate links to SDOH resulting in non-adherence to timely surgery.” (lines 272-276).

Comment 9. Lines 330-335: the previous issue is also reflected in the conclusion - there is little indication of how alerting plastic surgeons to these 'associations' could lead to useful clinical or related changes. Perhaps some consideration of how the service design and delivery might also contribute to the disparity between groups. In other words the paper implies causal conclusions when the design, data and analysis do not enable this. Some critical consideration of this, however brief, would help clinicians use of the findings as signpost rather than as conclusive evidence of where the issues lie.

Reply 9. The Conclusion was rewritten to, “This lends supports for the necessity to better understand the relation between SDOH and care received by surgery patients, thus allowing intervention at an earlier period to mitigate the delays.” (lines 281-283).