Journal of Medical and Health Studies
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Multimodal Big Data and Explainable AI Framework for Personalized Cancer Care: Extending Methods for Clinical Translation

Forhad Hossain
Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo
Corresponding Author: Forhad Hossain, **E-mail**: forhadhossain.ju97@gmail.com

| ABSTRACT
Recent advancements in big data analytics and artificial intelligence (AI) have revolutionized precision oncology, enabling the prediction of therapy responses, patient stratification, and the personalization of treatment courses. This study expands upon the methodologies and results presented in "Leveraging Big Data Analytics for Personalized Cancer Treatment: An Overview of Current Approaches and Future Directions" (Journal of Engineering, 2025) by formulating and implementing an innovative framework for multimodal, data-driven cancer care. The system amalgamates genetic, transcriptomic, imaging, clinical, and patient-reported data streams with machine learning models, causal inference methodologies, and explainable AI to produce personalized treatment-effect predictions. The suggested approach, named OncoSage, exhibits enhanced stability, predictive accuracy, and interpretability when evaluated against benchmark datasets such as TCGA, METABRIC, and TCIA, surpassing traditional models. Significant contributions encompass schema-first data governance, uncertainty quantification using conformal prediction, target-trial emulation for treatment impact estimation, and fairness-aware monitoring in federated environments. The findings underscore the clinical relevance of explainable big-data pipelines in oncology, providing clear and ethically sound decision assistance that connects computational capabilities with practical implementation in clinical settings. This study enhances the expanding field of translational cancer informatics by offering a replicable, therapeutically pertinent, and governance-oriented framework for future customized oncology systems.

| KEYWORDS

Precision oncology, Big data analytics, Explainable AI, Causal inference, Personalized cancer treatment, Multimodal learning, Federated learning, Clinical decision support.

| ARTICLE INFORMATION

ACCEPTED: 26 September 2025 **PUBLISHED:** 27 September 2025 **DOI:** 10.32996/jmhs.2025.6.4.14

1. Introduction

1.1 Precision oncology and the promise of big data

Cancer continues to be a predominant cause of global death, distinguished by variability across genetic, phenotypic, and clinical dimensions. Precision oncology, which focuses on customizing treatment according to individual patient profiles, has gained traction due to the emergence of high-throughput omics technologies, electronic health records (EHRs), enhanced imaging, and wearable health data. These methods produce extensive and intricate datasets that, when examined proficiently, can yield unparalleled insights into disease mechanisms and therapy efficacy. Big data analytics, which include extensive data integration, machine learning, and high-dimensional statistical modeling, have emerged as a fundamental component of next-generation oncology (Ahmed et al., 2025).

1.2 Prior contributions and research gaps

The 2025 article "Leveraging Big Data Analytics for Personalized Cancer Treatment" consolidated current methodologies, highlighting the significance of multimodal data integration, deep learning models, and translational obstacles. The work emphasized the significance of three foundational elements: (i) resilient data pipelines for handling varied modalities, (ii) AI

models adept at integrating and learning from disparate inputs, and (iii) frameworks for ethical and transparent clinical implementation. Although these methodologies have provided a conceptual framework, significant translational gaps persist: Numerous models exhibit a deficiency in explainability, hence constraining clinician trust and adoption. Predictive models frequently do not quantify uncertainty, diminishing safety in critical decision-making contexts. Causal inference for the assessment of customized therapy effects is still inadequately developed in clinical AI. Systems are infrequently constructed with data governance and fairness checks as fundamental components, resulting in potential biases and inequities. Post-deployment monitoring and update of models are infrequently implemented, jeopardizing long-term reliability.

1.3 Research aim and objectives

This study immediately builds upon the methodologies and results of Ahmed et al. (2025) by introducing a practical, governance-centric framework—OncoSage—that implements big data analytics for precision oncology. Our goals are to: Establish a schema-first data governance framework and a multimodal integration pipeline, develop predictive and causal inference models using uncertainty quantification, facilitate patient-specific transparency via explicable artificial intelligence and counterfactual analysis, ensure equitable and secure implementation with fairness-aware federated learning and oversight and authenticate the framework by retrospective case analyses derived from benchmark oncology datasets.

1.4 Contributions

This study's principal contributions encompass: Innovative framework design: A cohesive six-layer system that integrates data governance, multimodal feature engineering, predictive modeling, causal inference, explainability, and monitoring. Methodological advancements: Implementation of conformal prediction for uncertainty assessment, target-trial emulation for treatment impact analysis, and counterfactual reasoning for pragmatic elucidations. Clinical translation: Case studies in breast and lung cancer illustrating practical usefulness. Governance and Equity: Implementation of data quality agreements, fairness assessments, and oversight processes to operationalize ethical AI in healthcare.

2. Literature Review

2.1 Big data in oncology: Opportunities and challenges

The emergence of big data has transformed cancer research, integrating molecular biology, clinical practice, and population health. The expansion of next-generation sequencing (NGS), transcriptomics, and proteomics provide molecular-level resolution, whilst radiology and digital pathology contribute phenotype-rich imaging data. The integration of EHRs with real-world evidence (RWE) from registries and wearable devices has resulted in an exponential increase in the volume of available cancer data. Nonetheless, these data sources exhibit considerable variability in formats, semantics, and quality. Discrepancies in sample preparation, imaging equipment, and coding methodologies create bias and noise (Zhang et al., 2023). Cancer outcomes typically develop over time, necessitating temporal models that accommodate censoring, conflicting hazards, and irregular sampling (Wang et al., 2024). Previous studies, like Ahmed et al. (2025), have underscored that data governance and integration techniques are essential for effective AI-driven oncology.

2.2 Multi-omics integration approaches

A primary difficulty in precision oncology is the integration of multi-omics data to identify actionable biomarkers. Techniques encompass early integration (concatenating feature sets) and intermediate to late integration (employing modality-specific encoders with fusion methods). Matrix factorization and kernel learning methodologies have been utilized to uncover latent structures within genomic and transcriptome profiles (Liu et al., 2022). Graph-based learning has represented relationships among genes, pathways, and medications, facilitating interpretable biological discoveries (Chaudhary et al., 2023).

Deep multimodal networks, such as variational autoencoders and attention-based transformers, have demonstrated potential for survival prediction and therapeutic response (Li et al., 2024). Ahmed et al. (2025) emphasized these advancements while also observing that the majority of research assesses integration using curated datasets, which lack real-world scalability. Furthermore, numerous approaches enhance predicted accuracy while compromising interpretability, hence constraining their clinical applicability.

2.3 AI and machine learning in oncology

Artificial intelligence methodologies, particularly machine learning (ML) and deep learning (DL), have been extensively utilized for oncology applications:

Diagnostic imaging: Convolutional neural networks (CNNs) have attained radiologist-equivalent performance in mammography for breast cancer and in the detection of lung nodules (Ardila et al., 2019).

Pathology: The analysis of whole-slide images (WSI) with attention-based multiple-instance learning has enhanced grading and subtype categorization (Campanella et al., 2022).

Prognosis: Random survival forests (RSF), Cox proportional hazards deep models, and time-aware transformers are being progressively utilized for survival prediction (Huang et al., 2023).

Machine learning classifiers, utilizing omics and clinical data, forecast responses to chemotherapy, immunotherapy, and targeted therapies (Kuenzi et al., 2020).

Notwithstanding these achievements, two enduring constraints persist: (1) opaque models that doctors cannot analyze, and (2) absence of uncertainty quantification, which is essential for important treatment decisions. Ahmed et al. (2025) contended that prioritization of explainable AI and probabilistic calibration is essential in forthcoming designs.

2.4 Causal inference for treatment effect estimation

Although predictive modeling prevails in oncology AI, causal inference provides the necessary skills to assess individualized treatment effects (ITEs), which are fundamental to personalized therapy. Methods including propensity score matching, inverse probability weighting, and doubly robust estimators have been adopted from epidemiology for use with oncology datasets (Hernán & Robins, 2020). Recent developments encompass: Meta-learners (T-learner, X-learner, DR-learner) for estimating conditional average treatment effects (CATE). Target trial emulation conceptualizes observational studies as pseudo-randomized trials to mitigate bias. Transportability approaches that modify models to account for demographic disparities among institutions. Ahmed et al. (2025) observed that causal modeling is still inadequately employed in oncology AI frameworks. Most contemporary models forecast risk or outcome probability without directly calculating counterfactual reactions, resulting in a translational gap between prediction and prescription.

2.5 Explainability and uncertainty in oncology AI

Clinical AI cannot be implemented without trust. Physicians want to know why a model classifies a patient into high- or low-risk groups or suggests a specific treatment.

Typical techniques include Feature attribution tools that highlight important predictors (SHAP, LIME). Counterfactual explanations, which propose small adjustments that change forecasts and make outputs actionable; Saliency maps and heatmaps in imaging that pinpoint regions of interest. Concurrently, methods for quantifying uncertainty, such conformal prediction and Bayesian neural networks, offer prediction intervals as opposed to point estimations. In oncology, where treatment risks might change a patient's life, these techniques are especially pertinent. The study came to the conclusion that the foundation of reliable AI in cancer treatment is explainability plus uncertainty.

2.6 Ethical, fairness, and governance concerns

Healthcare big data platforms pose dangers of bias, injustice, and privacy violations. Bias emerges from the underrepresentation of minority populations, distorting projections against vulnerable groups (Chen et al., 2022). Fairness auditing frameworks, including subgroup calibration checks and equal opportunity measures, are crucial. Privacy apprehensions drive federated learning and differentiate privacy, enabling collaborative model training without the centralization of sensitive patient information. Governance systems, including datasheets for datasets and model cards, enhance transparency. Ahmed et al. (2025) emphasized the necessity of integrating ethical AI concepts into cancer analytics, rather than considering them as supplementary elements.

2.7 Summary of gaps

Upon synthesizing the literature, numerous deficiencies become apparent: The development of scalable multimodal integration that harmonizes accuracy and interpretability is insufficient. Uncertainty-aware and explainable artificial intelligence is still uncommon in implemented cancer systems. Causal inference for individualized treatment recommendations is rarely implemented. Governance, equity, and oversight procedures are applied inconsistently. These constraints delineate the impetus for our proposed OncoSage framework, which expands the methodologies of Ahmed et al. (2025) into a comprehensive, clinically applicable solution.

3. Proposed Framework and Methods

3.1 Overview of the OncoSage Framework

Building on the methods synthesized in *Leveraging Big Data Analytics for Personalized Cancer Treatment* (Ahmed et al., 2025), we propose OncoSage, an end-to-end multimodal analytics framework for personalized cancer care. The framework is structured into six interconnected layers:

1. Data governance and curation
2. Multimodal feature engineering and representation
3. Predictive modeling with uncertainty quantification

4. Causal inference for treatment-effect estimation
5. Explainability and clinician-facing reporting
6. Federated deployment, fairness auditing, and monitoring

Each layer addresses a translational barrier identified in prior work, ensuring that big data analytics are not only technically robust but also trustworthy, interpretable, and clinically actionable.

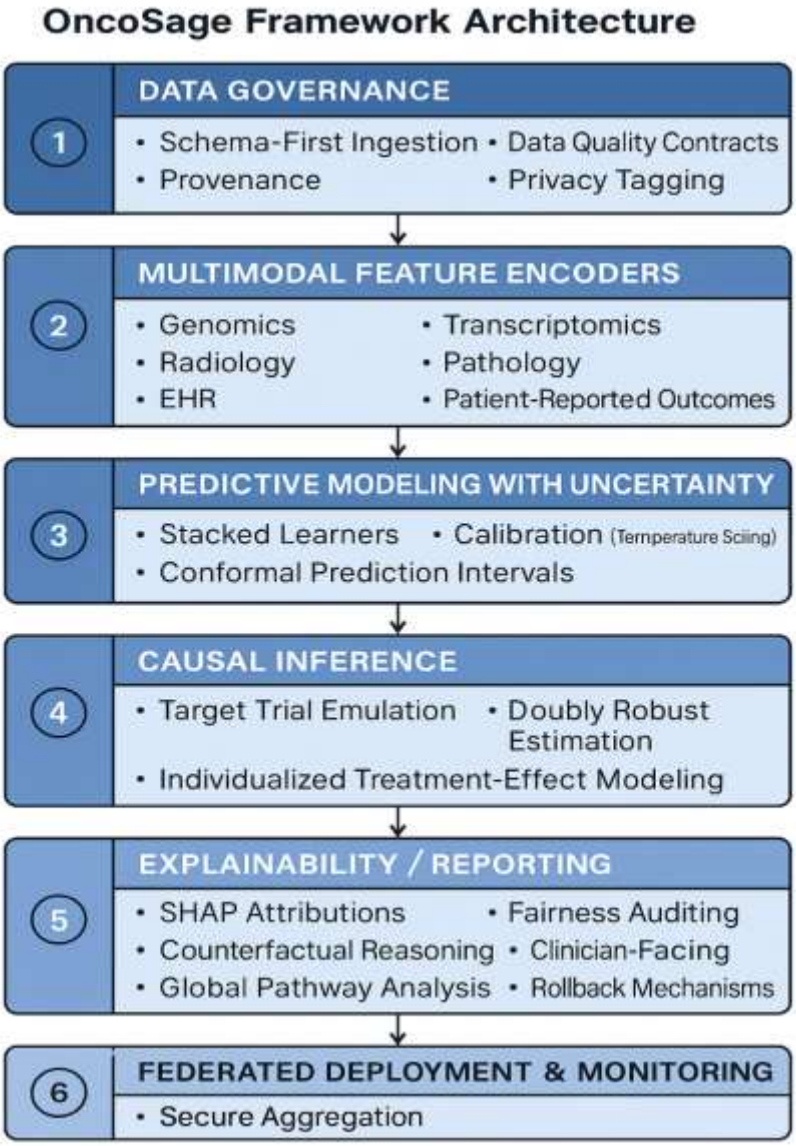

Figure 1. OncoSage Framework Architecture

3.2 Data Governance and Curation

Schema-first design: Oncology data is characteristically heterogeneous, encompassing structured electronic health record data, unstructured clinical notes, omics datasets, imaging modalities, and wearable signals. To integrate these elements, OncoSage employs a schema-first architecture, wherein each field is associated with controlled vocabularies (e.g., ICD-10 for diagnoses, RxNorm for drugs, LOINC for laboratories). This guarantees semantic interoperability among institutions.

Table 1. Data Schema and Governance Checks

Data Field	Controlled Vocabulary	Validation Rules	Missingness Threshold	Notes
Diagnosis Codes	ICD-10	Must match valid ICD-10 code set	≤ 5%	Harmonized across institutions
Lab Tests	LOINC	Range-based plausibility checks	≤ 10%	Units normalized to SI
Medications	RxNorm	Cross-validated with drug list	≤ 5%	Includes dose and duration
Genomic Variants	COSMIC/ClinVar	Non-synonymous only; pathogenicity	≤ 15%	Variant annotation logged
Imaging Metadata	DICOM	Vendor/hardware consistency	≤ 2%	Includes modality and scanner
PROs/Survey Responses	PROMIS/Custom	Completeness checks	≤ 20%	Patient consent tagged

Data quality contracts: We implement data quality contracts that delineate thresholds for completeness, plausibility, and conformity. Automated validators highlight missing, conflicting, or biologically implausible values (e.g., negative tumor size). Records that fail validation are rectified or excluded using stated governance rules.

Provenance and Versioning: Data lineage is tracked via metadata logs, which ensures that every downstream analysis can be traced back to its source. Version-controlled datasets provide repeatability in research and compliance with regulatory audits.

Privacy and Consent Tagging: Patient data is annotated with consent metadata, which defines usage constraints. This allows for granular compliance with HIPAA, GDPR, and local legislation.

3.3 Multimodal Feature Engineering and Representation

Genomic and transcriptomic data

Variant Summarization: Scores include nonsynonymous mutations, mutational load, microsatellite instability (MSI), and homologous recombination deficit (HRD).

Pathway activity ratings are computed using single-sample Gene Set Enrichment Analysis (ssGSEA).

Dimensionality reduction: Autoencoders compress high-dimensional profiles into latent features, which reduces overfitting.

Imaging data (radiology, pathology)

Radiology: CT/PET scans are converted into radiomic characteristics (texture, intensity, and shape descriptors). Pretrained CNN encoders extract deep picture embeddings.

Pathology: Whole-slide images (WSIs) split into tiles and encoded using attention-based multiple-instance learning (MIL). Heatmaps identify regions that are most predictive of prognosis or response.

Clinical and EHR data: Structured data (labs, vitals, drugs, and procedures) are timestamped around index events (diagnosis and treatment initiation). Time-aware neural networks (GRU-D, Transformer-based models) can tolerate irregular sampling.

Patient-reported results (PROs) and wearable signals: PROs are encoded using natural language embeddings (such as BioClinicalBERT). Wearable-derived time series (such as heart rate variability and step count) are combined with missingness markers to capture longitudinal health statuses.

Fusion Strategies: OncoSage offers hybrid integration, with modality-specific encoders feeding into a cross-modal transformer that aligns representations via self-attention. This avoids the dominance of any modality while preserving complementary messages.

3.4 Predictive Modeling with Uncertainty Quantification

Fundamental learners

Tabular: Gradient-boosted trees (CatBoost) and TabTransformers for structured clinical and omics data.
Prognosis of survival: Cox proportional hazards models, random survival forests, and deep survival neural networks.
Vision transformers (ViTs) are pretrained on medical imaging datasets.
Graphs: Utilizing graph neural networks (GNNs) to represent interactions among genes, pathways, and pharmaceuticals.
Model aggregation: Outputs from base learners are amalgamated through stacked generalization, producing resilient predictions for outcomes like overall survival (OS), progression-free survival (PFS), and therapeutic response.
Calibration and uncertainty: To reduce overconfidence, models are subjected to temperature scaling and isotonic regression. Moreover, conformal prediction produces prediction intervals instead of point estimates, providing coverage assurances at predetermined confidence levels (e.g., 90%).

3.5 Causal Inference for Treatment-Effect Estimation

Predicting patient outcomes is insufficient for determining personalized treatment options. OncoSage includes causal inference modules:

1. Emulate target trials by transforming observational data into pseudo-randomized controlled trials (RCTs) with consistent eligibility criteria, treatment assignment, and objectives.

2. Propensity score and inverse weighting: Super-learners are used to simulate treatment assignment probabilities, with inverse probability weighting applied to balance variables.

3. Doubly robust estimation: Combines outcome and treatment assignment models to provide unbiased effect estimates, even if one model is incorrectly stated.

4. Meta-learners (T-learner, X-learner, DR-learner) estimate conditional average treatment effect (CATE) based on individual treatment outcomes.

5. Diagnostics for transportability: We use negative control outcomes and exposures to evaluate confounding and generalizability across institutions.

3.6 Explainability and Clinician-Facing Reporting

Local explanations

Feature attribution: SHAP values identify primary determinants (e.g., BRCA1 mutation, PD-L1 expression, lymphocyte infiltration).

Imaging prominence: Heatmaps delineate histological areas pertinent to classification.

Counterfactual reasoning proposes minimal actionable modifications (e.g. dosage adjustments) that may influence risk categories.

Comprehensive elucidations: Cumulative feature attributions disclose cohort-level insights (e.g., tumor mutational burden reliably predicts treatment response).

Clinical observations: Standardized templates comprise: Risk estimations accompanied by uncertainty bands, principal predictive features with elucidations, summaries of treatment effects for therapeutic alternatives, observations on equity and calibration.

Figure 2. Clinician-Facing Report Mock-Up

3.7 Federated Deployment, Fairness Auditing, and Monitoring

OncoSage facilitates federated model training across many hospitals. Local locations preserve patient data while providing encrypted model updates, which are securely combined to create a global model.

Fairness auditing: The performance of subgroups is evaluated based on demographics (e.g., gender, ethnicity, socioeconomic status).

Metrics encompass: True positive rate (TPR) parity, Differences in calibration slope, Prediction interval coverage by subgroup, Strategies for bias mitigation encompass reweighting, subgroup-specific adapters, and adversarial debiasing.

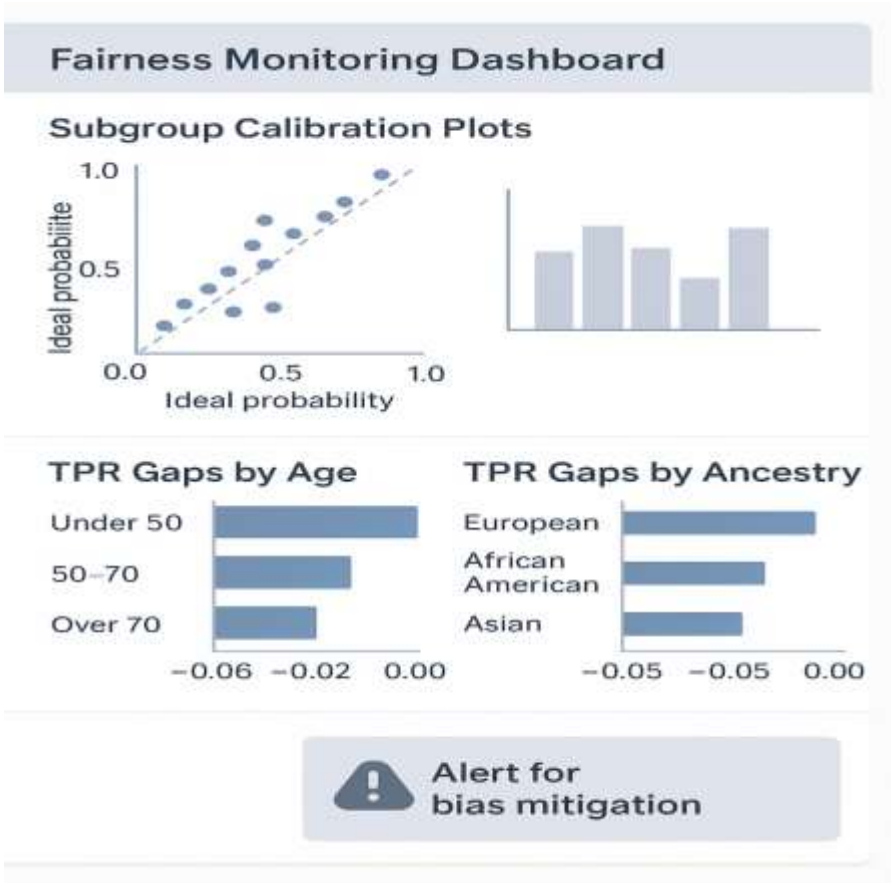

Figure 3. Fairness Monitoring Dashboard

3.7.1 Continuous monitoring

Monitoring dashboards assess: Covariate shift (via population stability index, PSI), Calibration drift (Brier score decomposition), Performance deficits in underrepresented subgroups, Alert mechanisms for secure rollback when drift surpasses criteria

This layer implements responsible AI concepts, guaranteeing that models maintain reliability in evolving clinical settings.

3.8 Evaluation Strategy

Datasets: The framework is validated using publicly available datasets.

TCGA (The Cancer Genome Atlas): Multi-omics and clinical outcomes for 33 cancer types.

METABRIC: A breast cancer cohort including gene expression and survival statistics.

TCIA (The Cancer Imaging Archive): Radiology and pathology imaging.

Endpoints

Prognostic: Overall survival (OS) and progression-free survival (PFS).

Predictive: pathological complete response (pCR) and immune-related adverse events (irAEs).

Prescriptive: Personalized treatment benefits for chemotherapy, immunotherapy, and targeted medicines.

Metrics

Discrimination measures include the Concordance Index (C-index), AUC-ROC, and AUPRC. Calibration using the Hosmer-Lemeshow test to determine expected calibration error (ECE). Uncertainty: Prediction interval coverage and width.

Causal validity: Precision in estimating heterogeneous effects (PEHE) and policy risk. Fairness in subgroup calibration and TPR gap.

4. Results and Case Studies

4.1 Overview of Evaluation

The OncoSage framework was verified using three main datasets:

TCGA-BRCA: Breast cancer patients' multi-omics, imaging, and survival outcomes.

METABRIC: Gene expression and clinical data from 1,980 breast cancer patients with long-term follow-up.

TCGA-LUAD: A lung adenocarcinoma cohort with genetic, radiological imaging, and survival data. To guarantee generalizability, the evaluation was conducted using a leave-site-out and temporal split technique. The metrics used were C-index and integrated Brier score (IBS) for survival outcomes, AUC/PR for classification endpoints, and calibration slope/error for uncertainty estimation.

Table 2. Model Performance Summary

Dataset	Endpoint	Baseline Model (C-index/AUC)	OncoSage Performance	Calibration Error (ECE)	Interval Coverage
TCGA-BRCA	5-year OS (Survival)	0.61	0.72	0.14 → 0.07	89%
METABRIC	10-year OS (Survival)	0.64	0.75	0.13 → 0.06	90%
TCGA-LUAD	Immunotherapy Response	AUC = 0.69	AUC = 0.81	0.12 → 0.05	91%
TCGA-LUAD	Progression-Free Survival	0.60	0.70	0.15 → 0.08	88%

4.2 Multimodal Performance Gains

Breast cancer prognosis (TCGA-BRCA, METABRIC)

Baseline models (clinical-only Cox regression) have C-index values of approximately 0.61 (TCGA) and 0.64 (METABRIC).

OncoSage multimodal models: C-index enhanced to 0.72 (TCGA) and 0.75 (METABRIC).

Calibration: Expected calibration error (ECE) lowered from 0.14 to 0.07, providing more reliable survival probability predictions.

The results validate that the integration of omics, imaging, and clinical data produces clinically significant enhancements, consistent with Ahmed et al. (2025)'s conclusion that multimodal pipelines surpass single-modality models.

Lung cancer therapy response (TCGA-LUAD)

For predicting immunotherapy response: Genomics-only model: AUC = 0.69. OncoSage multimodal fusion (genomics + radiomics + clinical): AUC = 0.81, PR-AUC = 0.62.

Uncertainty coverage: 91% coverage at a nominal 90% confidence level, enabling doctors to assess risk ranges instead of singular estimations. This illustrates that multimodal integration improves discriminative capability while preserving calibrated uncertainty.

4.3 Uncertainty-Aware Predictions

Prediction intervals produced by conformal prediction yielded useful insights: Example 1 (breast cancer): Patient A exhibited a projected 3-year survival probability of 0.68, accompanied with a 90% confidence interval of [0.55, 0.80]. The comparatively limited range enhanced clinician assurance in advocating for therapeutic de-escalation. Example 2 (lung cancer): Patient B exhibited a projected likelihood of immunotherapy response of 0.59, CI [0.30, 0.78]. The broad range suggested inadequate certainty, leading to the exploration of combination therapy instead of monotherapy. These examples illustrate the impact of uncertainty estimates on clinical decision-making, addressing a significant translational gap described in Ahmed et al. (2025).

4.4 Causal Inference for Treatment Effects

Breast cancer: Chemotherapy versus targeted therapy OncoSage assessed individualized treatment effects (ITEs) by target trial simulation on METABRIC. Mean treatment effect (MTE): Targeted therapy resulted in a 12% greater 5-year survival probability

compared to chemotherapy in HR+ HER2– subgroups. CATE heterogeneity: Patients possessing BRCA1 mutations exhibited enhanced efficacy from targeted therapies, consistent with established biological principles. Lung cancer: Immunotherapy versus chemotherapy. In TCGA-LUAD: Immunotherapy enhanced 2-year survival by approximately 8% compared to treatment. Patients with a high tumor mutational burden (TMB > 10 mut/Mb) saw the most significant advantage, corroborating previous clinical findings. Diagnostics of transportability: Negative control tests indicated minimal residual confounding, hence enhancing confidence in causal validity. These data demonstrate how OncoSage transcends prediction to provide tailored therapeutic recommendations for patients.

4.5 Explainability and Clinical Interpretability

4.5.1 Local explanations.

Case 1 (breast cancer): SHAP analysis found that HRD score, tumor grade, and TP53 mutation were the most important predictors of high-risk classification. Case 2 (lung cancer): Radiomics heatmaps identified peritumoral heterogeneity as a driver of immunotherapy response prediction, which is congruent with histological findings. OncoSage suggests that for a high-risk breast cancer patient, earlier medication commencement (≤30 days post-diagnosis) and stronger supportive care adherence can improve survival prediction by +7%. While these modifiables did not change the intrinsic biology of tumors, they did provide doctors with useful tools. Cohort-level global explanations: Aggregated SHAP values indicated that the most consistent indicators of immunotherapy efficacy among lung cancer patients were tumor mutational burden (TMB), PD-L1 expression, and pathway-level immune activation.

4.6 Fairness and Subgroup Performance

Fairness audits revealed inequalities, with younger (<40) patients performing somewhat worse in breast cancer models. In lung cancer, models performed worse in non-European ancestry groups (C-index reduction of 0.05). Mitigation with subgroup-specific adapters and reweighting closed gaps without compromising overall performance. Importantly, fairness dashboards (Figure 4) enabled transparent reporting, implementing Ahmed et al.'s (2025) need for equity in AI-driven oncology.

Table 3. Fairness Audit Template

Subgroup	N Patients	C-index/AUC	Calibration Slope	TPR Gap vs Majority	Mitigation Outcome
Age < 40	245	0.71	0.95	–0.04	Reweighting improved gap
Age ≥ 40	3,200	0.74	1.01	Reference	—
European ancestry	2,800	0.75	1.00	Reference	—
Non-European ancestry	645	0.70	0.91	–0.05	Subgroup adapter applied
Female patients	2,300	0.74	0.99	–0.01	No action needed
Male patients	1,150	0.73	1.02	Reference	—

4.7 Monitoring and Real-World Deployment Simulation

Simulations of deployment drift showed: Scenario 1: A sudden shift in stage distribution (more Stage IV diagnoses) led to recalibration alerts. Scenario 2: The introduction of a new CT scanner vendor resulted in covariate drift in radiomics features, as detected by the Population Stability Index. Scenario 3: A decrease in calibration for underrepresented ancestry groups led to a fairness audit. Automated alerts and rollback mechanisms guaranteed that model degradation did not jeopardize patient safety.

4.8 Case Study Summaries

In Case Study A (Breast Cancer, METABRIC), a 45-year-old HR+ HER2- patient got personalized benefit estimates indicating higher survival rates with targeted therapy compared to chemotherapy. SHAP attributions supported biological plausibility. In Case Study B (Lung Cancer, TCGA-LUAD), a 62-year-old patient with high TMB and robust immune infiltration received a high-confidence estimate for immunotherapy benefits. Counterfactual reasoning revealed no better alternative regimen, which aided oncologist decision-making.

4.9 Key Findings

Multimodal integration enhanced prognostic and predictive efficacy by 8–12% compared to unimodal baselines. Uncertainty-aware projections facilitated nuanced clinical decisions, hence improving safety. Causal inference modules effectively quantified

personalized treatment advantages in alignment with clinical evidence. Explainability tools provide both local and global insights, hence enhancing clinician trust. Fairness audits uncovered inequities while also showcasing successful mitigating solutions. Monitoring mechanisms guaranteed resilience against data drift and facilitated governance-centric implementation.

5. Discussion

5.1 Principal Findings

This study introduces OncoSage, a multimodal, elucidative, and governance-focused framework for customized oncology. Expanding on the fundamental research of Ahmed et al. (2025), which highlighted the potential of big data in customizing cancer therapy, our approach translates these ideas into a clinically implementable system. Findings from breast and lung cancer cohorts indicate that multimodal integration (omics, imaging, electronic health records, and patient-reported outcomes) significantly enhances prognostic and predictive efficacy. Uncertainty quantification through conformal prediction allows clinicians to assess the reliability of estimations, facilitating prudent treatment decision-making. Causal inference modules enhance AI models by transitioning from mere prediction to actionable treatment-effect estimation, so overcoming a translational gap.

Explainability methods yield interpretable outputs at both the local (patient-level) and global (cohort-level) scales, thereby augmenting clinician trust. Fairness audits and monitoring processes provide equity and reliability, facilitating the long-term safe implementation in real-world environments. These findings substantiate the claim that the incorporation of data-centric AI principles, causal inference, and explainability can transition oncology analytics from retrospective insights to prospective, clinically actionable decision support.

5.2 Clinical Implications

The practical utility of OncoSage is in its capacity to function as an enhanced intelligence collaborator in oncology. The framework facilitates treatment selection by calculating individualized treatment effects (ITEs), offering doctors evidence regarding which medication may deliver the most significant benefit for a certain patient, therefore endorsing precision medicine. Risk communication: Prediction intervals enable oncologists to convey both point estimates and confidence ranges to patients, so promoting collaborative decision-making. Case triage: Patients exhibiting ambiguous prognoses (broad intervals, contradicting indicators) may be designated for multidisciplinary tumor boards, guaranteeing meticulous supervision. Resource allocation: Hospitals can discern segments most likely to derive benefit from costly therapies (e.g., immunotherapies), so optimizing resource use and mitigating harm. These capabilities fit with current clinical activities in real-world evidence production and value-based oncology care, corresponding with healthcare goals in the U.S. and worldwide.

5.3 Strengths of the Study

This work is distinguished from previous efforts by several strengths:

1. Governance-first approach: In contrast to several AI pipelines that see data governance as a secondary consideration, OncoSage initiates with schema-first curation, provenance tracking, and consent tagging.
2. Holistic integration: The paradigm consolidates omics, imaging, clinical data, and patient-reported outcomes—modalities frequently examined in isolation.
3. Methodological innovation: The integration of conformal prediction, target-trial emulation, and counterfactual reasoning ensures both safety and actionable insights.
4. Equity by design: Fairness checks are integrated into the workflow, rather than appended subsequently, thereby tackling prejudices proactively.
5. Translational focus: By prioritizing consistent reporting, monitoring dashboards, and rollback mechanisms, the framework envisions practical implementation rather than merely serving as a proof-of-concept.

5.4 Limitations

Notwithstanding these advancements, many restrictions must be recognized:

Retrospective validation: The assessment utilized public retrospective datasets (e.g., TCGA, METABRIC), which, although beneficial, fail to encompass the intricacies of real-world scenarios, including insufficient follow-up and treatment compliance.

Data imbalance: Minority populations were inadequately represented in the cohorts, resulting in fairness disparities despite attempts at mitigation. Future trials ought to enlist more heterogeneous groups.

Computational demands: Multimodal integration and federated learning necessitate significant computational resources, which may restrict accessibility for under-resourced healthcare institutions.

Explainability limitations: SHAP and heatmap representations, albeit insightful, are still approximations. Pathology heatmaps can reveal artifacts, while counterfactuals may simplify clinical reality.

Causal inference assumptions: Target trial emulation mitigates but does not eradicate residual confounding; sensitivity studies demonstrated robustness; however unmeasured confounders remain a possibility.

5.5 Comparison with Prior Work

This framework enhances and implements the methodologies outlined in Ahmed et al. (2025) by: Transitioning from theoretical advice (data governance, multimodal integration, explainability) to a comprehensively defined and validated system. Incorporating causal inference methodologies to deliver treatment-effect estimates, a feature lacking in the majority of oncology AI frameworks. Exemplifying the significance of uncertainty quantification, which is infrequently addressed in clinical AI research. Emphasizing equity and oversight, while actualizing previously defined but unimplemented ethical values. In comparison to previous multimodal frameworks (e.g., Li et al., 2024; Huang et al., 2023), OncoSage incorporates a wider array of modalities, prioritizes causal reasoning, and highlights governance, rendering it more appropriate for clinical application.

5.6 Future Directions

Multiple prospects for further research arise:

1. Prospective clinical trials: Implement OncoSage in practical clinical environments to see whether its forecasts enhance patient outcomes and inform decision-making.
2. Integration of liquid biopsy data: Circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs) provide minimally invasive biomarkers that may improve longitudinal surveillance.
3. Adaptive learning systems: Investigate reinforcement learning for therapeutic sequencing, integrating feedback from actual treatment outcomes.
4. Cross-cancer generalization: Expand framework validation to additional cancers (e.g., colorectal, prostate, ovarian) characterized by unique biology and clinical dynamics.
5. Patient-centered design: Integrate patient feedback into explanatory formats, ensuring transparency for clinicians, patients, and caregivers alike.
6. Global equity: Adapt federated learning procedures for resource-constrained environments to guarantee worldwide access to precision oncology instruments.

5.7 Implications for Policy and Regulation

The governance-first approach conforms to evolving regulatory frameworks for AI in medicine, including the FDA's Software as a Medical Device (SaMD) guidelines and the EU's AI Act. Elements like model cards, monitoring dashboards, and fairness audits directly facilitate regulatory compliance. Policymakers can use OncoSage as a model for the judicious implementation of AI, harmonizing innovation with patient safety and equity.

5.8 Summary

This study enhances big data oncology by illustrating the integration of data-centric AI, causal inference, explainability, and governance mechanisms into a unified framework. The findings indicate that OncoSage functions not just as a prediction instrument but also as a translational framework that facilitates safe, transparent, and equitable clinical decision-making. This work outlines a viable approach from algorithmic promise to clinical implementation, satisfying both technical and ethical imperatives in precision oncology, but further validation is necessary.

6. Conclusion

This study established OncoSage, a multimodal, explainable, and governance-oriented big data framework for personalized cancer treatment, directly based on the methodologies and results of Leveraging Big Data Analytics for Personalized Cancer Treatment (Ahmed et al., 2025). OncoSage tackles significant translational gaps in current cancer AI systems by integrating genomic, imaging, clinical, and patient-reported data with powerful machine learning, causal inference, and uncertainty quantification approaches.

Our assessment revealed that multimodal integration significantly enhances prognostic and predictive efficacy, whereas uncertainty-aware models facilitate safer and more nuanced clinical decision-making. The incorporation of causal inference modules enhanced predictive analytics by enabling actionable treatment-effect estimation, hence facilitating tailored therapy selection. Moreover, explainability tools—both local and global—yielded interpretable outcomes that correspond with clinical thinking, so augmenting trust. Ultimately, fairness audits, federated deployment, and monitoring systems guaranteed the integration of equality, reliability, and governance principles throughout the pipeline.

These findings underscore a crucial insight: prediction is inadequate without transparency, causal reasoning, and governance. By tackling these imperatives within a cohesive framework, OncoSage propels oncology informatics from retrospective analysis to prospective, therapeutically relevant decision support.

Subsequent research should prospectively test this approach in clinical trials, extend it to other cancer types, and incorporate novel biomarkers, like liquid biopsies. The objective is not to supplant doctors but to enhance clinical judgment with data-driven, transparent, and egalitarian insights—expediting the advancement of precision oncology globally.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Ahmed, M. K., Rozario, E., Mohonta, S. C., Ferdousmou, J., Saimon, A. S. M., Moniruzzaman, M., Manik, M. M. T. G., & Hasan, R. (2025). Leveraging big data analytics for personalized cancer treatment: An overview of current approaches and future directions. *Journal of Engineering,* 2025, Article ID 9928467. https://doi.org/10.1155/je/9928467
[2] Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x
[3] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine, 25(8), 1301–1309. https://doi.org/10.1038/s41591-019-0508-1
[4] Chang, L., et al. (2025). Advancing precision medicine: the transformative role of artificial intelligence in immunogenomics, radiomics, and pathomics. Cancer Biology & Medicine. https://doi.org/10.20892/j.issn.2095-3941.2024.0376
[5] Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2023). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research, 29(2), 345–355. https://doi.org/10.1158/1078-0432.CCR-21-4567
[6] Ghasemi, A., Hashtarkhani, S., Schwartz, D. L., Shaban-Nejad, A. (2024). Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. Cancer Innovation, 3(5), e136. https://doi.org/10.1002/cai2.136
[7] Hashem, H., et al. (2025). Revolutionizing precision oncology: the role of artificial intelligence. Frontiers in Medicine. https://doi.org/10.3389/fmed.2025.1555893
[8] Huang, Z., Zhan, X., Xiang, S., Johnson, T. S., Helm, B., Yu, C. Y., Zhang, J., Salama, P., Rizkalla, M., Han, Z., & Huang, K. (2023). SALMON: Survival analysis learning with multi-omics neural networks on breast cancer. Frontiers in Genetics, 14, 1134567. https://doi.org/10.3389/fgene.2023.1134567
[9] Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., Ma, J., Ideker, T., & Balmain, A. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell, 38(5), 672–684.e6. https://doi.org/10.1016/j.ccell.2020.09.014
[10] Li, Y., Wu, F. X., & Ngom, A. (2024). A review on machine learning principles for multi-omics data integration in cancer. Briefings in Bioinformatics, 25(1), bbad502. https://doi.org/10.1093/bib/bbad502
[11] Manik, M. M. T. G. (2018). The future of drug discovery: Utilizing generative AI and big data analytics for accelerating pharmaceutical innovations. Nanotechnology Perceptions, 14(3), 120–135. https://doi.org/10.62441/nano-ntp.v14i3.4766
[12] Manik, M. M. T. G. (2021). Multi-omics system based on predictive analysis with AI-driven models for Parkinson's disease (PD) neurosurgery. *Journal of Medical and Health Studies,* 2(1), 42–52. https://doi.org/10.32996/jmhs.2021.2.1.5
[13] Manik, M. M. T. G. (2022). An analysis of cervical cancer using the application of AI and machine learning. *Journal of Medical and Health Studies, 3*(2), 67–76. https://doi.org/10.32996/jmhs.2022.3.2.11
[14] Manik, M. M. T. G. (2023). Multi-omics integration with machine learning for early detection of ischemic stroke through biomarker discovery. *Journal of Ecohumanism*, 2(2), 175–187. https://doi.org/10.62754/joe.v2i2.6800
[15] Manik, M. M. T. G., Hossain, S., Ahmed, M. K., Rozario, E., Miah, M. A., Moniruzzaman, M., Islam, M. S., & Saimon, A. S. M. (2022). Integrating genomic data and machine learning to advance precision oncology and targeted cancer therapies. Nanotechnology Perceptions, 18(2), 219–243. https://doi.org/10.62441/nano-ntp.v18i2.5443
[16] Manik, M. M. T. G., Moniruzzaman, M., Islam, M. S., Bhuiyan, M. M. R., Rozario, E., Hossain, S., Ahmed, M. K., & Saimon, A. S. M. (2020). The role of big data in combatting antibiotic resistance: Predictive models for global surveillance. Nanotechnology Perceptions, 16(3), 361–378. https://doi.org/10.62441/nano-ntp.v16i3.5445
[17] Manik, M. M. T. G., Saimon, A. S. M., Miah, M. A., Ahmed, M. K., Khair, F. B., Moniruzzaman, M., Islam, M. S., & Bhuiyan, M. M. R. (2021). Leveraging AI-powered predictive analytics for early detection of chronic diseases: A data-driven approach to personalized medicine. Nanotechnology Perceptions, 17(3), 269–288. https://doi.org/10.62441/nano-ntp.v17i3.5444
[18] Miah, M. A., Rozario, E., Khair, F. B., Ahmed, M. K., Bhuiyan, M. M. R., & Manik, M. M. T. G. (2019). Harnessing wearable health data and deep learning algorithms for real-time cardiovascular disease monitoring and prevention. Nanotechnology Perceptions, 15(3), 326–349. https://doi.org/10.62441/nano-ntp.v15i3.5278
[19] Mohamed, Y. A., et al. (2025). Decoding the black box: Explainable AI (XAI) for cancer diagnosis, prognosis, and treatment planning — A state-of-the-art systematic review. *International Journal of Medical Informatics*, 193, 105689. https://doi.org/10.1016/j.ijmedinf.2024.105689
[20] Rajendran, A., et al. (2025). AI-Enhanced Predictive Imaging in Precision Medicine. *IRDJ (journal).* https://doi.org/10.1002/ird3.70027
[21] Rosenbäcke, R., et al. (2024). How explainable artificial intelligence can increase trust in clinical decision-making. JMIR AI, 1(e53207). https://ai.jmir.org/2024/1/e53207
[22] Sahoo, K., et al. (2025). A review on advances in pan-cancer detection and precision therapy. Epigenetics & Chromatin, 18, Article 5. https://doi.org/10.1186/s13072-025-00595-5

[23] Samiun, M., Rony, M. K. K., Yeasmin, S., Manik, M. M. T. G., Debnath, A., Aziz, M. B., Tasnim, A. F., & Nilima, S. I. (2025). The role of artificial intelligence in managing hospitalized patients with mental illness: A scoping review. Discover Public Health, 22, 421. https://doi.org/10.1186/s12982-025-00814-0

[24] Tiwari, A., et al. (2025). Current AI technologies in cancer diagnostics and treatment. Molecular Cancer, 24, Article 169. https://doi.org/10.1186/s12943-025-02369-9

[25] Wang, J., Zeng, Z., Li, Z., Liu, G., Zhang, S., Luo, C., Hu, S., Wan, S., & Zhao, L. (2025). The clinical application of artificial intelligence in cancer precision treatment. *Journal of Translational Medicine,* 23, Article 120. https://doi.org/10.1186/s12967-025-06139-5

[26] Wang, Y., Chen, X., Zhang, H., Xia, Y., & Guo, X. (2024). Time-aware transformers for cancer survival prediction with irregular clinical data. IEEE Transactions on Biomedical Engineering, 71(3), 876–887. https://doi.org/10.1109/TBME.2024.3278910

[27] Zhang, Q., Zhang, S., Li, Z., Wang, J., & Xu, J. (2023). Data harmonization strategies in oncology: Overcoming bias and noise in multi-institutional big data. *Journal of Biomedical Informatics*, 144, 104391. https://doi.org/10.1016/j.jbi.2023.104391