

Supplementary Materials

1. Development and test sets of internal validation

The final study sample was divided into the development set (underwent surgery from January 1, 2007, to October 23, 2017, n=101) and the test set (underwent surgery from October 24, 2017, to August 13, 2020, n=44) (S1 Fig.).

2. Results of internal validation

Clinical, surgical, and pathological characteristics, and oncologic outcomes of the development and test sets are compared in S5 Table. Computed tomography (CT) findings of the development and test sets are compared in S6 Table.

3. Pre- and perioperative predictors of survival

In univariable analysis, carbohydrate antigen 19-9 (CA19-9) response to neoadjuvant therapy (NAT), change in tumor size, change in tumor-artery contact, presence of suspicious node at the preoperative stage, invasion to the stomach or duodenum at baseline and preoperative stage were significant predictors. R1 resection was not a significant predictor ($p=0.377$). In multivariable analysis, elevated preop. CA19-9 ($p=0.046$) and stable or increased tumor size on CT ($p < 0.001$) were significant independent poor predictors of overall survival (OS) (S7 Table).

4. Development and validation of prognosis stratification criteria

We developed three-tier stratification criteria, “post-NAT response criteria,” by combining the two significant independent predictors of OS: low-risk=decreased tumor size and normal preop. CA19-9; intermediate-risk=(decreased tumor size and elevated preop. CA19-9) OR (stable or increased tumor size and normal preop. CA19-9); high-risk=stable or increased tumor size and elevated preop. CA19-9. Considering that not-elevated CA19-9 was not a significantly poor prognostic factor compared to normalized CA19-9 in Cox survival analysis (S7 Table), it was included in the criteria for the low-risk group. As a result, 40.6% (41/101), 37.6% (38/101), and 21.8% (22/101) of patients in the development set; 70.5% (31/44), 27.3% (12/44), and 2.3% (1/44) of patients in the test set were categorized into the low, intermediate, and high-risk groups, respectively.

In the development set, Harrell’s C-indices of the post-NAT response criteria and ypTNM staging were 0.707 (0.653 to 0.759) and 0.642 (0.572 to 0.707) showing no significant difference (0.065 [−0.006 to 0.143]). In the test set, C-indices of the post-NAT response criteria and ypTNM staging were 0.688 (0.436 to 0.839) and 0.708 (0.474 to 0.872) showing no significant difference (−0.020 [−0.290 to 0.262]) for OS; 0.676 (0.408 to 0.766), 0.628 (0.498 to 0.778), and 0.048 (−0.204 to 0.171) for recurrence-free survival (RFS). Log-rank test showed significant differences in OS ($p < 0.001$ for development set and $p=0.003$ for test set) and RFS ($p < 0.001$ for both development and test

sets) between the groups stratified by post-NAT response criteria. Calibration plots of the post-NAT response criteria demonstrated good correlation between the predicted and actual probabilities of a 3-year OS.