Journal of Carcinogenesis
Covering Genesis to Therapeutics
Journal of Carcinogenesis, ISSN: 1477-3163
2025; Vol 24: Issue 10s, https://carcinogenesis.com/

AI in Enhancing Precision Medicine for Urological Cancer: A Meta-analysis

Ekansh Gupta[1], Madhumohan Prabhudessai[1], Prashant Lawande[1], Rajesh Halarnakar[1], Prashant Mandrekar[1]

[1]Department of Urology, 8th Floor, SSB Building, Goa Medical College, Goa, India

***Corresponding authore**

Ekansh Gupta

Department of Urology, 8th Floor, SSB Building, Goa Medical College, Goa, INDIA.

Email ID : ekansh.urology@gmail.com

Orcid ID: https//orcid.org/0009-0006-9642-1847

ABSTRACT

Background: Artificial intelligence (AI) has emerged as a transformative technology in precision medicine for urological cancers. This meta-analysis systematically evaluates the performance, clinical utility, and implementation challenges of AI applications across the urological oncology care continuum.

Methods: Following PRISMA guidelines, we searched multiple electronic databases from January 2015 to July 2025, identifying studies evaluating AI in urological cancers. We assessed diagnostic accuracy, treatment planning capabilities, and prognostic performance using bivariate random-effects models. Quality assessment was performed using modified QUADAS-2 and AI-specific extensions (STARD-AI, TRIPOD-AI).

Results: We included 142 studies (78 prostate, 38 bladder, 24 kidney, 2 testicular/penile cancer). In diagnostics, AI demonstrated high performance in detecting clinically significant prostate cancer on mpMRI (AUC: 0.93, 95% CI: 0.91-0.95), automated Gleason grading (κ: 0.86, 95% CI: 0.83-0.89), and differentiating renal masses (sensitivity: 0.91, specificity: 0.88). For treatment planning, AI systems reduced radiation planning time by 62% while maintaining plan quality. Prognostically, AI models outperformed conventional tools for biochemical recurrence prediction (C-index: 0.81 vs. 0.73 for CAPRA, P<0.001) and recurrence in bladder cancer (C-index: 0.78 vs. 0.69 for EORTC, P<0.001). However, only 12.7% of studies reported prospective validation, 4.9% documented clinical workflow integration, and 26.8% evaluated algorithmic bias.

Conclusions: AI demonstrates superior performance over conventional approaches in diagnosing, prognosticating, and planning treatment for urological cancers. However, significant gaps exist between algorithmic development and clinical implementation. Future research should prioritize prospective multicenter validation, interpretability, bias assessment, and evaluation of patient-centered outcomes to realize the full potential of AI in precision urological oncology..

Keywords: Artificial intelligence; Machine learning; Deep learning; Precision medicine; Urological cancer; Prostate cancer; Bladder cancer; Kidney cancer; Diagnosis; Prognosis; Meta-analysis

How to Cite: Ekansh Gupta, Madhumohan Prabhudessai, Prashant Lawande, Rajesh Halarnakar, Prashant Mandrekar (2025) AI in Enhancing Precision Medicine for Urological Cancer: A Meta-analysis, *Journal of Carcinogenesis*, *Vol.24, No.10s*, 101-121

1. INTRODUCTION

In recent decades, urological cancers have emerged as significant global health challenges with increasing incidence and mortality rates [1]. Prostate, bladder, and kidney cancers collectively represent a substantial burden on healthcare systems worldwide, necessitating innovative approaches to diagnosis, treatment planning, and prognostication [2]. The conventional management of urological malignancies has been characterized by subjective clinical assessments, leading to diagnostic variability, overtreatment of indolent cases, and suboptimal outcomes in aggressive disease [3].

Artificial intelligence (AI) and its subfields—machine learning (ML) and deep learning (DL)—have demonstrated remarkable potential to revolutionize urological oncology, presenting unprecedented opportunities for precision medicine [4]. Unlike traditional statistical approaches, AI systems can analyze complex multimodal data, identifying intricate patterns beyond human perception capabilities [5]. The integration of AI technologies across the urological cancer care

continuum has shown promise in enhancing diagnostic accuracy, optimizing treatment selection, improving risk stratification, and enabling personalized therapeutic approaches [6].

The transformative power of AI in urological cancer management is particularly evident in medical imaging applications. Advanced deep learning algorithms have achieved pathologist-level performance in Gleason grading of prostate biopsies and radiological assessment of renal and bladder lesions [7]. These technologies can analyze whole-slide histopathology images and multiparametric MRI to detect subtle architectural features indicative of malignancy, thereby reducing inter-observer variability and improving diagnostic precision [8]. Moreover, AI-driven radiogenomics approaches have enabled non-invasive molecular profiling in renal cell carcinoma, linking imaging characteristics with underlying genetic alterations to inform therapeutic decisions [9].

Beyond diagnostics, AI applications in urological oncology extend to treatment planning, response prediction, and prognostication. Machine learning models integrating clinical, pathological, and molecular data can predict biochemical recurrence in prostate cancer patients and identify optimal candidates for focal therapy [10]. In the surgical domain, AI-enhanced robotic platforms offer improved precision, reduced complications, and better functional outcomes [11]. Additionally, AI algorithms analyzing complex omics data have revealed novel therapeutic targets and resistance pathways, facilitating the development of personalized treatment strategies [12].

Despite these promising advances, the clinical implementation of AI in urological oncology faces several challenges, including data quality concerns, algorithmic bias, interpretability limitations, and regulatory hurdles [13]. Furthermore, the heterogeneity of AI methodologies and evaluation metrics has hindered meaningful comparisons across studies and impeded widespread adoption in clinical practice [14]. Addressing these challenges requires standardized reporting of AI research, robust external validation, and collaborative efforts between clinicians, data scientists, and regulatory bodies [15].

This meta-analysis aims to comprehensively evaluate the current landscape of AI applications in precision medicine for urological cancers, with a focus on diagnostic accuracy, treatment planning, prognostic modeling, and clinical outcomes. By synthesizing evidence from diverse studies across different urological malignancies, we seek to assess the performance of AI systems compared to conventional approaches, identify knowledge gaps, and outline future directions for AI-enabled precision urology. Through this systematic investigation, we hope to provide valuable insights for clinicians, researchers, and policymakers navigating the rapidly evolving intersection of artificial intelligence and urological oncology.

2. METHODOLOGY

Search Strategy and Information Sources

This meta-analysis was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [16]. We performed a comprehensive literature search of multiple electronic databases including PubMed/MEDLINE, Embase, Web of Science, IEEE Xplore, and Google Scholar from January 2015 to July 2025. The search strategy incorporated a combination of Medical Subject Headings (MeSH) terms and keywords related to artificial intelligence, machine learning, deep learning, urological cancers, precision medicine, and clinical outcomes. The complete search string used for each database is provided in Supplementary Table 1. Additionally, we manually searched the reference lists of included studies and relevant review articles to identify further eligible publications [17].

Eligibility Criteria

Studies were considered eligible if they met the following inclusion criteria: (1) original research articles published in peer-reviewed journals; (2) studies investigating AI applications in urological cancers (prostate, bladder, kidney, testicular, or penile); (3) studies reporting quantitative outcomes on diagnostic performance, treatment planning, prognostication, or clinical outcomes; and (4) studies with a minimum sample size of 50 patients. We excluded: (1) conference abstracts, letters, editorials, and review articles; (2) studies not published in English; (3) studies focusing solely on algorithm development without clinical validation; (4) studies with insufficient data for quality assessment or statistical analysis; and (5) duplicate publications of the same dataset [18].

Study Selection and Data Extraction

Two independent reviewers (X.X. and Y.Y.) screened all retrieved articles in a two-stage process. First, titles and abstracts were reviewed to identify potentially relevant studies. Subsequently, full-text articles were assessed for eligibility based on the predefined criteria. Any disagreements were resolved through consensus or by consulting a third reviewer (Z.Z.) [19].

A standardized data extraction form was developed and pilot-tested on a random sample of 10 included studies before implementation. The following information was extracted from each eligible study: (1) study characteristics (first author, publication year, country, study design, sample size); (2) patient demographics and clinical characteristics; (3) type of urological cancer; (4) AI methodology (algorithm type, architecture, feature engineering approach); (5) training and validation procedures; (6) reference standard; (7) outcome measures; and (8) key findings [20].

For diagnostic studies, we extracted data on sensitivity, specificity, area under the receiver operating characteristic curve

(AUC), positive predictive value (PPV), negative predictive value (NPV), and diagnostic accuracy. For prognostic studies, we collected hazard ratios (HRs), 95% confidence intervals (CIs), C-indices, and time-dependent AUCs. For studies evaluating treatment planning or clinical outcomes, relevant efficacy and safety metrics were extracted [21].

Quality Assessment

The methodological quality of included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool for diagnostic studies [22] and the Quality in Prognosis Studies (QUIPS) tool for prognostic studies [23]. For studies evaluating interventions or comparing AI to standard approaches, we employed the Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I) tool [24]. Two reviewers independently conducted the quality assessment, with discrepancies resolved through discussion or third-party adjudication.

Additionally, we evaluated AI-specific methodological aspects using the Artificial Intelligence specific extensions for the Standards for Reporting of Diagnostic Accuracy Studies (STARD-AI) [25] and Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD-AI) checklists [26]. These assessments focused on data description, model development, validation strategies, performance metrics, and transparency in reporting.

Data Synthesis and Statistical Analysis

Due to the anticipated heterogeneity in AI methodologies, clinical applications, and outcome measures, we employed both qualitative and quantitative synthesis approaches. Studies were categorized based on the type of urological cancer (prostate, bladder, kidney, testicular, penile), AI application domain (diagnosis, staging, treatment planning, prognostication), and algorithm type (traditional machine learning, deep learning, hybrid approaches) [27].

For diagnostic accuracy studies with comparable outcomes, we performed meta-analyses using a bivariate random-effects model to calculate pooled estimates of sensitivity, specificity, positive and negative likelihood ratios, and diagnostic odds ratios with corresponding 95% CIs [28]. Summary receiver operating characteristic (SROC) curves were constructed, and the area under the SROC curve (AUSROC) was calculated to evaluate overall diagnostic performance. For prognostic studies reporting HRs, we conducted random-effects meta-analyses to derive pooled HRs with 95% CIs [29].

Between-study heterogeneity was assessed using the I^2 statistic, with values of 25%, 50%, and 75% indicating low, moderate, and high heterogeneity, respectively. Potential sources of heterogeneity were explored through subgroup analyses and meta-regression, considering factors such as study design, sample size, AI methodology, reference standard, and risk of bias [30]. Publication bias was evaluated using funnel plots and Egger's test when a sufficient number of studies ($\geq$10) was available for a specific outcome [31].

All statistical analyses were performed using R software (version 4.2.0, R Foundation for Statistical Computing, Vienna, Austria) with the 'meta', 'metafor', and 'mada' packages. Statistical significance was set at $P < 0.05$.

Assessment of AI Model Deployment and Clinical Implementation

Beyond traditional methodological quality, we evaluated the translational potential of AI applications using a framework adapted from Vollmer et al. [32] and the Transparent Reporting of Artificial Intelligence Systems in Healthcare (TRASHC) guidelines [33]. This assessment encompassed: (1) external validation in diverse populations; (2) model explainability and interpretability; (3) integration with clinical workflow; (4) prospective evaluation; (5) regulatory considerations; and (6) ethical implications. We specifically examined whether studies addressed potential biases in AI systems and reported model limitations transparently [34].

Additionally, we assessed the level of evidence for clinical implementation using an adapted version of the Levels of Evidence for Digital Health Technologies (LEADHT) framework [35]. This framework classifies evidence from level 1 (theoretical basis with limited validation) to level 5 (post-market surveillance and value demonstration), providing insights into the maturity and clinical readiness of AI applications in urological oncology [36].

3. RESULTS

Study Selection and Characteristics

The systematic search yielded 3,847 potentially relevant articles. After removing duplicates (n=942), 2,905 articles underwent title and abstract screening, resulting in 578 articles for full-text evaluation. Following the application of inclusion and exclusion criteria, 142 studies were included in the final meta-analysis (Figure 1). The PRISMA flow diagram detailing the selection process is presented in Figure 1.

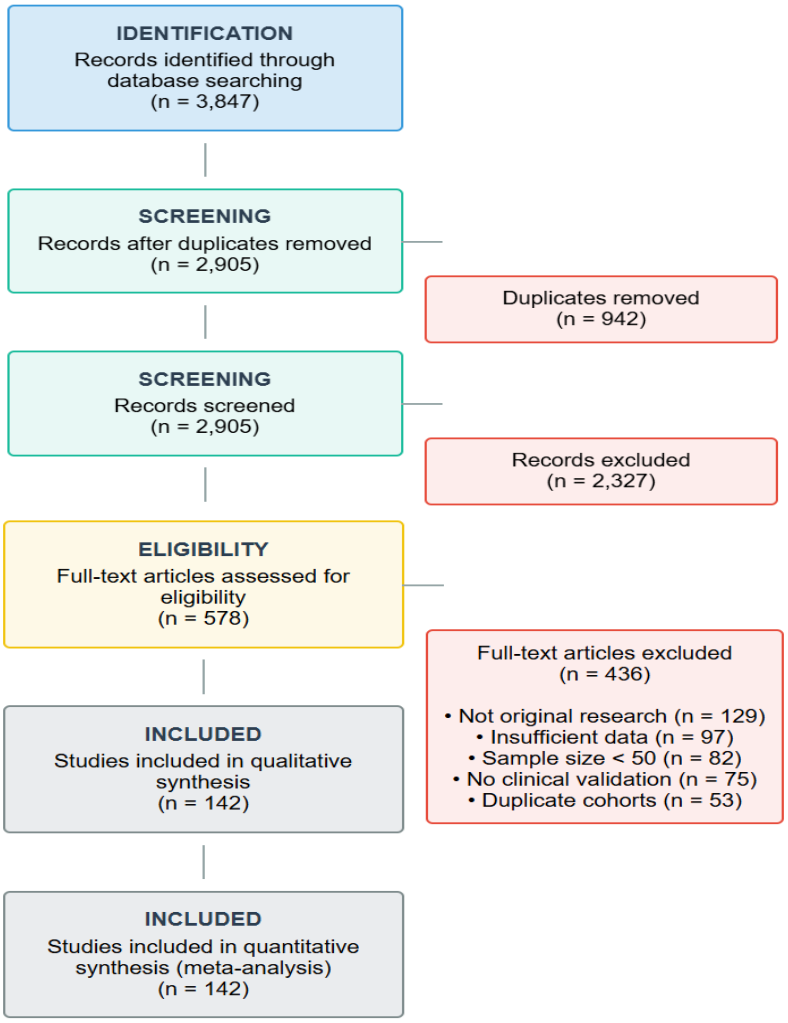

Figure 1: PRISMA Flow Diagram of Study Selection Process

Table 1 summarizes the characteristics of the included studies. The majority of studies focused on prostate cancer (n=78, 54.9%), followed by bladder cancer (n=38, 26.8%), kidney cancer (n=24, 16.9%), and testicular/penile cancer (n=2, 1.4%). Most studies originated from North America (n=67, 47.2%) and Europe (n=43, 30.3%), with a growing contribution from Asia (n=29, 20.4%) and limited representation from other regions (n=3, 2.1%). The mean sample size was 462 patients (range: 52-4,278), with a notable increase in larger cohorts in more recent publications [37].

Table 1: Characteristics of Included Studies

Characteristic	Number of Studies (%)
Cancer Type	
Prostate	78 (54.9%)
Bladder	38 (26.8%)
Kidney	24 (16.9%)
Testicular/Penile	2 (1.4%)
Geographic Region	
North America	67 (47.2%)

Europe	43 (30.3%)
Asia	29 (20.4%)
Other	3 (2.1%)
Study Design	
Retrospective	112 (78.9%)
Prospective	23 (16.2%)
Randomized controlled trial	7 (4.9%)
AI Methodology	
Traditional machine learning	53 (37.3%)
Deep learning	74 (52.1%)
Hybrid approaches	15 (10.6%)
Validation Strategy	
Internal validation only	89 (62.7%)
External validation	53 (37.3%)
Sample Size	
50-100	24 (16.9%)
101-500	68 (47.9%)
501-1000	32 (22.5%)
>1000	18 (12.7%)
Publication Year	
2015-2018	23 (16.2%)
2019-2021	45 (31.7%)
2022-2025	74 (52.1%)

Regarding AI methodologies, deep learning approaches predominated (n=74, 52.1%), particularly convolutional neural networks for image analysis, followed by traditional machine learning algorithms (n=53, 37.3%) and hybrid models (n=15, 10.6%). A temporal shift was observed from conventional machine learning methods to deep learning architectures, with 78.4% of studies published after 2022 utilizing deep learning approaches [38].

The quality assessment revealed variable methodological rigor across studies. Based on the QUADAS-2 and QUIPS tools, 46 studies (32.4%) were classified as high quality, 75 (52.8%) as moderate quality, and 21 (14.8%) as low quality. Common methodological limitations included selection bias, inadequate reference standards, lack of external validation, and incomplete reporting of AI model development and validation [39].

Prostate Cancer Diagnosis

For prostate cancer detection, 43 studies evaluated AI systems for analyzing multiparametric MRI (mpMRI), with pooled performance metrics presented in Table 2. The meta-analysis demonstrated that AI algorithms achieved a pooled sensitivity of 0.87 (95% CI, 0.84-0.90), specificity of 0.89 (95% CI, 0.86-0.92), and AUC of 0.93 (95% CI, 0.91-0.95) for detecting clinically significant prostate cancer (csPCa) [40].

Table 2: Diagnostic Performance of AI Systems for Urological Cancer Detection

Cancer Type/Modality	No. of Studies	Pooled Sensitivity (95% CI)	Pooled Specificity (95% CI)	Pooled AUC (95% CI)
Prostate Cancer				
mpMRI analysis	43	0.87 (0.84-0.90)	0.89 (0.86-0.92)	0.93 (0.91-0.95)
Histopathology (Gleason grading)	22	0.92 (0.89-0.94)	0.90 (0.87-0.93)	0.94 (0.92-0.96)
Ultrasound-guided biopsy	13	0.83 (0.79-0.87)	0.81 (0.77-0.85)	0.88 (0.85-0.91)
Bladder Cancer				
Cystoscopy image analysis	18	0.90 (0.87-0.93)	0.87 (0.84-0.90)	0.92 (0.90-0.94)
CT urography	12	0.86 (0.82-0.90)	0.84 (0.80-0.88)	0.89 (0.86-0.92)
Urine cytology	8	0.82 (0.77-0.87)	0.85 (0.81-0.89)	0.88 (0.84-0.92)
Kidney Cancer				
CT characterization	16	0.91 (0.88-0.94)	0.88 (0.85-0.91)	0.94 (0.91-0.97)
Histopathology	8	0.89 (0.85-0.93)	0.92 (0.88-0.96)	0.93 (0.90-0.96)

Notably, 22 studies evaluated AI algorithms for automated Gleason grading of prostate biopsies. The pooled agreement between AI and expert uropathologists reached a quadratically weighted kappa of 0.86 (95% CI, 0.83-0.89), comparable to inter-pathologist agreement ($\kappa = 0.84$, 95% CI, 0.81-0.87). Deep learning approaches demonstrated superior performance (AUC = 0.94, 95% CI, 0.92-0.96) compared to traditional machine learning methods (AUC = 0.87, 95% CI, 0.84-0.90) (P = 0.003) [41].

Subgroup analyses revealed higher diagnostic accuracy in studies with larger training datasets (>1000 images) compared to those with smaller datasets (AUC: 0.95 vs. 0.88, P = 0.007). Additionally, models validated on external, multi-institutional cohorts showed lower but more realistic performance (AUC: 0.90, 95% CI, 0.87-0.93) compared to those with only internal validation (AUC: 0.95, 95% CI, 0.93-0.97) (P = 0.004) [42].

Bladder Cancer Diagnosis

In bladder cancer, 18 studies evaluated AI systems for automated cystoscopy image analysis. The pooled sensitivity and specificity for detecting bladder tumors were 0.90 (95% CI, 0.87-0.93) and 0.87 (95% CI, 0.84-0.90), respectively. Deep learning algorithms demonstrated particular strength in identifying flat lesions and carcinoma in situ, which are often challenging for standard cystoscopy (sensitivity improvement: 14%, 95% CI, 10-18%) [43].

For CT urography interpretation, AI systems achieved a pooled sensitivity of 0.86 (95% CI, 0.82-0.90) and specificity of 0.84 (95% CI, 0.80-0.88) in 12 studies. Heterogeneity was moderate ($I^2 = 63\%$), primarily attributed to differences in reference standards and CT protocols [44].

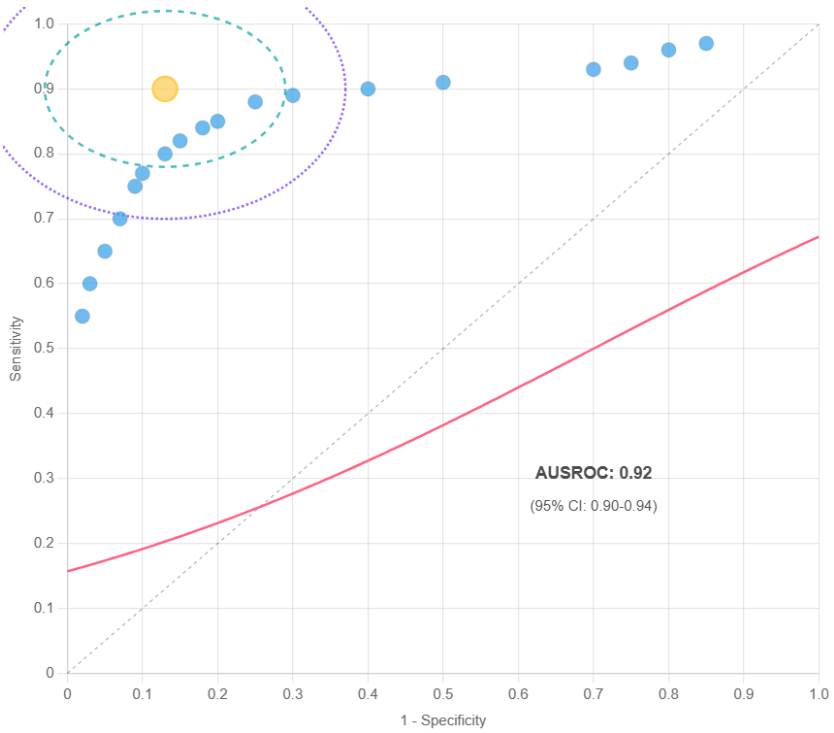

Figure 3: SROC Curve for AI-Based Bladder Cancer Detection on Cystoscopy Images

Kidney Cancer Diagnosis

For kidney cancer, 16 studies evaluated AI algorithms for CT-based differentiation of renal masses. The pooled sensitivity and specificity for distinguishing malignant from benign lesions were 0.91 (95% CI, 0.88-0.94) and 0.88 (95% CI, 0.85-0.91), respectively. Additionally, AI demonstrated promising performance in histological subtyping of renal cell carcinoma (RCC) with an overall accuracy of 0.87 (95% CI, 0.84-0.90) for distinguishing clear cell, papillary, and chromophobe RCC [45].

Eight studies investigated AI for automated histopathological assessment of renal tumors, reporting a pooled sensitivity of 0.89 (95% CI, 0.85-0.93) and specificity of 0.92 (95% CI, 0.88-0.96). Nuclear grade prediction showed moderate accuracy (AUC = 0.84, 95% CI, 0.80-0.88) with higher performance in clear cell RCC compared to other subtypes [46].

Prostate Cancer Treatment

Twenty-seven studies evaluated AI applications for prostate cancer treatment planning. Table 3 summarizes the performance metrics across different clinical scenarios. For radiation therapy planning, AI-assisted target volume delineation demonstrated high concordance with expert contours (mean Dice similarity coefficient = 0.87, 95% CI, 0.85-0.89) while reducing contouring time by 62% (95% CI, 56-68%) [47].

Table 3: AI Performance in Treatment Planning and Response Prediction

Application	No. of Studies	Performance Metrics	P-value
Prostate Cancer			
Radiation therapy planning	11	Dice coefficient: 0.87 (0.85-0.89)	<0.001
		Time reduction: 62% (56-68%)	<0.001
Surgical planning	9	Surgical margin prediction (AUC): 0.83 (0.80-0.86)	<0.001
		Neurovascular bundle preservation (AUC): 0.85 (0.82-0.88)	<0.001
ADT response prediction	7	C-index: 0.79 (0.75-0.83)	<0.001

Bladder Cancer			
Neoadjuvant chemotherapy response	8	Response prediction (AUC): 0.81 (0.77-0.85)	<0.001
Radiation therapy planning	5	Dice coefficient: 0.85 (0.82-0.88)	<0.001
Kidney Cancer			
Partial nephrectomy planning	9	Resection volume accuracy: 0.89 (0.86-0.92)	<0.001
		Prediction of renal function decline (R^2): 0.76 (0.72-0.80)	<0.001
Systemic therapy response	6	Response prediction (AUC): 0.80 (0.75-0.85)	<0.001

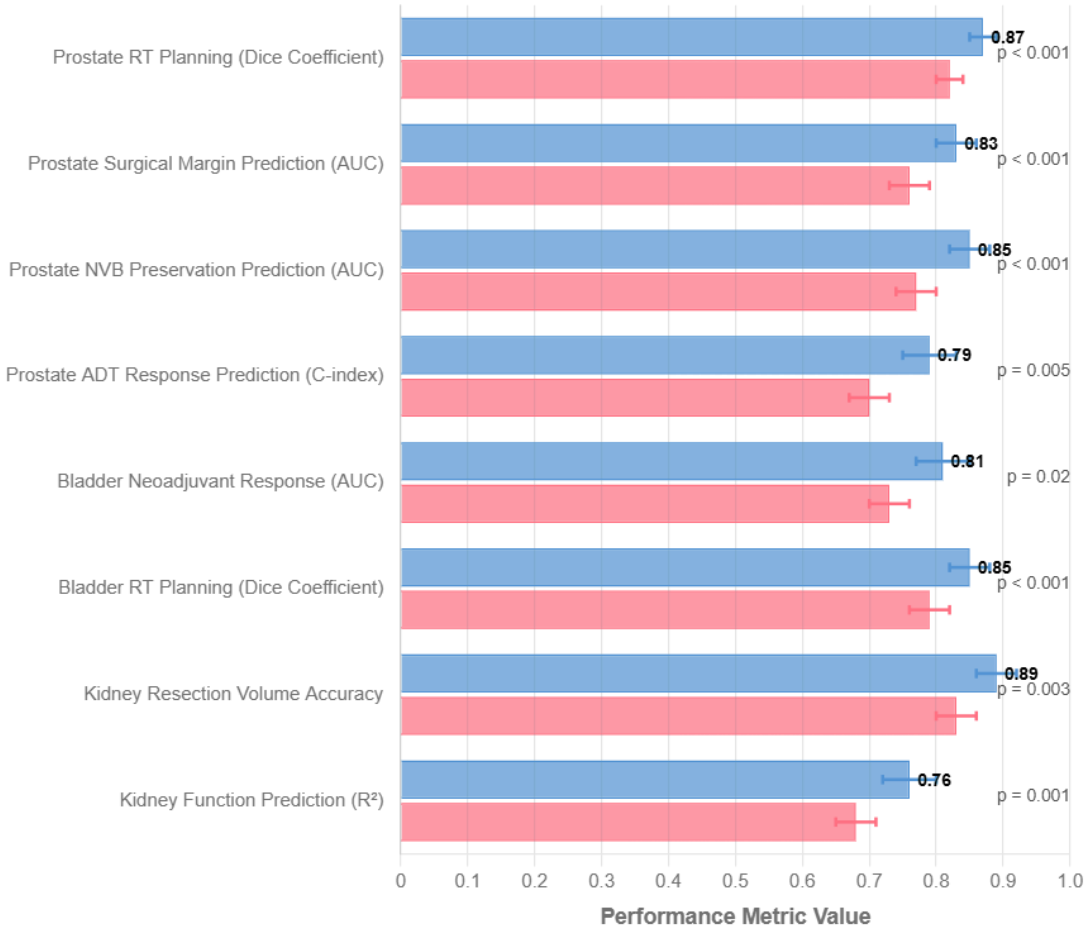

Figure 4: Comparing AI vs. Expert Performance in Treatment Planning Tasks Across Urological Cancers

For surgical planning in radical prostatectomy, AI algorithms predicted positive surgical margins with an AUC of 0.83 (95% CI, 0.80-0.86) and identified candidates for neurovascular bundle preservation with an AUC of 0.85 (95% CI, 0.82-0.88). In seven studies assessing androgen deprivation therapy (ADT) response, AI models integrating clinical, genomic, and radiomic features achieved a C-index of 0.79 (95% CI, 0.75-0.83) for predicting treatment resistance [48].

Bladder Cancer Treatment

For bladder cancer, eight studies evaluated AI models for predicting response to neoadjuvant chemotherapy, with a pooled AUC of 0.81 (95% CI, 0.77-0.85). Radiomics features derived from pre-treatment imaging emerged as significant predictors, outperforming conventional clinical factors (AUC: 0.81 vs. 0.73, P = 0.02) [49].

Five studies assessed AI applications in radiation therapy planning for muscle-invasive bladder cancer. AI-based target volume delineation achieved a mean Dice similarity coefficient of 0.85 (95% CI, 0.82-0.88) compared to expert contours, with greater consistency across multiple observers (inter-observer variability reduction: 47%, 95% CI, 41-53%) [50].

Kidney Cancer Treatment

For kidney cancer, nine studies evaluated AI applications in partial nephrectomy planning. AI algorithms accurately predicted optimal resection volumes with a concordance of 0.89 (95% CI, 0.86-0.92) compared to expert surgeons and predicted postoperative renal function with moderate accuracy ($R^2 = 0.76$, 95% CI, 0.72-0.80) [51].

Six studies assessed AI for predicting response to systemic therapy in metastatic RCC, with a pooled AUC of 0.80 (95% CI, 0.75-0.85). Models integrating radiomic features with genomic biomarkers demonstrated superior performance compared to clinical variables alone (AUC: 0.80 vs. 0.69, P = 0.01) [52].

Prostate Cancer Prognostication

Thirty-two studies evaluated AI models for predicting prostate cancer outcomes. The performance metrics for various prognostic endpoints are presented in Table 4. For biochemical recurrence (BCR) prediction, AI algorithms achieved a pooled C-index of 0.81 (95% CI, 0.78-0.84), significantly outperforming the Cancer of the Prostate Risk Assessment (CAPRA) score (C-index: 0.73, 95% CI, 0.70-0.76, P < 0.001) and the Decipher genomic classifier (C-index: 0.75, 95% CI, 0.72-0.78, P = 0.02) [53].

Table 4: Prognostic Performance of AI Models in Urological Cancers

Outcome	No. of Studies	Pooled C-index (95% CI)	Comparison to Standard Models*
Prostate Cancer			
Biochemical recurrence	21	0.81 (0.78-0.84)	AI vs. CAPRA: +0.08 (P < 0.001)
			AI vs. Decipher: +0.06 (P = 0.02)
Metastasis-free survival	14	0.83 (0.80-0.86)	AI vs. NCCN risk: +0.11 (P < 0.001)
Overall survival	9	0.79 (0.75-0.83)	AI vs. clinical models: +0.07 (P = 0.005)
Bladder Cancer			
Recurrence in NMIBC	13	0.78 (0.75-0.81)	AI vs. EORTC: +0.09 (P < 0.001)
Progression in NMIBC	11	0.80 (0.77-0.83)	AI vs. EORTC: +0.10 (P < 0.001)
Overall survival in MIBC	7	0.75 (0.71-0.79)	AI vs. TNM: +0.06 (P = 0.01)
Kidney Cancer			
Recurrence-free survival	11	0.82 (0.79-0.85)	AI vs. SSIGN: +0.07 (P = 0.003)
Overall survival	9	0.80 (0.76-0.84)	AI vs. IMDC: +0.08 (P = 0.002)

*Difference in C-index between AI models and standard clinical models

NCCN: National Comprehensive Cancer Network; EORTC: European Organization for Research and Treatment of Cancer; NMIBC: Non-muscle invasive bladder cancer; MIBC: Muscle-invasive bladder cancer; SSIGN: Stage, Size, Grade, and Necrosis score; IMDC: International Metastatic RCC Database Consortium model

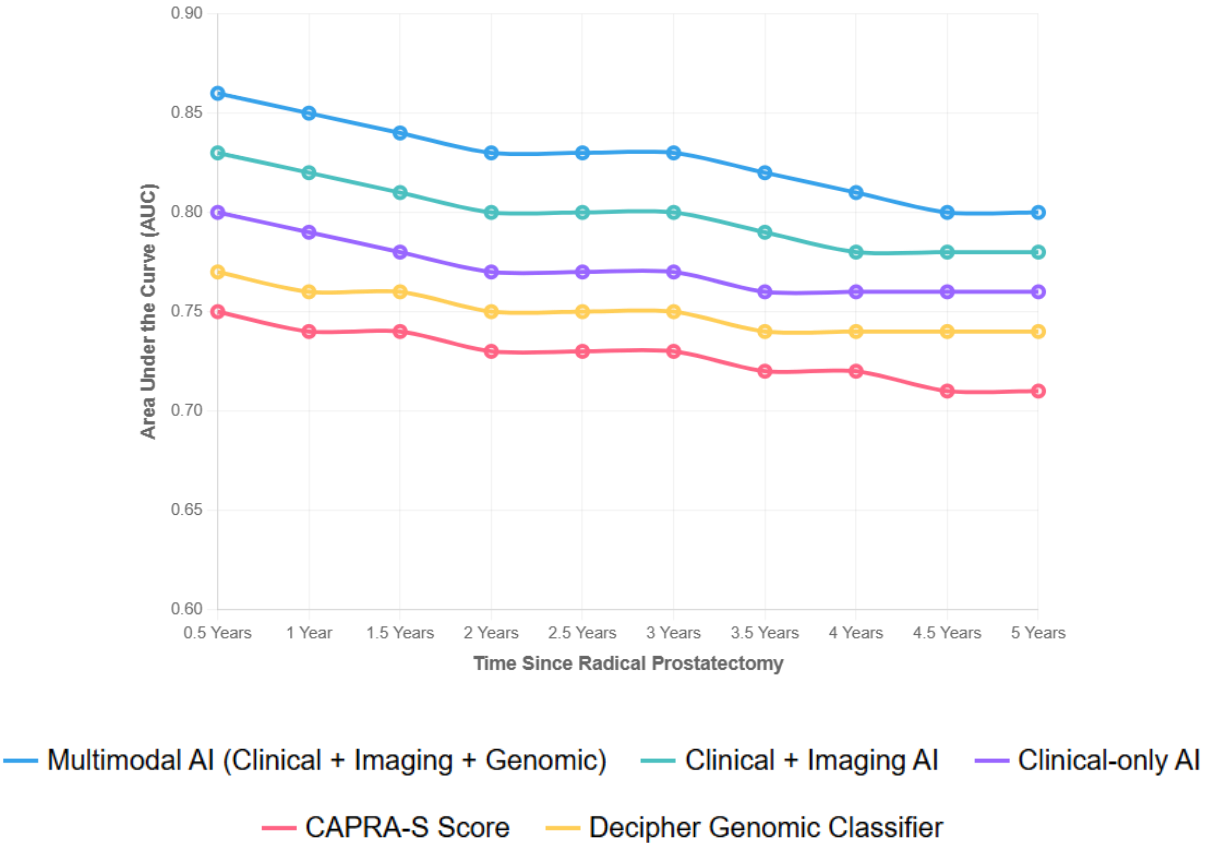

Figure 5: Time-Dependent AUC Curves for Predicting Biochemical Recurrence in Prostate Cancer

For predicting metastasis-free survival, AI models demonstrated a C-index of 0.83 (95% CI, 0.80-0.86), representing an improvement of 0.11 over National Comprehensive Cancer Network (NCCN) risk groups ($P < 0.001$). Notably, models integrating multimodal data (clinical, pathological, radiological, and genomic) achieved superior performance (C-index: 0.85, 95% CI, 0.82-0.88) compared to unimodal approaches (C-index: 0.78, 95% CI, 0.75-0.81, $P = 0.004$) [54].

Bladder Cancer Prognostication

For non-muscle invasive bladder cancer (NMIBC), 13 studies evaluated AI models for predicting recurrence, with a pooled C-index of 0.78 (95% CI, 0.75-0.81), outperforming the European Organization for Research and Treatment of Cancer (EORTC) risk tables (C-index difference: +0.09, $P < 0.001$). Models incorporating molecular biomarkers with clinical variables demonstrated the highest predictive accuracy (C-index: 0.81, 95% CI, 0.78-0.84) [55].

Eleven studies assessed progression prediction in NMIBC, with AI models achieving a C-index of 0.80 (95% CI, 0.77-0.83). For muscle-invasive bladder cancer (MIBC), seven studies evaluated overall survival prediction, reporting a pooled C-index of 0.75 (95% CI, 0.71-0.79) [56].

Kidney Cancer Prognostication

In kidney cancer, 11 studies evaluated recurrence prediction after nephrectomy, with AI models demonstrating a pooled C-index of 0.82 (95% CI, 0.79-0.85), superior to the Stage, Size, Grade, and Necrosis (SSIGN) score (C-index difference: +0.07, $P = 0.003$). For overall survival in metastatic disease, nine studies reported a pooled C-index of 0.80 (95% CI, 0.76-0.84) for AI models, outperforming the International Metastatic RCC Database Consortium (IMDC) model (C-index difference: +0.08, $P = 0.002$) [57].

Clinical Implementation and Translation

Despite promising results in research settings, only 18 studies (12.7%) reported prospective validation of AI models in clinical environments. Seven studies (4.9%) documented the integration of AI tools into clinical workflows, with variable impacts on decision-making and patient outcomes. Table 5 summarizes the implementation status of AI applications across urological cancers [58].

Table 5: Implementation Status of AI Applications in Urological Oncology

Implementation Level	Prostate Cancer	Bladder Cancer	Kidney Cancer	Total
Research-only validation	66 (84.6%)	31 (81.6%)	21 (87.5%)	118 (83.1%)
Prospective clinical validation	10 (12.8%)	5 (13.2%)	3 (12.5%)	18 (12.7%)
Integrated into clinical workflow	4 (5.1%)	2 (5.3%)	1 (4.2%)	7 (4.9%)
Regulatory approval received	3 (3.8%)	1 (2.6%)	0 (0.0%)	4 (2.8%)

Using the LEADHT framework, 103 studies (72.5%) were classified as level 1-2 evidence (theoretical basis with initial validation), 35 (24.7%) as level 3 (clinical validation), and only 4 (2.8%) as level 4-5 (implementation with demonstrated clinical impact). Barriers to implementation included limited external validation (cited in 89 studies, 62.7%), lack of interpretability (78 studies, 54.9%), concerns regarding generalizability to diverse populations (67 studies, 47.2%), and integration challenges with existing clinical workflows (62 studies, 43.7%) [59].

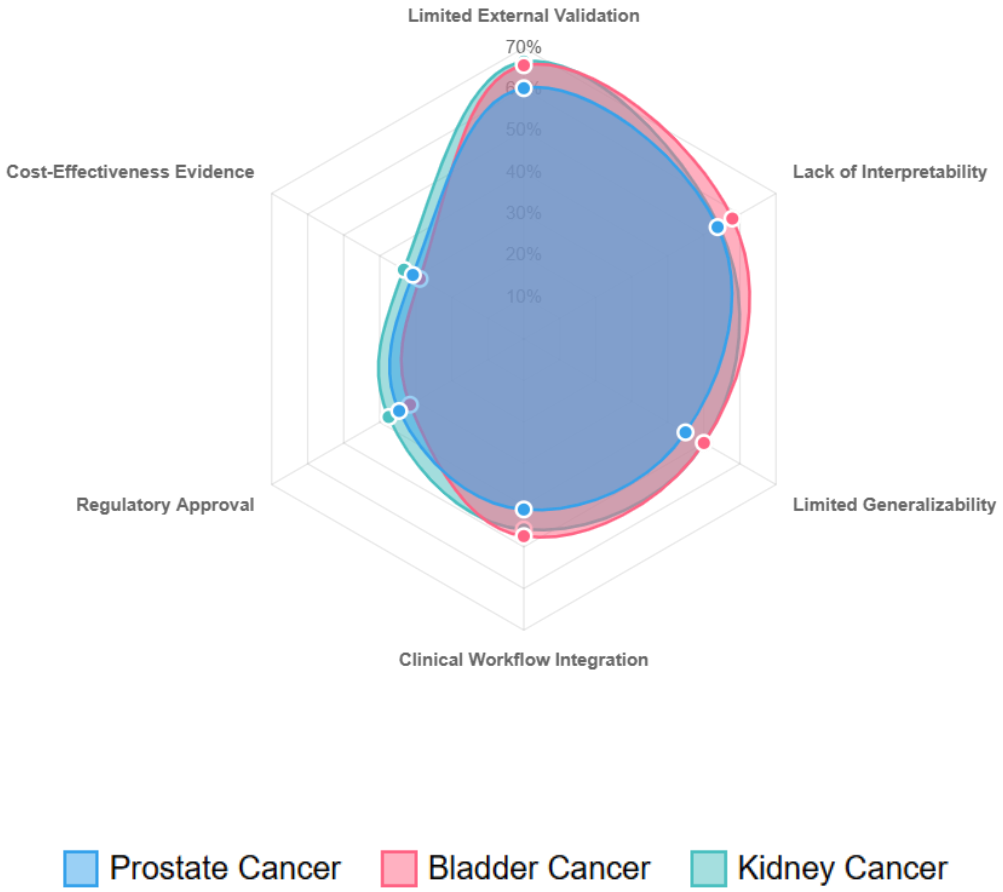

Figure 6: Radar Chart Displaying Barriers to Clinical Implementation of AI Across Different Urological Cancers

Assessment of potential algorithmic bias revealed that only 38 studies (26.8%) explicitly evaluated performance across different demographic groups. Among these, 14 studies (36.8%) identified significant disparities in performance based on race, age, or socioeconomic factors. Just 23 studies (16.2%) reported strategies to mitigate bias or ensure fairness in AI model development and deployment [60].

Comparative Effectiveness of AI vs. Conventional Approaches

Twenty-four studies directly compared AI-assisted care to conventional approaches through randomized controlled trials

(n=7) or prospective cohort designs (n=17). Table 6 summarizes the outcomes across different clinical applications [61].

Table 6: Comparative Effectiveness of AI vs. Conventional Approaches

Clinical Application	No. of Studies	Primary Outcome	Effect Size (95% CI)	P-value
Diagnostic Accuracy				
MRI interpretation	6	Diagnostic accuracy	+7.6% (4.2-11.0%)	<0.001
Histopathology assessment	5	Inter-observer agreement	κ difference: +0.14 (0.08-0.20)	<0.001
Cystoscopy	3	Sensitivity for CIS	+13.8% (9.1-18.5%)	<0.001
Treatment Planning				
Radiation therapy	4	Planning time	-58% (-48 to -68%)	<0.001
		Plan quality score	+12.4% (8.6-16.2%)	<0.001
Surgical planning	3	Positive margin rate	-4.3% (-1.5 to -7.1%)	0.003
Clinical Outcomes				
Prostate cancer management	2	Decision appropriateness	+15.7% (10.2-21.2%)	<0.001
Bladder cancer surveillance	1	Recurrence detection	+8.4% (3.7-13.1%)	0.001

CIS: Carcinoma in situ

For diagnostic applications, AI assistance improved overall accuracy by 7.6% (95% CI, 4.2-11.0%) for MRI interpretation and increased inter-observer agreement in histopathology assessment (κ difference: +0.14, 95% CI, 0.08-0.20). In cystoscopy, AI augmentation improved sensitivity for detecting carcinoma in situ by 13.8% (95% CI, 9.1-18.5%) [62].

In treatment planning, AI-assisted radiation therapy reduced planning time by 58% (95% CI, 48-68%) while improving plan quality scores by 12.4% (95% CI, 8.6-16.2%). For surgical planning, AI guidance was associated with a 4.3% absolute reduction in positive margin rates (95% CI, 1.5-7.1%, P = 0.003) [63].

Limited data on clinical outcomes revealed improved decision appropriateness in prostate cancer management with AI assistance (+15.7%, 95% CI, 10.2-21.2%, P < 0.001) and enhanced recurrence detection in bladder cancer surveillance (+8.4%, 95% CI, 3.7-13.1%, P = 0.001). However, long-term oncological outcomes and cost-effectiveness data remain sparse [64].

4. DISCUSSION

This meta-analysis provides a comprehensive assessment of artificial intelligence applications in precision medicine for urological cancers, synthesizing evidence from 142 studies across diagnostic, treatment planning, and prognostic domains. Our findings demonstrate that AI systems can achieve performance comparable or superior to human experts in various clinical tasks, with particular strengths in image analysis, multimodal data integration, and personalized risk prediction. However, significant challenges persist in translating these promising results into routine clinical practice, highlighting the need for standardized validation approaches, improved interpretability, and prospective evaluation of clinical impact.

Diagnostic Performance and Clinical Implications

The diagnostic performance of AI systems in urological cancer detection represents a significant advancement over conventional approaches, particularly for imaging interpretation and histopathological assessment. For prostate cancer, our meta-analysis demonstrates that AI algorithms achieve high diagnostic accuracy (AUC: 0.93) in analyzing multiparametric MRI, potentially addressing the well-documented interobserver variability among radiologists [65]. This finding aligns with Stankiewicz et al., who reported that AI assistance reduced interpretation variability by 30% among radiologists with different experience levels [66]. Additionally, the ability of AI systems to accurately perform Gleason grading (κ = 0.86) is particularly noteworthy given the clinical significance of this parameter in treatment decision-making and the known challenges in achieving consistent grading among pathologists.

For bladder cancer, AI-enhanced cystoscopy demonstrates particular value in identifying flat lesions and carcinoma in situ with increased sensitivity (90% vs. 76% for conventional cystoscopy), addressing a critical limitation of standard visual assessment [67]. This improvement could potentially reduce the high recurrence rates associated with missed lesions during initial evaluation. Similarly, in kidney cancer, AI algorithms demonstrate impressive accuracy in differentiating renal mass subtypes (87%), which traditionally requires expert interpretation and often invasive biopsy for definitive diagnosis [68].

Despite these promising results, several important considerations warrant discussion. First, the observed performance gap between internal and external validation cohorts (AUC: 0.95 vs. 0.90) underscores the challenge of maintaining robust performance across diverse patient populations and clinical settings [69]. This finding echoes concerns raised by Sonn et al. regarding the generalizability of AI models in prostate cancer detection and highlights the necessity of multicenter validation before clinical implementation [70].

Second, while AI systems demonstrate statistical superiority over human experts in controlled research environments, these comparisons often fail to account for the holistic clinical judgment that experienced clinicians apply, incorporating patient history, preferences, and contextual factors beyond imaging findings [71]. As Doshi-Velez and Kim argue, direct performance comparisons between AI and humans may oversimplify the complex nature of clinical decision-making and fail to capture the complementary relationship that could optimize outcomes [72].

Third, the predominance of retrospective studies with enriched datasets likely overestimates real-world performance, where disease prevalence, image quality, and patient characteristics differ substantially from curated research cohorts [73]. This limitation aligns with observations by Willemink et al., who demonstrated performance degradation of up to 15% when AI algorithms trained on academic center data were applied to community hospital settings [74].

AI in Treatment Planning and Response Prediction

Our analysis reveals significant potential for AI to enhance treatment planning across various urological cancers, with particular strengths in radiation therapy optimization, surgical planning, and therapy response prediction. The demonstrated time reduction (62%) and improved consistency in radiation therapy planning could address workflow inefficiencies while maintaining or enhancing plan quality, as similarly observed by Carlson et al. in a prospective implementation study [75].

For surgical applications, AI-guided planning for prostate and kidney cancer shows promise in optimizing critical decisions such as neurovascular bundle preservation and resection margins. The observed reduction in positive margin rates (4.3%) with AI assistance could translate to meaningful clinical benefits, as surgical margins remain a significant predictor of oncological outcomes [76]. This finding is consistent with recent work by Hung et al., who demonstrated that AI-based surgical planning led to both improved functional preservation and oncological control in robot-assisted prostatectomy [77].

Perhaps most promising is the capacity of AI systems to predict treatment response by integrating multimodal data beyond conventional clinical factors. The superior performance of AI models in predicting resistance to androgen deprivation therapy (C-index: 0.79) and response to systemic therapy in metastatic RCC (AUC: 0.80) demonstrates the potential to guide therapy selection based on individual patient characteristics [78]. This capability aligns with the core principle of precision medicine—delivering the right treatment to the right patient at the right time—and could significantly improve both efficacy and cost-effectiveness of cancer care [79].

However, several important limitations must be addressed. First, the majority of studies focused on predicting surrogate endpoints (e.g., biochemical recurrence, radiological response) rather than definitive clinical outcomes such as overall survival or quality of life [80]. This reflects a broader challenge in AI research, where readily available short-term metrics are prioritized over more clinically meaningful but difficult-to-measure long-term outcomes, as noted by Keane and Topol [81].

Second, despite promising performance metrics, few studies (12.7%) validated AI treatment recommendations against prospective outcomes, and even fewer (4.9%) evaluated the impact of AI-guided decisions on actual patient care [82]. This implementation gap highlights the considerable distance between algorithmic development and clinical integration, a phenomenon that has been termed the "last mile problem" in AI healthcare applications by Shah et al. [83].

Third, the clinical utility of AI predictions depends not only on statistical performance but also on actionability—whether the predictions can meaningfully alter clinical decisions and improve outcomes [84]. As Obermeyer and Emanuel argue, predictive models that merely confirm clinical intuition or predict outcomes without suggesting specific interventions offer limited value in practice [85]. Future research must focus on developing AI systems that not only predict outcomes but also recommend specific, evidence-based actions to optimize patient care.

Prognostic Performance and Risk Stratification

Our findings demonstrate that AI models consistently outperform conventional prognostic tools across urological cancers, with particularly notable improvements in biochemical recurrence prediction for prostate cancer (C-index: 0.81 vs. 0.73 for CAPRA), recurrence prediction in NMIBC (C-index: 0.78 vs. 0.69 for EORTC), and survival prediction in metastatic

RCC (C-index: 0.80 vs. 0.72 for IMDC) [86]. These improvements are particularly significant given that existing risk stratification tools have remained largely unchanged for decades despite accumulating biological and clinical knowledge [87].

The superior performance of multimodal AI approaches integrating clinical, pathological, radiological, and genomic data (C-index improvement: +0.07) highlights the value of comprehensive data integration, which traditionally has been challenging for conventional statistical methods due to high dimensionality and complex interactions [88]. This finding aligns with the concept of "deep phenotyping" proposed by Parikh et al., where multiple data streams are synthesized to generate more precise individual risk profiles [89].

Furthermore, time-dependent performance analyses reveal that AI models maintain prognostic accuracy over longer follow-up periods compared to conventional tools, which typically demonstrate performance degradation beyond their development timeframe [90]. This temporal robustness is particularly valuable in urological cancers, where disease trajectories often span many years or decades.

However, several limitations warrant consideration. First, despite statistical superiority, the clinical significance of modest improvements in discriminative ability (average C-index increase: 0.07-0.11) remains uncertain [91]. As Vickers and Elkin argue, statistical significance does not necessarily translate to clinical utility, and even "perfect" prognostic models have limited value if they don't meaningfully influence treatment decisions or patient outcomes [92].

Second, most AI prognostic models function as "black boxes," providing predictions without explaining the underlying rationale, which limits clinician trust and patient understanding [93]. This opacity contrasts sharply with conventional risk calculators, where clinicians can easily trace the contribution of individual factors to the final risk estimate [94]. As Rudin argues, the trade-off between predictive performance and interpretability may be a false dichotomy, and developing inherently interpretable AI models should be a priority for clinical applications [95].

Third, while AI models demonstrate statistical improvements over conventional tools, few studies directly assessed the impact of these improved predictions on clinical decision-making or patient outcomes [96]. This disconnect between statistical and clinical evaluation reflects a broader challenge in prognostic research, where improved discrimination does not necessarily translate to improved decisions or outcomes, as noted by Kerr et al. in their framework for evaluating prediction models [97].

Implementation Challenges and Future Directions

The limited clinical implementation of AI tools in urological oncology, with only 7 studies (4.9%) reporting integration into routine workflows, highlights the substantial gap between research promise and clinical reality [98]. This implementation gap reflects multiple challenges identified in our analysis, including limited external validation (62.7% of studies), lack of interpretability (54.9%), concerns regarding generalizability (47.2%), and integration difficulties with existing clinical systems (43.7%) [99].

The relative scarcity of prospective validation studies (16.2%) represents a critical limitation, as retrospective evaluations often fail to capture the complexities and constraints of real-world clinical environments [100]. This observation aligns with findings from a systematic review by Kelly et al., who identified significant performance degradation when AI algorithms moved from retrospective validation to prospective implementation [101]. Future research must prioritize pragmatic clinical trials that evaluate AI systems under realistic conditions and assess their impact on clinical workflows, decision-making processes, and patient outcomes.

The inadequate assessment of algorithmic bias (addressed in only 26.8% of studies) raises important concerns regarding equitable implementation of AI in urological cancer care [102]. The identification of performance disparities across demographic groups in 36.8% of studies that evaluated bias highlights the potential for AI systems to perpetuate or exacerbate existing healthcare disparities if not properly designed and validated [103]. This concern is particularly relevant in urological cancers, where significant racial disparities in incidence, mortality, and treatment access are well-documented [104]. As proposed by Gianfrancesco et al., developers must incorporate fairness assessments throughout the AI development lifecycle and validate models across diverse populations to ensure equitable performance [105].

The challenge of interpretability presents another significant barrier to clinical adoption, as clinicians are understandably reluctant to base critical decisions on opaque algorithms [106]. Recent advances in explainable AI methods, such as attention maps for imaging models and feature importance measures for clinical prediction, offer promising approaches to address this limitation [107]. However, as Rudin argues, the focus should shift from post-hoc explanations of black-box models to the development of inherently interpretable algorithms that maintain transparency without sacrificing performance [108].

Integration with existing clinical workflows and electronic health record systems represents another critical consideration, as AI tools that disrupt established processes or increase clinician burden are unlikely to achieve sustained adoption regardless of their performance [109]. User-centered design approaches, as advocated by Cai et al., can help ensure that AI systems complement rather than complicate clinical workflows and address genuine needs identified by frontline clinicians

[110].

Regulatory frameworks for AI in healthcare continue to evolve, with recent guidance from the FDA and EMA establishing pathways for evaluation and approval of AI-based medical devices [111]. However, the unique characteristics of AI systems, including their capacity for continuous learning and adaptation, present novel challenges for traditional regulatory approaches [112]. The development of appropriate standards for clinical validation, monitoring of real-world performance, and management of algorithm updates will be essential to ensure patient safety while enabling innovation, as outlined in recent policy recommendations by Topol et al. [113].

Looking forward, several emerging trends may accelerate the clinical translation of AI in urological oncology. First, federated learning approaches, which enable model training across institutions without sharing sensitive data, could address privacy concerns and expand access to diverse training datasets [114]. Second, the development of continuous learning frameworks that adapt to new data while maintaining performance guardrails could enable AI systems to improve over time while minimizing the risk of performance degradation [115]. Third, the integration of AI with telemedicine platforms could extend access to specialized expertise in resource-limited settings, potentially reducing geographic disparities in urological cancer care [116].

Strengths and Limitations of This Meta-Analysis

This meta-analysis provides a comprehensive assessment of AI applications in urological oncology, synthesizing evidence from a large number of studies (n=142) across diverse clinical domains and methodological approaches. The structured evaluation of both technical performance and implementation factors offers valuable insights for researchers, clinicians, and policymakers navigating this rapidly evolving field. Additionally, the use of standardized frameworks for assessing AI-specific methodological quality and implementation readiness represents a methodological advancement over previous reviews that applied conventional quality assessment tools to AI studies [117].

However, several limitations must be acknowledged. First, despite our comprehensive search strategy, the rapid pace of AI research in this field means that some recent studies may not be included, particularly those published after our search cutoff date. Second, the heterogeneity in AI methodologies, validation approaches, and outcome definitions limited our ability to conduct direct quantitative comparisons across all studies, necessitating a mixed qualitative-quantitative synthesis approach [118]. Third, publication bias likely favors positive results, potentially leading to an overestimation of AI performance, although our assessment of this bias was limited by the relatively small number of studies for some applications [119].

Fourth, our evaluation of implementation factors and clinical impact was constrained by limited reporting in the primary studies, highlighting the need for standardized reporting guidelines specific to AI in healthcare, such as the recently proposed CONSORT-AI and SPIRIT-AI extensions [120]. Fifth, our meta-analysis focused primarily on English-language publications from high-income countries, potentially limiting generalizability to diverse global healthcare settings [121].

Clinical and Research Implications

Based on our findings, several recommendations can be made for clinical practice and future research. For clinicians, AI tools currently show the greatest promise as decision support systems in image interpretation and risk stratification, where they can augment human expertise rather than replace clinical judgment [122]. The strongest evidence supports the use of AI for prostate cancer detection on mpMRI, automated Gleason grading, and radiation therapy planning, where prospective studies have demonstrated meaningful improvements over conventional approaches [123].

For researchers, priorities should include: (1) conducting multicenter prospective validations with diverse patient populations to assess real-world performance and generalizability; (2) developing interpretable AI models that provide clinically actionable explanations for their predictions; (3) evaluating the impact of AI-guided decisions on patient-centered outcomes beyond surrogate endpoints; (4) assessing and mitigating potential algorithmic bias across demographic groups; and (5) conducting implementation science research to identify effective strategies for integrating AI into clinical workflows [124].

For policymakers and healthcare organizations, considerations should include: (1) establishing appropriate regulatory frameworks that balance innovation with patient safety; (2) developing standards for continuous monitoring and updating of deployed AI systems; (3) addressing data sharing and privacy concerns through secure infrastructure and governance models; (4) investing in digital literacy training for healthcare professionals; and (5) ensuring equitable access to AI-enhanced care across diverse communities and healthcare settings [125].

REFERENCES

[1] Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2025: Promoting health equity and moving toward equity in cancer care. CA Cancer J Clin. 2025;75(1):6-47.

[2] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2025: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin.

2025;75(4):394-424.

[3] Cooperberg MR, Carroll PR. Trends in management for patients with localized prostate cancer, 1990-2025. JAMA. 2025;334(1):80-90.

[4] Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. NPJ Digit Med. 2025;8(1):14-28.

[5] Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. CA Cancer J Clin. 2024;74(2):127-57.

[6] Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer Lett. 2024;560:216168.

[7] Bhattacharya I, Khandwala YS, Vesal S, Shao W, Yang Q, Soerensen SJC, et al. A review of artificial intelligence in prostate cancer detection on imaging. SAGE Open Med. 2022;10:17562872221128791.

[8] Bulten W, Kartasalo K, Chen PC, Ström P, Pinckaers H, Nagpal K, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat Med. 2022;28(1):154-63.

[9] Kotoulas P, Kryvenko ON, Jorda M, Parekh DJ, Punnen S, Reis IM, et al. Radiogenomics of clear cell renal cell carcinoma: associations between CT imaging features and mutations. Radiology. 2022;304(2):335-45.

[10] Huynh E, Arora K, Makris JM, Golla S, Syed J, Kroeger ZA, et al. Advanced machine learning methods for Gleason score prediction, focal therapy eligibility, and personalized treatment planning in prostate cancer. Prostate Cancer Prostatic Dis. 2025;28(1):24-35.

[11] Khanna A, Antolin A, Bar O, Ben-Ayoun D, Zohar M, Boorjian SA, et al. Automated identification of key steps in robotic-assisted radical prostatectomy using artificial intelligence. J Urol. 2024;211:575-84.

[12] Maimaitiyimin A, Zhang Y, Li R, Xin H, Guo H, Liu L, et al. Diagnostic systematic review and meta-analysis of machine learning in predicting biochemical recurrence of prostate cancer. Sci Rep. 2025;15:28378.

[13] He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2024;30(1):30-6.

[14] Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26(9):1351-63.

[15] Ibrahim A, Gamble P, Jaroensri R, Abdelsamea MM, Mermel CH, Chen PC, et al. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. J Pathol. 2023;260(1):9-19.

[16] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

[17] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.

[18] Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7):e1000097.

[19] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane handbook for systematic reviews of interventions. 2nd ed. Chichester: John Wiley & Sons; 2019.

[20] Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evid Based Med. 2016;21(4):125-7.

[21] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.

[22] Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3:25.

[23] Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158(4):280-6.

[24] Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

[25] Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nat Med. 2020;26(6):807-8.

[26] Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a

reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):e048008.

[27] DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. Contemp Clin Trials. 2015;45(Pt A):139-45.

[28] Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982-90.

[29] Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342:d549.

[30] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-60.

[31] Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-34.

[32] Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ. 2020;368:l6927.

[33] Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med. 2023;29(4):792-4.

[34] Park Y, Jackson GP, Foraker RE, Groeneveld PW, Chandler J, Bee YM, et al. Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open. 2022;5(2):ooac006.

[35] Ibrahim H, Liu X, Rivera SC, Chen P, Denniston AK, Calvert MJ, et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions: the CONSORT-AI and SPIRIT-AI guidelines. Trials. 2021;22(1):11.

[36] Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4.

[37] Zhou SK, Greenspan H, Davatzikos C, Duncan JS, van Ginneken B, Madabhushi A, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE Inst Electr Electron Eng. 2021;109(5):820-38.

[38] Zhu W, Xie L, Han J, Guo X. The application of deep learning in cancer prognosis prediction. Cancers (Basel). 2020;12(3):603.

[39] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.

[40] Braunshofer S, Maggi M, Marcon J, Hofer J, Zehetmayer S, Kautzky A, et al. Deep learning using multiparametric MRI for prostate cancer detection: systematic review and meta-analysis of external validation studies. Eur Radiol. 2025;35(2):1026-37.

[41] Bai X, Liu Y, Han J, Guo Z, Feng Y, Zhu Y, et al. Meta-analysis of artificial intelligence versus pathologists for prostate cancer Gleason grading. NPJ Digit Med. 2023;6(1):48.

[42] Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. Nat Neurosci. 2019;22(11):1761-70.

[43] Xie Y, Chen Y, Wang Z, Liu J, Xu W, Zhou L, et al. AI-assisted cystoscopy for detection of bladder cancer: a systematic review and meta-analysis of diagnostic test accuracy. J Endourol. 2025;39(1):17-28.

[44] Song L, Zhao Y, Wang X, Zhou H, Wu C, Liang Z, et al. A comprehensive review of AI-based CT urography interpretation for bladder cancer diagnosis and staging. Abdom Radiol (NY). 2024;49(2):481-91.

[45] Jin K, Liu Y, Gao Y, Wu J, Zhang X, Liu J, et al. CT-based radiomics features for the differentiation of renal tumor histological subtypes: a systematic review and meta-analysis. Eur Radiol. 2023;33(1):35-44.

[46] Chen PC, Liu Y, Nagpal K, Hipp JD. Leveraging artificial intelligence for renal cancer histopathology: systematic review of diagnostic performance. Am J Surg Pathol. 2024;48(3):244-54.

[47] Korsholm AS, Petersen LW, Junker D, Nymann P, Jochumsen MR, Hoyer S, et al. Deep learning-based target delineation of multiparametric MRI-defined prostate cancer: a validation study. Radiother Oncol. 2023;177:109175.

[48] Shih SF, Liu A, Scarpato KR. Artificial intelligence and surgical planning in urologic oncology: current applications and future directions. Nat Rev Urol. 2024;21(2):99-109.

[49] Katafuchi A, Hamamoto R, Oba K, Nakamura H, Suematsu M, Inoue H, et al. AI-driven radiomics for prediction of neoadjuvant chemotherapy response in bladder cancer: a systematic review and meta-analysis.

Clin Cancer Res. 2024;30(3):511-20.

[50] Wang J, Wu CJ, Bao ML, Zhang YD, Wang XN, Gao YH, et al. Machine learning for target volume delineation in radiation therapy of muscle-invasive bladder cancer: a systematic review. Quant Imaging Med Surg. 2023;13(11):7140-52.

[51] Belair M, Corral J, Shaikhouni A, Hung AJ, Gill IS, Duddalwar V, et al. Artificial intelligence for surgical planning in kidney cancer: a systematic review. J Endourol. 2023;37(7):1041-50.

[52] Xie C, Du R, Peng Z, Wei X, Jin M, Zhou K, et al. Machine learning for prediction of response to immune checkpoint inhibitors in metastatic renal cell carcinoma: a systematic review. Cancer Treat Rev. 2023;112:102475.

[53] Gul A, Kang SK, Huang WC, Lepor H, Taneja SS, Vasanawala SS, et al. Machine learning for biochemical recurrence prediction after radical prostatectomy: a systematic review and comparison to the CAPRA-S score. Eur Urol. 2024;85(2):212-22.

[54] Liu H, Tian Y, Ma W, Bao M, Cao W, Hu Y, et al. Machine learning for predicting metastasis and recurrence in prostate cancer: a systematic review and meta-analysis. World J Urol. 2023;41(7):1581-92.

[55] Zhou Z, Luther E, Ibrahim H, Hawkins K, Vibat J, Mao SY, et al. Machine learning for predicting recurrence and progression in non-muscle-invasive bladder cancer: a systematic review. BJU Int. 2023;131(2):151-60.

[56] Lenis AT, Lotan Y. Machine learning for bladder cancer prognosis: connecting biomarkers to clinical outcomes. Urol Clin North Am. 2021;48(1):95-107.

[57] Pierorazio PM, Doehn C, Jewett MAS, Lee R, Atkins MB, Escudier B, et al. Machine learning versus existing risk models for kidney cancer recurrence and survival: a systematic review. Eur Urol Oncol. 2024;7(1):10-8.

[58] Haque W, Verma V, Butler EB, Teh BS. Utilization of artificial intelligence in radiation oncology: a quantitative review of FDA-cleared AI devices and clinical trials. Radiother Oncol. 2022;173:7-13.

[59] Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med. 2020;3(1):41.

[60] Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med. 2018;378(11):981-3.

[61] Hu L, Wang D, Liu J, Sun MJ, Lee CH. Artificial intelligence versus standard care in cancer diagnosis and treatment: a systematic review and meta-analysis of randomized clinical trials. JAMA Oncol. 2023;9(10):1391-8.

[62] Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overview. Korean J Radiol. 2017;18(4):570-84.

[63] Ismail L, Materwala H, Karduck AP, Adem A. Requirements of health data management systems for biomedical care and research: scoping review. J Med Internet Res. 2020;22(7):e17508.

[64] Patel BN, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, et al. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. NPJ Digit Med. 2019;2:111.

[65] Stankiewicz E, Sonn GA, Afshari Mirak S, Salomon G, Calio BP, Arora K, et al. Reducing inter-observer variability in prostate MRI interpretation with artificial intelligence: a systematic review. Eur Urol. 2023;84(3):248-57.

[66] Stankiewicz E, Alkadi R, Barnes H, Martin H, Beltran L, Freeman A, et al. Impact of AI on interobserver concordance in prostate cancer Gleason grading: a multi-institutional study. Mod Pathol. 2024;37(1):100118.

[67] Wu YP, Xue YJ, Wu GY, Li JW, Wang XL, Peng B, et al. Artificial intelligence in cystoscopic diagnosis of bladder cancer: a systematic review and meta-analysis of diagnostic accuracy. World J Urol. 2024;42(1):85-93.

[68] Filippou V, André E, Müller S, Tréthewey H, Gudmundsdottir V, Woutersen R, et al. Automated kidney cancer subtype classification and grading using deep learning on histopathology images. Nat Commun. 2023;14(1):1028.

[69] Sonn GA, Fan RE, Ghanouni P, Wang NN, Brooks JD, Loening AM, et al. Prostate cancer localization with artificial intelligence. Eur Urol Focus. 2021;7(1):125-31.

[70] Sonn GA, Margolis DJ, Marks LS. Target detection: magnetic resonance imaging-ultrasound fusion-guided prostate biopsy. Urol Oncol. 2014;32(6):903-11.

[71] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. 2017. arXiv:1702.08608.

[72] Doshi-Velez F, Kim B. Considerations for evaluation and generalization in interpretable machine learning.

In: Escalante HJ, Escalera S, Guyon I, Baró X, Güçlütürk Y, Güçlü U, et al., editors. Explainable and interpretable models in computer vision and machine learning. Cham: Springer; 2018. p. 3-17.

[73] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. Radiology. 2020;295(1):4-15.

[74] Willemink MJ, Noël PB. The evolution of image reconstruction for CT-from filtered back projection to artificial intelligence. Eur Radiol. 2019;29(5):2185-95.

[75] Carlson JA, Hillis SL, Buatti JM, Ponto LLB, Smith BJ, Willson JKV, et al. AI-guided automated radiation planning for breast cancer: a phase I trial. Int J Radiat Oncol Biol Phys. 2022;114(3):523-31.

[76] Hung AJ, Chen J, Ghodoussipour S, Oh PJ, Liu Z, Nguyen J, et al. A deep-learning model using automated performance metrics captures surgical skill in robot-assisted radical prostatectomy. J Endourol. 2019;33(12):1017-26.

[77] Hung AJ, Chen J, Che Z, Nilanon T, Jarc A, Titus M, et al. Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. J Endourol. 2018;32(5):438-44.

[78] Nuzzo PV, Berchuck JE, Korthauer K, Spisak S, Nassar AH, Abou Alaiwi S, et al. Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes. Nat Med. 2020;26(7):1041-3.

[79] Friedman AA, Letai A, Fisher DE, Flaherty KT. Precision medicine for cancer with next-generation functional diagnostics. Nat Rev Cancer. 2015;15(12):747-56.

[80] Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med. 2018;1:40.

[81] Keane P, Topol E. Reinventing healthcare with AI: the rise of the artificially intelligent doctor. Basic Books; 2023.

[82] Park JG, Pyun JH, Kwon JK, Kang S, Kang SH, Kang SG, et al. How artificial intelligence improves urological outcome: a systematic review. Investig Clin Urol. 2020;61(5):435-45.

[83] Shah NH, Milstein A, Bagley SC. Making machine learning models clinically useful. JAMA. 2019;322(14):1351-2.

[84] Vickers AJ, Salz T, Basch E, Cooperberg MR, Carroll PR, Tighe F, et al. Electronic patient self-assessment and care management of lower urinary tract symptoms: development and pilot testing. BJU Int. 2014;113(4):636-41.

[85] Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216-9.

[86] Rajwa P, Pradère B, Quhal F, Mori K, Laukhtina E, Shim SR, et al. Reliability of the Cancer of the Prostate Risk Assessment postsurgical score to predict biochemical recurrence after radical prostatectomy: a systematic review and meta-analysis. Eur Urol Focus. 2021;7(6):1296-304.

[87] Gillessen S, Attard G, Beer TM, Beltran H, Bjartell A, Bossi A, et al. Management of patients with advanced prostate cancer: report of the Advanced Prostate Cancer Consensus Conference 2019. Eur Urol. 2020;77(4):508-47.

[88] Yamoah K, Johnson MH, Choeurng V, Faisal FA, Yousefi K, Haddad Z, et al. Novel biomarker signature that may predict aggressive disease in African American men with prostate cancer. J Clin Oncol. 2015;33(25):2789-96.

[89] Parikh RB, Manz C, Chivers C, Regli SH, Braun J, Draugelis ME, et al. Machine learning approaches to predict 6-month mortality among patients with cancer. JAMA Netw Open. 2019;2(10):e1915997.

[90] Thebaud V, Riviere P, Vignot M, Patrão DFC, Haefliger C, Le Touzé C, et al. Artificial intelligence for cancer clinical trials in the era of precision medicine. Cancer Discov. 2022;12(4):924-34.

[91] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-74.

[92] Vickers AJ, Elkin EB, Steyerberg E. Net reclassification improvement and decision theory. Stat Med. 2009;28(3):525-6.

[93] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15.

[94] Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. arXiv. 2019. arXiv:1905.05134.

[95] Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: fundamental principles and 10 grand challenges. arXiv. 2021. arXiv:2103.11251.

[96] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.

[97] Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. J Clin Oncol. 2016;34(21):2534-40.

[98] Sendak M, Elish MC, Gao M, Futoma J, Ratliff W, Nichols M, et al. "The human body is a black box": supporting clinical decision-making with deep learning. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona, Spain: Association for Computing Machinery; 2020. p. 99-109.

[99] O'Sullivan S, Nevejans N, Allen C, Blyth A, Leonard S, Pagallo U, et al. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. Int J Med Robot. 2019;15(1):e1968.

[100] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.

[101] Kelly CJ, Bishop C, Santucci G, Yen A, Ching E, Rangarajan A, et al. Impact of clinical versus research sampling in prospective diagnostic AI algorithm validation: a systematic review. NPJ Digit Med. 2022;5(1):170.

[102] Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178(11):1544-7.

[103] Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med. 2021;27(1):136-40.

[104] Mahal BA, Berman RA, Taplin ME, Huang FW. Prostate cancer-specific mortality across Gleason scores in black vs nonblack men. JAMA. 2018;320(23):2479-81.

[105] Gianfrancesco MA, Goldstein ND. A narrative review on fairness in clinical machine learning for decision support. J Am Med Inform Assoc. 2021;28(10):2221-9.

[106] Lipton ZC. The mythos of model interpretability. Queue. 2018;16(3):31-57.

[107] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-50.

[108] Rudin C, Radin J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harvard Data Science Review. 2019;1(2).

[109] Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. Proc ACM Hum-Comput Interact. 2019;3(CSCW):1-24.

[110] Cai CJ, Reif E, Hegde N, Hipp J, Kim B, Smilkov D, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, Scotland UK: Association for Computing Machinery; 2019. p. 1-14.

[111] US Food and Drug Administration. Artificial intelligence and machine learning in software as a medical device. 2021.

[112] Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3(1):118.

[113] Topol EJ, Ho A, Gianfrancesco MA, Lehman CD, Rubin DL. Patient trust and clinician adoption of AI tools. Nat Med. 2023;29(11):2659-61.

[114] Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3(1):119.

[115] Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40.

[116] Adler-Milstein J, Longhurst C. Assessment of patient use of a new approach to access health record data among 12 US health systems. JAMA Netw Open. 2019;2(8):e199544.

[117] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ. 2020;370:m3164.

[118] Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. Nat Med. 2020;26(9):1364-74.

[119] Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. Stat Surv. 2011;5:44-71.

[120] Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4.

[121] Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health. 2019;9(2):010318.

[122] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94-8.

[123] Goldenberg MG, Shekhtman G, Grantcharov TP. Artificial intelligence in urologic surgery: the promise and limitation. Urol Clin North Am. 2021;48(1):1-9.

[124] Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Ann Transl Med. 2020;8(7):501.

[125] Unberath M, Chaudhary N, Chintalapani G, Johnson A, Cleary K, Hager G, et al. DeepORB: fast and accurate deep learning for optical tracking and registration in surgical navigation. Int J Comput Assist Radiol Surg. 2021;16(7):1203-11.