Predicting Lymph Node Metastasis in Non-small Cell Lung Cancer: Prospective External and Temporal Validation of the HAL and HOMER Models

Gabriela Martinez-Zayas
Francisco A. Almeida
Lonny Yarmus
Daniel Steinfort
Donald R. Lazarus

See next page for additional authors

Follow this and additional works at: https://scholarlycommons.henryford.com/pulmonary_articles
Authors
Predicting Lymph Node Metastasis in Non-small Cell Lung Cancer
Propective External and Temporal Validation of the HAL and HOMER Models

Gabriela Martinez-Zayas, MD; Francisco A. Almeida, MD; Lonny Yarmus, DO; Daniel Steinfort, PhD; Donald R. Lazarus, MD; Michael J. Simoff, MD; Timothy Saettele, MD; Septimiu Murgu, MD; Tarek Dammad, MD; D. Kevin Duong, DO; Lakshmi Mudambi, MD; Joshua J. Filner, MD; Sofia Molina, MD; Carlos Aravena, MD; Jeffrey Thiboutot, MD; Asha Bonney, MD; Adriana M. Rueda, MD; Labib G. Debiane, MD; D. Kyle Hogarth, MD; Harmeet Bedi, MD; Mark Deffebach, MD; Ala-Eddin S. Sagar, MD; Joseph Cicenia, MD; Diana H. Yu, MD; Avi Cohen, MD; Laura Frye, MD; Horiana B. Grosu, MD, MPH; Thomas Gildea, MD; David Feller-Kopman, MD; Roberto F. Casal, MD; Michael Machuzak, MD; Muhammad H. Arain, MBBS; Sonali Sethi, MD; George A. Eapen, MD; Louis Lam, MD; Carlos A. Jimenez, MD; Manuel Ribeiro, MD; Laila Z. Noor, MD; Atul Mehta, MD; Juhee Song, PhD; Humberto Choi, MD; Junsheng Ma, PhD; Liang Li, PhD; and David E. Ost, MD, MPH

BACKGROUND: Two models, the Help with the Assessment of Adenopathy in Lung cancer (HAL) and Help with Oncologic Mediastinal Evaluation for Radiation (HOMER), were recently developed to estimate the probability of nodal disease in patients with non-small cell lung cancer (NSCLC) as determined by endobronchial ultrasound-transbronchial needle aspiration (EBUS-TBNA). The objective of this study was to prospectively externally validate both models at multiple centers.

RESEARCH QUESTION: Are the HAL and HOMER models valid across multiple centers?

STUDY DESIGN AND METHODS: This multicenter prospective observational cohort study enrolled consecutive patients with PET-CT clinical-radiographic stages T1-3, N0-3, M0 NSCLC undergoing EBUS-TBNA staging. HOMER was used to predict the probability of N0 vs N1 vs N2 or N3 (N2j3) disease, and HAL was used to predict the probability of N2j3 (vs N0 or N1) disease. Model discrimination was assessed using the area under the receiver operating characteristics curve (ROC-AUC), and calibration was assessed using the Brier score, calibration plots, and the Hosmer-Lemeshow test.

RESULTS: Thirteen centers enrolled 1,799 patients. HAL and HOMER demonstrated good discrimination: HAL ROC-AUC = 0.873 (95%CI, 0.856-0.891) and HOMER ROC-AUC = 0.837 (95%CI, 0.814-0.859) for predicting N1 disease or higher (N1j2j3) and 0.876 (95%CI, 0.855-0.897) for predicting N2j3 disease. Brier scores were 0.117 and 0.349, respectively. Calibration plots demonstrated good calibration for both models. For HAL, the difference
between forecast and observed probability of N2|3 disease was +0.012; for HOMER, the difference for N1|2|3 was −0.018 and for N2|3 was +0.002. The Hosmer-Lemeshow test was significant for both models ($P = .034$ and .002), indicating a small but statistically significant calibration error.

**INTERPRETATION:** HAL and HOMER demonstrated good discrimination and calibration in multiple centers. Although calibration error was present, the magnitude of the error is small, such that the models are informative.

**KEY WORDS:** endobronchial ultrasound; lung cancer; lung cancer staging; mediastinal adenopathy

Correct mediastinal staging is critical for decision-making in non-small cell lung cancer (NSCLC) because treatment options and prognosis vary according to stage. Surgical resection with mediastinal lymph node dissection is warranted for patients without nodal involvement (N0) and for select patients with ipsilateral hilar nodal involvement (T1-2, N1). For patients who cannot tolerate surgery because of comorbidities and those who refuse surgery, stereotactic ablative radiotherapy (SABR) may be an option, provided there is no mediastinal or hilar node involvement. Once mediastinal lymph nodes are involved, multimodality treatment with chemotherapy, radiation, or targeted therapy is usual.

Because lymph node involvement markedly impacts treatment decisions, the diagnostic approach to NSCLC revolves around assessment of the probability of lymph node involvement. Current American College of Chest Physicians (CHEST) and National Comprehensive Cancer Network (NCCN) guidelines suggest using PET-CT to look for evidence of lymph node metastasis. In patients with a sufficiently high probability of nodal metastasis, mediastinal lymph node sampling with a needle technique such as endobronchial ultrasound guided-transbronchial needle aspiration (EBUS-TBNA) is warranted. However, determining the probability of nodal disease is difficult because PET-CT, while useful, has limited sensitivity and specificity for nodal involvement.

The CHEST and NCCN guidelines have suggested using prediction models to estimate the probability of cancer to manage solitary pulmonary nodules (SPNs). Similarly, models exist to estimate the probability of cancer to manage solitary pulmonary nodules (SPNs).

The Help with the Assessment of Adenopathy in Lung cancer (HAL) and Help with Oncologic Mediastinal Evaluation for Radiation (HOMER) models were developed retrospectively and externally validated on three retrospective independent datasets. HAL is a binary logistic regression model, which predicts the probability of N0 or N1 disease (prN0|1) vs the probability of N2 or N3 disease (prN2|3) as determined by EBUS-TBNA. HOMER is an ordinal logistic regression model that predicts prN0 vs prN1...
Take-home Points

**Study Question:** Are the HAL and HOMER models for predicting malignant mediastinal lymph node involvement valid across multiple centers?

**Results:** HAL and HOMER demonstrated good discrimination (ROC ≥ 0.83) in 1,799 patients across 13 centers, and they demonstrated good calibration between forecast and observed probabilities of nodal disease (difference, +0.012 for HAL and −0.018 and +0.002 for HOMER).

**Interpretation:** In multiple centers, HAL and HOMER demonstrated good discrimination and calibration.

Study Design and Methods

This multicenter international prospective observational cohort study involved 13 different hospitals. All consecutive patients with untreated PET-CT clinical radiographic stage T1-3, N0-3, M0 NSCLC that underwent EBUS-TBNA staging were included. Patients who had CT-proven mediastinal invasion, distant metastases, synchronous primaries, recurrent disease, or small cell lung cancer were excluded (e-Table 1).10,17 The study was approved by the Institutional Review Board, Committee 4, Protocol PA16-0107 at the University of Texas MD Anderson on June 10, 2016, and enrollment closed on March 1, 2019 (e-Appendix 1).

All variable definitions were developed before data abstraction, provided to all sites, and are identical to previously published criteria for HAL and HOMER (see e-Appendix 1, Methods, for further details).10,17 PET-CT was used to define N stage and tumor location (central vs peripheral), using previously established definitions.10,17,19 Lymph node N stage was based on the radiology report and supplemented by interventional pulmonologist review. CT N stage was based on the enlargement of lymph nodes to ≥1 cm on their short-axis diameter. PET N stage was based on the radiologist’s assessment of fluorodeoxyglucose avidity in mediastinal and hilar lymph nodes. Radiographic N stage by PET and by CT were defined as the highest abnormal nodal station using The Eighth Edition Lung Cancer Stage Classification.20 Central tumors were defined as those in the inner one third of the hemithorax (e-Fig 1).19

The predicted outcome variable was the presence of nodal involvement as determined by EBUS-TBNA. All EBUS-TBNA procedures used a systematic approach, defined as sampling of N3 nodes first, then N2 nodes, and finally N1 nodes.21 All lymph nodes ≥ 0.5 cm by EBUS were sampled, independent of PET-CT status.21

Statistical Analysis

**Prediction Models**

HAL was used to calculate the predicted probability of the highest nodal station with malignant lymph node involvement for each patient. Highest nodal station was determined by EBUS-TBNA and was categorized in an ordinal fashion as N0 vs N1 vs N2/3.17 HOMER was used to calculate the predicted probability of the highest nodal stage and tumor extent by EBUS-TBNA (prN1 vs prN2/3 as determined by EBUS-TBNA).17 HOMER performed well over long periods at the model development institution (ie, temporal validation), but performance over time was not measured at other hospitals. HAL has not been temporally validated at any site.

The objective of this study was to prospectively and externally validate HAL and HOMER at multiple new centers to determine whether these clinical prediction rules are suitable for generalized clinical use. The secondary objective was to temporally validate HAL (at the model development and external validation institutions) and HOMER (at the external validation institutions).

**Temporal and External Validation**

For HAL validation, prospective data from 13 hospitals (nine new plus original institutions) were used. For HOMER validation, prospective data from 12 hospitals (nine new plus original external validation institutions) were used, because the previously published temporal validation of HOMER used a portion of the model development institution data.17

**Assessment of Model Performance**

Model performance was assessed for discrimination using the area under the receiver operating characteristics (ROC) curve (AUC). Calibration was assessed using the Hosmer-Lemeshow goodness-of-fit test, Brier scores, and observed vs predicted graphs. We prespecified that the primary analysis would assess model performance for the entire cohort, and a secondary analysis would assess model performance stratified by institution.

**Exploratory Analyses**

Hospital-level variables have been previously shown to impact bronchoscopic diagnostic yield22,23 and might impact outcome and model performance. We used random effects models to test for center-level effects (e-Appendix 1).

In the previously published HAL and HOMER reports, model discrimination was good in all centers, but calibration was off in two centers.10,17 In these studies, institution-specific calibrated models were created (e-Table 2).10,17,24 We performed an exploratory analysis to temporally validate the calibrated models at their respective institutions (e-Appendix 1).

**Sample Size**

For the entire cohort, sample size was calculated so that the lower bound of the 95% CI for ROC-AUC would be 0.70 or higher. We chose the threshold value of 0.70 based on currently used clinical prediction models for SPNs (e-Table 3).25 The rationale was that if the ROC-AUC lower limit of the 95% CI was 0.70 or higher, the model would potentially be clinically useful. Based on prior data and conservative assumptions, we estimated the ROC-AUC would be 0.73 and the prevalence of

...
N2/3 disease would be 23%, which resulted in a sample size of 1,252 patients \( (e\text{-Appendix 1}) \).\footnote{For the secondary analysis stratified by institution, we estimated that 100 patients per institution would be sufficient to obtain a stable ROC-AUC for that center. Therefore, we prespecified that centers with fewer than 100 patients would be excluded from institution-level calibration assessment, and those with fewer than 50 patients would be excluded from institution-level discrimination assessment, because estimates would be unreliable \( (e\text{-Appendix 1 for details}) \). We therefore decided to enroll a minimum of 1,300 consecutive patients across 13 centers. Institutions were allowed to enroll more than 100 patients, but the goal was to have as many institutions reach the 100-patient threshold as possible.

All statistical analyses used SAS 9.4, STATA 15.1, and R 3.5.1.

### Results

A total of 1,799 patients in 13 hospitals with clinical-radiographic T1-3, N0-3, M0 NSCLC as determined by PET-CT who underwent EBUS-TBNA were enrolled \( (Table 2) \).

### Prospective External and Temporal Validation of HAL for Predicting N0/1 vs N2/3 Disease

Clinical characteristics according to nodal status \( (N0/1 vs N2/3) \) are shown in \( e\text{-Table 4} \). Predictions using HAL demonstrated good discrimination and calibration in the validation cohort \( (n = 1,799) \). ROC-AUC was 0.873 \( (95\% CI, 0.856-0.891; Fig 1A) \). Calibration as assessed by observed vs predicted plots \( (Fig 2A) \) and Brier score \( (0.117) \) was good. However, the Hosmer-Lemeshow \( P \) value was significant \( (P = .034) \). The forecast vs observed prN2/3 disease was 0.251 and 0.239, respectively. Although the small Hosmer-Lemeshow \( P \) value suggests lack of calibration beyond random variation, the absolute magnitude of the calibration error is small \( (see Fig 2 and Brier scores) \) and may not be consequential from a clinical perspective; also, the impact of lack of calibration depends on the clinical context.\footnote{Three centers enrolled fewer than 50 patients per center \( (total n = 98) \). Excluding these patients from the combined analysis did not significantly impact discrimination \( (n = 1,701, \text{ROC-AUC} = 0.875; Fig \)
1B) or calibration (Brier score, 0.112; Hosmer-Lemeshow P = .675; forecast prN2|3 disease = 0.242, observed prN2|3 = 0.220). The Hosmer-Lemeshow P value is dependent on sample size. Therefore, an increase in P value with reduced sample size does not necessarily imply better calibration.

Model performance stratified by institution is provided in Table 3. Discrimination on a per-hospital basis remained good, with ROC-AUC ranging from 0.810 to 0.917 and Brier scores from 0.083 to 0.141.

Exploratory testing demonstrated presence of center-level effects. The random effects model using a random intercept is provided in e-Appendix 1. Although it did outperform the baseline HAL model (P < .001), the absolute magnitude of discrimination improvement was small (ROC-AUC improved from 0.873 to 0.887).

Details of the exploratory analysis comparing the previously published institution-specific calibrated models are summarized in e-Table 5. The baseline HAL model outperformed the institution-specific calibrated HAL models in every institution (e-Appendix 1).

TABLE 2  Enrolled Patients Across 13 Different Hospitals

<table>
<thead>
<tr>
<th>Variable</th>
<th>Overall (N = 2,127)</th>
<th>Excluded (n = 328)</th>
<th>Included (n = 1,799)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institution, No. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Christus St. Vincent Santa Fea</td>
<td>38 (1.8)</td>
<td>5 (1.7)</td>
<td>33 (1.8)</td>
</tr>
<tr>
<td>Cleveland Clinic Foundation</td>
<td>361 (17.0)</td>
<td>53 (13.0)</td>
<td>308 (17.6)</td>
</tr>
<tr>
<td>Henry Ford Hospital</td>
<td>112 (5.3)</td>
<td>4 (1.4)</td>
<td>108 (5.9)</td>
</tr>
<tr>
<td>Michael E. DeBakey Hospital</td>
<td>117 (5.5)</td>
<td>0 (0.0)</td>
<td>117 (6.4)</td>
</tr>
<tr>
<td>Johns Hopkins</td>
<td>239 (11.2)</td>
<td>46 (15.8)</td>
<td>193 (10.5)</td>
</tr>
<tr>
<td>Kaiser Permanentea</td>
<td>51 (2.4)</td>
<td>17 (5.8)</td>
<td>34 (1.9)</td>
</tr>
<tr>
<td>MD Anderson Cancer Center</td>
<td>679 (31.9)</td>
<td>124 (37.7)</td>
<td>555 (30.8)</td>
</tr>
<tr>
<td>Royal Melbourne Hospital</td>
<td>146 (6.9)</td>
<td>13 (4.5)</td>
<td>133 (7.2)</td>
</tr>
<tr>
<td>VA Portland Healthcare System</td>
<td>61 (2.9)</td>
<td>7 (1.4)</td>
<td>54 (3.1)</td>
</tr>
<tr>
<td>St. Luke’s of Kansas City</td>
<td>103 (4.8)</td>
<td>1 (0.3)</td>
<td>102 (5.6)</td>
</tr>
<tr>
<td>Stanford University</td>
<td>67 (3.1)</td>
<td>3 (0.3)</td>
<td>64 (3.6)</td>
</tr>
<tr>
<td>University of Chicago</td>
<td>85 (4.0)</td>
<td>18 (5.8)</td>
<td>67 (3.7)</td>
</tr>
<tr>
<td>University of New Mexicoa</td>
<td>68 (3.2)</td>
<td>37 (12.3)</td>
<td>31 (1.7)</td>
</tr>
</tbody>
</table>

*Institutions with fewer than 50 patients that were included for the combined cohort analysis but excluded from the stratified by institution analysis. The reasons of patients’ exclusion are found in e-Table 1.

Figure 1 – Receiver operating characteristic curves of the HAL model in the combined validation cohort. The figure plots the area under the curve (AUC) for A, N2 or N3 (vs N0 or N1) disease (AUC = 0.873) for all patients that met inclusion criteria from all 13 institutions (n = 1,799), and for B, N2 or N3 (vs N0 or N1) disease (AUC = 0.875) for patients that met inclusion criteria from institutions that had more than 50 eligible patients (n = 1,701). HAL = Help with the Assessment of Adenopathy in Lung cancer model.
Prospective External and Temporal Validation of HOMER for Predicting N0 vs N1 vs N2/3 Disease

Clinical characteristics according to nodal status (N0 vs N1 vs N2/3) are shown in Table 6. Predictions using HOMER demonstrated good discrimination and calibration in the validation cohort (n = 1,244). ROC-AUC was 0.837 (95% CI, 0.814-0.859; Fig 3A) for predicting prN1|2|3 (vs prN0) disease and 0.876

Figure 2 – Observed vs predicted plots for the HAL model in the combined validation cohort for A, all patients that met inclusion criteria from all 13 institutions (n = 1,799) and B, for patients that met inclusion criteria from institutions that had more than 50 eligible patients (n = 1,701). The figure plots the probability of N2 or N3 (vs N0 or N1) disease by decile of expected risk as a function of the actual observed risk in that group. The observed probability for each decile is on the vertical axis; the predicted probability for each decile is on the horizontal axis. A perfect model, where observed = predicted, is shown by the line. HAL = Help with the Assessment of Adenopathy in Lung cancer model.

### Table 3: HAL Model Performance at Different Institutions

| Population | ROC-AUC (95% CI) | Brier Score | Observed prN2|3 | Forecast prN2|3 | Hosmer-Lemeshow P |
|------------|------------------|-------------|---------------|---------------|---------------|-----------------|
| Entire cohort (n = 1,799) | 0.873 (0.856-0.891) | 0.117 | 0.239 | 0.251 | .034 |
| Cohort excluding centers with <50 patients (n = 1,701) | 0.875 (0.856-0.894) | 0.112 | 0.220 | 0.242 | .675 |
| MD Anderson Cancer Centera (n = 555) | 0.871 (0.835-0.949) | 0.109 | 0.193 | 0.228 | .0467 |
| Cleveland Clinic Foundationb (n = 308) | 0.913 (0.880-0.945) | 0.092 | 0.182 | 0.216 | .704 |
| Henry Ford Hospitalb (n = 108) | 0.848 (0.772-0.924) | 0.136 | 0.269 | 0.290 | .293 |
| Johns Hopkinsb (n = 193) | 0.838 (0.776-0.899) | 0.141 | 0.285 | 0.304 | .280 |
| Royal Melbourne Hospitalb (n = 133) | 0.917 (0.862-0.973) | 0.110 | 0.338 | 0.230 | .010 |
| Michael E. DeBakey VA Hospitalb (n = 117) | 0.912 (0.847-0.977) | 0.083 | 0.188 | 0.165 | .768 |
| St. Luke’s of Kansas Cityb (n = 102) | 0.887 (0.813-0.962) | 0.120 | 0.245 | 0.204 | .459 |
| VA Portland Healthcare Systemb (n = 54) | 0.909 (0.828-0.990) | NA | NA | NA | NA |
| Stanford Universityc (n = 64) | 0.839 (0.742-0.937) | NA | NA | NA | NA |
| University of Chicagoc (n = 67) | 0.810 (0.692-0.989) | NA | NA | NA | NA |

Brier scores range from 0 (perfect) to 1 (worst) for a binary outcome. EBUS-TBNA = endobronchial ultrasound guided-transbronchial needle aspiration; NA = Not applicable because enrollment did not reach the required n ≥ 100 for these centers; N0|1 = N0 or N1; prN2|3 = probability of N2 or N3 disease; ROC-AUC = area under the receiver operator characteristic curve.

*aHospital where HAL was developed; data tests temporal validation.

*bHospital where HAL was previously tested; new data tests temporal external validation of HAL.

*cHospital where HAL was never tested before; external validation of HAL.
(95% CI, 0.855–0.897; Fig 3B) for predicting prN2|3 (vs prN0|1) disease. Calibration as assessed by observed vs predicted plots (Figs 4A, 4B) and Brier score (0.349) was good. However, the Hosmer-Lemeshow P value was significant (P = .002). The forecast vs observed prN1|2|3 disease was 0.339 and 0.357, respectively. The forecast vs observed prN2|3 disease was 0.261 and 0.259, respectively.

Excluding centers that enrolled fewer than 50 patients did not significantly change discrimination for prN1|2|3 (n = 1,146; ROC-AUC = 0.837; Fig 3C) and prN2|3 (ROC-AUC = 0.878, Fig 3D), or calibration (Brier score, 0.341; Hosmer-Lemeshow P < .001; observed vs predicted plots, Figs 4C and 4D, forecast prN1|2|3 vs observed prN1|2|3 was 0.325 and 0.335, respectively, and forecast prN2|3 disease vs observed prN2|3 was 0.249 and 0.233, respectively).

Model performance stratified by institution demonstrated good discrimination in eight of nine hospitals, with ROC-AUC ranging from 0.711 to 0.912 (Table 4) for prN1|2|3 (vs N0) disease, and diminished discrimination in one hospital (ROC-AUC = 0.672; Table 4). Discrimination was good in all hospitals for prN2|3 (vs N0|1) disease, with ROC-AUC ranging from 0.809 to 0.919 (Table 4). Calibration on a per-hospital basis was similar, with Brier scores ranging from 0.283 to 0.370 (Table 4).

Exploratory testing using a random effects model demonstrated the presence of center-level effects (e-Appendix 1). Although the random effects model did
outperform the baseline HOMER model ($P < .001$ for predicting prN1[2|3] vs N0 disease and $P < .001$ for predicting prN2[3] vs N0[1] disease), the absolute magnitude of discrimination improvement was small (ROC-AUC improved from 0.837 to 0.854 for predicting prN1[2|3] vs N0 disease and from 0.876 to 0.891 for predicting prN2[3] vs N0[1] disease).

Details of the exploratory analysis comparing the previously published institution-specific calibrated models are summarized in e-Table 7. The baseline HOMER model outperformed the institution-specific calibrated HOMER models in every institution (e-Appendix 1).

**Discussion**

In this study, we prospectively and externally validated the HOMER and HAL models, which estimate the probability of metastatic nodal disease in NSCLC patients as determined by EBUS-TBNA. We found that both models performed well in multiple outside institutions. External validation confirmed good discrimination. Although the Hosmer-Lemeshow test
| Population | ROC-AUC prN1|2|3 (95% CI) | ROC-AUC prN2|3 (95% CI) | Brier Score | Observed prN1|2|3 (vs N0) | Forecast prN1|2|3 (vs N0) | Observed prN2|3 (vs N0|1) | Forecast prN2|3 (vs N0|1) | HL | P |
|------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-----|---|
| Entire non-MDACC cohort (n = 1,244) | 0.837 (0.814-0.859) | 0.876 (0.855-0.897) | 0.349 | 0.357 | 0.339 | 0.259 | 0.261 | .002 |
| Non-MDACC data excluding centers with <50 patients (n = 1,146) | 0.837 (0.813-0.860) | 0.878 (0.86-0.900) | 0.341 | 0.335 | 0.325 | 0.233 | 0.249 | <.001 |
| Cleveland Clinic Foundation\(a\) (n = 308) | 0.851 (0.803-0.912) | 0.912 (0.879-0.945) | 0.283 | 0.263 | 0.290 | 0.182 | 0.218 | .098 |
| Henry Ford Hospital\(a\) (n = 108) | 0.812 (0.729-0.897) | 0.849 (0.773-0.924) | 0.345 | 0.389 | 0.361 | 0.269 | 0.289 | .008 |
| Johns Hopkins\(a\) (n = 193) | 0.803 (0.739-0.867) | 0.838 (0.777-0.900) | 0.363 | 0.363 | 0.380 | 0.285 | 0.302 | .038 |
| Royal Melbourne Hospital\(a\) (n = 133) | 0.878 (0.819-0.938) | 0.919 (0.864-0.974) | 0.370 | 0.504 | 0.402 | 0.338 | 0.288 | .027 |
| Michael E. DeBakey VA\(b\) (n = 117) | 0.912 (0.857-0.966) | 0.911 (0.845-0.978) | 0.263 | 0.265 | 0.208 | 0.188 | 0.169 | .148 |
| St. Luke’s of Kansas City\(b\) (n = 102) | 0.867 (0.798-0.937) | 0.887 (0.813-0.962) | 0.332 | 0.333 | 0.276 | 0.245 | 0.205 | .354 |
| VA Portland\(b\) (n = 54) | 0.886 (0.786-0.985) | 0.911 (0.832-0.990) | NA | NA | NA | NA | NA | NA |
| Stanford University\(b\) (n = 64) | 0.711 (0.581-0.841) | 0.834 (0.736-0.932) | NA | NA | NA | NA | NA | NA |
| University of Chicago\(b\) (n = 67) | 0.672 (0.535-0.809) | 0.809 (0.691-0.927) | NA | NA | NA | NA | NA | NA |

Brier scores range from 0 (perfect) to 2 (worst) when there are three possible outcomes, so the interpretation is different for HOMER and HAL. The lower Brier scores in HAL do not imply the HAL is better than HOMER. HL = Hosmer-Lemeshow; MDACC = MD Anderson Cancer Center; N0|1 = N0 or N1; NA = not applicable because enrollment did not reach the required n ≥ 100 for these centers; prN1|2|3 = probability of N1 disease or higher; prN2|3 = probability of N2 or N3 disease; ROC-AUC = area under the receiver operator characteristic curve.\(a\)Hospital where HOMER was previously validated, new data here tests temporal external validation of HOMER.\(b\)Hospital where HOMER was never tested before, external validation of HOMER.
suggested some calibration error beyond random variation, the calibration error was small and not evident as assessed by observed vs predicted plots and Brier scores.27 The forecast vs observed probabilities of disease for HAL (25.1% vs 23.9% for N2(3)) and HOMER (33.9% vs 35.7% for N1(2)3 and 26.1% vs 25.9% for N2(3)) suggest that HAL slightly overestimates the prN2(3) disease, whereas HOMER overestimates the prN0 disease.

Graphically this means that the curve in the observed vs predicted plots for HAL (Fig 2) is slightly but statistically significantly shifted to the right of the perfect prediction line. For HOMER, in the plot of prN1(2)3 (vs N0) disease (Fig 4A), the curve is shifted slightly to the left. This shift is a representation of the Hosmer-Lemeshow goodness-of-fit test, which tests overall calibration error. However, the absolute magnitude of the shift is small, as reflected in the difference between forecast and observed probabilities (+1.2% difference for HAL and −1.8% and +0.2% for HOMER for prN1(2)3 and prN2(3), respectively. Therefore, the data suggest that although calibration error is present, the magnitude of the difference is relatively modest, such that the models are informative.27

How well clinical prediction models perform over extended periods is important if they are to be used in everyday practice. The original HAL model10 used data from 2009, and the last patients in this study were from 2019. Both HAL and HOMER demonstrate good discrimination and calibration over a 10-year period at the four centers that performed both studies, providing evidence of temporal validation and stability over time.

Improving decision-making during the evaluation of lung cancer is fundamental to improve the value of patient care.28 ACCP and NCCN guidelines suggest that mediastinal lymph node sampling with EBUS be performed as the first invasive test in patients with sufficiently high probability of nodal disease. Inaccurate estimates of the probability of nodal disease can in turn lead to ineffective staging. For example, examination of SEER-Medicare data from 2004 to 2013 shows that only 23% to 34% of patients with lung cancer stage T1-3, N1-3, M0 disease received mediastinal sampling as part of their first invasive test, as recommended by guidelines.29

Instead of guideline-consistent care, most patients had a biopsy of the peripheral nodule or mass first, usually using a percutaneous CT-guided approach. Mediastinal sampling was delayed and sometimes never done. Unfortunately, this guideline-inconsistent pattern of care led to more biopsies and more complications. One factor contributing to this quality gap in cancer care is the difficulty in quantifying the probability of mediastinal and hilar nodal involvement.

Clinical prediction rules such as HAL and HOMER can help address this problem by better informing providers as to the probability of nodal involvement. HAL and HOMER demonstrated adequate discrimination and fairly good calibration. We believe that on balance HAL and HOMER remain informative and can improve decision-making during the evaluation of lung cancer analogous to the VA and Mayo models for SPNs.

HAL and HOMER can also inform decisions in scenarios in which the guidelines’ recommendations for the next step in management is not clear, such as determining whether a PET-scan is required after CT staging when EBUS is being considered, whether performing EBUS-TBNA is warranted after PET-CT imaging is done, and whether confirmatory mediastinoscopy after a negative EBUS-TBNA is required.12,4,8,10,17 For example, consider a 60-year-old patient with a peripherally located, CT N0 stage adenocarcinoma. Should this patient have a PET scan? According to HOMER, the prN1 ranges from 1.5% to 3.9% and a prN2(3) of 5.2% to 36.6%, depending on the PET N stage. The European Society of Thoracic Surgery (ESTS) suggests that proceeding to thoracotomy is reasonable, provided the probability of finding N2(3) disease at the time of surgery is less than 10%.30 A PET scan would be useful in this patient, because a PET N0 result would lead to surgery (prN2(3) < 10% threshold), whereas a PET showing nodal uptake would change management to EBUS (prN2(3) >10% threshold).

Conversely, if a patient instead had CT N2 disease, the prN1 ranges from 4.6% to 6.0%, and the prN2(3) ranges from 12.7% to 69.8%, depending on the PET N stage. In this case, we see that regardless of the PET N stage, the lowest the prN2(3) will be is 12.7%, which is higher than the suggested ESTS 10% threshold.30 Therefore, EBUS sampling in this case will be warranted no matter what, so proceeding directly to EBUS rather than waiting for a PET scan is warranted. In both patients, calculating the range of probabilities is useful.

The second decision that HOMER can inform is determining whether EBUS should be done.17 For
instance, consider two cases in which SABR is a treatment option. In the first case (A), a 60-year-old patient with a peripheral undifferentiated NSCLC, PET-CT N0, who is surgically fit but prefers SABR, is evaluated. HOMER predicts a 93% prN0 disease. In the second case (B), an 80-year-old patient with severe COPD who cannot tolerate lobectomy has a peripheral squamous cell carcinoma PET-CT N0. HOMER predicts a 98% prN0 disease. In both cases, finding N2 disease would lead to multimodal chemoradiation over SABR. However, in case A, finding N1 disease leads to surgery, whereas in case B it leads to definitive radiation therapy (alone or with chemotherapy). In case A, when weighing the risk of complications at approximately 1.15% vs the probability of occult nodal involvement at 7%, EBUS seems warranted because nodal involvement will change treatment from SABR to surgery with adjuvant chemotherapy or chemoradiation. Conversely, in case B, where there is only a 2% probability of finding malignant nodal disease vs a 1.15% risk of complications, proceeding directly to SABR seems reasonable. Although the absolute difference in probability of occult nodal metastasis between the two scenarios is small (2% vs 7%), the decision context is complex and involves consideration of the probability of EBUS being positive.

The third decision the models inform is whether confirmatory mediastinoscopy is warranted in patients that have a negative EBUS. Guidelines recommend that if EBUS is negative for malignancy in the N2/3 nodes but the suspicion of mediastinal nodal involvement remains high, mediastinoscopy be performed. But the guidelines do not identify which patients with a negative EBUS are high risk for nodal disease and which patients have a sufficiently low risk that proceeding directly to lobectomy is warranted. HAL can inform this decision. Because we know the sensitivity and specificity of EBUS, HAL can be used to approximate the pretest prN2/3 disease that can, in turn, be used to estimate the posterior probability of N2/3 disease given a negative EBUS (see e-Appendix 1 for details). The posterior probability of N2/3 disease then informs the decision on whether a confirmatory mediastinoscopy is necessary; if it is less than the surgical decision threshold (eg, ESTS threshold of 10%), then proceeding directly to lobectomy is warranted, whereas if it is greater than the decision threshold probability, confirmatory mediastinoscopy is warranted. HAL and HOMER do not dictate what the surgical decision threshold probability or what the next step in management is, but merely inform and improve the decision process.

HAL and HOMER also refine and challenge existing paradigms with large lesions (>3cm) that are PET-CT N0. The ACCP lung cancer guidelines recommend that for patients with a peripheral clinical stage IA tumor (PET-CT N0), invasive preoperative evaluation of the mediastinal node is not required (Grade 2B). The ACCP guidelines remain silent on whether peripheral tumors > 3 cm that are PET-CT N0 should have invasive preoperative evaluation of the mediastinal nodes. However, indirectly they imply that it would be reasonable to do mediastinal staging in such lesions. Similarly, the NCCN guidelines do recommend that patients with tumors > 3 cm that are PET-CT N0 undergo invasive preoperative evaluation of the mediastinal nodes. However, HAL and HOMER suggest that the probability of malignant nodal disease in patients with PET-CT N0 disease is low enough that foregoing EBUS-TBNA would be reasonable even when the tumor is large. Indeed, in this study, among 198 patients with peripheral lesions measuring >3 cm who were PET-CT N0, only two (1.0%; 95% CI, 0.1%-3.6%; e-Table 8) had N2/3 disease by EBUS. Note that the upper bound of the 95% credible interval of the proportion of patients having N2/3 disease by EBUS is significantly less than the 10% surgical threshold suggested by the ESTS (P < .001, binomial exact test). Therefore, although the guidelines indirectly suggest that patients with tumors measuring > 3 cm warrant EBUS, our data suggest that foregoing EBUS in select cases might be reasonable.

This study is the first multicenter prospective external and temporal validation of a clinical prediction model for mediastinal and hilar nodal involvement as determined by EBUS. Future studies will be needed to test the utility of the models and determine whether they impact care. Integration of HAL and HOMER into electronic decision aids may facilitate these endeavors. A webpage is now available that can be used to calculate nodal metastasis probabilities using HOMER, and prediction tables are in e-Tables 9 and 10. However, although HAL and HOMER have both been shown to have overall good performance, it is important to acknowledge their limitations, especially as applied to clinical practice. First, HAL and HOMER demonstrated...
some calibration issue in one institution (The Royal Melbourne Hospital). Also, performance might vary significantly in other settings if the type of EBUS staging procedure were changed. In this study, all patients had systematic staging, with sampling of all lymph nodes measuring ≥ 5 mm by EBUS. If this were changed, HAL and HOMER might not perform as well. In addition, HAL and HOMER predict EBUS results, and EBUS has limitations in terms of its sensitivity, which must be accounted for as previously described. Finally, HAL and HOMER are dependent on the cytopathologist interpreting the samples. If there is significant variability between centers in terms of specimen handling or cytopathologist interpretation, this will impact model performance.

**Interpretation**

In summary, we found that both HAL and HOMER performed well across multiple institutions in terms of their ability to predict mediastinal and hilar lymph node metastasis as identified by EBUS. The data provide external validation as well as temporal validation for the models. Discrimination and calibration were good, as assessed by observed vs predicted plots and Brier scores. The Hosmer-Lemeshow test suggested some degree of calibration error. The magnitude of the calibration error was small, such that the models are acceptable for clinical use. Future studies will need to assess their utility and hopefully use these models to improve lung cancer care.

**Acknowledgments**


**Other contributions:** The authors thank Linda Conry, MA, for editorial assistance in the preparation of this manuscript.

**Additional information:** The e-Appendix, e-Figures, and e-Tables can be found in the Supplemental Materials section of the online article.

**References**


