Supplementary

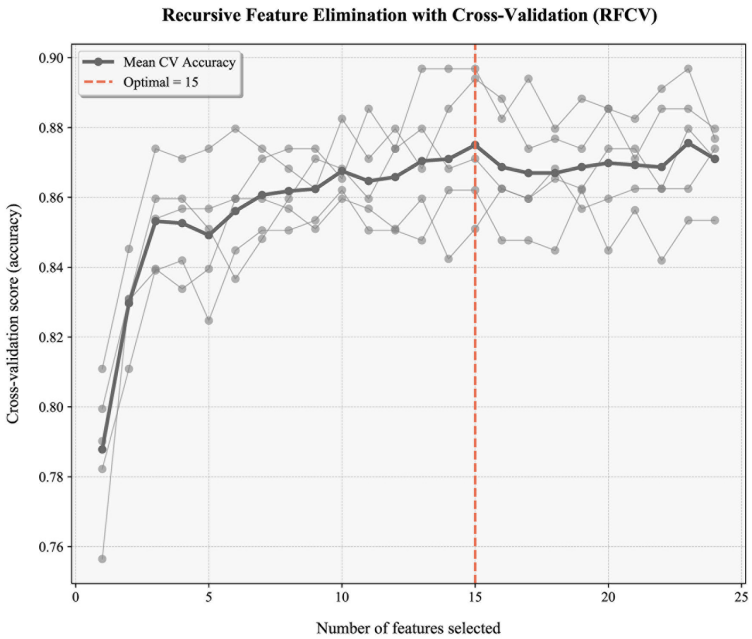


**Figure S1** Feature selection of 15 features using recursive feature elimination with cross-validation and 5-fold cross-validation, with accuracy as the selection criterion.


https://dx.doi.org/10.21037/qims-24-2016

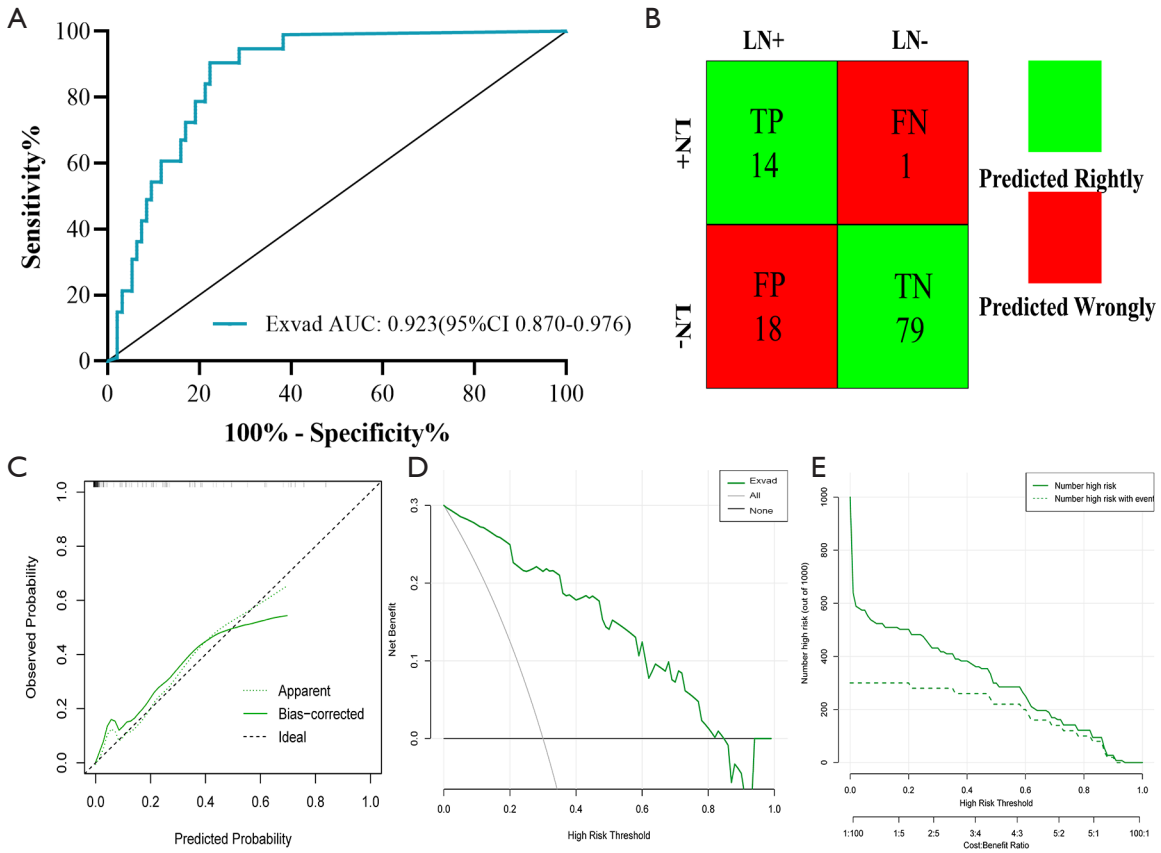


**Figure S2** The model of the Exvad dataset: (A) ROC curve; (B) confusion matrix; (C) calibration curve; (D) DCA curve; (E) CIC curve. AUC, area under the curve; CI, confidence interval; CIC, clinical impact curve; DCA, decision curve analysis; Exvad, external verification; FN, false negative; LN, lymph node; ROC, receiver operating characteristic; TP, true positive.


https://dx.doi.org/10.21037/qims-24-2016

## Appendix 1

To analyze the correlation between the identified risk factors and both TNM staging and histological subtypes, we conducted statistical analysis.

There is a significant correlation between pathological T staging and various risk factors [short diameter (SD), emphysema/bullae, obstructive inflammation, and mixed ground-glass opacity (mGGO)]. As the pathological T stage increases, the mean SD significantly increases ($F=759.55$, $P<0.001$), with T1 having a mean SD of 12.18 and T4 having a mean SD of 45.42, indicating that the lesion size increases with higher T staging, consistent with previous research suggesting that higher T staging may correspond to larger tumors. The difference in heterogeneous ventilation or perfusion (HVP) between T staging groups is not significant ($\chi^2=0.82$, $P=0.844$), suggesting a weak correlation with pathological T staging. The incidence of emphysema/bullae is higher in higher T stage groups (T3: 31.67%, T4: 29.03%), with significant inter-group differences ($\chi^2=40.34$, $P<0.001$), implying that the risk of emphysema or bullae around the tumor increases with higher T staging, potentially related to local tumor growth. The occurrence of obstructive inflammation significantly increases with higher T staging ($\chi^2=200.24$, $P<0.001$), with T1 having a 7.38% incidence and T4 reaching 38.71%, indicating that higher T stages are associated with an increased risk of inflammation, likely due to tumor invasiveness and its effect on surrounding tissues. The distribution of mGGO significantly varies between T staging groups ($\chi^2=519.77$, $P<0.001$), with T1 showing the highest proportion of mGGO 0 (35.15%), while T3 and T4 show higher proportions of mGGO 5 (70.83% and 80.65%, respectively), suggesting that the pattern of mGGO changes with higher T staging, reflecting alterations in tumor histology. Overall, pathological T staging is strongly correlated with SD, emphysema/bullae, obstructive inflammation, and mGGO, making T staging an important predictor of these risk factors, while HVP does not show significant correlation (*Table S1*).

The relationship between risk factors and pathological N staging shows significant associations with several factors. The SD significantly increases with the progression of pathological N stage, with the *F*-value of 120.40 ($P<0.001$), indicating a strong correlation. As N stage increases, the tumor size (measured by SD) also increases, suggesting that tumors with lymph node metastasis tend to be larger and more aggressive. HVP showed no significant correlation with N stage ($\chi^2=5.67$, $P=0.129$), indicating that HVP is not closely related to lymph node involvement. The incidence of emphysema/bullae increases significantly with N stage, with tumors at higher stages showing more local tissue damage. Obstructive inflammation also shows a strong correlation with N stage ($\chi^2=138.68$, $P<0.001$), with an increased incidence in higher N stages, reflecting tumor invasion of adjacent structures. mGGO (a potential pathological biomarker) shows a significant correlation with N stage ($\chi^2=503.79$, $P<0.001$), with higher proportions of mGGO 5 in advanced N stages, suggesting changes in tumor biology and increased invasiveness. Overall, SD and mGGO are closely associated with lymph node involvement, while emphysema/bullae and obstructive inflammation also correlate with higher N stages. However, HVP has limited relevance for predicting lymph node metastasis. These findings highlight the importance of pathological N staging in understanding tumor progression and related risk factors (*Table S2*).

The relationship between various risk factors and different tumor histological subtypes (invasive adenocarcinoma, squamous cell carcinoma, and other types) is as follows: Squamous cell carcinoma has a significantly larger SD (27.55±13.03) compared to invasive adenocarcinoma (15.68±8.50) and other types (10.21±6.42), indicating faster tumor growth. The HVP shows no significant differences across subtypes, with the majority (86.88%) of cases having normal ventilation/perfusion (HVP 0). However, squamous cell carcinoma is associated with a higher incidence of emphysema/bullae (49.43%) compared to adenocarcinoma (17.27%) and other types (8.89%), suggesting a stronger link to chronic obstructive pulmonary diseases (COPD). Obstructive inflammation is notably higher in squamous cell carcinoma (39.08%) than in adenocarcinoma (13.33%) and other types (4.76%), reflecting its tendency to cause airway invasion and inflammation. The distribution of mGGO markers shows that squamous cell carcinoma predominantly exhibits mGGO 5 (95.40%), while adenocarcinoma is concentrated in mGGO 0 and 1 stages. These findings suggest that mGGO expression is a significant biomarker distinguishing adenocarcinoma from squamous cell carcinoma, with


https://dx.doi.org/10.21037/qims-24-2016

higher expression in squamous cell carcinoma indicating more aggressive biological behavior. In conclusion, squamous cell carcinoma is characterized by larger tumor size, more frequent emphysema, obstructive inflammation, and higher mGGO 5 expression, correlating with chronic pulmonary changes and advanced biological markers. Invasive adenocarcinoma tends to be smaller, with lower mGGO expression and less association with emphysema or obstructive inflammation. Other tumor types show distinct features, suggesting different biological behaviors and progression patterns. These findings emphasize the biological and clinical differences between lung cancer subtypes, highlighting the potential for risk factors to aid in subtype differentiation and prognosis assessment (*Table S3*).

**Table S1** Training set T distribution and correlation of features

| Characteristics | Total (n=2,180) | Pathological T staging | | | | Statistic | P |
|---|---|---|---|---|---|---|---|
| | | 1 (n=1,653) | 2 (n=376) | 3 (n=120) | 4 (n=31) | | |
| SD | 15.36±9.03 | 12.18±5.16 | 21.84±7.76 | 31.23±9.53 | 45.42±17.57 | F=759.55 | <0.001 |
| HVP | | | | | | $\chi^2$=0.82 | 0.844 |
| 0 | 1,894 (86.88) | 1,434 (86.75) | 331 (88.03) | 103 (85.83) | 26 (83.87) | | |
| 1 | 286 (13.12) | 219 (13.25) | 45 (11.97) | 17 (14.17) | 5 (16.13) | | |
| Emphysema/bullae | | | | | | $\chi^2$=40.34 | <0.001 |
| 0 | 1,802 (82.66) | 1,412 (85.42) | 286 (76.06) | 82 (68.33) | 22 (70.97) | | |
| 1 | 378 (17.34) | 241 (14.58) | 90 (23.94) | 38 (31.67) | 9 (29.03) | | |
| Obstructive inflammation | | | | | | $\chi^2$=200.24 | <0.001 |
| 0 | 1,894 (86.88) | 1,531 (92.62) | 264 (70.21) | 80 (66.67) | 19 (61.29) | | |
| 1 | 286 (13.12) | 122 (7.38) | 112 (29.79) | 40 (33.33) | 12 (38.71) | | |
| mGGO | | | | | | $\chi^2$=519.77 | <0.001 |
| 0 | 585 (26.83) | 581 (35.15) | 3 (0.80) | 1 (0.83) | 0 | | |
| 1 | 474 (21.74) | 429 (25.95) | 44 (11.70) | 1 (0.83) | 0 | | |
| 2 | 142 (6.51) | 109 (6.59) | 25 (6.65) | 8 (6.67) | 0 | | |
| 3 | 147 (6.74) | 102 (6.17) | 35 (9.31) | 9 (7.50) | 1 (3.23) | | |
| 4 | 173 (7.94) | 103 (6.23) | 49 (13.03) | 16 (13.33) | 5 (16.13) | | |
| 5 | 659 (30.23) | 329 (19.90) | 220 (58.51) | 85 (70.83) | 25 (80.65) | | |

Data were presented as mean ± standard deviation or n (%). F, analysis of variance; HVP, heterogeneous ventilation or perfusion; mGGO, mixed ground-glass opacity; SD, short diameter; $\chi^2$, Chi-squared test.


https://dx.doi.org/10.21037/qims-24-2016

**Table S2** Training set n distribution and correlation of features

| Characteristics | Total (n=2,180) | Pathological N staging | | | | Statistic | P |
|---|---|---|---|---|---|---|---|
| | | N0 (n=1,854) | N1 (n=138) | N2 (n=139) | N1–2 (n=49) | | |
| SD | 15.36±9.03 | 13.94±7.93 | 23.13±8.96 | 23.14±11.71 | 25.24±11.04 | F=120.40 | <0.001 |
| HVP | | | | | | $\chi^2$=5.67 | 0.129 |
| 0 | 1,894 (86.88) | 1,598 (86.19) | 125 (90.58) | 125 (89.93) | 46 (93.88) | | |
| 1 | 286 (13.12) | 256 (13.81) | 13 (9.42) | 14 (10.07) | 3 (6.12) | | |
| Emphysema/bullae | | | | | | $\chi^2$=12.58 | 0.006 |
| 0 | 1,802 (82.66) | 1,549 (83.55) | 99 (71.74) | 114 (82.01) | 40 (81.63) | | |
| 1 | 378 (17.34) | 305 (16.45) | 39 (28.26) | 25 (17.99) | 9 (18.37) | | |
| Obstructive inflammation | | | | | | $\chi^2$=138.68 | <0.001 |
| 0 | 1,894 (86.88) | 1,676 (90.40) | 88 (63.77) | 99 (71.22) | 31 (63.27) | | |
| 1 | 286 (13.12) | 178 (9.60) | 50 (36.23) | 40 (28.78) | 18 (36.73) | | |
| mGGO | | | | | | $\chi^2$=503.79 | <0.001 |
| 0 | 585 (26.83) | 585 (31.55) | 0 | 0 | 0 | | |
| 1 | 474 (21.74) | 470 (25.35) | 0 | 2 (1.44) | 2 (4.08) | | |
| 2 | 142 (6.51) | 135 (7.28) | 1 (0.72) | 6 (4.32) | 0 | | |
| 3 | 147 (6.74) | 129 (6.96) | 3 (2.17) | 14 (10.07) | 1 (2.04) | | |
| 4 | 173 (7.94) | 130 (7.01) | 21 (15.22) | 16 (11.51) | 6 (12.24) | | |
| 5 | 659 (30.23) | 405 (21.84) | 113 (81.88) | 101 (72.66) | 40 (81.63) | | |

Data were presented as mean ± standard deviation or n (%). F, analysis of variance; HVP, heterogeneous ventilation or perfusion; mGGO, mixed ground-glass opacity; SD, short diameter; $\chi^2$, Chi-squared test.


https://dx.doi.org/10.21037/qims-24-2016

**Table S3** Correlation between pathological types and features of the training set

| Characteristics | Total (n=2,180) | invasive adenocarcinomas (n=1,778) | squamous cell carcinomas (n=87) | other histological types (n=315) | Statistic | P |
|---|---|---|---|---|---|---|
| SD | 15.36±9.03 | 15.68±8.50 | 27.55±13.03 | 10.21±6.42 | F=149.76 | <0.001 |
| HVP | | | | | $\chi^2$=13.32 | 0.001 |
| 0 | 1,894 (86.88) | 1,523 (85.66) | 78 (89.66) | 293 (93.02) | | |
| 1 | 286 (13.12) | 255 (14.34) | 9 (10.34) | 22 (6.98) | | |
| Emphysema/bullae | | | | | $\chi^2$=78.19 | <0.001 |
| 0 | 1,802 (82.66) | 1,471 (82.73) | 44 (50.57) | 287 (91.11) | | |
| 1 | 378 (17.34) | 307 (17.27) | 43 (49.43) | 28 (8.89) | | |
| Obstructive inflammation | | | | | $\chi^2$=70.82 | <0.001 |
| 0 | 1,894 (86.88) | 1,541 (86.67) | 53 (60.92) | 300 (95.24) | | |
| 1 | 286 (13.12) | 237 (13.33) | 34 (39.08) | 15 (4.76) | | |
| mGGO | | | | | $\chi^2$=353.00 | <0.001 |
| 0 | 585 (26.83) | 404 (22.72) | 0 | 181 (57.46) | | |
| 1 | 474 (21.74) | 433 (24.35) | 0 | 41 (13.02) | | |
| 2 | 142 (6.51) | 131 (7.37) | 0 | 11 (3.49) | | |
| 3 | 147 (6.74) | 139 (7.82) | 2 (2.30) | 6 (1.90) | | |
| 4 | 173 (7.94) | 155 (8.72) | 2 (2.30) | 16 (5.08) | | |
| 5 | 659 (30.23) | 516 (29.02) | 83 (95.40) | 60 (19.05) | | |

Data were presented as mean ± standard deviation or n (%). F, analysis of variance; HVP, heterogeneous ventilation or perfusion; mGGO, mixed ground-glass opacity; SD, short diameter; $\chi^2$, Chi-squared test.


https://dx.doi.org/10.21037/qims-24-2016