

Supplementary

Appendix 1 Methods

Missing data handling and multiple-imputation sensitivity analysis. Missingness before complete-case selection was summarized at the screened-population level. Missing predictor values alone did not constitute a study-level clinical exclusion. The primary analysis used the 1,019 complete cases, whereas the MICE sensitivity-analysis cohort comprised these 1,019 patients plus 26 additional patients with adjudicable primary outcomes and missing predictor values (total $n=1,045$). Patients meeting the 50 predefined clinical exclusion criteria and the 11 patients with non-adjudicable primary outcomes were not included in the imputation procedure. Multiple imputation by chained equations (MICE) was performed under a missing-at-random assumption, and 20 imputed datasets were generated. The primary outcome was fully observed among MICE-eligible patients, was included as a conditioning variable to preserve associations with predictor missingness, and was never imputed. Cohort assignment based on treatment period was preserved across all imputed datasets and was not imputed. Missing base-variable values were imputed in Python 3.11 using the statsmodels MICEData/MICE implementation (statsmodels 0.14.6). Continuous variables with missingness (including BMI, neutrophil count, lymphocyte count, and albumin) were updated using linear conditional models followed by predictive mean matching (PMM). Ordinal variables with missingness (including ECOG performance status and the FRAIL total score) were imputed on their observed integer support using PMM-based donor draws. Fully observed baseline variables, including age, sex, TNM stage, Charlson Comorbidity Index, treatment regimen, and treatment year, were included as conditioning variables where appropriate. After completion of each imputed dataset, derived variables were recalculated deterministically from their completed components: FRAIL category from the FRAIL total score, NLR from neutrophil and lymphocyte counts, PNI from albumin and lymphocyte count, PLR from platelet and lymphocyte counts, and SII from platelet, neutrophil, and lymphocyte counts. A 10-cycle burn-in preceded 20 retained imputations, and the default MICE skip interval ($n_skip=3$) was used between successive retained analysis samples. Convergence was assessed from the stability of imputed-variable summary statistics and chained-equation history/trace diagnostics across iterations. For comparability with the primary analysis, NLR and PNI were standardized in every imputed dataset using

the locked complete-case development-cohort constants (NLR mean 2.9091, SD 0.9628; PNI mean 48.8724, SD 4.9444). In each completed dataset, Model C was fitted in the imputed development cohort and evaluated in the corresponding same-center temporal validation cohort. Regression coefficients and standard errors were combined across datasets using Rubin rules; AUC, Brier score, calibration intercept, and calibration slope were summarized descriptively as the mean and range across the 20 imputed datasets.

Treatment-era and regimen sensitivity analysis. Treatment era was defined according to the first chemotherapy-based treatment year. Treatment exposure was summarized into model-specified chemotherapy-based categories: single-agent chemotherapy, doublet chemotherapy, and immuno-chemotherapy combination. Individual drug-level combinations were sparse and were not included in the locked model specification. These categories were used to describe treatment-era heterogeneity and for prediction-model adjustment, not to compare individual regimen safety.

Toxicity-spectrum classification. CTCAE grade ≥ 3 treatment-related toxicities were summarized according to prespecified categories. Primary toxicity type classified each event patient into one main category, whereas any-type toxicity allowed more than one category per patient. Multiple grade ≥ 3 toxicity categories were separately flagged.

Detailed model-wise grouped calibration plots were generated for Models A-C to visually compare predicted and observed risks in the development and same-center temporal validation cohorts. Smoothed calibration plots for final Model C are provided in the main manuscript as Figure 3.

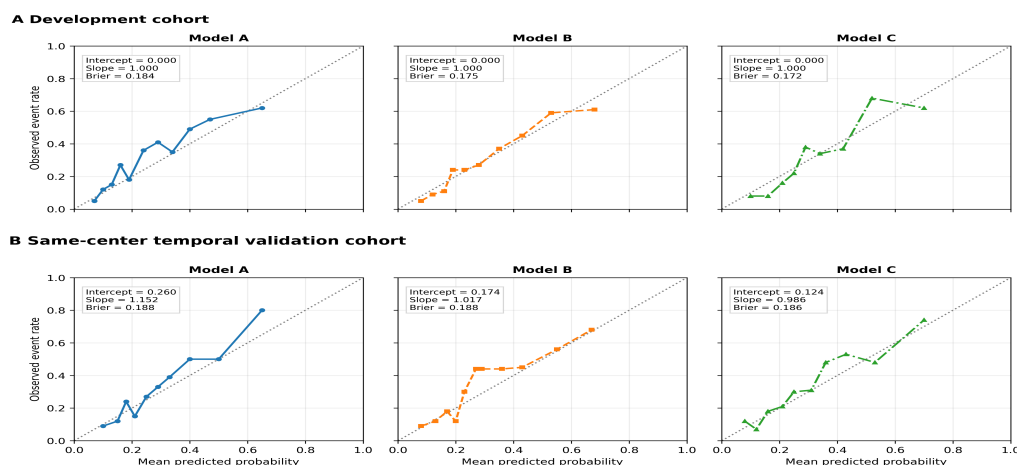


Figure S1 Detailed grouped calibration plots for Models A–C in the development and same-center temporal validation cohorts. Grouped calibration plots compare mean predicted probability with observed event rates for Model A, Model B, and Model C. The dashed diagonal line indicates ideal calibration. (A) Development cohort; (B) same-center temporal validation cohort.

Table S1 Missingness before complete-case selection and multiple-imputation eligibility

Variable	Unavailable at screening, n (%)	Not applicable due to predefined exclusion, n (%)	Missing after predefined exclusions, n (%)	MICE-eligible missing, n
Age	0 (0.0%)	0 (0.0%)	0 (0.0%)	0
Sex	0 (0.0%)	0 (0.0%)	0 (0.0%)	0
Body mass index	10 (0.9%)	0 (0.0%)	10 (0.9%)	10
Charlson Comorbidity Index	0 (0.0%)	0 (0.0%)	0 (0.0%)	0
ECOG performance status	10 (0.9%)	0 (0.0%)	10 (0.9%)	10
TNM stage	0 (0.0%)	0 (0.0%)	0 (0.0%)	0
Treatment regimen	50 (4.5%)	50 (4.5%)	0 (0.0%)	0
FRAIL total score	4 (0.4%)	0 (0.0%)	4 (0.4%)	4
FRAIL category	4 (0.4%)	0 (0.0%)	4 (0.4%)	4
Neutrophil count	11 (1.0%)	0 (0.0%)	11 (1.0%)	11
Lymphocyte count	18 (1.6%)	0 (0.0%)	18 (1.7%)	18
Albumin	10 (0.9%)	0 (0.0%)	10 (0.9%)	10
NLR	18 (1.6%)	0 (0.0%)	18 (1.7%)	18
PNI	18 (1.6%)	0 (0.0%)	18 (1.7%)	18
PLR	18 (1.6%)	0 (0.0%)	18 (1.7%)	18
SII	18 (1.6%)	0 (0.0%)	18 (1.7%)	18
Grade ≥ 3 treatment-related toxicity	61 (5.5%)	50 (4.5%)	11 (1.0%)	0

Percentages for screened missingness used the 1,106 screened patients as the denominator. Percentages for missingness after predefined exclusions used the 1,056 patients remaining after the 50 predefined clinical exclusions as the denominator. Patients with synchronous targeted therapy or end-of-life care/palliative sedation before chemotherapy were predefined exclusions; therefore, treatment regimen and primary outcome were recorded as not applicable rather than ordinary missingness. Eleven patients with non-adjudicable primary outcomes were not imputed. Twenty-six patients with adjudicable outcomes and missing predictor values were included with the 1,019 complete cases in the MICE sensitivity-analysis cohort. BMI, body mass index; ECOG,

Eastern Cooperative Oncology Group; FRAIL, Fatigue, Resistance, Ambulation, Illnesses, and Loss of weight; MICE, multiple imputation by chained equations; NLR, neutrophil-to-lymphocyte ratio; PLR, platelet-to-lymphocyte ratio; PNI, prognostic nutritional index; SII, systemic immune-inflammation index; TNM, tumor-node-metastasis.

Table S2 Model formulas, variable coding, standardization constants, and coefficients

(A) Model formulas

Model	Formula
Model B	$\text{logit}(p) = \alpha + \beta_1 \times \text{Age5} + \beta_2 \times \text{Female} + \beta_3 \times \text{BMI} + \beta_4 \times \text{Charlson} + \beta_5 \times \text{ECOG} + \beta_6 \times \text{TNM II} + \beta_7 \times \text{TNM III} + \beta_8 \times \text{TNM IV} + \beta_9 \times \text{Doublet chemotherapy} + \beta_{10} \times \text{Immuno-chemotherapy combination} + \beta_{11} \times \text{Pre-frail} + \beta_{12} \times \text{Frail}$
Model B + PNI	Model B + $\beta_{13} \times \text{zPNI}$
Model B + NLR	Model B + $\beta_{13} \times \text{zNLR}$
Model C	Model B + $\beta_{13} \times \text{zNLR} + \beta_{14} \times \text{zPNI}$

(B) Variable coding

Model term	Predictor	Coding
Intercept	Model intercept	Estimated constant term
Age5	Age	Age5 = age (years)/5; continuous and not mean-centered; OR corresponds to each 5-year increase
Female	Sex	1=female; 0=male (reference)
BMI	Body mass index	Continuous; each 1 kg/m ² increase
Charlson	Charlson Comorbidity Index	Continuous; each 1-point increase
ECOG	ECOG performance status	Continuous ordinal score; each 1-point increase
TNM II	TNM stage II	1=stage II; 0=stage I reference
TNM III	TNM stage III	1=stage III; 0=stage I reference
TNM IV	TNM stage IV	1=stage IV; 0=stage I reference
Doublet chemotherapy	Treatment regimen	1=doublet chemotherapy; 0=single-agent chemotherapy reference
Immuno-chemotherapy combination	Treatment regimen	1=immuno-chemotherapy combination; 0=single-agent chemotherapy reference
Pre-frail	FRAIL category	1=pre-frail; 0=robust reference
Frail	FRAIL category	1=frail; 0=robust reference
zNLR	Neutrophil-to-lymphocyte ratio	$(\text{NLR} - 2.9091)/0.9628$; standardized using the development cohort
zPNI	Prognostic nutritional index	$(\text{PNI} - 48.8724)/4.9444$; standardized using the development cohort

(C) Standardization constants and risk cut-offs

Parameter	Value
NLR mean	2.9091
NLR SD	0.9628
PNI mean	48.8724
PNI SD	4.9444
Risk stratification lower cut-off	0.1725
Risk stratification upper cut-off	0.3433
Low-risk group	Predicted probability ≤ 0.1725
Intermediate-risk group	$0.1725 < \text{predicted probability} \leq 0.3433$
High-risk group	Predicted probability > 0.3433

(D) Final Model C coefficients

Model term	Coefficient β	OR (95% CI)	P value
Intercept	-3.3068	–	0.040
Age5	0.0327	1.03 (0.85–1.25)	0.74
Female	-0.5267	0.59 (0.38–0.91)	0.02
BMI	0.0250	1.03 (0.95–1.10)	0.51
Charlson	0.1293	1.14 (0.91–1.43)	0.26
ECOG	0.4830	1.62 (1.26–2.08)	<0.001
TNM II	0.5558	1.74 (0.79–3.87)	0.17
TNM III	0.5473	1.73 (0.85–3.50)	0.13
TNM IV	0.7791	2.18 (1.23–3.87)	0.008
Doublet chemotherapy	-0.1666	0.85 (0.56–1.28)	0.43
Immuno-chemotherapy combination	-0.5158	0.60 (0.34–1.04)	0.07
Pre-frail	0.2990	1.35 (0.80–2.29)	0.27
Frail	1.1930	3.30 (1.79–6.06)	<0.001
zNLR	0.0270	1.03 (0.84–1.26)	0.80
zPNI	-0.3448	0.71 (0.56–0.89)	0.004

Model B added FRAIL category to the clinical base model. Model B + PNI and Model B + NLR were incremental comparison models; their relative performance is summarized in Table S3. Model C added both standardized NLR and standardized PNI to Model B and was the prespecified final model. Age5 was calculated as age in years divided by 5 and was not mean-centered. NLR and PNI were standardized using the mean and standard deviation from the development cohort and were applied unchanged to the same-center temporal validation cohort. Risk cut-offs were based on tertiles of Model C predicted probability in the development cohort and then applied unchanged to the validation cohort. BMI, body mass index; CI, confidence interval; ECOG, Eastern Cooperative Oncology Group; FRAIL, Fatigue, Resistance, Ambulation, Illnesses, and Loss of weight; NLR, neutrophil-to-lymphocyte ratio; OR, odds ratio; PNI, prognostic nutritional index; SD, standard deviation; TNM, tumor-node-metastasis.

Table S3 Incremental value of NLR beyond Model B and PNI

Model	Predictors added	Development AUC	Validation AUC	Development Brier	Validation Brier	Validation calibration slope	AIC	Likelihood ratio test
Model B	Reference model: clinical variables + FRAIL category	0.758	0.749	0.175	0.188	1.017	739.4	Reference
Model B + PNI	Standardized PNI	0.770	0.752	0.172	0.186	0.978	729.9	vs. Model B: P<0.001
Model B + NLR	Standardized NLR	0.762	0.756	0.175	0.186	1.058	738.5	vs. Model B: P=0.09
Model C	Standardized NLR + standardized PNI	0.770	0.753	0.172	0.186	0.986	731.9	vs. Model B + PNI: P=0.80

Model B was used as the reference for incremental comparisons. Model B + PNI and Model B + NLR separately added standardized PNI or standardized NLR to Model B. Model C added both standardized NLR and standardized PNI. The likelihood ratio test was performed in the development cohort. AIC, Akaike information criterion; AUC, area under the receiver operating characteristic curve; Brier, Brier score; FRAIL, Fatigue, Resistance, Ambulation, Illnesses, and Loss of weight; NLR, neutrophil-to-lymphocyte ratio; PNI, prognostic nutritional index.

Table S4 Treatment-era and regimen distribution

Analysis population	Treatment era	N	Single-agent chemotherapy, n (%)	Doublet chemotherapy, n (%)	Immuno-chemotherapy combination, n (%)	Grade toxicity, ≥ 3 n/N (%)
Complete-case analytic cohort	2015–2017	150	101 (67.3%)	49 (32.7%)	0 (0.0%)	42/150 (28.0%)
Complete-case analytic cohort	2018–2021	529	252 (47.6%)	172 (32.5%)	105 (19.8%)	160/529 (30.2%)
Complete-case analytic cohort	2022–2025	340	181 (53.2%)	114 (33.5%)	45 (13.2%)	114/340 (33.5%)
MICE sensitivity-analysis cohort	2015–2017	150	101 (67.3%)	49 (32.7%)	0 (0.0%)	42/150 (28.0%)
MICE sensitivity-analysis cohort	2018–2021	533	254 (47.7%)	174 (32.6%)	105 (19.7%)	162/533 (30.4%)
MICE sensitivity-analysis cohort	2022–2025	362	181 (50.0%)	114 (31.5%)	67 (18.5%)	126/362 (34.8%)

Treatment era was defined according to the first chemotherapy-based treatment year. The 2022–2025 era corresponded to the same-center temporal validation cohort. No eligible patient treated in 2015–2017 was classified as receiving immuno-chemotherapy combination; this category was first observed in the cohort after 2018, but the proportions should not be interpreted as a monotonic temporal increase. Individual drug-level combinations were sparse and were not included in the locked model specification; therefore, treatment exposure is summarized according to the model-specified regimen categories. These categories were used to describe treatment heterogeneity and for prediction-model adjustment, not to compare individual regimen safety. MICE, multiple imputation by chained equations.

Table S5 Treatment-heterogeneity sensitivity analyses

(A) Model C performance in treatment-heterogeneity sensitivity analyses

Analysis	Development N/events	Validation N/events	Development AUC	Validation AUC	Development Brier	Validation Brier	Validation calibration slope
Primary Model C, complete-case cohort	679/202	340/114	0.770	0.753	0.172	0.186	0.986
Model C after excluding immuno-chemotherapy	574/176	295/102	0.778	0.745	0.172	0.191	0.962
Model C plus treatment year term	679/202	340/114	0.770	0.753	0.171	0.185	0.984

(B) Treatment-era stratified performance of the locked Model C

Treatment era	N	Grade ≥ 3 toxicity, n/N (%)	AUC	Brier score	Calibration slope
2015–2017	150	42/150 (28.0%)	0.733	0.177	0.854
2018–2021	529	160/529 (30.2%)	0.780	0.170	1.040
2022–2025	340	114/340 (33.5%)	0.753	0.186	0.986

(C) Observed toxicity rates stratified by treatment regimen

Treatment regimen	Overall N	Overall Grade ≥ 3 toxicity, n/N (%)	Development N	Development Grade ≥ 3 toxicity, n/N (%)	Validation N	Validation Grade ≥ 3 toxicity, n/N (%)
Single-agent chemotherapy	534	173/534 (32.4%)	353	111/353 (31.4%)	181	62/181 (34.3%)
Doublet chemotherapy	335	105/335 (31.3%)	221	65/221 (29.4%)	114	40/114 (35.1%)
Immuno-chemotherapy combination	150	38/150 (25.3%)	105	26/105 (24.8%)	45	12/45 (26.7%)

(D) Additional calendar-period adjustment

Additional term	Adjusted OR (95% CI)	P value
Treatment year, per 1-year increase	1.04 (0.94–1.16)	0.44

All analyses used the 1,019-patient complete-case analytic cohort unless otherwise specified. Model C included clinical variables, FRAIL category, standardized NLR, and standardized PNI. In the non-immuno-chemotherapy sensitivity analysis, Model C was refitted in the development cohort after excluding immuno-chemotherapy patients and then evaluated in the corresponding validation subset. Era-stratified performance used the locked primary Model C predictions and is descriptive because the 2015–2017 and 2018–2021 periods were development-era subgroups, whereas 2022–2025 corresponded to the same-center temporal validation cohort. Observed regimen-stratified toxicity rates are descriptive and should not be interpreted as comparative safety estimates across treatment regimens. AUC, area under the receiver operating characteristic curve; BMI, body mass index; CI, confidence interval; ECOG, Eastern Cooperative Oncology Group; FRAIL, Fatigue, Resistance, Ambulation, Illnesses, and Loss of weight; NLR, neutrophil-to-lymphocyte ratio; OR, odds ratio; PNI, prognostic nutritional index.

Table S6 Distribution of CTCAE grade ≥ 3 treatment-related toxicity types in the complete-case analytic cohort

Toxicity category	Primary type among events, n=316	Any type among events, n=316	Any type in analytic cohort, n=1,019	Development cohort events, n=202	Validation cohort events, n=114
Hematologic toxicity	186 (58.9%)	220 (69.6%)	220/1019 (21.6%)	119 (58.9%)	67 (58.8%)
Gastrointestinal toxicity	68 (21.5%)	95 (30.1%)	95/1019 (9.3%)	45 (22.3%)	23 (20.2%)
Infectious toxicity	33 (10.4%)	59 (18.7%)	59/1019 (5.8%)	17 (8.4%)	16 (14.0%)
Hepatic/renal/metabolic toxicity	12 (3.8%)	13 (4.1%)	13/1019 (1.3%)	10 (5.0%)	2 (1.8%)
Immune-related toxicity	5 (1.6%)	12 (3.8%)	12/1019 (1.2%)	2 (1.0%)	3 (2.6%)
Other toxicity	12 (3.8%)	12 (3.8%)	12/1019 (1.2%)	9 (4.5%)	3 (2.6%)
Multiple grade ≥ 3 toxicity categories	–	90 (28.5%)	90/1019 (8.8%)	52 (25.7%)	38 (33.3%)

The table summarizes the 1,019-patient complete-case analytic cohort, including 316 patients with CTCAE grade ≥ 3 treatment-related toxicity during the first two chemotherapy-based treatment cycles. Primary toxicity type classifies each event patient into one main category. Any-type toxicity columns allow more than one category per patient. One patient could have more than one grade ≥ 3 toxicity category; therefore, counts for any-type toxicities may exceed the number of event patients. Development cohort refers to patients treated from September 2015 through December 2021, and validation cohort refers to patients treated from January 2022 through September 2025. For the first six toxicity categories, the development- and validation-cohort columns report primary toxicity types; for the final row, these columns report the proportion of event patients with multiple grade ≥ 3 toxicity categories. CTCAE, Common Terminology Criteria for Adverse Events.

Table S7 Multiple-imputation sensitivity analysis of final Model C

(A) Complete-case and multiple-imputation model performance

Analysis	Dev. N/events	Val. N/events	Dev. AUC	Val. AUC	Dev. Brier	Val. Brier	Val. intercept	Val. slope
Complete-case Model C, n=1,019	679/202	340/114	0.770	0.753	0.172	0.186	0.124	0.986
MICE datasets M1–M20, mean, n=1,045	683/204	362/126	0.769	0.743	0.173	0.193	0.151	0.942
MICE datasets M1–M20, range	–	–	0.767–0.771	0.733–0.748	0.172–0.173	0.191–0.196	0.100–0.171	0.868–0.968

(B) Rubin-rules pooled coefficients of Model C across imputed datasets

Model term	Pooled coefficient β	Pooled OR (95% CI)
Intercept	–3.2545	–
Age5	0.0291	1.03 (0.85–1.25)
Female	–0.5250	0.59 (0.39–0.91)
BMI	0.0256	1.03 (0.95–1.11)
Charlson	0.1274	1.14 (0.91–1.42)

ECOG	0.4965	1.64 (1.28–2.11)
TNM II	0.5603	1.75 (0.79–3.88)
TNM III	0.5939	1.81 (0.90–3.65)
TNM IV	0.7802	2.18 (1.23–3.88)
Doublet chemotherapy	−0.1502	0.86 (0.57–1.30)
Immuno-chemotherapy combination	−0.5078	0.60 (0.34–1.05)
Pre-frail	0.2593	1.30 (0.77–2.19)
Frail	1.1400	3.13 (1.71–5.72)
zNLR	0.0363	1.04 (0.84–1.28)
zPNI	−0.3432	0.71 (0.56–0.90)

The complete-case analysis used 1,019 patients. The MICE sensitivity analysis used 1,045 patients with eligible treatment exposure and adjudicable primary outcomes, including the 1,019 complete cases and 26 additional patients with missing predictor values. The 50 patients meeting predefined clinical exclusion criteria and the 11 patients with non-adjudicable primary outcomes were not included in the imputation procedure. Across the 20 imputed datasets, Model C was fitted in each imputed development cohort and evaluated in the corresponding same-center temporal validation cohort. Performance metrics (AUC, Brier score, calibration intercept, and calibration slope) were summarized descriptively as the mean and range across imputed datasets; Rubin rules were used only for pooling regression coefficients and standard errors. The MICE sensitivity analysis showed generally consistent results, with a mild decrease in validation AUC and a mild increase in validation Brier score compared with the complete-case analysis. AUC, area under the receiver operating characteristic curve; BMI, body mass index; Brier, Brier score; CI, confidence interval; ECOG, Eastern Cooperative Oncology Group; FRAIL, Fatigue, Resistance, Ambulation, Illnesses, and Loss of weight; MICE, multiple imputation by chained equations; NLR, neutrophil-to-lymphocyte ratio; OR, odds ratio; PNI, prognostic nutritional index; TNM, tumor-node-metastasis.

Table S8 Bootstrap optimism-corrected performance estimates for final Model C

Performance measure	Apparent development	Mean bootstrap optimism	Optimism-corrected development	Same-center temporal validation
AUC	0.770	0.023	0.747	0.753
Brier score	0.172	0.008	0.180	0.186
Calibration intercept	0.000	0.081	−0.081	0.124
Calibration slope	1.000	0.115	0.885	0.986

Bootstrap optimism correction was performed in the 679-patient development cohort with 202 events using 1,000 successful resamples. In each resample, final Model C was

refitted and evaluated in the bootstrap sample and in the original development cohort. For AUC, calibration intercept, and calibration slope, optimism-corrected estimates were calculated by subtracting mean optimism from apparent development performance. For Brier score, because lower values indicate better performance, the optimism-corrected estimate was calculated by adding prediction-error optimism to the apparent Brier score. Same-center temporal validation estimates are shown for comparison and were not optimism-corrected. AUC, area under the receiver operating characteristic curve; Brier, Brier score.