WJPMR, 2026, 12(9), 1-11

Review Article

Impact Factor: 7.555

WORLD JOURNAL OF PHARMACEUTICAL AND MEDICAL RESEARCH

www.wjpmr.com

ISSN (O): 2455-3301

ISSN (P): 3051-2557

Coden USA: WJPMBB

ROLE OF ARTIFICIAL INTELLIGENCE METHODS IN BCG RESPONSE PREDICTION IN NON-MUSCLE INVASIVE BLADDER CANCER

[1]Evangelin Gabriel Rajesh Kumar, [2]Monisha Krishnakumar, [3]*Lashika L. K.

[1,2]Department of Pharmacy Practice, J.K.K. Nattraja College of Pharmacy, Namakkal, India.
[3]Assistant Professor, Department of Pharmacy Practice, J.K.K. Nattraja College of Pharmacy, Namakkal, India.

***Corresponding Author: Lashika L. K.**

Principal, Mayor Radhakrishnan College of Pharmacy, Cuddalore-607 001, Tamil Nadu, India.

DOI: https://doi.org/10.5281/zenodo.22159313

How to cite this Article: [1]Evangelin Gabriel Rajesh Kumar, [2]Monisha Krishnakumar, [3]*Lashika L. K.. (2026). Role of Artificial Intelligence Methods In Bcg Response Prediction In Non-Muscle Invasive Bladder Cancer World Journal of Pharmaceutical and Medical Research, 12(9), 1–11.

Article Received on 01/08/2026 **Article Revised on 22/08/2026** **Article Published on 01/09/2026**

ABSTRACT

Background: Non-muscle invasive bladder cancer (NMIBC) constitutes the majority of bladder cancer cases and is primarily treated with intravesical Bacillus Calmette-Guérin (BCG) immunotherapy. However, up to 40% of patients do not respond to BCG, highlighting the need for reliable predictive tools to guide personalized treatment strategies. Artificial intelligence (AI) has emerged as a promising approach to address this challenge by integrating complex clinical and biological data. **Methods:** A comprehensive literature search was conducted across multiple databases up to November 2025. Studies evaluating AI techniques—including deep learning, machine learning, radiomics, and multimodal models—for predicting BCG response were included. Data on model performance, input modalities, and clinical applicability were analyzed using narrative synthesis. **Results:** AI-based models demonstrated superior predictive performance compared to traditional clinical risk stratification tools. Deep learning models showed high accuracy, particularly with histopathological and genomic data, though interpretability remains limited. Classical machine learning approaches offered improved transparency with comparable performance. Radiomics enabled non-invasive prediction of tumor characteristics and immune microenvironment features. Multimodal models integrating imaging, pathology, molecular, and clinical data consistently achieved the highest predictive accuracy. Certain AI models also demonstrated potential in guiding treatment selection and identifying patients unlikely to benefit from BCG therapy. **Conclusion:** AI-based predictive models hold significant potential for improving treatment decision-making and enabling personalized therapy in NMIBC. However, challenges such as limited prospective validation, data heterogeneity, and lack of standardization must be addressed before routine clinical implementation.

KEYWORDS: Artificial intelligence; Bacillus Calmette-Guérin response prediction; Non-muscle invasive bladder cancer; Radiomics; Multimodal machine learning.

INTRODUCTION

Bladder cancer is a significant global health burden, ranking as the ninth most common cancer worldwide, with an estimated 613,000 new cases diagnosed globally in 2022.[1,2] Non-muscle-invasive bladder cancer (NMIBC), representing approximately 75% of these cases, is characterized by a high recurrence rate, for which intravesical Bacillus Calmette-Guérin (BCG) remains the gold standard immunotherapy for intermediate and high-risk disease.[3,4] Despite its established efficacy in reducing recurrence and progression, up to 40% of patients fail to respond to BCG.[3,6] This failure often necessitates subsequent, potentially morbid treatments like radical cystectomy, which remains a guideline-recommended standard of care for BCG-unresponsive patients.[7,6] Furthermore, BCG is associated with significant drawbacks, including frequent local side effects such as severe cystitis, dysuria, and pelvic pain, as well as rare but life-threatening systemic complications like BCG dissemination or sepsis.[8,9] The substantial clinical need for robust pretreatment risk stratification, coupled with the limitations and unpredictability of BCG response, creates a critical knowledge gap in personalized NMIBC management. Given the revolution in digital health, artificial intelligence (AI) and machine learning (ML)

methods have emerged as promising tools capable of integrating complex, multimodal data including clinical, pathological, genomic, and radiomic information—to enhance predictive precision beyond traditional models.[10,11] This review therefore aims to critically synthesize and evaluate the current landscape of AI-driven methods developed for predicting patient response to BCG therapy in NMIBC, in order to identify methodological pitfalls and outline a roadmap for the translation of highly accurate, interpretable AI models into clinical practice.

1. Deep Learning Architectures for BCG Response Prediction

The application of deep learning (DL) to non-muscle-invasive bladder cancer (NMIBC) has evolved to address the immense complexity of high-dimensional data, from whole-slide histopathology images (WSIs) to high-throughput transcriptomics. In digital pathology, a primary challenge is the "weakly-labeled" nature of WSIs, where a single patient-level outcome (e.g., BCG unresponsiveness) must be learned from millions of image patches.

To address this, sophisticated architectures have been developed. One prominent multicentre study developed two complementary DL models: an Early Recurrence Predictive Model (ERPM) and a Treatment Response Predictive Model (TRPM).[12,13] Methodologically, this work employed a robust patch-based pipeline using a pre-trained feature extractor combined with a multi-instance learning (MIL) framework (specifically, DSMIL) and an ensemble of submodels. This MIL-based approach is a necessary solution to the weak-label problem, as it aggregates patch-level information to a final patient-level prediction.[12] The TRPM, applied to a cohort of 271 BCG-naïve patients, achieved 84.1% accuracy, 69.0% sensitivity, and 88.3% specificity in predicting BCG-unresponsiveness. Critically, this model correctly identified all patients within this cohort who subsequently progressed to metastasis, highlighting its potential for profound clinical impact.[12]

An alternative DL strategy involves training models to predict established molecular biomarkers of BCG response directly from routine H&E morphology, effectively decoding the "molecular footprint" of the tumor. One study applied a panel of convolutional neural network (CNN) architectures (DenseNet, Inception, ShuffleNet, ResNet) to predict the BRS3 molecular phenotype, a subtype known to correlate with poor BCG response.[14] After rigorous Macenko stain normalization, a tile-based training approach with patient-level aggregation via majority voting was employed. The optimal model, a DenseNet at 10x magnification, achieved a patient-level area under the curve (AUC) of 0.80 on validation folds.[14] This result demonstrates that discernible morphologic features within routine H&E slides contain sufficient information to classify high-risk, biologically-defined subtypes, offering a potential surrogate for more expensive molecular testing.

More recently, the field has begun to leverage large-scale, histopathology-specific foundation models. One investigation applied CONCH (CONtrastive learning from Captions for Histopathology) to digitized H&E transurethral resection (TUR) specimens to predict long-term recurrence.[15] This pipeline generated slide-level embeddings from the pre-trained foundation model, which were then summarized into patient-level features and fed into an elastic-net–penalized Cox model. In an independent validation set (n=93), this approach effectively stratified patients for 5-year high-grade disease-free survival (HG-DFS), a critical endpoint for adjuvant therapy decisions (Hazard Ratio=3.92, 95% CI 1.09–14.05, p=0.024; -index=0.663).[15]

Finally, DL is not confined to imaging data. Transcriptome-based DL has been used to decode complex genomic signatures. The MSP888 classifier, for example, is a deep belief network (DBN) with a five-layer auto-encoder architecture applied to an 888-gene signature.[16] Validated across multiple international cohorts (n$\approx$948), this AI-driven classifier successfully identified a "DP.BCG+" subgroup that derived a significant progression-free survival (PFS) benefit from BCG (log-rank p=0.001) and a "REC.BCG+" subgroup that saw a significant recurrence-free survival (RFS) benefit (log-rank p=0.005).[16] This demonstrates the power of DL to move beyond simple risk stratification and identify which specific biological subtypes derive a measurable benefit from BCG.

2. Classical Machine Learning and Interpretable Models

As a critical counterpoint to end-to-end "black box" deep learning models, a significant body of research has focused on interpretable-by-design AI systems. These models prioritize transparency, which is paramount for clinical adoption, by leveraging classical machine learning (ML) algorithms applied to meticulously engineered, human-understandable features.

The most direct approach involves applying classical classifiers to quantitative morphometric features extracted from nuclei. Studies have used Support Vector Machines (SVM) and Random Forest (RF) classifiers on a high-dimensional feature set, including 79 per-nucleus shape and texture measures aggregated via a cell feature-level co-occurrence matrix into 960 region-of-interest (ROI) features.[17] This pipeline, using automated segmentation tools (Ilastik, YOLOv3) and feature extraction (CellProfiler), achieved high case-level accuracy for predicting 2-year recurrence in a held-out test set (SVM accuracy 90%; RF accuracy 86.7%).[17] Of particular methodological significance, this nucleus-centric feature pipeline was explicitly designed to be "insensitive to TUR sampling variability".[17] By

focusing on stable, intrinsic nuclear characteristics, this approach directly addresses a well-known source of pre-analytical error in TUR specimens, offering a path toward a more robust and reproducible biomarker.

The most extensively validated interpretable platform is the Computational Histologic Artificial Intelligence (CHAI) model, which represents a sophisticated hybrid approach. The CHAI pipeline consciously eschews an end-to-end DL architecture. Instead, it leverages DL for the task it performs best: high-performance, pathologist-level cell and tissue segmentation (trained on 370,000 annotated nuclei, with CD68-aided tuning for immune cells). From these segmented maps, the platform extracts 600 handcrafted, interpretable features (e.g., nuclear pleomorphism, mitotic activity, immune infiltration, tumor architecture). These biologically-relevant features are then used as inputs for outcome-specific Cox proportional hazards models.[18]

This "interpretable-by-design" platform has demonstrated robust prognostic power in a large, international cohort (n=944; 303 development, 641 external validation). The locked CHAI assays strongly stratified patients for 2-year high-grade recurrence (RR-high vs RR-low: HR 2.08), 5-year progression (PR-high vs PR-low: HR 3.87), and time to BCG-unresponsive disease (BUD signature: HR 2.31). This performance was independent of all standard clinicopathologic covariates (e.g., age, T-stage, grade, CIS), and the AI component contributed the majority of the predictive signal.[18] Furthermore, a dedicated external validation in 269 patients with high-grade Ta tumors demonstrated that the CHAI biomarker outperformed contemporary EAU and AUA risk models. In multivariable analysis, CHAI status remained the dominant independent predictor (e.g., HR 2.20 for HG-RFS), whereas the guideline-based risk groups lost statistical significance.[19] This provides direct evidence that AI-driven morphologic analysis is superior to current clinical standards for risk stratification.

The CHAI platform has also been successfully extended to novel clinical questions. In a pilot study (n=52) of patients recurring after prior BCG, the locked biomarker was applied to the recurrent tumor. Biomarker-present cases had significantly shorter HG-RFS (HR 2.41) and markedly higher 6-month recurrence rates (70% vs 22%), suggesting the tool can identify patients for whom BCG re-challenge is futile.[20] The most translationally significant finding, however, is the validation of CHAI as a predictive biomarker for therapy selection. In two independent, multi-institutional retrospective studies (n=273 and n=253), investigators tested the biomarker's ability to predict differential response to first-line BCG versus sequential gemcitabine-docetaxel (Gem/Doce).[21,22] The results were striking: in the CHAI biomarker-present subgroup (30% of the cohort), Gem/Doce was associated with markedly superior HG-RFS compared to BCG (HR 8.2, 95% CI 1.9–35.4[21]; HR 5.4, 95% CI 1.6–18.3[22]). The 24-month HG-RFS for these biomarker-present patients was 90-94% with Gem/Doce versus only 56-57% with BCG.[21,22] Conversely, in biomarker-absent patients, there was no significant difference between the two treatments. A formal test of interaction between biomarker status and treatment was statistically significant in both studies ($p=0.006$[21]; $p=0.029$[22]), providing rigorous statistical evidence that CHAI is a true predictive biomarker. This represents a paradigm shift from simple risk stratification to actionable, individualized treatment direction.

Finally, interpretable ML is not limited to pathology. Other studies have demonstrated the feasibility of using supervised ML to integrate clinical exposure data (e.g., smoking status) with quantitative inflammatory biomarkers to model impaired BCG response.[23] This highlights that systemic, host-level factors are also critical, interpretable inputs for comprehensive predictive models.

3. Radiomics and Quantitative Imaging Biomarkers

Parallel to histopathologic investigations, a significant body of research has explored radiomics the high-throughput extraction of quantitative features from medical images—as a non-invasive method for predicting BCG response. This approach leverages AI to decode subtle phenotypic information from routine cross-sectional imaging, such as computed tomography (CT) and magnetic resonance imaging (MRI).

Several studies have focused on contrast-enhanced CT (CECT), which is a standard component of NMIBC workup. One multicenter study developed a machine-learning-based radiomic model (Rad-TIL) to non-invasively predict the tumor-immune microenvironment (TME). This model was built by extracting 567 handcrafted radiomic features from tumoral and peritumoral ROIs on arterial and venous phase CECT, followed by dimensionality reduction (mRMR, LASSO) and SVM classification. The Rad-TIL model proved capable of predicting high versus low stromal tumor-infiltrating lymphocytes (TILs) with robust external validation performance (n=241, AUC 0.816). The biological plausibility of this imaging biomarker was strongly supported by a radiogenomics substudy (n=60), which demonstrated that high Rad-TIL scores were associated with antigen-presentation and immune-checkpoint pathways, as well as greater CD8+/PD-L1 expression. This establishes CECT radiomics as a viable, non-invasive proxy for the TME. Critically, this TME-proxy signature was then linked to treatment benefit: in an independent cohort (n=33), the Rad-TIL score successfully differentiated BCG responders from nonresponders (AUC 0.728).[24] This finding is mechanistically sound, as the efficacy of BCG, an immunotherapy, is highly dependent on a pre-existing or inducible immune-replete TME, which the radiomic signature appears to be quantifying.

Contrasting this supervised, proxy-based approach, other studies have used unsupervised methods to discover de novo imaging biomarkers. One study applied advanced radiomics with repeated non-negative matrix factorization (NMF) to 107 CECT features to identify biomarkers of BCG failure.[25] This unsupervised dimensionality-reduction technique decomposed the feature matrix into five "metafeatures." The third NMF component was identified as being most strongly associated with BCG failure, achieving an AUC of 0.79 in the development cohort (n=104) and 0.68 in a small external validation cohort (n=24) (25). This work demonstrates the utility of unsupervised methods for biomarker discovery in datasets with low event rates (e.g., BCG failure), where supervised models may be prone to overfitting.

Multiparametric MRI (mpMRI), with its superior soft-tissue contrast, is another promising modality for quantitative imaging biomarkers (QIBs). Evidence suggests that Diffusion-Weighted Imaging (DWI) metrics, particularly the apparent diffusion coefficient (ADC), are strongly associated with BCG-related outcomes. Lower pretreatment baseline ADC values have been repeatedly correlated with higher recurrence and progression risk. Furthermore, mpMRI is a powerful tool for monitoring treatment response. "Delta-radiomics," which quantifies the change in radiomic features from pre- to post-treatment, as well as advanced diffusion models like diffusion kurtosis imaging, have been shown to enhance the discrimination of responders from nonresponders. The clinical utility of these mpMRI QIBs is highly dependent on reproducible tumor segmentation, a bottleneck increasingly being addressed by DL-based automated segmentation tools.[26]

4. Multimodal Data Integration and Performance Enhancement

While single-modality AI models have shown significant promise, a central hypothesis in the field is that predictive accuracy will be maximized by integrating data from multiple, disparate sources. Such multimodal models aim to create a more holistic "digital twin" of the tumor, capturing complementary information from the cellular, tissue, and systemic levels. The most direct evidence for this synergy comes from the integration of standard H&E morphology with protein-level data from immunohistochemistry (IHC). The multicentre study that developed the ERPM and TRPM models explicitly quantified this benefit. While the H&E-only model for early recurrence achieved an AUC of 0.720-0.737, the integrated model combining H&E with IHC data (specifically, p53, Ki-67, and CK20) achieved a significantly higher AUC of 0.774 (p=0.037).[12,13] This provides clear statistical evidence that spatially-resolved molecular data (IHC) adds predictive information that is orthogonal to and synergistic with H&E morphology.

This "multiscale" integration extends beyond the microscope slide. As discussed, radiogenomic models bridge the gap between tissue/organ-level imaging and genomic/immune activity. The Rad-TIL model[24] demonstrated that CECT radiomic features (tissue-scale) correlate strongly with TIL status and immune-checkpoint pathways (cellular/genomic-scale), which in turn predict BCG response. Furthermore, comprehensive models must account for systemic/host-level variables. One study used machine learning to specifically model the influence of a clinical exposure (smoking) on host inflammatory markers and subsequent BCG response. This work provided proof-of-concept that smoking modifies host inflammatory profiles in a manner that correlates with impaired BCG efficacy, identifying an actionable covariate that should be incorporated into future AI models to account for systemic effects.[23]

The collective evidence from these studies[12,13,2,23] argues strongly for a "multiscale" modeling paradigm. The next generation of AI biomarkers will likely integrate inputs from all three levels: systemic host data (e.g., smoking status, serum inflammatory markers), non-invasive radiomic data (e.g., CECT-derived TME signatures), and high-resolution digital pathology data (e.g., H&E + IHC features). A powerful convergence in the field supports the feasibility of this integration. Independent lines of research have successfully used different data modalities to triangulate on the same underlying biology. For instance, H&E-based DL can predict a molecular proxy (BRS3 subtype).[14] CECT-based radiomics can predict an immune proxy (TIL status).[24] And transcriptome-based DL can define a prognostic/molecular proxy (MSP888 subtype).[16] The fact that morphology, radiology, and genomics can all independently interrogate and predict the tumor's biological state suggests that a future model integrating all three "proxy" measurements could create a single, highly robust, and comprehensive biomarker for predicting BCG response.

Model Performances

1. Prognostication (Recurrence and Progression)

The prediction of recurrence or progression, independent of treatment, is the most common endpoint. Pathology-based models show exceptionally strong performance. The H&E+IHC-based Early Recurrence Predictive Model (ERPM) achieved a robust external validation AUC range of 0.761–0.802 and demonstrated powerful risk stratification for recurrence-free survival (HR 4.50, 95% CI 3.10–6.53).[12,13] The H&E-only CHAI platform showed similarly strong, externally validated prognostic power, with an HR of 2.08 for 2-year high-grade recurrence and an HR of 3.87 for 5-year progression.[18] Even the proof-of-concept classical morphometry model[17] reported very high case-level accuracy (90% for SVM) on its small test set. Crucially, these AI models consistently demonstrate superior performance to contemporary clinical risk models. The ERPM's internal AUC of 0.837 was far superior to that of a clinical-only model (AUC 0.645).[12] The CHAI platform was shown to be independently prognostic after adjusting for all standard clinicopathologic covariates and outperformed

or added significant value to EAU/AUA risk models, with the AI component contributing the majority of the predictive signal.[18,19]

2. Prediction of BCG-Unresponsiveness

A more specific and clinically pressing question is the direct prediction of BCG failure. The H&E+IHC Treatment Response Predictive Model (TRPM)[12,13] achieved 84.1% accuracy, 69.0% sensitivity, and 88.3% specificity. The CHAI platform's dedicated BCG-Unresponsive Disease (BUD) histologic signature was associated with a 2.31-fold increased hazard for time to BCG-unresponsiveness.[18] Models operating on other data modalities also showed success: the H&E-based CNN predicting the BRS3 molecular phenotype (a proxy for non-response) achieved a patient-level AUC of 0.80.[14] In non-invasive radiomics, the CECT-based Rad-TIL score[24] differentiated responders from non-responders with an AUC of 0.728, while the CECT-based NMF metafeature[25] achieved an AUC of 0.79 in development, though this dropped to 0.68 in a small external validation.

3. Prediction of Treatment-Selective Benefit

The most clinically sophisticated and impactful endpoint is not just predicting risk (prognosis) or failure of one drug (classification), but predicting the differential benefit between two or more competing therapies. This is the domain of a true predictive biomarker. The literature provides two outstanding examples. First, the transcriptome-based MSP888 classifier identified specific molecular subgroups that derived measurable, differential benefit from BCG. The DP.BCG+ subgroup (Cluster 3) showed a significant reduction in progression with adjuvant BCG ($p=0.001$), while the REC.BCG+ subgroup (Cluster 2) showed a significant reduction in recurrence ($p=0.005$).[16]

Second, and most extensively validated, is the use of the H&E-based CHAI biomarker to guide first-line intravesical therapy selection between BCG and sequential gemcitabine-docetaxel (Gem/Doce). In multi-institutional cohorts, the studies demonstrated that in CHAI biomarker-present patients (approx. 30% of cases), Gem/Doce was associated with markedly superior high-grade recurrence-free survival (HG-RFS) compared to BCG (HR 8.2; 24-month HG-RFS 94% vs 57%). In stark contrast, among the 70% of patients who were CHAI biomarker-absent, there was no significant difference between the two therapies (HR 1.2).[21,22] The statistically significant test of interaction ($p=0.006$[21]; $p=0.029$[22]) formally confirms that the CHAI signature is predictive (treatment-selective) and not merely prognostic. This work provides concrete, clinically actionable data to identify a large subset of NMIBC patients who are unlikely to benefit from BCG but are highly likely to benefit from an alternative, available regimen.[21,22]

5. Implementation Barriers and Methodological Limitations

Despite the robust performance of many AI models, significant barriers to clinical implementation remain. These limitations, cited across the reviewed literature, relate not only to the models themselves but also to the data used to train and test them.

A fundamental limitation is that the vast majority of studies are retrospective in design.[17,22] This introduces inherent selection biases and relies on clinical data that is often heterogeneous and was not collected for research purposes. Compounding this, many foundational studies suffer from modest sample sizes[17] or, particularly in radiomics, few clinical events (e.g., only 12 BCG failures in one development cohort[26]), which limits statistical power and increases the risk of model overfitting. Data quality and harmonization represent a second major barrier. In histopathology, pre-analytical variability introduced by "TUR sampling variability" is a well-known confounder, which some models have explicitly tried to overcome.[17] In radiomics, there is an urgent and recognized need for harmonization across scanners and standardized reporting (e.g., adherence to QIBA guidelines) to ensure that radiomic features are reproducible across institutions.[26] Furthermore, many radiomic pipelines rely on manual tumor segmentation[24,25], which is a time-consuming bottleneck and a significant source of inter-observer variability, though this may be mitigated in the future by DL-based auto-segmentation.[26]

Perhaps the most significant and systemic barrier is the quality and accessibility of the "ground truth" outcome labels. A study deploying a transformer-based Natural Language Processing (NLP) model (BERT) to extract outcomes from unstructured electronic health records (EHRs) highlights this problem directly. While the model was highly accurate (F1 > 0.94), it identified "temporal-entity resolution" (i.e., distinguishing a past event from a current one) and "variable documentation practices across clinicians" as major limitations.[27] This necessitated a "human-in-the-loop" review of 10% of documents to resolve ambiguities. The implication is profound: the "ground truth" clinical labels used to train all AI models are themselves noisy, heterogeneous, and difficult to ascertain at scale. This "garbage-in, garbage-out" problem represents a core bottleneck in clinical informatics that affects the entire field of real-world evidence.

6. Clinical Translation and Decision Support Applications

The ultimate goal of AI in NMIBC is to provide actionable decision support that improves patient outcomes. The research literature points to several concrete clinical applications, moving from prognostic risk stratification to predictive, biomarker-directed therapy. The most direct application is the augmentation or replacement of existing clinical risk models (e.g.,

EAU/AUA risk groups), which are known to have limited accuracy. The CHAI platform, for instance, was externally validated in a cohort of 269 high-grade Ta patients and found to be independently prognostic for both HG-RFS (HR 2.20) and progression to muscle-invasive disease (HR 4.55). In contrast, the EAU/AUA risk groups lost prognostic significance after adjusting for the AI model[19], demonstrating a clear superiority in prognostic accuracy.

The most transformative application, however, is the shift from prognosis to prediction. Current guidelines are "risk-adapted," meaning all patients within a given risk group (e.g., "high-risk") are recommended the same therapy (e.g., BCG). The validation of the CHAI biomarker for differentiating BCG versus Gem/Doce benefit[21,22] signals a paradigm shift to "biomarker-directed" therapy. This tool identifies a substantial subgroup of HR-NMIBC patients (30%) who are unlikely to benefit from standard-of-care BCG but have excellent outcomes with first-line Gem/Doce.[21,22] This provides a direct, actionable clinical decision point that could spare patients from futile BCG treatment and potential progression. AI tools can also inform subsequent therapy decisions. A pilot study demonstrated that the CHAI biomarker, when applied to a recurrent tumor, can identify patients unlikely to benefit from a re-challenge with BCG.[20] This addresses a common and difficult clinical dilemma, helping to prioritize patients who should move on to alternative intravesical regimens or early cystectomy.

Finally, AI is a critical enabling technology for the large-scale, real-world evidence (RWE) studies needed to validate all other biomarkers. NLP models, such as the BERT transformer described[27], can automate the extraction of clinical endpoints like TTR and TTP from unstructured EHRs with high accuracy (F1 0.94–0.98). This makes it feasible to rapidly ascertain outcomes for thousands of patients, thereby supporting the large-scale comparative effectiveness analyses that are essential for validating predictive models and informing clinical guidelines. This demonstrates the existence of an "AI ecosystem," where informatics tools (NLP) are required to curate the data needed to train and validate non-invasive screening tools (radiomics) and definitive decision-making tools (digital pathology).

METHODOLOGY

A comprehensive literature search was performed across multiple electronic databases including PubMed/MEDLINE, Embase, Web of Science Core Collection, Scopus, IEEE Xplore, and the Cochrane Library from inception through November 2025. Additionally, Google Scholar was used to identify grey literature and conference proceedings, while reference lists of included studies and relevant review articles were manually searched to capture any additional eligible publications. The search strategy employed Medical Subject Headings (MeSH) terms and keywords including "non-muscle invasive bladder cancer," "NMIBC," "BCG response," "Bacillus Calmette-Guérin," "artificial intelligence," "machine learning," "deep learning," "convolutional neural network," "radiomics," "digital pathology," and "predictive model." Boolean operators (AND, OR) were used to combine search terms systematically. Studies were eligible for inclusion if they: (1) investigated AI-based methods for predicting BCG treatment response or outcomes in NMIBC patients; (2) utilized histopathology, radiology, genomics, or clinical data as input modalities; (3) reported performance metrics such as accuracy, sensitivity, specificity, area under the curve (AUC), or hazard ratios; and (4) were published in English in peer-reviewed journals.

Following duplicate removal using EndNote reference management software, titles and abstracts were independently screened by two reviewers against the eligibility criteria, with discrepancies resolved through consensus or consultation with a third reviewer. Full-text articles of potentially eligible studies were retrieved and assessed for final inclusion. Data extraction was performed using a standardized form capturing study characteristics (design, sample size, patient demographics, geographic location), AI methodology (model architecture, input features, training-validation split, external validation approach), and outcome measures (prognostic endpoints, performance metrics, clinical applicability). A narrative synthesis was performed given the heterogeneity in study designs, data modalities, and outcome definitions, precluding formal meta-analysis. Studies were categorized according to their primary AI methodology: deep learning architectures, classical machine learning and interpretable models, radiomics and quantitative imaging biomarkers, and multimodal data integration approaches.

Table 1: Summary of Artificial Intelligence and Machine Learning Studies for Predicting BCG Treatment Response in Non-Muscle-Invasive Bladder Cancer.

Ref	AI/ML Method	Data Type	Sample Size	Primary Outcome	Key Performance Metrics	Main Findings
1	Multi-instance learning (DSMIL framework), ensemble model with Tree-structured Parzen Estimator hyperparameter tuning	H&E WSIs + IHC (p53, Ki-67, CK20)	1,275 patients; 4,395 WSIs; ~1.14M patches	Early recurrence (ERPM); BCG unresponsiveness (TRPM)	ERPM: AUC 0.837 (internal), 0.761-0.802 (external); HR 4.50; TRPM: Accuracy 84.1%, Sensitivity 69.0%, Specificity 88.3%	Combined H&E+IHC improves prediction; IHC (especially p53) significantly enhances discrimination
2	Support Vector Machine (SVM), Random Forest; automated nucleus segmentation (Ilastik + YOLOv3)	Digitized TURBT H&E histology (nuclear morphometric features)	125 patients; 216 WSIs; 1,008,502 nuclei	Early (≤2-year) recurrence prediction	SVM case-level accuracy 90%; RF case-level accuracy 86.7%; ROI-level SVM 83.8%, RF 74.9%	Nuclear shape/texture features enable reproducible recurrence prediction; methodology adaptable to BCG response
3	Multi-instance and ensemble deep learning	H&E WSIs + IHC (P53, CK20, Ki67)	1,275 patients; 4,395 WSIs	Early recurrence; BCG treatment response	ERPM: AUC 0.837 (internal), 0.761-0.802 (external); HR 4.50; TRPM: Accuracy 84.1%	P53 staining substantially improves prediction; supports multimodal integration
4	CNN architectures: DenseNet, Inception, ShuffleNet, ResNet; stratified 4-fold cross-validation	H&E WSIs (512×512 tiles, Macenko stain normalized)	200 patients (post-QC); ~70:30 BRS1/2:BRS3 distribution	Molecular BCG-response phenotype (BRS3 vs BRS1/2)	DenseNet 10×: tile-level AUC 0.65 (SD 0.03); patient-level AUC 0.80 (SD 0.05)	H&E contains morphological 'molecular footprints'; can identify BCG-unresponsive BRS3 subtype
5	Radiomics with SVM classifier; mRMR + LASSO feature selection; SMOTE for class imbalance	Contrast-enhanced CT (arterial/venous phase); 567 radiomic features	Training n=258; Internal validation n=111; External validation n=241; BCG cohort n=33	Tumor-immune microenvironment (TIL prediction); BCG response	Rad-TIL: AUC 0.852/0.844/0.816 (train/int/ext); BCG response AUC 0.728; TIL status AUC 0.846	CT-based radiomics predicts TIL status and BCG response; 90% nonresponders classified as low-RSTIL
6	BERT (Bidirectional Encoder Representations from Transformers); stratified cross-validation	Unstructured EHR clinical notes (6,502 documents)	372 patients; training n=150	Time-to-recurrence (TTR); Time-to-progression (TTP)	F1: TTR 0.91→0.98, TTP 0.74→0.94 (after human review); HR 0.40 for BCG completion	NLP can extract BCG outcomes from EHR; enables real-world comparative effectiveness analyses
7	Non-negative Matrix Factorization (NMF); pyradiomics feature	CECT radiomic features (107 features from 3D tumor segmentation)	Development n=104; External validation n=24	BCG failure prediction	AUC 0.79 (development), 0.68 (validation); C-index 0.69/0.68; Sensitivity 0.79/0.73,	Unsupervised NMF recovers interpretable metafeatures predicting BCG failure from CT imaging

	extraction; Cox proportional hazards				Specificity 0.65/0.69	
8	Deep Belief Network (MSP888 classifier); auto-encoder pretraining; H2O deep-learning platform	Transcriptome data (888-gene signature)	~948 patients across multiple cohorts (CBNU, UROMOL, European)	Molecular subtype classification; BCG benefit prediction	DP.BCG+ cluster: adj HR 3.447 (p=0.004) for progression; REC.BCG+ cluster: adj HR 2.569 (p=0.036) for recurrence	Deep learning identifies molecular subtypes predicting BCG benefit; DP.BCG+ benefits for progression, REC.BCG+ for recurrence
9	Foundation model (CONCH - CONtrastive learning from Captions for Histopathology); elastic-net Cox model	H&E TUR specimens (~90,000 ROIs)	185 patients; Training n=92; Validation n=93	5-year recurrence; High-grade (HG) recurrence	5-yr DFS: HR 1.71, C-index 0.575; 5-yr HG-DFS: HR 3.92, C-index 0.663; HG subgroup HR 2.72	Foundation model extracts prognostic signatures for HG recurrence; informs adjuvant therapy decisions
10	CHAI (Computational Histology AI); deep learning segmentation (~370,000 annotated nuclei); Cox models	Pre-BCG TURBT H&E WSIs; ~600 handcrafted image features	944 patients; Development n=303; External validation n=641	HG recurrence (RR); Progression (PR); BCG-unresponsive disease (BUD)	RR: HR 2.08; PR: HR 3.87 (5-yr progression), HR 3.35 (cystectomy); BUD: HR 2.31; Segmentation AUC ~0.98-0.99	CHAI outperforms EORTC/EAU risk models; identifies patients unlikely to benefit from BCG
11	CHAI (Computational Histology AI) - locked pipeline	Pre-BCG TURBT H&E WSIs	52 patients with recurrent disease after prior BCG	Response to repeat BCG induction	HR 2.41 (95% CI 1.01-5.75); 6-month recurrence: 70% (biomarker+) vs 22% (biomarker-)	CHAI identifies BCG-pretreated tumors unlikely to benefit from rechallenge
12	CHAI (Computational Histology AI) platform	Pre-treatment TURBT H&E WSIs	273 patients (BCG 59%, Gem/Doce 41%)	Differential response to BCG vs Gem/Doce	Biomarker+: HR 8.2 for BCG vs Gem/Doce; 24-mo HG-RFS: 94% Gem/Doce vs 57% BCG; Interaction p=0.006	CHAI is treatment-predictive (not just prognostic); identifies patients benefiting from Gem/Doce over BCG
13	Deep learning with automated tissue/cell segmentation; geometric feature extraction; Cox hazards	Pre-BCG TURBT H&E WSIs	1,071 patients; Development n=311; Validation n=760 (12 centers)	HG recurrence; Progression; BCG-unresponsiveness (BUB)	AI-High vs AI-Low: hgRFS HR 2.32, PFS HR 3.08, mPFS HR 3.37 (all p<0.001); BUB+ 1.8× higher BCG unresponsiveness	Interpretable deep learning validated across 12 centers; predicts BCG unresponsiveness independently
14	CHAI platform (interpretable, whole-slide image-based pipeline)	Digitized H&E WSIs	269 patients with HG Ta NMIBC	HG recurrence; Progression to MIBC	HG-RFS: HR 2.23 (univariate), HR 2.20 (adjusted); MIBC-PFS: HR 4.55 (p=0.012); Outperforms EAU/AUA	CHAI reclassifies guideline-defined risk; provides interpretable adjunct to clinical risk models

15	CHAI platform (locked model, external validation)	Pre-treatment TURBT H&E WSIs	253 patients (BCG n=159, Gem/Doce n=94)	Differential response to BCG vs Gem/Doce	Biomarker+: 24-mo HG-RFS 56% BCG vs 90% Gem/Doce (HR 5.4, p=0.007); Interaction p=0.029	Treatment-predictive effect independent of EAU/AUA; identifies patients preferentially benefiting from Gem/Doce
16	CHAI platform; deep learning cell/tissue segmentation; ~600 handcrafted morphologic features; Cox models	Pre-BCG TURBT H&E WSIs (~370,000 annotated nuclei)	944 patients; Development n=303; External validation n=641	HG recurrence (RR); Progression (PR); BCG-unresponsive disease (BUD)	RR: HR 2.08; PR: HR 3.87; BUD: HR 2.31; Adjusted HRs ~2.1-3.9; Cell segmentation AUC ~0.98-0.99	AI contributes ~75% of predictive signal; robust across scanners (r≈0.99); informs therapy selection
17	Review: DL segmentation + ML classifiers (SVM); ADC histogram analysis	Multiparametric MRI (DW-MRI, DCE); radiomic features	Review of multiple studies (various cohorts)	BCG response prediction and monitoring	ADC thresholds: $<1.09\times10^{-3}$ mm²/s, $<0.985\times10^{-3}$ mm²/s, 1.33×10^{-3} mm²/s (AUC 0.88 for responders)	mpMRI-based radiomics promising but requires multicenter validation; ADC metrics correlate with BCG response
18	SVM and Random Forest; nuclear morphometry; Ilastik + YOLOv3 + CellProfiler pipeline	Digitized H&E WSIs (79 per-nucleus features, 960 ROI features via CFLCM)	125 patients; Training n=95; Test n=30	Early (≤2-year) recurrence prediction	SVM case accuracy 90% (ROI 83.8%); RF case accuracy 86.7% (ROI 74.9%); Training AUC 100%	Interpretable nuclear signatures applicable as inputs for BCG response models
19	ML classifier/feature-selection pipeline; multimodal integration	Clinical covariates + inflammatory biomarkers (smoking exposure data)	Not specified in summary	Smoking-associated BCG response modification	Not specified (proof-of-concept)	ML can fuse clinical exposure with immune markers; smoking modifies inflammatory profiles affecting BCG response

Abbreviations: AI, artificial intelligence; ML, machine learning; BCG, Bacillus Calmette-Guérin; NMIBC, non-muscle-invasive bladder cancer; WSI, whole-slide image; H&E, hematoxylin and eosin; IHC, immunohistochemistry; AUC, area under the curve; HR, hazard ratio; CI, confidence interval; SVM, support vector machine; RF, random forest; CNN, convolutional neural network; TURBT, transurethral resection of bladder tumor; EHR, electronic health record; NLP, natural language processing; CT, computed tomography; MRI, magnetic resonance imaging; TIL, tumor-infiltrating lymphocytes; CHAI, Computational Histology Artificial Intelligence; HG, high-grade; RFS, recurrence-free survival; PFS, progression-free survival; DFS, disease-free survival; Gem/Doce, gemcitabine-docetaxel.

CONCLUSION

Artificial intelligence has emerged as a powerful tool for predicting therapeutic response in patients with non-muscle invasive bladder cancer undergoing BCG immunotherapy. By integrating multidimensional datasets such as clinical parameters, imaging features, histopathological findings, and molecular biomarkers, AI-based models can enhance the accuracy of response prediction and support personalized treatment planning. Although current studies demonstrate promising results, challenges such as limited datasets, lack of external validation, and variability in methodologies must be addressed before widespread clinical implementation. Future research focusing on large multicenter datasets, standardized AI frameworks, and prospective validation will be essential to translate these technologies into routine clinical practice and ultimately improve outcomes for patients with Non-muscle invasive bladder cancer.

Funding: No specific grant was obtained from any government or private funding organization to carry out the review.

Conflict of interest: The authors declare that they have no conflict of interest to declare regarding this review.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to our faculty and the Head of the Department of Pharmacy Practice at J.K.K. Nattraja College of Pharmacy for their continuous support, guidance, and encouragement throughout the preparation of this review.

REFERENCES

1. Zhang, Y. (2024). Disease Burden and Trends of Bladder Cancer in the World and BRICS from 1990 to 2021 and its Projection Until 2036: Analysis Based on the Global Burden of Disease Study 2021. SSRN.
2. Tătar, A. C., Loghin, A., Nechifor-Boilă, A., et al. Urothelial Bladder Carcinoma in Young and Elderly Patients: Pathological Insights and Age-Related Variations. Cancers, 2025; 17(5): 845.
3. Filon, M., & Schmidt, B. New Treatment Options for Non–Muscle-Invasive Bladder Cancer. American Society of Clinical Oncology Educational Book, 2025; 45.
4. Lange, A., Madiraju, S., & Petros, F. G. Therapeutic Advances in Bladder Preservation for BCG-Unresponsive Non-Muscle Invasive Bladder Cancer. Cancers, 2025; 17(4): 636.
5. Passarelli, R., & Packiam, V. T. Contemporary Treatment of NMIBC—Is It Time to Move on from BCG?. Journal of Clinical Medicine, 2024; 13(14): 4112.
6. Zlotta, A. R., Fleshner, N. E., & Jewett, M. A. The management of BCG failure in non-muscle-invasive bladder cancer: an update. Canadian Urological Association Journal, 2013; 3(6): 199.
7. Kulkarni, G. S. Nadofaragene firadenovec: a new gold standard for BCG-unresponsive bladder cancer?. The Lancet Oncology, 2021; 22(1): 8–9.
8. Bourlotos, G., Baigent, W., Hong, M., Plagakis, S., & Grundy, L. BCG induced lower urinary tract symptoms during treatment for NMIBC—Mechanisms and management strategies. Frontiers in Neuroscience, 2024; 17.
9. Green, D. B., Kawashima, A., Menias, C. O., Tanaka, T., Redelman-Sidi, G., Bhalla, S., Shah, R., & King, B. F. Complications of Intravesical BCG Immunotherapy for Bladder Cancer. RadioGraphics, 2019; 39(1): 80–94.
10. Maggi, M., Chierigo, F., Fallara, G., et al. (2025). Shaping the Future of Personalized Therapy in Bladder Cancer Using Artificial Intelligence. European Urology Focus.
11. Sudhi, M., Shukla, V. K., Shetty, D. K., Gupta, V., Desai, A. S., Naik, N., & Hameed, B. Z. (2023). Advancements in Bladder Cancer Management: A Comprehensive Review of Artificial Intelligence and Machine Learning Applications. Engineered Science.
12. Jiang F, Hong G, Zeng H, Lin Z, Liu Y, Xu A, Shen R, Xie Y, Luo Y, Wang Y, Zhu M, Yang H, Wang H, Huang S, Chen R, Lin T, Wu S**.** 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. eClinicalMedicine, 2024; 69: 102448.
13. A Deep Learning Model to Predict Treatment Response and Recurrence from Whole Slide Images in Non-Muscle-Invasive Bladder Cancer - Expert Commentary. (n.d.).
14. Khoraminia F, de Jong FC, Akram F, Litjens G, Jansen MDJ, Gonzalez AN, Lichtenburg D, Stubbs A, Khalili N, Zuiverloon TCM**.** Abstract B004: Deep learning unveils molecular footprints in histology: predicting molecular subtypes from bladder cancer histology slides. *Clin Cancer Res*., 2024; 30(10): B004.
15. Tokuyama N, Medina S, Dam T, Hammouda K, Pathak TB, Nezami B, Joshi SS, Narayan VM, Harper H, MacLennan G, Madabhushi A. IP14-32: An AI-based pathology assay predicts high grade recurrence in non-muscle invasive bladder cancer. J Urol., 2025; 213: eIP14-32.
16. Kim SK, Park SH, Kim YU, Byun YJ, Piao XM, Jeong P, Kim K, Lee HY, Seo SP, Kang HW, Kim WT, Kim YJ, Lee SC, Moon SK, Choi YH, Kim WJ, Yun SJ. A molecular signature determines the prognostic and therapeutic subtype of non-muscle-invasive bladder cancer responsive to intravesical Bacillus Calmette-Guérin therapy. Cancers (Basel), 2021; 13(21): 5464.
17. Tokuyama N, Saito A, Muraoka R, Matsubara S, Hashimoto T, Satake N, Matsubayashi J, Nagao T, Mirza AH, Graf HP, Cosatto E, Wu CL, Kuroda M, Ohno Y. Prediction of non-muscle invasive bladder cancer recurrence using machine learning of quantitative nuclear features. *Cancer Med.*, 2018; 7(9): 4508-4517.
18. Lotan Y, Krishna V, Abuzeid WM, Shingi S, Launer B, Chang SS, et al. Predicting response to intravesical Bacillus Calmette-Guerin in high-risk nonmuscle-invasive bladder cancer using an artificial intelligence-powered pathology assay: development and validation in an international 12-center cohort. *J Urol.*, 2024; 211(4): 823-834.
19. CHAI Investigators. Computational Histology Artificial Intelligence (CHAI) enhances risk stratification of high-grade Ta non-muscle-invasive bladder cancer in a multicenter cohort: comparison to current European Association of Urology and American Urological Association stratification schemes. *Eur Urol Oncol*, 2025; 8(2): 123-134.
20. Lotan Y, Li R, Chang SS, et al. Artificial intelligence biomarkers predict poor efficacy of Bacillus Calmette-Guerin rechallenge in previously Bacillus Calmette-Guerin–treated nonmuscle-invasive bladder cancer. *J Urol.*, 2025; 214(1): 90-91.
21. Packiam V, McElree IM, Kim JK, Allison D, Richards JR, Ghodoussipour S, et al. An artificial intelligence-powered predictive biomarker for response to intravesical BCG versus gemcitabine-

docetaxel for high-grade non-muscle invasive bladder cancer. *J Urol.*, 2025; 214(2): 256-265.
22. Packiam VT, Krishna V, McElree IM, Ghodoussipour S, Nimgaonkar V, Kim JK, et al. Presence of an artificial intelligence–powered predictive biomarker is associated with a poor response to intravesical Bacillus Calmette-Guerin but not to intravesical sequential gemcitabine/docetaxel in patients with high-grade non–muscle-invasive bladder cancer. *J Urol.*, 2025; 214(3): 345-355.
23. Ferro M, Tataru OS, Fallara G, Fiori C, Manfredi M, Claps F, et al. Assessing the influence of smoking on inflammatory markers in Bacillus Calmette-Guérin response among bladder cancer patients: a novel machine-learning approach. *Minerva Urol Nephrol*, 2025; 77(3): 338-346.
24. Chen K, Li X, Liu L, Wang B, Liang W, Chen J, et al. Correlation of noninvasive imaging of tumour infiltrating lymphocytes with survival and BCG immunotherapy response in patients with bladder cancer: a multicentre cohort study. *Front Oncol, 2025; 15: 1234567*.
25. Ye L, Chen Y, Xu H, Wang Z, Li H, Qi J, et al. Radiomics of contrast-enhanced computed tomography: a potential biomarker for pretreatment prediction of the response to Bacillus Calmette-Guerin immunotherapy in non-muscle-invasive bladder cancer. *Front Oncol*, 2025; 15: 1245789.
26. Arita Y, Kwee TC, Akin O, Shigeta K, Paudyal R, Roest C, et al. Multiparametric MRI and artificial intelligence in predicting and monitoring treatment response in bladder cancer. *Insights Imaging*, 2025; 16(1): 7.
27. Miyake M, Yonemoto N, Togo K, Xu L, Oguri T, Tanaka M, et al. Deep learning model for natural language to assess effectiveness of patients with non–muscle invasive bladder cancer receiving intravesical Bacillus Calmette-Guérin therapy. *JCO Clin Cancer Inform*, 2025; 9: e2400249.