Deep learning-based model for prediction of early recurrence and therapy response on whole slide images in non-muscle-invasive bladder cancer: a retrospective, multicentre study

Fan Jiang,[a,i] Guibin Hong,[a,i] Hong Zeng,[b,i] Zhen Lin,[c,i] Ye Liu,[d,i] Abai Xu,[e,i] Runnan Shen,[a] Ye Xie,[a] Yun Luo,[f] Yun Wang,[a] Mengyi Zhu,[a] Hongkun Yang,[a] Haoxuan Wang,[a] Shuting Huang,[c] Rui Chen,[c] Tianxin Lin,[a,g,h,**] and Shaoxu Wu[a,g,h,*]

[a]Department of Urology, Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou, China
[b]Department of Pathology, Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou, China
[c]CellsVision Medical Technology Services Co., Ltd., Guangzhou, China
[d]Department of Pathology, The Fifth Affiliated Hospital, Sun Yat-sen University, Guangdong, China
[e]Department of Urology, Zhujiang Hospital, Southern Medical University, Guangdong, China
[f]Department of Urology, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[g]Guangdong Provincial Key Laboratory of Malignant Tumour Epigenetics and Gene Regulation, Guangdong-Hong Kong Joint Laboratory for RNA Medicine, Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou, China
[h]Guangdong Provincial Clinical Research Centre for Urological Diseases, Guangdong, China

eClinicalMedicine
2025;81: 103125

Published Online xxx

https://doi.org/10.1016/j.eclinm.2025.103125

Summary

Background Accurate prediction of early recurrence is essential for disease management of patients with non-muscle-invasive bladder cancer (NMIBC). We aimed to develop and validate a deep learning-based early recurrence predictive model (ERPM) and a treatment response predictive model (TRPM) on whole slide images to assist clinical decision making.

Methods In this retrospective, multicentre study, we included consecutive patients with pathology-confirmed NMIBC who underwent transurethral resection of bladder tumour from five centres. Patients from one hospital (Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou, China) were assigned to training and internal validation cohorts, and patients from four other hospitals (the Third Affiliated Hospital of Sun Yat-sen University, and Zhujiang Hospital of Southern Medical University, Guangzhou, China; the Fifth Affiliated Hospital of Sun Yat-sen University, Zhuhai, China; Shenshan Medical Centre, Shanwei, China) were assigned to four independent external validation cohorts. Based on multi-instance and ensemble learning, the ERPM was developed to make predictions on haematoxylin and eosin (H&E) staining and immunohistochemistry staining slides. Sharing the same architecture of the ERPM, the TRPM was trained and evaluated by cross validation on patients who received Bacillus Calmette–Guérin (BCG). The performance of the ERPM was mainly evaluated and compared with the clinical model, H&E-based model, and integrated model through the area under the curve. Survival analysis was performed to assess the prognostic capability of the ERPM.

Findings Between January 1, 2017, and September 30, 2023, 4395 whole slide images of 1275 patients were included to train and validate the models. The ERPM was superior to the clinical and H&E-based model in predicting early recurrence in both internal validation cohort (area under the curve: 0.837 vs 0.645 vs 0.737) and external validation cohorts (area under the curve: 0.761–0.802 vs 0.626–0.682 vs 0.694–0.723) and was on par with the integrated model. It also stratified recurrence-free survival significantly ($p < 0.0001$) with a hazard ratio of 4.50 (95% CI 3.10–6.53). The TRPM performed well in predicting BCG-unresponsive NMIBC (accuracy 84.1%).

Interpretation The ERPM showed promising performance in predicting early recurrence and recurrence-free survival of patients with NMIBC after surgery and with further validation and in combination with TRPM could be used to guide the management of NMIBC.

Funding National Natural Science Foundation of China, the Science and Technology Planning Project of Guangdong Province, the National Key Research and Development Programme of China, the Guangdong Provincial Clinical Research Centre for Urological Diseases, and the Science and Technology Projects in Guangzhou.

*Corresponding author. Department of Urology, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510120, China.
**Corresponding author. Department of Urology, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510120, China.
E-mail addresses: wushx29@mail.sysu.edu.cn (S. Wu), lintx@mail.sysu.edu.cn (T. Lin).

[i]These authors contributed equally.

Keywords: Bladder cancer; Deep learning; Early recurrence; Therapy response; Multicentre study

Research in context

Evidence before this study

We perform a PubMed search for relevant articles published in any language from the database inception to Oct 30, 2024, using the search string: ("artificial intelligence" OR "deep learning" OR "machine learning") AND ("recurrence" OR "therapy response" OR "prognosis") AND ("whole slide image" OR "pathology image"). We reviewed the 97 search results and identified 7 articles relevant to bladder cancer. Only 1 study developed an artificial intelligence model on haematoxylin and eosin (H&E)-stained slides to detect recurrence in non-invasive bladder cancer. However, this study had limited clinical impact due to small sample size and limited type of input data. No study has reported an artificial intelligence model that uses both H&E and immunohistochemistry stains for predicting recurrence and therapy response in non-muscle-invasive bladder cancer.

Added value of this study

In this multicentre study, we developed and validated a deep learning-based model called the early recurrence predictive model (ERPM) to predict early recurrence in non-muscle-invasive bladder cancer, which used whole slide images of H&E and immunochemistry stains. The ERPM showed satisfactory and robust performance in cross-centre cohorts, outperforming the clinical model, H&E-based model, and integrated model.

Implications of all the available evidence

To our knowledge, this is the first study that used H&E and immunochemistry stains to train and validate models to predict early recurrence and therapy response in non-muscle-invasive bladder cancer. By combining the two models, we developed a novel system that can stratify the recurrence risk and with further validation might instruct the postoperative management of non-muscle-invasive bladder cancer.

Introduction

Bladder cancer (BCa) is the ninth most common cancer globally.[1] Accounting for approximately 75% of newly diagnosed BCa, non-muscle-invasive BCa (NMIBC) recurs frequently.[2] Despite transurethral resection of bladder tumour (TURBT), NMIBC patients still have a reported recurrence rate of 40% in 1 year,[3] and over 20% of recurrent patients would progress to muscle-invasive BCa.[4] Patients with early recurrence were reported to have more frequent recurrence, poorer cystectomy-free survival, and overall survival than patients with late recurrence.[5] NMIBC patients at high risk of early recurrence should be offered more aggressive management.

Guidelines, like European Association of Urology (EAU) guidelines, recommended treatments, including intravesical instillation and radical cystectomy, for decreasing the recurrence, and follow-up for monitoring.[6,7] However, approximately 30% of high-risk NMIBC still relapsed in 2 years after the start of maintenance Bacillus Calmette-Guérin (BCG) therapy.[8] For BCG-unresponsive patients, further intravesical BCG may be useless.[9] In addition, the above managements will reduce quality of life due to the side effects and economic burden, especially radical cystectomy.[10] It is therefore important to accurately stratify the NMIBC patients based on their recurrence risk and treatment response for an individualized strategy of treatment and follow-up.

Among the previous studies,[7,11–13] the scoring models of the Spanish Urological Club for Oncological Treatment (CUETO) and the European Organization for Research and Treatment of Cancer (EORTC) are the most widely used models to predict recurrence of NMIBC. Both models comprise clinical and pathological features, and are aimed at predicting the risk of early and late recurrence. However, their capability was proven to be insufficient by recent studies.[14,15] Mateusz Jobczyk et al. recalibrated the above models and showed an improvement in predicting progression, whereas the ability to predict recurrence remained inadequate.[15] This was also reported by other studies, especially when predicting the recurrence of high-grade NMIBC.[16] There is a high demand for a more efficient model to predict post-TURBT recurrence of NMIBC.

In the past decade, artificial intelligence has shown great potential in the diagnosis and prognosis of tumours.[17–19] Several machine-learning-based models have been developed for predicting the recurrence of NMIBC, but their performance varies with study cohorts.[20,21] Deep learning becomes a better choice for its ability to model non-linear parameters and robustness across large datasets.[22] Marit Lucas et al. combined digital haematoxylin and eosin (H&E) staining slides with clinical data to develop a deep-learning-based model to predict recurrence-free survival (RFS) of NMIBC patients, but still showed moderate capability in predicting early recurrence (area under the curve [AUC] of 0.62).[14] Generally, the above studies only used pathological features from H&E staining slides and got insufficient performance. Previous studies have shown that immunohistochemistry (IHC) status was related to the recurrence of BCa,[23,24] such as P53 (TP53), CK20 (cytokeratin 20), Ki67 (MKI67), etc. By combining the H&E and IHC stains, we hypothesized that we could achieve a more precise prediction of early recurrence and therapy response.

In this study, we developed an early recurrence predictive model (ERPM) in NMIBC. We compared the ERPM with conventional models and estimated its ability to stratify RFS. Moreover, we explored its potential in predicting the response to BCG therapy in NMIBC.

Methods

Participants

In this retrospective, multicentre study, we included consecutive patients with NMIBC in five hospitals (Sun Yat-sen Memorial Hospital of Sun Yat-sen University [SYSMH]; The Fifth Affiliated Hospital of Sun Yat-sen University [SYUFH]; Shenshan Medical Centre [SSMC]; Zhujiang Hospital of Southern Medical University [ZJH]; the Third Affiliated Hospital of Sun Yat-sen University [SYUTH]) from January 1, 2017 to September 30, 2023. The inclusion criteria were as follows: (1) receiving TURBT and immediate postoperative installation of chemotherapy. (2) NMIBC confirmed by pathological diagnosis; (3) complete clinicopathological data; (4) available H&E staining slides along with or without IHC staining slides. The exclusion criteria were as follows: (1) clinical suspicions of nodal or distant metastasis; (2) concurrent other malignancies; (3) immediate RC after the first TURBT; (4) endpoint (recurrence within 2 years or RFS longer than 2 years) was not reached.

Ethics

This study has been approved by the Sun Yat-sen Memorial Hospital Institutional Review Board (number SYSKY-2024-353-01) and the requirement for informed consent was waived because of the observational design.

Procedures

The baseline characteristics of the patients, including age, sex, prior recurrence status, number of tumours, tumour size, T stage, concomitant carcinoma in situ (CIS), histologic grade, and follow-up data, were collected from medical record archives. T stage and histologic grade were confirmed by pathologists based on the AJCC/UICC 8th Edition TNM staging criteria and the 2004/2016 WHO histological classification system. Standard follow-up was defined as cystoscopy and urinary cytology every 3–6 months for the first 2 years, every 6 months for the following 3 years, and once a year thereafter. Recurrence was pathologically confirmed by the presence of BCa in the specimen from subsequent TURBT or cystectomy after 3 months. The reappearing BCa within 3 months was considered as the residual part of primary tumours. Early recurrence was defined as recurrence within 2 years, due to the delayed effect of intravesical therapy on recurrence.[25] RFS was defined as the period from the time they received the TURBT to the time of recurrence, the last follow-up, RC, or recurrence-free death.

Slides of the resected tumour tissue in TURBT were collected from the pathology department's archive, which were stained by H&E along with or without IHC (including P53, CK20, and Ki67). The whole slide images (WSIs) were generated by a digital slide scanner (SQS-40P, Shengqiang Technology, Shenzhen, China) with 20 × magnification. We excluded low-quality slides owing to improper staining or unremovable obstruction. Patients from the south and north branches of SYSMH were assigned to the 'training cohort' and the 'internal validation cohort', respectively. Patients from ZJH, SYUTH, SYUFH, and SSMC were assigned to four independent external validation cohorts.

The development of ERPM consisted of three sequential stages. In the first stage, the WSIs were first cropped into valid patches. Then, a pre-trained feature extraction model was applied to extract the corresponding features of the patches. All the extracted patch-level features from the same case would be merged according to the demand (H&E or H&E along with IHC).

The ERPM is an ensemble model consisting of two different sub-models. The sub-model mainly used the multi-instance part of DSMIL (available at https://github.com/binli123/dsmil-wsi), while the patch classification part of the output was modified from top1 to the average of top N. The selection of hyperparameter was accomplished through a handcrafted method and Neural Network Intelligence in a strategy of Tree-structured Parzen Estimator. Being set different hyperparameters, the two different sub-models were selected from two different 5-fold cross-validation.

In the final stage, the merged features of one case were input into two sub-models to obtain the bags prediction confidence and the patches confidence. The top 1 patch's confidence and the bags prediction confidence were averaged to obtain the final prediction of the single sub-model. The average of the prediction of two sub-models is the final prediction of this case. More details about the training process are provided in the Appendix (pp 2–4).

The H&E-based model shared the same network architecture with the ERPM, while the only difference was the training and validation data. It only used features from H&E staining WSIs for training and validation. The variables used to construct clinical model and integrated model were selected through LASSO (Least Absolute Shrinkage and Selection Operator) regression (Appendix p 4). Receiver operating characteristic (ROC) curves, decision curve analysis (DCA), and calibration curves were used to compare the performances of four models.

Survival analysis was also performed to estimate the prognostic capability of the ERPM. The binary classification of the ERPM was based on the cut-off value with optimal Youden index. We conducted univariable and multivariable Cox regression analysis in the training cohort, and selected clinical factors with statistical significance to build a Cox regression model (clinical Cox model). The integrated Cox model was built by combining the prediction of the ERPM and clinical Cox model. The C indexes of four models (the ERPM, H&E-based model, clinical Cox model, and integrated Cox model) in each validation cohort were further compared.

To enhance the explainability of the ERPM, we conduct comparisons between different combinations of IHC stains by adjusting the strategy of merging data in the first stage of the ERPM. We also extracted quantitative features from the top1 patch through CellProfiler (Version 4.2.6) and QuPath(Version 0.5.1), and compared the above features between the high-risk and low-risk groups defined by the ERPM (Appendix p 4).

Based on the architecture of ERPM, we developed a treatment response predictive model (TRPM). We included the BCG-naïve patients treated by intravesical BCG from SYSMH (training cohort and internal validation cohort of the ERPM). According to the US Food and Drug Administration's criteria,[26] the BCG-unresponsive NMIBC included persistent or recurrent CIS within 12 months of completion of adequate BCG therapy, recurrent high-grade Ta/T1 tumour within 6 months of completion of adequate BCG therapy, and high-grade T1 disease at the first evaluation following BCG induction. Adequate BCG therapy is defined as at least five of six doses of an initial induction course with more than two additional doses in maintenance therapy or a second induction course. The endpoints of patients were defined as confirmed BCG-unresponsive NMIBC or the last follow-up after 12 months since completion of adequate BCG therapy. The patients who hadn't reached endpoints were excluded. The training process of the TRPM was similar to the ERPM, and details are available in the Appendix (p 3).

Statistics

All statistical analyses and data visualization were performed using R software (Version 4.3.1) and Prism 9 (GraphPad Software). The glmnet package and stats package were used to select relevant clinical factors and build the clinical as well as the integrated model. The ROC curve, calibration curves, DCA, and area under the curve (AUC) were employed to evaluate the performance of the models (pROC package, rmda package, and rms package). The Delong test was used to compare the two ROCs. The Kaplan–Meier method, log-rank test, C index, and Cox proportional hazards model were used to estimate the prognostic capability of the ERPM (survival package and survminer package). Shapiro–Wilk test was used to test the distribution of the values for normality. Mann–Whitney U test was used to analyse non-normally distributed data, and Student's t-test was used to test normally distributed data for continuous variables. The chi-square test was used to test categorical variables. One-way ANOVA and Dunnett's test were used when comparing more than two groups with normal distribution. All statistical tests were two-sided and $p < 0.05$ was considered statistically significant.

Role of the funding source

The funders of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Between January 1, 2017, and September 30, 2023, 1620 consecutive patients received TURBT and immediate postoperative installation of chemotherapy, and were pathology-confirmed NMIBC (Fig. 1). We included 1275 patients, and excluded 345 patients based on the predefined exclusion criteria. 572 patients were assigned to the training cohort, 133 patients were assigned to the internal validation cohort, and 570 patients were assigned to four external validation cohorts. Table 1 shows the baseline characteristics of participants. There were no missing data in all five cohorts. A total of 1,135,413 image patches were generated from 4395 WSIs of 1275 patients to train and validate the models.

In Lasso regression, sex, T stage, histologic grade, concomitant CIS, number of tumours, tumours size, and prior recurrence were selected to construct the clinical model (Appendix p 5). The prediction of the ERPM was added to these factors to develop the integrated model.

As is shown in Fig. 2A, the ERPM achieved the best AUC (0.837, 95% CI 0.757–0.918) in the internal validation cohort, which is significantly higher than the clinical model and H&E-based model (Delong test, $p < 0.0001$ and $p < 0.01$, respectively). The calibration plot of all models in the internal validation cohort is presented in Fig. 2B. The Brier scores of four models were 0.173 (ERPM), 0.197 (HE-based model), 0.219(Clinical model), and 0.162 (Integrated model). DCA plot indicated that the ERPM gained greater net benefit than other models over the most range of threshold probability (Fig. 2C). Fig. 2D–I and Supplementary Fig. S2 (Appendix p 6) show robust performances of the ERPM in four external validation cohorts, outperforming H&E-based model and clinical model. There is no statistically significant difference between the ERPM and the integrated model in every validation cohort.

We further assess the prognostic capability of the ERPM. Patients were stratified into "high risk" and "low risk" groups (training cohort 322 vs 250, internal validation cohort 87 vs 46, and external validation cohorts

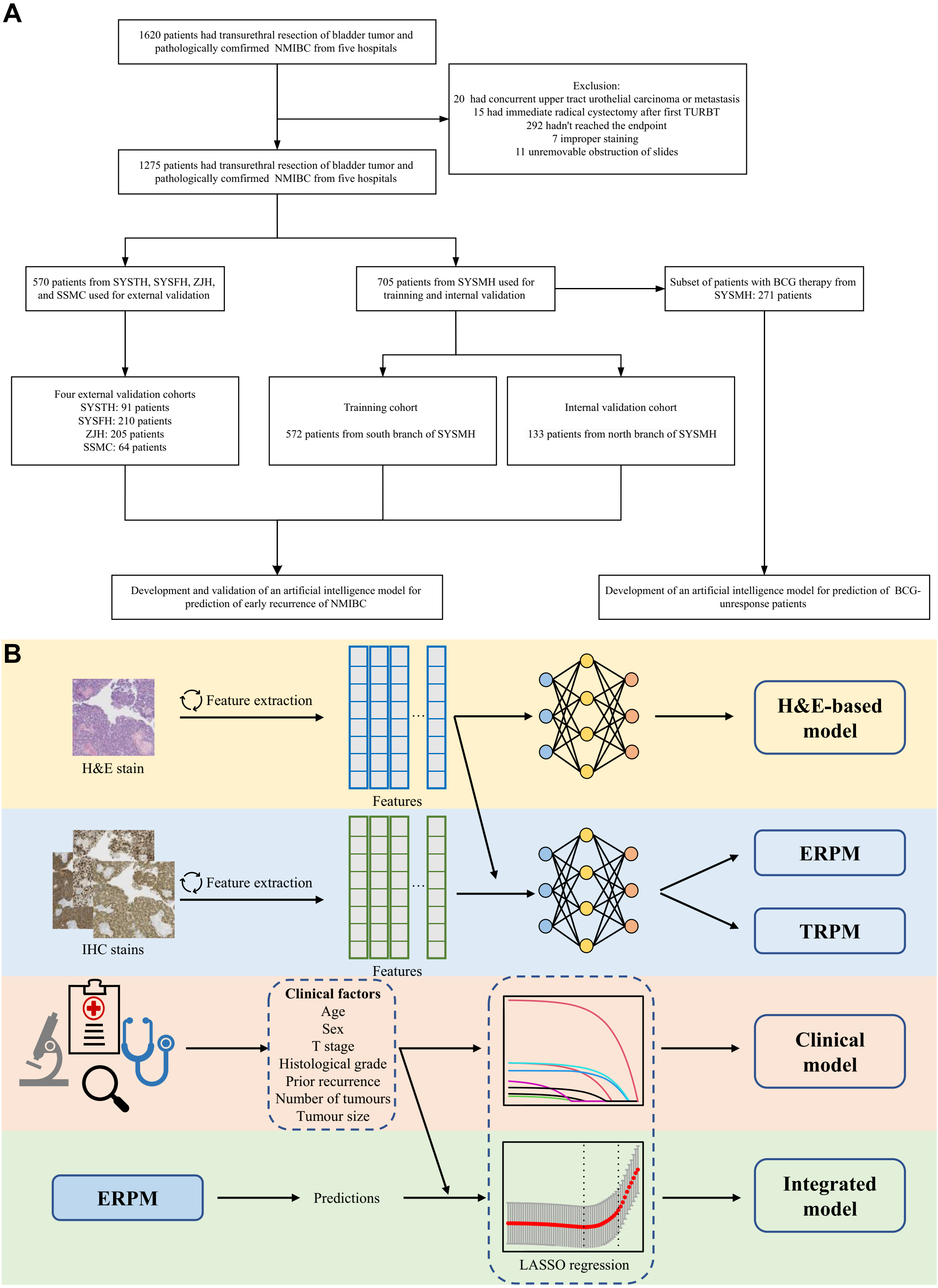

A
1620 patients had transurethral resection of bladder tumor and pathologically comfirmed NMIBC from five hospitals
Exclusion:
20 had concurrent upper tract urothelial carcinoma or metastasis
15 had immediate radical cystectomy after first TURBT
292 hadn't reached the endpoint
7 improper staining
11 unremovable obstruction of slides
1275 patients had transurethral resection of bladder tumor and pathologically comfirmed NMIBC from five hospitals
570 patients from SYSTH, SYSFH, ZJH, and SSMC used for external validation
705 patients from SYSMH used for trainning and internal validation
Subset of patients with BCG therapy from SYSMH: 271 patients
Four external validation cohorts
SYSTH: 91 patients
SYSFH: 210 patients
ZJH: 205 patients
SSMC: 64 patients
Trainning cohort
572 patients from south branch of SYSMH
Internal validation cohort
133 patients from north branch of SYSMH
Development and validation of an artificial intelligence model for prediction of early recurrence of NMIBC
Development of an artificial intelligence model for prediction of BCG-unresponse patients
B
Feature extraction
H&E stain
Features
H&E-based model
Feature extraction
IHC stains
Features
ERPM
TRPM
Clinical factors
Age
Sex
T stage
Histological grade
Prior recurrence
Number of tumours
Tumour size
Clinical model
ERPM
Predictions
LASSO regression
Integrated model

Fig. 1: **Flowchart of the study.** (A) Flowchart of included patients. (B) Framework of our study. NMIBC = non-muscle-invasive bladder cancer. TURBT = transurethral resection of bladder tumour. SYSMH = Sun Yat-sen Memorial Hospital of Sun Yat-sen University. SYSFH = The Fifth

	SYSMH training cohort (N = 572)	SYSMH validation cohort (N = 133)	SYSFH validation cohort (N = 210)	ZJH validation cohort (N = 205)	SYSTH validation cohort (N = 91)	SSMC validation cohort (N = 64)
Age, years	62.00 (55.00,71.00)	65.00 (56.00,72.00)	67.00 (58.00,72.00)	66.00 (57.00,74.00)	67.00 (57.00,73.00)	64.00 (57.00,73.00)
Sex						
Female	87 (15.2%)	27 (20.3%)	35 (16.7%)	39 (19.0%)	11 (12.1%)	7 (10.9%)
Male	485 (84.8%)	106 (79.7%)	175 (83.3%)	166 (81.0%)	80 (87.9%)	57 (89.1%)
Number of tumours						
Single	282 (49.3%)	110 (82.7%)	103 (49.0%)	160 (78.0%)	69 (75.8%)	34 (53.1%)
Multiple	290 (50.7%)	23 (17.3%)	107 (51.0%)	45 (22.0%)	22 (24.2%)	30 (46.9%)
Tumour size						
>3 cm	159 (27.8%)	34 (25.6%)	48 (22.9%)	41 (20.0%)	19 (20.9%)	22 (34.4%)
≤3 cm	413 (72.2%)	99 (74.4%)	162 (77.1%)	164 (80.0%)	72 (79.1%)	42 (65.6%)
Tumour stage						
T1	212 (37.1%)	61 (45.9%)	104 (49.5%)	44 (21.5%)	21 (23.1%)	16 (25.0%)
Ta	353 (61.7%)	70 (52.6%)	100 (47.6%)	160 (78.0%)	69 (75.8%)	46 (71.9%)
Tis	7 (1.2%)	2 (1.5%)	6 (2.9%)	1 (0.5%)	1 (1.1%)	2 (3.1%)
Tumour grade						
High grade	356 (62.2%)	97 (72.9%)	120 (57.1%)	69 (33.7%)	40 (44.0%)	24 (37.5%)
Low grade	204 (35.7%)	35 (26.3%)	86 (41.0%)	131 (63.9%)	49 (53.8%)	38 (59.4%)
PUNLMP	12 (2.1%)	1 (0.8%)	4 (1.9%)	5 (2.4%)	2 (2.2%)	2 (3.1%)
Concomitant CIS						
No	566 (99.0%)	127 (95.5%)	201 (95.7%)	203 (99.0%)	89 (97.8%)	62 (96.9%)
Yes	6 (1.0%)	6 (4.5%)	9 (4.3%)	2 (1.0%)	2 (2.2%)	2 (3.1%)
Prior recurrence						
No	445 (77.8%)	115 (86.5%)	144 (68.6%)	162 (79.0%)	77 (84.6%)	58 (90.6%)
Yes	127 (22.2%)	18 (13.5%)	66 (31.4%)	43 (21.0%)	14 (15.4%)	6 (9.4%)
EAU risk stratification						
Low	146 (25.5%)	29 (21.8%)	39 (18.6%)	93 (45.4%)	38 (41.8%)	20 (31.3%)
Intermediate	150 (26.2%)	34 (25.6%)	55 (26.2%)	62 (30.2%)	23 (25.3%)	15 (23.4%)
High	261 (45.6%)	64 (48.1%)	111 (52.9%)	49 (23.9%)	29 (31.9%)	26 (40.6%)
Very high	15 (2.6%)	6 (4.5%)	5 (2.4%)	1 (0.5%)	1 (1.1%)	3 (4.7%)
Recurrence						
No	400 (69.9%)	88 (66.2%)	113 (53.8%)	113 (55.1%)	63 (69.2%)	32 (50.0%)
Yes	172 (30.1%)	45 (33.8%)	97 (46.2%)	92 (44.9%)	28 (30.8%)	32 (50.0%)
Early recurrence						
No	442 (77.3%)	93 (69.9%)	119 (56.7%)	125 (61.0%)	66 (72.5%)	44 (68.8%)
Yes	130 (22.7%)	40 (30.1%)	91 (43.3%)	80 (39.0%)	25 (27.5%)	20 (31.2%)
RFS, months	37.00 (26.00,48.00)	26.00 (21.00,39.00)	27.00 (13.00,45.00)	34.00 (13.00,48.00)	29.00 (22.00,36.00)	26.00 (14.00,30.00)

Data are n (%) or median (IQR). SYSMH = Sun Yat-sen Memorial Hospital of Sun Yat-sen University. SYSFH = The Fifth Affiliated Hospital of Sun Yat-sen University. ZJH = Zhujiang Hospital of Southern Medical University. SYUTH = The Third Affiliated Hospital of Sun Yat-sen University. SSMC = The Medical Centre of Shenshan. PUNLMP = Papillary Urothelial Neoplasms of Low Malignant Potential. CIS = carcinoma in Situ. EAU = European Association of Urology. RFS = recurrence-free survival.

Table 1: **Baseline characteristics of the training and validation cohorts.**

287 vs 283) based on the classification defined by the ERPM. In the Kaplan–Meier analysis of each cohort, patients in high-risk group had shorter RFS than patients in low-risk group (Fig. 3A–C and Appendix p 8, log-rank test, p < 0.0001). Through the univariable and multivariable Cox regression analysis in the training cohort, T stage (T1), histologic grade (high grade), and prior recurrence were selected to construct clinical and integrated Cox models (Appendix p 13). Fig. 3D–E shows the multivariable Cox regression analysis in training and validation cohorts (hazard ratio [HR] 4.50 [3.10–6.53], 6.53 [4.86–8.79], respectively), demonstrating that the prediction of the ERPM is the independent prognostic factor. In the validation cohorts, the

Affiliated Hospital of Sun Yat-sen University. ZJH = Zhujiang Hospital of Southern Medical University. SYUTH = The Third Affiliated Hospital of Sun Yat-sen University. SSMC = The Medical Centre of Shenshan. ERPM = Early Recurrence Predictive Model. H&E = Hematoxylin and Eosin. BCG = Bacillus Calmette–Guérin. TRPM = Treatment Response Predictive Model. LASSO = least absolute shrinkage and selection operator.

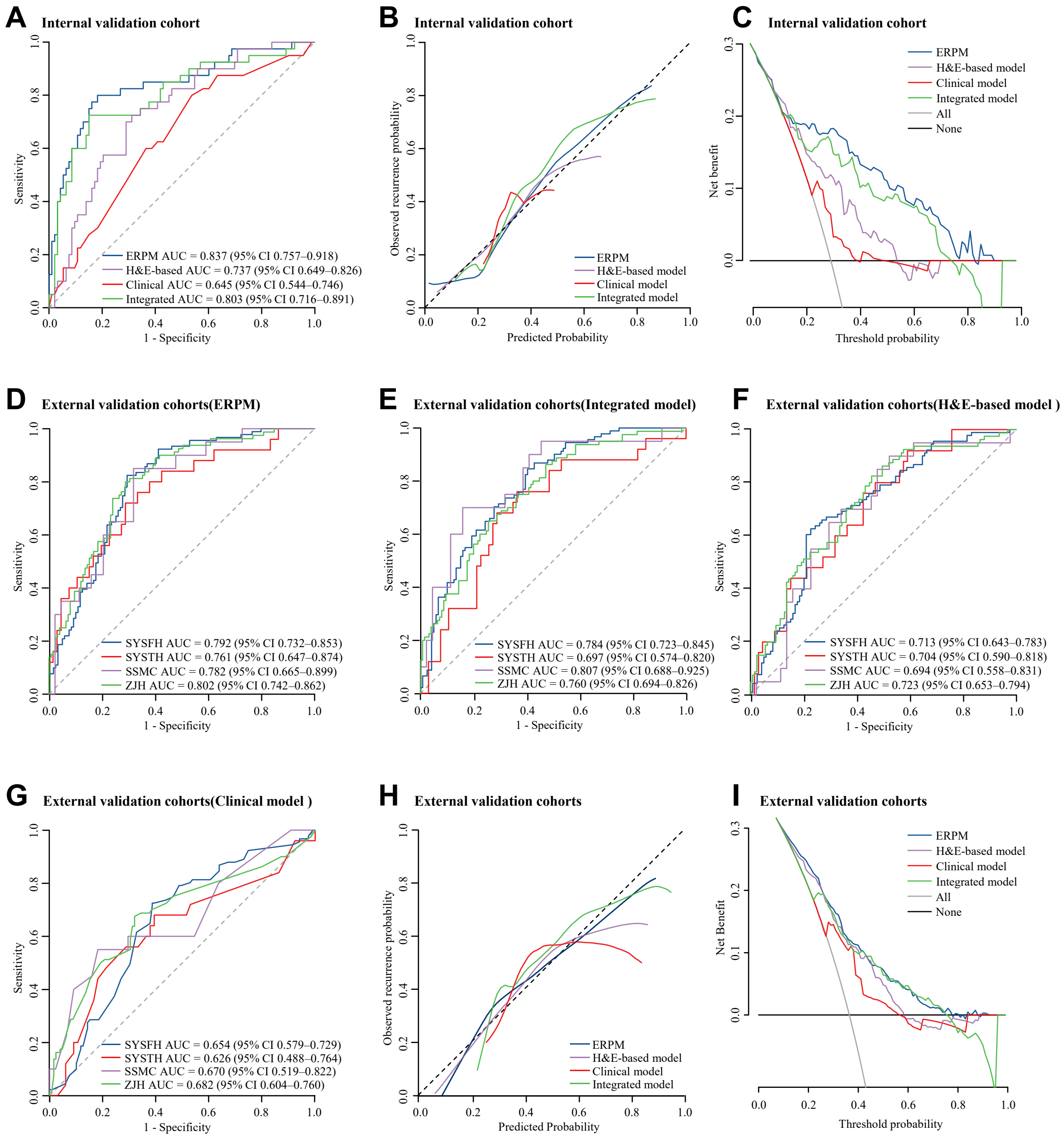


***Fig. 2:* Performance of models in validation cohorts.** (A–C) ROC curves (A), calibration plot (B), and DCA plot (C) of different models in predicting early recurrence in internal validation cohort. (D–G) ROC curves of the ERPM (D), integrated model (E), H&E-based model (F), and clinical model (G) in four external validation cohorts. (H and I) Calibration plot (H) and DCA plot (I) of different models in predicting early recurrence in overall external validation cohort. AUC = area under the curve. CI = confidence interval. ERPM = Early Recurrence Predictive Model. H&E = Haematoxylin and Eosin. SYSFH = The Fifth Affiliated Hospital of Sun Yat-sen University. ZJH = Zhujiang Hospital of Southern Medical University. SYUTH = The Third Affiliated Hospital of Sun Yat-sen University. SSMC = The Medical Centre of Shenshan.

ERPM achieved significantly higher C indexes than the clinical Cox model and H&E-based model (0.715–0.80 vs 0.596–0.714, $p < 0.05$; 0.715–0.80 vs 0.654–0.717, $p < 0.001$), and still showed no significant difference between the ERPM and the integrated Cox model (Fig. 3 F). We further assess its clinical utility by predicting RFS in different EAU risk groups in validation cohorts. The result shows that the ERPM can significantly stratified recurrence risk in every EAU risk group (Appendix p 9).

To understand how IHC assists in prediction, we included 362 patients with all three IHC stains from

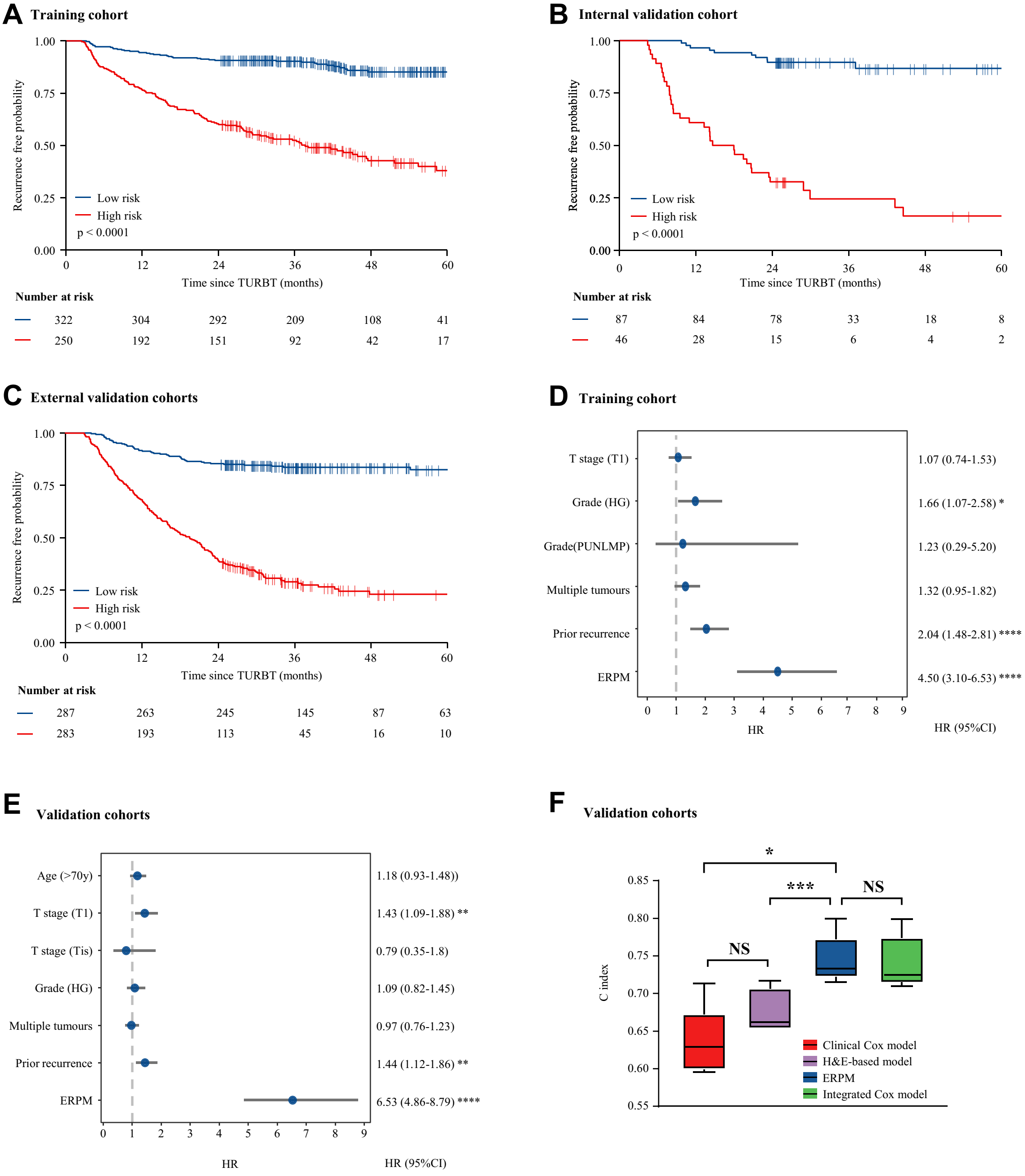
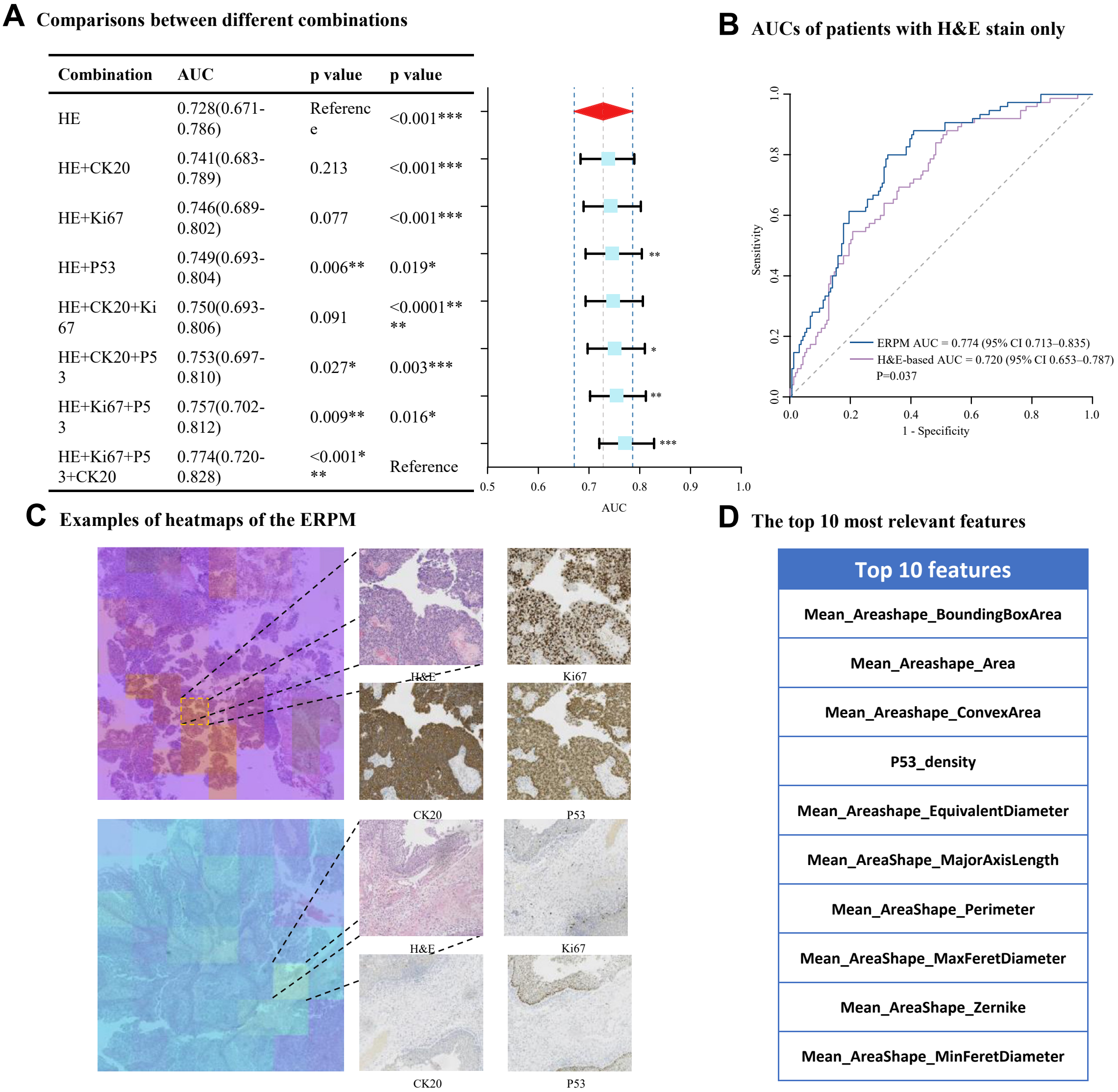


***Fig. 3:* Survival analysis in validation cohorts.** (A–C) Kaplan–Meier survival curves for recurrence free survival between high risk and low risk patients defined by the ERPM in training, internal validation and external validation cohorts. p values were calculated through the log-rank test. (D and E) HR of RFS for patients stratified by clinical factors and the ERPM in training and validation cohorts using multivariable Cox regression analysis and forest plot. (F) The C indexes of the ERPM, H&E-based model, clinical Cox model, and integrated Cox model in 5 validation cohorts. The comparison of C index was assessed by One-way ANOVA and Dunnett's test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001. TURBT = transurethral resection of bladder tumour. HR = hazard ratio. ERPM = Early Recurrence Predictive Model. H&E = Haematoxylin and Eosin. RFS = recurrence free survival. HG = high grade. PUNLMP = papillary urothelial neoplasm of low malignant potential.

validation cohorts, and used the ERPM to make predictions based on the different combinations of IHC stains. The result shows that the P53 stain can significantly enhance the predictive capability of model, and the combination of all three IHC stains achieves the best performance (Fig. 4A). For 239 patients from validation cohorts who only have H&E stain, the AUC of the ERPM is significantly higher than H&E-based model's



Combination	AUC	p value	p value
HE	0.728(0.671-0.786)	Reference	<0.001***
HE+CK20	0.741(0.683-0.789)	0.213	<0.001***
HE+Ki67	0.746(0.689-0.802)	0.077	<0.001***
HE+P53	0.749(0.693-0.804)	0.006**	0.019*
HE+CK20+Ki67	0.750(0.693-0.806)	0.091	<0.0001****
HE+CK20+P53	0.753(0.697-0.810)	0.027*	0.003***
HE+Ki67+P53	0.757(0.702-0.812)	0.009**	0.016*
HE+Ki67+P53+CK20	0.774(0.720-0.828)	<0.001***	Reference

Top 10 features
Mean_Areashape_BoundingBoxArea
Mean_Areashape_Area
Mean_Areashape_ConvexArea
P53_density
Mean_Areashape_EquivalentDiameter
Mean_AreaShape_MajorAxisLength
Mean_AreaShape_Perimeter
Mean_AreaShape_MaxFeretDiameter
Mean_AreaShape_Zernike
Mean_AreaShape_MinFeretDiameter

Fig. 4: **Explainability of the ERPM.** (A) Comparisons between different combinations of H&E and IHC stains in 362 patients. p values were calculated through Delong test. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001. (B) AUCs of the ERPM and H&E-based model on patients with H&E stains only. p values were calculated through Delong test. (C) Example of heatmaps of H&E WSIs and the original input along with IHC stains from high risk (upper) and low risk (lower) patients predicted by the ERPM. (D) The top 10 most relevant features were identified based on their statistical difference between patients from high risk and low risk. AUC = area under the curve. ERPM = Early Recurrence Predictive Model. H&E = Haematoxylin and Eosin. IHC = immunochemistry. WSIs = whole slide images.

(0.774 [95% CI 0.713–0.835] vs 0.720 [95% CI 0.653–0.787], p = 0.037; Fig. 4B), which indicates that the IHC stains may guide the model to focus on more relevant features in training. To interpret the ERPM, we visualized the focused region in H&E staining WSIs through heat maps (Fig. 4C) and extracted relevant features from the top 1 patch. The top 10 features with statistical significance are listed in Fig. 4D. Except for area shape, texture and intensity of nuclei and cells, the density of each IHC stain is also significantly differed between low-risk and high-risk groups (p values of p53, ck20, ki67 are <0.001, 0.016, 0.003, respectively).

To test whether the architecture of ERPM can be applied in predicting the BCG unresponsive patients, we developed the TRPM on the basis of the ERPM. 271 patients with intravesical BCG from SYSMH cohort were included, and their characteristics are presented in Appendix (p 15). After 5-fold cross-validation, the model achieved an accuracy of 84.1%, a sensitivity of 69.0%, and a specificity of 88.3%. Fig. 5A shows the confusion

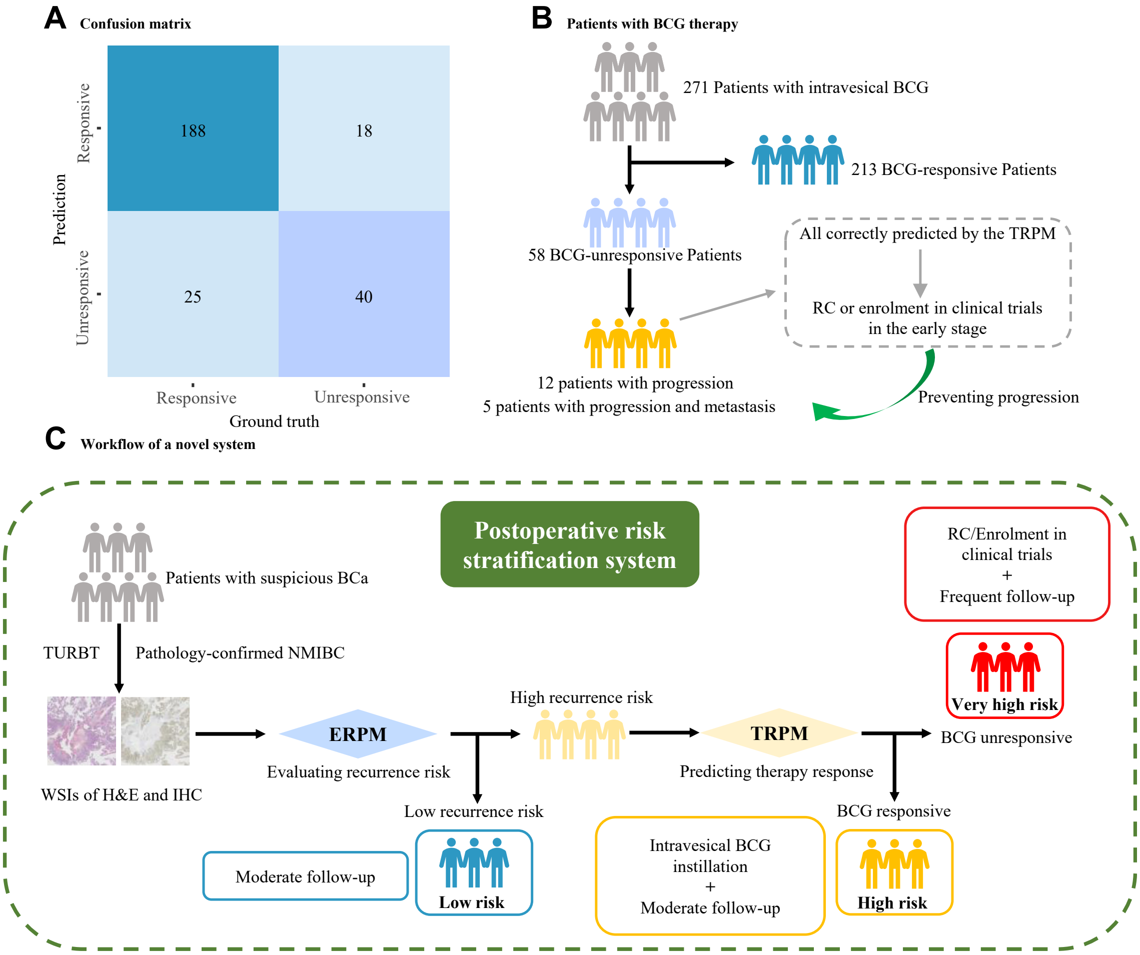

A Confusion matrix
Prediction
Responsive
Unresponsive
188
18
25
40
Responsive
Unresponsive
Ground truth
B Patients with BCG therapy
271 Patients with intravesical BCG
213 BCG-responsive Patients
58 BCG-unresponsive Patients
All correctly predicted by the TRPM
RC or enrolment in clinical trials in the early stage
12 patients with progression
5 patients with progression and metastasis
Preventing progression
C Workflow of a novel system
Postoperative risk stratification system
Patients with suspicious BCa
TURBT
Pathology-confirmed NMIBC
WSIs of H&E and IHC
ERPM
Evaluating recurrence risk
High recurrence risk
Low recurrence risk
TRPM
Predicting therapy response
BCG unresponsive
BCG responsive
RC/Enrolment in clinical trials + Frequent follow-up
Very high risk
Moderate follow-up
Low risk
Intravesical BCG instillation + Moderate follow-up
High risk

Fig. 5: **Performance of the TRPM and workflow of novel risk stratification system.** (A) Confusion matrix of the 5-fold cross validation from the TRPM in 271 patients. (B) Predictions of the TRPM on Patients with progression. (C) Workflow of a novel system which stratifies patients into low risk, high risk, and very high risk groups based on their risk of recurrence and BCG-unresponsive NMIBC. BCa = bladder cancer. NMIBC = non-muscle-invasive bladder cancer. TURBT = transurethral resection of bladder tumour. H&E = Haematoxylin and Eosin. ERPM = Early Recurrence Predictive Model. TRPM = Treatment Response Predictive Model. IHC = immunochemistry. BCG = Bacillus Calmette–Guérin. RC = radical cystectomy.

matrix of the TRPM. Of the 58 BCG-unresponsive patients, 12 patients progressed and 5 patients showed progression and metastasis after adequate intravesical BCG, who were all correctly predicted by the TRPM (Fig. 5B). It also significantly stratified the RFS with $p < 0.0001$ (Appendix p 10). By combining the ERPM and TRPM, we developed a novel system for stratification of NMIBC patients, which stratified patients into three groups and offered recommended management for each group (Fig. 5C).

Discussion

To our knowledge, this is the first study to develop a deep learning-based model for predicting early recurrence of NMIBC on H&E and IHC staining slides. The ERPM achieved superior performance in cross-centre cohorts, outperforming the conventional clinical model and the H&E-based model. It also showed good prognostic capability in predicting RFS. In the end, the TRPM was developed to predict the response of BCG therapy and showed promising performance.

The high incidence of recurrence and adverse impacts of disease management make it hard to manage NMIBC patients.[27] Conventional models (such as the EORTC and CUETO models) and previous H&E-based AI models were reported to have limited ability to predict early recurrence.[4,14,15,28] As conventional clinical methods and H&E-based models still showed insufficient performance in our cohorts, the ERPM exhibited superior capability in predicting early recurrence, indicating the remarkable assistance provided by the IHC staining slides. Previous studies always took IHC stains as a categorial variable (negative or positive, low or high, etc.).[24,29] However, IHC staining can provide additional information of tumour and peripheral tissue,[19,30] if the inputs were WSIs. We selected the p53, Ki-67, and CK20 for their relations to the outcomes of NMIBC reported by considerable studies.[23] Since each centre may adopt different choices of IHC to evaluate individual NMIBC, the ERPM is designed to analyse combinations of H&E and 0–3 IHC stains of each patient, which makes further validation and application more practical. To be noticed, there is unavoidable interobserver variability in evaluating the clinical risk factors and pathological features (including T stage, CIS, grade, etc.).[6] As no statistically significant difference between the AUC values of ERPM and integrated models, ERPM is an independent prognostic factor of NMIBC and is spared affection from the interobserver variability. Overall, ERPM achieves promising performance with the assistance of the IHC stains in predicting early recurrence.

In survival analysis, we demonstrated the important prognostic value of the ERPM. ERPM's predictions significantly stratified the RFS across two risk groups defined by the ERPM. It indicates that the ERPM can also evaluate the recurrence risk after 2 years. According to EAU's prognostic factor risk groups for NMIBC, it is recommended that, after 2 years, intermediate and low-risk patients receive cystoscopy annually, while high-risk patients receive cystoscopy and cytology every 6 months up to 5 years.[6] However, the recurrence rate after 2 years cannot be ignored.[4] Since the strong prognostic capability of the ERPM, we recommend that high-risk patients predicted by the ERPM should take extra surveillance during two cystoscopies, such as radiology, and cytology. Meanwhile, 63 patients with false positive prediction actually relapsed after 2 years. A review conducted an evidence synthesis and found that both inadequate and adequate intravesical BCG would delay the recurrence of NMIBC patients.[25] Therefore, we inferred that ERPM may actually make correct predictions, but the real outcome was influenced by postoperative therapies. In the future, we will include more centres and better categorize patients to eliminate the interference of different treatments on predictions. Besides, according to the EAU guidelines, stratification of EAU risk groups is based on the probability of progression to muscle-invasive disease.[6] The ERPM can further stratify patients in each EAU risk group significantly, helping patients at high risk of recurrence adjust their strategies of therapy and follow-up (Appendix p 11). Patients who were predicted to be at high risk of recurrence by the ERPM could be offered more aggressive management, while the rest can follow the original management of their EAU risk groups. This needs to be further validated, especially in group with few patients like very high-risk group.

According to the performance of ERPM on different combinations of H&E and IHC stains, the use of P53 significantly improved the capability of the model and the combination of H&E along with 3 IHC stains was superior to any other combination. Therefore, we recommended combination of these three IHC satins should be routinely used in the pathologic workflow of NMIBC. Furthermore, the features that differed most between the risk groups stratified by the ERPM were quantified and distinguished, including the density of each IHC stain, which emphasized the importance of three IHC stains again. Meanwhile, the density of P53 and Ki67 positivity were significantly higher in high recurrence risk group, suggesting that our model may make predictions based on tumour malignancy and rate of proliferation.

BCG-unresponsive NMIBC is reported to have a low likelihood of responding to further BCG treatment.[9] For non-responders, repeated BCG instillation may be harmful for putting them in danger of recurrence and progression. Y Lotan et al. developed artificial intelligence-based histologic assays, and successfully predict recurrence, progression, BCG unresponsive disease, and cystectomy across an international cohort of patients with high-risk NMIBC.[31] In addition to the high-risk group, there are patients in the intermediate- and very high-risk groups who have received intravesical BCG. The TRPM is designed to identify non-responders from intermediate-, high- and very high-risk groups before the postoperative treatment and prevent them from receiving inappropriate treatments. It performed well in our cohort with an accuracy of 84.1%, and predicted the BCG-unresponsive patients with progression precisely with accuracy of 100%. After TURBT, immediate RC instead of intravesical BCG may significantly improve the outcomes of BCG-unresponsive patients with progression, which highlight the function of the TRPM. By combining the ERPM and the TRPM, we aim to create a novel system that can stratify the recurrence risk and instruct the disease management of NMIBC patients. The workflow of our system is shown in Fig. 5C. The ERPM will screen for the patients with a low risk of recurrence, labelled as the "low risk" group. The rest patients will be subsequently classified by the TRPM into "high risk" and "very high risk" groups. Guided by our novel risk stratification system, low-risk patients can be offered moderate follow-up plans, while high-risk patients should receive BCG therapy in addition. Very high-risk patients are predicted to be BCG-unresponsive and need more aggressive treatment to prevent the recurrence as well as progression, such as RC or enrolment in clinical trials assessing new treatment strategies.[6] Frequent follow-up monitoring is also needed to detect the recurrence early and re-plan the therapy in time. However, our system hasn't considered the risk of progression, which need further study.

Our study has some limitations. First, this study was retrospective and included cohorts from 5 centres in China. The inclusion and exclusion criteria were strictly implemented, which further reduced the number of cases, especially in EAU high- and very-high risk group. More cohorts from prospective studies or other ethnicities are needed to confirm the model's capability and generalizability. Second, we only chose three IHC markers (P53, Ki67, CK20) based on the reported association between these markers and relapse, as well as their routine use in clinic. However, other markers may also contribute to prediction, such as programmed death 1 (PD-1) and programmed death ligand 1 (PD-L1),[32] CK5/6 and CD44,[29] etc. Further study will discover the predictive role of other markers on the recurrence of NMIBC and integrate them into ERPM. Third, the explainability of our models needs further enhancement. Although we tried explaining the models by the biological meanings of IHCs, other key features and the interactions between the features remained unclear. Fourth, the risk of progression to muscle-invasive BCa is another important factor guiding postoperative management of NMIBC. In the future, we will further improve the model to measure both the risk of recurrence and progression to achieve a more accurate stratification of patients.

In conclusion, we developed and validated a deep learning-based model for predicting early recurrence and therapy response of NMIBC, which has promising potential in improving postoperative management.

Contributors

SW, TL, FJ, GH, and HZ conceived and designed the study. FJ, YW, YX, RS, HY, SH and HW did the data collection. FJ, GH, MZ, RS, and YL screened patients and images. AX and YL curated the pathological examinations. FJ, ZL, RC and GH trained and developed the AI model. SW, FJ, and GH did the data analysis, data interpretation, and wrote the original draft manuscript. SW, TL, and FJ accessed and verified the underlying data. SW and TL provided study supervision. All authors reviewed the manuscript, approved the submitted version, had full access to all the data reported in the study, and had final responsibility for the decision to submit for publication.

Data sharing statement

To protect patient privacy, pathology image datasets and other patient-related data are not publicly accessible, but all data are available upon reasonable request emailed to the corresponding author. To gain access, data requestors will need to sign a data-access agreement.

Declaration of interests

We declare no competing interests.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (grant numbers U21A20383, 81825016, 82341018, and 82003151), the National Key Research and Development Programme of China (2018YFA0902803), the Science and Technology Planning Project of Guangdong Province (2023B1212060013, 2018B010109006), the Guangdong Provincial Clinical Research Centre for Urological Diseases (2020B1111170006), and the Science and Technology Projects in Guangzhou (2024A03J1190).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2025.103125.

References

1 Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229–263.

2 van Hoogstraten LMC, Vrieling A, van der Heijden AG, Kogevinas M, Richters A, Kiemeney LA. Global trends in the epidemiology of bladder cancer: challenges for public health and clinical practice. *Nat Rev Clin Oncol*. 2023;20(5):287–304.

3 Soukup V, Čapoun O, Cohen D, et al. Risk stratification tools and prognostic models in non-muscle-invasive bladder cancer: a critical assessment from the European association of Urology non-muscle-invasive bladder cancer guidelines panel. *Eur Urol Focus*. 2020;6(3):479–489.

4 Cambier S, Sylvester RJ, Collette L, et al. EORTC nomograms and risk groups for predicting recurrence, progression, and disease-specific and overall survival in non-muscle-invasive stage Ta-T1 urothelial bladder cancer patients treated with 1-3 Years of maintenance Bacillus Calmette-Guérin. *Eur Urol*. 2016;69(1):60–69.

5 Jeong SH, Han JH, Jeong CW, et al. Clinical determinants of recurrence in pTa bladder cancer following transurethral resection of bladder tumor. *BMC Cancer*. 2022;22(1):631.

6 Gontero P, Birtle A, Capoun O, et al. European Association of Urology guidelines on non-muscle-invasive bladder cancer (TaT1 and carcinoma in situ)-A summary of the 2024 guidelines update. *Eur Urol*. 2024;86(6):531–549.

7 Flaig TW, Spiess PE, Abern M, et al. NCCN Guidelines® insights: bladder cancer, version 2.2022. *J Natl Compr Canc Netw*. 2022;20(8):866–878.

8 Kamat AM, Sylvester RJ, Böhle A, et al. Definitions, end points, and clinical trial designs for non-muscle-invasive bladder cancer: recommendations from the international bladder cancer group. *J Clin Oncol*. 2016;34(16):1935–1944.

9 Li R, Tabayoyong WB, Guo CC, et al. Prognostic implication of the United States Food and Drug administration-defined BCG-unresponsive disease. *Eur Urol*. 2019;75(1):8–10.

10 Dyrskjøt L, Hansel DE, Efstathiou JA, et al. Bladder cancer. *Nat Rev Dis Primers*. 2023;9(1):58.

11 Chang SS, Boorjian SA, Chou R, et al. Diagnosis and treatment of non-muscle invasive bladder cancer: AUA/SUO guideline. *J Urol*. 2016;196(4):1021–1029.

12 Fernandez-Gomez J, Madero R, Solsona E, et al. Predicting non-muscle invasive bladder cancer recurrence and progression in patients treated with bacillus Calmette-Guerin: the CUETO scoring model. *J Urol*. 2009;182(5):2195–2203.

13 Sylvester RJ, van der Meijden AP, Oosterlinck W, et al. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *Eur Urol*. 2006;49(3):466–477.

14 Lucas M, Jansen I, van Leeuwen TG, Oddens JR, de Bruin DM, Marquering HA. Deep learning-based recurrence prediction in patients with non-muscle-invasive bladder cancer. *Eur Urol Focus*. 2022;8(1):165–172.

15 Jobczyk M, Stawiski K, Kaszkowiak M, et al. Deep learning-based recalibration of the CUETO and EORTC prediction tools for recurrence and progression of non-muscle-invasive bladder cancer. *Eur Urol Oncol*. 2022;5(1):109–112.

16 Ślusarczyk A, Garbas K, Pustuła P, Zapała Ł, Radziszewski P. Assessing the predictive accuracy of EORTC, CUETO and EAU risk stratification models for high-grade recurrence and progression after Bacillus Calmette-Guérin therapy in non-muscle-invasive bladder cancer. *Cancers*. 2024;16(9).

17 Jiang Y, Zhang Z, Yuan Q, et al. Predicting peritoneal recurrence and disease-free survival from CT images in gastric cancer with multitask deep learning: a retrospective study. *Lancet Digit Health*. 2022;4(5):e340–e350.

18 Wu S, Hong G, Xu A, et al. Artificial intelligence-based model for lymph node metastases detection on whole slide images in bladder cancer: a retrospective, multicentre, diagnostic study. *Lancet Oncol*. 2023;24(4):360–370.

19 Mi H, Bivalacqua TJ, Kates M, et al. Predictive models of response to neoadjuvant chemotherapy in muscle-invasive bladder cancer using nuclear morphology and tissue architecture. *Cell Rep Med*. 2021;2(9):100382.

20 Lee J, Choo MS, Yoo S, Cho MC, Son H, Jeong H. Intravesical prostatic protrusion and prognosis of non-muscle invasive bladder cancer: analysis of long-term data over 5 Years with machine-learning algorithms. *J Clin Med*. 2021;10(18).

21 Tokuyama N, Saito A, Muraoka R, et al. Prediction of non-muscle invasive bladder cancer recurrence using machine learning of quantitative nuclear features. *Mod Pathol*. 2022;35(4):533–538.

22 Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med*. 2021;13(1):152.

23 Bertz S, Otto W, Denzinger S, et al. Combination of CK20 and Ki-67 immunostaining analysis predicts recurrence, progression, and cancer-specific survival in pT1 urothelial bladder cancer. *Eur Urol*. 2014;65(1):218–226.

24 George B, Datar RH, Wu L, et al. p53 gene and protein status: the role of p53 alterations in predicting outcome in patients with bladder cancer. *J Clin Oncol*. 2007;25(34):5352–5358.

25 Roumiguié M, Kamat AM, Bivalacqua TJ, et al. International bladder cancer group consensus statement on clinical trial design for patients with Bacillus Calmette-Guérin-exposed high-risk non-muscle-invasive bladder cancer. *Eur Urol*. 2022;82(1):34–46.

26 Li R, Hensley PJ, Gupta S, et al. Bladder-sparing therapy for Bacillus Calmette-Guérin-unresponsive non-muscle-invasive bladder cancer: international bladder cancer group recommendations for optimal sequencing and patient selection. *Eur Urol*. 2024;86(6):516–527.

27 Compérat E, Amin MB, Cathomas R, et al. Current best practice for bladder cancer: a narrative review of diagnostics and treatments. *Lancet*. 2022;400(10364):1712–1721.

28 Pi J, Xiong Y, Liu C, et al. A nomogram model to predict recurrence of non-muscle invasive bladder urothelial carcinoma after resection based on clinical parameters and immunohistochemical markers. *J Invest Surg*. 2022;35(5):1186–1194.

29 Jung M, Kim B, Moon KC. Immunohistochemistry of cytokeratin (CK) 5/6, CD44 and CK20 as prognostic biomarkers of non-muscle-invasive papillary upper tract urothelial carcinoma. *Histopathology*. 2019;74(3):483–493.

30 Foersch S, Glasner C, Woerl AC, et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat Med*. 2023;29(2):430–439.

31 Lotan Y, Krishna V, Abuzeid WM, et al. Predicting response to intravesical BCG in high-risk NMIBC using an artificial intelligence-powered pathology assay: development and validation in an international 12-center cohort. *J Urol*. 2024. https://doi.org/10.1097/JU.0000000000004278.

32 Taber A, Prip F, Lamy P, et al. Immune contexture and differentiation features predict outcome in bladder cancer. *Eur Urol Oncol*. 2022;5(2):203–213.