Supplementary Table 2: Result metrics explained

|  |  |
| --- | --- |
| **Metric** | **Explanation** |
| AUC | The ability of the algorithm to discriminate between two classes ranging from 0 to 1. Excellent performance of a model has an AUC near of 1.0, which means it has a good measure of separability. |
| - ROC | Is a graphical plot illustrating the sensitivity as a function of “1 – specificity” in a binary outcome with varying thresholds. The area under the curve corresponds to how well the algorithm discriminates between the binary outcome, which ranges from 0.5 (no better than a coin-toss) to 1.0 (perfect discrimination). |
| - PRC | Is a graphical plot of recall (sensitivity) and precision. This is helpful in illustrating the algorithm performance on imbalanced datasets (outcomes that occur less than 1% in the dataset). A high number (close to 1.0) corresponds with high recall and high precision. |
| Brier | Computes the average mean squared difference between the predicted and observed outcomes; it ranges from 0 (excellent prediction) to 1 (worst prediction). |
| Calibration | Compares the observed to the predicted proportion of outcomes. A perfect model has a calibration slope of 1 and calibration intercept of 0. |
| Intercept | Indicates the extent that predictions are systematically too low or too high: negative calibration intercept suggests overestimation and a positive intercept suggests underestimation of the outcome. |
| Slope | Indicates the level of overfitting (>1) or underfitting (<1). For example, a slope of <1 indicates that predicted risks are on average too high for patients with high probabilities and too low for patients with low probabilities. |
| Sensitivity | Proportion of multiple bone metastases *classified* cases among the total of multiple bone metastases cases, ranging from 0% to a 100%. |
| Specificity | Proportion of single bone metastasis *classified* cases among the total of single bone metastasis cases, ranging from 0% to a 100%. |
| Negative predictive value | Proportion of single bone metastasis cases among the total number of single bone metastasis *classified* cases, ranging from 0% to a 100%. |
| Positive predictive value | Proportion of multiple bone metastases cases among the total number of single bone metastases classified cases, ranging from 0% to a 100%. |
| F1-score | Measures the overall accuracy of the algorithm, which ranges between 0 (total failure of an algorithm) and 1 (perfect algorithm). In other words, it is the harmonic mean of precision and recall. |
| LLR (+) | Indicates the *increase* in probability of multiple bone metastases if the algorithm is positive, ranging from 0 to infinity. |
| LLR (-) | Indicates the *decrease* in probability of multiple bone metastases if the algorithm is positive, ranging from 0 to 1. |
| Threshold | The threshold chosen by users may depend on the specific task at hand. If the goal is to identify all patients who had a multiple bone metastasis for building a cohort of only multiple bone metastases, the NLP algorithm may be used for initial screening followed by targeted review of those patients. On the other hand, if the goal is to identify specifically the patients with a multiple bone metastasis, the higher threshold may be appropriate. Regardless, no CPT codes are available for this clinical feature. So even at the high specificity threshold of 0.9 for the NLP algorithm, the algorithm provides better sensitivity. |
| AUC: area under the curve; PRC: precision-recall curve; ROC: receiver operating curve; LLR+: positive likelihood ratio; LLR-: negative likelihood ratio. | |