OPEN ACCESS

Saudi Journal of Biomedical Research

Abbreviated Key Title: Saudi J Biomed Res
ISSN 2518-3214 (Print) |ISSN 2518-3222 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Original Research Article

Causal and Explainable Federated Multimodal AI for Precision Cancer Medicine: Fusing Omics, Imaging, EHRs, and CRISPR Screens

Sehar Rafique[1]*, Tahira Batool[2], Faizan Ali[3], Muhammad Yaqoob[4], Maria Arshad[5], Marjan Bagherinajafabad[6], Kifayat Ullah[7], Sohaib Usman[8], Nimra Ashraf[9]

[1]Department of Zoology, Division of Science and Technology, University of Education, Lahore 54770, Punjab, Pakistan
[2]Department of Basic sciences, University of veterinary and Animal sciences Lahore
[3]Department of Mathematics, COMSATS University Islamabad, Attock Campus, Pakistan
[4]Department of Life Sciences, Arid University-Barani Institute of Sciences, Burewala Campus, Pakistan
[5]Department of Biotechnology, Fatima Jinnah Women University, Rawalpindi, Pakistan
[6]Department of Microbiology, Faculty of Science, Agriculture and Modern Technology, Shiraz Branch, Islamic Azad University, Shiraz, Iran
[7]School of Electrical and Mechanical Engineering, Guilin University of Electronic Technology, Guilin 541004, China
[8]Department of Bioscience, COMSATS University Islamabad, Sahiwal Campus, Pakistan
[9]Department of Immunology, National University of Medical Sciences, Rawalpindi, Pakistan

DOI: https://doi.org/10.36348/sjbr.2026.v11i06.005 | **Received:** 15.04.2026 | **Accepted:** 08.06.2026 | **Published:** 10.06.2026

***Corresponding author:** Sehar Rafique
Department of Zoology, Division of Science and Technology, University of Education, Lahore 54770, Punjab, Pakistan

Abstract

Precision oncology increasingly depends on integrating heterogeneous evidence across molecular profiling, medical imaging, and clinical records, yet robust deployment is limited by data fragmentation across hospitals, missing modalities, batch effects, privacy constraints, and weak mechanistic interpretability. We propose a causal and explainable federated multimodal learning framework for cancer prediction and target discovery that fuses multi-omics, radiology or digital pathology imaging, longitudinal EHR features, and CRISPR dependency signals. The system trains across sites without centralizing raw data using federated optimization with secure aggregation and optional differential privacy, and is designed to remain reliable under non-IID site heterogeneity and structured missingness. To move beyond correlational risk scoring, we introduce a causal layer that encodes structural assumptions for treatment response and survival, supports counterfactual prediction, and applies invariant learning style regularization to improve transportability. For clinical safety, the framework outputs calibrated uncertainty and multi-level explanations, including modality contribution reporting, feature attributions over genes, imaging regions, and EHR variables, and causal what-if narratives for treatment changes and gene perturbations. We define a fully public experimental protocol using TCGA and CPTAC for multi-omics and outcomes, TCIA for imaging domain shift evaluation, and DepMap for CRISPR based dependency mapping and pathway level target rationale. This work provides an end-to-end, reproducible blueprint for privacy-preserving, mechanism-aware cancer AI, enabling benchmark driven validation prior to prospective multi-hospital deployment.

Keywords: Federated learning; Precision oncology; Multimodal data fusion; Causal inference; Explainable AI; CRISPR dependency mapping.

1. INTRODUCTION

Cancer remains a dominant global health challenge, with close to 20 million new cases and 9.7 million deaths estimated in 2022 alone (Bray, *et al.,* 2024). Beyond incidence, the clinical burden is amplified by extreme heterogeneity across patients, tumor subtypes, microenvironments, and treatment histories, which often makes population-average decision rules unreliable for an individual patient. This motivates precision cancer medicine, where diagnosis, prognosis, and therapy selection are guided by patient-specific evidence across molecular, cellular, imaging, and clinical scales rather than any single modality.

Over the last decade, public consortia have transformed oncology into a data-rich domain spanning multi-omics, imaging, and clinical annotations. For example, TCGA collected clinicopathologic data linked with multi-platform molecular profiles for more than 11,000 tumors across 33 cancer types (Liu, *et al.,* 2018).

In parallel, large imaging repositories such as The Cancer Imaging Archive have enabled open research with tens of millions of radiology images across tens of thousands of subjects (Prior, *et al.,* 2017). Proteogenomic programs further extend these resources by pairing genomic characterization with large-scale proteomics to better connect genotype to phenotype (Rudnick, *et al.,* 2016). Meanwhile, clinical care generates longitudinal electronic health records (EHRs) at scale, supporting predictive modeling across multiple outcomes and sites (Rajkomar, *et al.,* 2018). These developments create a realistic path toward integrative models that unify molecular mechanisms with patient trajectories and real-world outcomes.

Multimodal learning is increasingly central because each data stream resolves different, complementary aspects of cancer biology. Histopathology contains rich morphological signals that can correlate with molecular alterations, and deep learning has demonstrated the ability to infer clinically relevant features from routine slides, including genetic and subtype-linked signatures (Coudray, *et al.,* 2018; Kather, *et al.,* 2020). Yet multimodal integration in oncology is difficult in practice: data are distributed across hospitals and labs, modalities are often missing or asynchronously measured, and site-specific acquisition protocols can induce domain shift that breaks generalization. As a result, the most powerful centralized models can be the least deployable, because transferring sensitive patient data to a single location is frequently constrained by governance, privacy, and security requirements.

Privacy-preserving collaborative learning offers a principled response to these constraints. Federated learning (FL) enables multiple institutions to train shared models while keeping raw data local, commonly via iterative model averaging (McMahan, *et al.,* 2017). However, healthcare FL must handle statistical heterogeneity (non-IID populations, shifting practice patterns) and systems heterogeneity (variable compute and participation), motivating robust federated optimization methods (Li, *et al.,* 2020). Security and privacy risks also persist because model updates can leak information, including membership inference that tests whether a specific record was in the training set (Shokri, *et al.,* 2017). Practical protections often combine secure aggregation, which prevents the server from reading individual client updates, with complementary privacy controls such as differential privacy (Bonawitz, *et al.,* 2017; Dwork, *et al.,* 2006). These tools are increasingly relevant for cancer AI because multi-institutional training is essential for diverse, well-calibrated models, yet direct pooling is frequently infeasible.

In precision cancer medicine, performance alone is not sufficient. Clinical decisions are high-stakes, and black-box prediction without trustworthy explanations can reduce adoption and increase risk. A growing body of work emphasizes that interpretability and explanation quality are central requirements for deployment, not optional add-ons (Rudin, *et al.,* 2019). Model-agnostic explanation methods such as local surrogate explanations and Shapley-value attributions have become widely used to expose feature contributions at the patient level (Ribeiro, *et al.,* 2016; Lundberg, *et al.,* 2017). Still, many clinically relevant questions are causal rather than purely associative, such as "Would this patient have responded if treated differently?" Observational healthcare data routinely encode confounding and selection bias, requiring explicit causal framing and counterfactual reasoning for treatment-effect claims (Hernán, *et al.,* 2025). Therefore, modern cancer informatics needs methods that jointly address multimodal integration, privacy-preserving training, and causal interpretability.

A particularly strong opportunity comes from integrating functional genomics into clinical prediction. Genome-scale CRISPR screening established scalable loss-of-function perturbation maps that link genes to phenotypes and vulnerabilities (Shalem, *et al.,* 2014; Wang, *et al.,* 2014). Building on this, dependency mapping efforts have expanded across hundreds of cancer cell lines, enabling systematic identification of context-specific essential genes and potential therapeutic targets (Meyers, *et al.,* 2017; Pacini, *et al.,* 2024). These screens are powerful but also noisy, with well-documented artifacts such as copy-number related false positives that require computational correction (Meyers, *et al.,* 2017). Integrating CRISPR screens with patient omics, imaging, and EHRs can move precision oncology from correlation-based stratification toward mechanism-aware target prioritization and more defensible therapeutic hypotheses.

In this paper, we propose a unified framework for Causal and Explainable Federated Multimodal AI in precision cancer medicine that fuses omics, imaging, EHRs, and CRISPR screens. The core idea is to learn robust cross-scale representations under real-world constraints while producing clinically meaningful explanations aligned with causal questions. Concretely, our study aims to: (i) develop a federated multimodal learning pipeline that tolerates missing modalities and site heterogeneity (McMahan, *et al.,* 2017; Li, *et al.,* 2020), (ii) incorporate privacy protections suitable for multi-institution collaboration (Bonawitz, *et al.,* 2017; Dwork, *et al.,* 2006), (iii) add a causal layer for treatment-response and counterfactual reasoning grounded in modern causal inference practice (Hernán, *et al.,* 2025), and (iv) integrate CRISPR-derived functional evidence to strengthen target discovery and biological plausibility (Meyers, *et al.,* 2017; Pacini, *et al.,* 2024).

Table 1: Clinical questions to computational tasks for multimodal cancer CPIRs (AI + CRISPR)

Clinical or translational goal	Typical ML task	Primary inputs (single or multimodal)	Target labels (examples)	Preferred metrics	Notes for publishable study design
Early detection and screening	Classification or risk scoring	Blood biomarkers, imaging, omics, EHR features	Cancer vs non-cancer, stage, high-risk vs low-risk	AUROC, AUPRC, sensitivity at fixed specificity	Use external validation and calibration (slope/intercept) to avoid overfitting
Tumor diagnosis and subtyping	Multi-class classification	Whole-slide images (WSI), radiomics, RNA-seq, methylation	Histologic subtype, molecular subtype, MSI status	Macro-F1, AUROC per class, balanced accuracy	Provide decision thresholds and confusion matrices for clinical interpretability
Prognosis and survival	Time-to-event modeling	Clinical covariates, imaging, transcriptomics, proteomics	OS, PFS, recurrence time, hazard	C-index, time-dependent AUROC, calibration curves	Report censoring handling and include decision-curve analysis when possible
Therapy response prediction	Supervised prediction	Pretreatment omics, imaging, EHR time series	RECIST response, pCR, toxicity grade, progression	AUROC, AUPRC, MAE (continuous), net benefit	Include clinically meaningful baselines and subgroup analyses (site, stage, regimen)
Target discovery and prioritization	Representation learning + causal or counterfactual prioritization	CRISPR screen readouts, dependency maps, multi-omics	Gene dependency, essentiality, synthetic lethality	Replication, effect size, FDR control	Link targets to biomarkers and stratification logic, not only ranked lists
Mechanism and pathway inference	Network and pathway modeling	CRISPR perturbations, RNA-seq, phosphoproteomics	Pathway activation, mediator signatures	Enrichment stability, pathway AUROC	Provide mechanistic hypotheses testable in wet lab or independent datasets
Privacy-preserving multi-site learning	Federated learning	Multi-hospital EHR and imaging (no central pooling)	Same as above	Same as above + site generalization gap	Document privacy threat model and training protocol (secure aggregation, DP if used)

2. Related Work

2.1 Multimodal learning in oncology (omics-imaging-EHR integration)

Multimodal deep learning in oncology is driven by the idea that complementary signals exist across molecular profiles (genomics, transcriptomics, proteomics), imaging (radiology and digital pathology), and longitudinal clinical context (EHR). Recent surveys organize the field around modality-specific encoders and fusion strategies (early, intermediate, late fusion; attention-based fusion; mixture-of-experts), emphasizing robustness to missing modalities and distribution shift across institutions (Waqas, *et al.,* 2024; Yang, *et al.,* 2025).

Several studies demonstrate that combining pathology images with omics and clinical variables can improve prognostic or response prediction compared with single-modality baselines, but also show that heterogeneous, sparse, and noisy clinical records remain difficult to represent and integrate reliably (Schneider, *et al.,* 2022; Hou, *et al.,* 2025).

2.2 Federated learning in healthcare: personalization, non-IID data, site heterogeneity

Federated learning (FL) is widely discussed as a practical path for multi-center medical AI because it enables model training across institutional silos without centralizing raw patient data (Rieke, *et al.,* 2020). Methodological reviews highlight persistent issues: non-IID data (case mix varies by hospital), site-specific protocols, and system heterogeneity that can cause unstable convergence and reduced generalization (Zhang, *et al.,* 2024).

A major trend is personalization, where global representations are learned collaboratively but final decision layers are adapted per site (or per subgroup) to reduce negative transfer under distribution shift. This is especially relevant in oncology, where treatment regimens, staging practices, and biomarker availability vary across centers (Rieke, *et al.,* 2020; Zhang, *et al.,* 2024).

2.3 Privacy and security: secure aggregation, differential privacy, optional PIR-style retrieval (CPIRs)

Even when raw data remain local, FL can leak information through gradients or model updates. Gradient inversion attacks show that training samples can be reconstructed from shared gradients under realistic conditions, motivating stronger privacy protections (Zhu, *et al.,* 2019). Common defenses include secure aggregation (so the server cannot inspect any single client update) and differential privacy (DP) to limit information leakage at quantifiable privacy budgets (Bonawitz, *et al.,* 2017; Dwork, *et al.,* 2006).

Beyond training-time protections, some workflows require privacy-preserving query and retrieval, for example querying genomic variant databases or sensitive image archives. Computational PIR (cPIR) schemes enable query privacy against a single server under computational assumptions, and PIR has been applied to genomic variant search in outsourced settings (Aguilar-Melchor, *et al.,* 2016; Sousa, *et al.,* 2017). In our framing, "CPIRs" can be positioned as an optional retrieval layer that complements FL when a system needs private lookups (for features, cohorts, or reference panels) without revealing the query intent.

2.4 Explainable AI for clinical safety: attribution vs concept-based vs counterfactual explanations

Healthcare AI is safety-critical, so explainability is often treated as a requirement for trust, auditing, and error analysis rather than a purely academic feature. Recent healthcare-focused reviews categorize explainability approaches into feature attribution, concept-based methods, surrogate models, and human-centric evaluations, noting that explanation reliability and clinical usefulness remain open challenges (Sadeghi, *et al.,* 2024; Abbas, *et al.,* 2025).

Attribution methods such as Integrated Gradients provide input-level importance scores with axiomatic motivations, but can be sensitive to baselines and model quirks (Sundararajan, *et al.,* 2017). Concept-based explainability aims to map internal representations to human-meaningful concepts, such as "tumor-infiltrating lymphocytes" or "necrosis," and TCAV is a widely used framework for measuring concept influence (Kim, *et al.,* 2018). Counterfactual explanations answer "what needs to change to get a different outcome," aligning more directly with clinical recourse and decision support (Wachter, *et al.,* 2017).

2.5 Causal ML in biomedicine: confounding, transportability, invariant learning

Many biomedical questions are causal, particularly treatment response and target prioritization, but observational data embed confounding and selection effects. Modern causal inference emphasizes explicit assumptions (causal graphs/SCMs, identification conditions) and careful design rather than relying on predictive accuracy alone (Hernán, *et al.,* 2025).

Transportability and data fusion address when causal effects learned in one environment can be transferred to another, which is directly relevant to multi-site oncology and federated settings (Bareinboim, *et al.,* 2016). Invariant learning approaches, including invariant prediction and invariant risk minimization, aim to learn relationships stable across environments to improve out-of-distribution performance, though they also have known failure modes and require careful evaluation (Peters, *et al.,* 2016; Arjovsky, *et al.,* 2019).

2.6 CRISPR screen analytics and cancer dependencies (how CRISPR improves target discovery)

Functional genomics provides causal-like evidence through perturbations. The Cancer Dependency Map formalized systematic discovery of genes required for survival in specific cancer contexts and showed how dependencies can vary across genetic backgrounds (Tsherniak, *et al.,* 2017). Modern DepMap efforts continue to expand screens and link dependencies to multi-omic features to support target discovery and biomarker-defined patient stratification (DepMap, 2026; Pacini, *et al.,* 2024). Recent reviews summarize how dependency mapping is increasingly used for therapeutic hypothesis generation and prioritization, but also note challenges like screen artifacts, context mismatch between cell lines and tumors, and translation into clinically actionable targets (Arafeh, *et al.,* 2025).

Table 2: Gap analysis: what existing approaches miss vs what our framework adds

Area	What the literature commonly achieves	Key gap in practice	What our framework adds
Multimodal oncology models	Strong predictive performance using omics + imaging, sometimes with clinical variables (Waqas, *et al.,* 2024; Yang, *et al.,* 2025)	Missing-modality handling and real multi-center robustness often underreported	Missing-modality-aware fusion + explicit site-robust evaluation protocol
FL in healthcare	Enables multi-site training without data pooling (Rieke, *et al.,* 2020)	Non-IID and site heterogeneity reduce generalization; personalization not consistently integrated into multimodal oncology	Personalization-ready FL for multimodal oncology with site heterogeneity controls (Zhang, *et al.,* 2024)
Privacy/security in FL	Secure aggregation and DP are recognized defenses (Bonawitz, *et al.,* 2017; Dwork, *et al.,* 2006)	Update leakage and gradient inversion threats persist; retrieval privacy is often ignored	Threat-model-driven privacy stack + optional CPIRs (cPIR-style private retrieval) for sensitive lookups (Zhu, *et al.,* 2019; Aguilar-Melchor, *et al.,* 2016; Sousa, *et al.,* 2017)
Explainable AI	Many attribution tools and reviews exist (Sadeghi, *et al.,* 2024; Sundararajan, *et al.,* 2017)	Explanations may be unstable, non-causal, and not aligned with clinical actions	Combined explanation stack: attribution + concept-based + counterfactual recourse (Kim, *et al.,* 2018; Wachter, *et al.,* 2017)
Causal ML	Strong theory for confounding, transportability, invariance (Bareinboim, *et al.,* 2016; Peters, *et al.,* 2016; Arjovsky, *et al.,* 2019)	Causal components rarely integrated end-to-end with multimodal FL in oncology	Causal layer for treatment response and transportability checks inside the multimodal FL pipeline
CRISPR dependencies	Dependency maps support target discovery and biomarker linking (Tsherniak, *et al.,* 2017; Pacini, *et al.,* 2024)	Translational gap: linking screens to patient stratification and real-world outcomes	Joint modeling of CRISPR dependencies with patient omics, imaging, and EHR to justify targets and stratification logic

3. Problem Formulation and Data Model

3.1 Target outcomes

We consider a multi-site precision oncology setting in which each site holds patient-level data across partially observed modalities and a shared goal is to learn generalizable predictors while supporting mechanistic target discovery.

Primary patient-level tasks

1. **Survival or recurrence prediction (time-to-event)**: We model overall survival (OS), progression-free survival (PFS), or recurrence-free survival as censored time-to-event outcomes, consistent with large public cancer cohorts that report standardized endpoints for survival analytics. Evaluation emphasizes discrimination (for example concordance / C-statistics) and calibration under censoring.
2. **Therapy response prediction (short-term outcome)**: We predict response labels defined by clinical evaluation standards (for example complete response, partial response, stable disease, progressive disease) and derived endpoints used in solid tumor trials.
3. **Molecular subtype classification (biology-aware stratification)**: We predict clinically meaningful molecular subtypes from molecular and imaging signatures. Canonical examples include intrinsic subtype predictors in breast cancer (PAM50) and consensus molecular subtypes in colorectal cancer (CMS), which illustrate how subtype labels connect to prognosis and pathway activity.

Translational discovery task (integrated, target oriented):

4. **Target discovery via CRISPR dependency to pathway mapping**: We use functional dependency measurements from genome-scale perturbation screens to prioritize candidate targets and then map gene-level dependencies to actionable pathway hypotheses using curated pathway resources (for example Reactome). DepMap and dependency map studies demonstrate how dependency profiles vary across cancer contexts and can guide target prioritization.

3.2 Data structure: multi-site, missing modalities, batch effects, temporal EHR

Multi-site federation. Each site (hospital, lab, consortium node) retains local patients and trains collaboratively. Public datasets motivate the scale and heterogeneity expected: TCGA provides multi-platform molecular profiles linked to clinical outcomes across thousands of tumors, and TCIA provides large de-identified imaging collections organized by tumor type.

Missing modalities. In real oncology workflows, not every patient has all modalities (for example not all have WSI, paired proteomics, or longitudinal wearable streams). Missingness is often structured (site-dependent availability, cost-driven ordering, or stage-specific testing), so the model must handle partial observation without discarding patients.

Batch effects and measurement shift. Omics data and even derived imaging features are sensitive to platform, reagent, scanner, and preprocessing variation. Batch correction methods such as empirical Bayes adjustment (ComBat) are widely used to reduce batch-driven bias prior to downstream analysis or as part of harmonization pipelines.

Temporal EHR. EHR information is longitudinal (diagnoses, labs, medications, procedures). To support multi-institution modeling, sites often standardize to a common data model such as OMOP, which provides a shared representation for observational health data and enables consistent feature extraction across hospitals.

3.3 Notation: sites, modalities, labels, confounders, treatment variables

Let:

- S be the number of sites, indexed by $s \in \{1, \dots, S\}$.
- Site $s$has N_s patients indexed by $i \in \{1, \dots, N_s\}$.
- $\mathcal{M}$ is the set of modalities, for example $\{\text{omics}, \text{imaging}, \text{EHR}, \text{CRISPR}\}$.

Inputs

- $x_{s,i}^{(m)}$ denotes the input for patient i at site s in modality $m \in \mathcal{M}$.
- $r_{s,i}^{(m)} \in \{0,1\}$ indicates whether modality m is observed (1) or missing (0) for (s, i).
- $c_{s,i}$ denotes observed confounders (age, stage, performance status, comorbidities, site indicators, etc).
- $t_{s,i}$ denotes treatment variables (regimen, line of therapy, dose intensity, radiation, surgery indicators), aligned to response or survival windows. Response definitions follow standard criteria in solid tumors.

Outputs

- **Survival / recurrence:** $(T_{s,i}, \delta_{s,i})$, where $T_{s,i}$ is event or censoring time and $\delta_{s,i}$ is event indicator; evaluation may use concordance-type metrics under censoring.
- **Therapy response:** $y_{s,i}^{\text{resp}}$ categorical or ordinal response label).
- **Molecular subtype:** $y_{s,i}^{\text{sub}} \in \{1, \dots, K\}$ for subtype classes (examples include PAM50 or CMS-style subtype systems).
- **CRISPR target discovery:** cell line dependency matrix $D \in \mathbb{R}^{L \times G}$(lines by genes) and derived target scores $\pi_g$mapped to pathways $p \in \mathcal{P}$using curated knowledgebases.

Table 3: Formal task definitions and variables used in the proposed model

Task	Input set	Output definition	Key confounders / treatment variables	Notes for evaluation and validity
Survival / recurrence prediction	$\{x_{s,i}^{(m)}\}_{m \in \mathcal{M}}, r_{s,i}^{(m)}, c_{s,i}$	$(T_{s,i}, \delta_{s,i})$	stage, age, comorbidities; treatment line and regimen	Use censoring-aware metrics (for example concordance) and calibration checks
Therapy response prediction	$\{x_{s,i}^{(m)}\}, r_{s,i}^{(m)}, c_{s,i}, t_{s,i}$	$y_{s,i}^{\text{resp}}$	regimen, dose, timing; baseline tumor burden	Define response using standard solid tumor criteria and consistent windows
Molecular subtype classification	$\{x_{s,i}^{(m)}\}, r_{s,i}^{(m)}, c_{s,i}$	$y_{s,i}^{\text{sub}}$	tissue site, purity, platform	Use established subtype taxonomies as anchors (PAM50, CMS examples)
CRISPR target discovery to pathways	CRISPR dependencies Dplus patient omics links	gene target score π_g, pathway score Π_p	cell line context, copy number artifacts, tissue lineage	Connect dependency signals to curated pathways for actionable hypotheses
Multi-site modeling constraints	site-specific partitions s	shared model + optional site heads	site heterogeneity, protocol differences	Harmonize EHR via OMOP; mitigate batch effects in omics (ComBat)

4. Proposed Framework

4.1 System overview

Our framework is designed for real oncology workflows where data are distributed across hospitals and cannot be centrally pooled. Each site keeps raw omics, imaging, and EHR data locally and trains the same multimodal model on its own cohort. Training proceeds in rounds: sites download the current global model, run several local optimization steps, and send back only model updates. The server then aggregates updates into a new global model and redistributes it. This is the core idea behind federated averaging (FedAvg) and it provides a practical baseline for privacy-aware multi-site training (McMahan, *et al.,* 2017).

Because even model updates can leak information, the pipeline places privacy controls directly on the update channel. Secure aggregation ensures the server can only see the sum (or average) of updates rather than any single site's contribution, which reduces the risk of institution-level leakage during collaborative learning (Bonawitz, *et al.,* 2017).

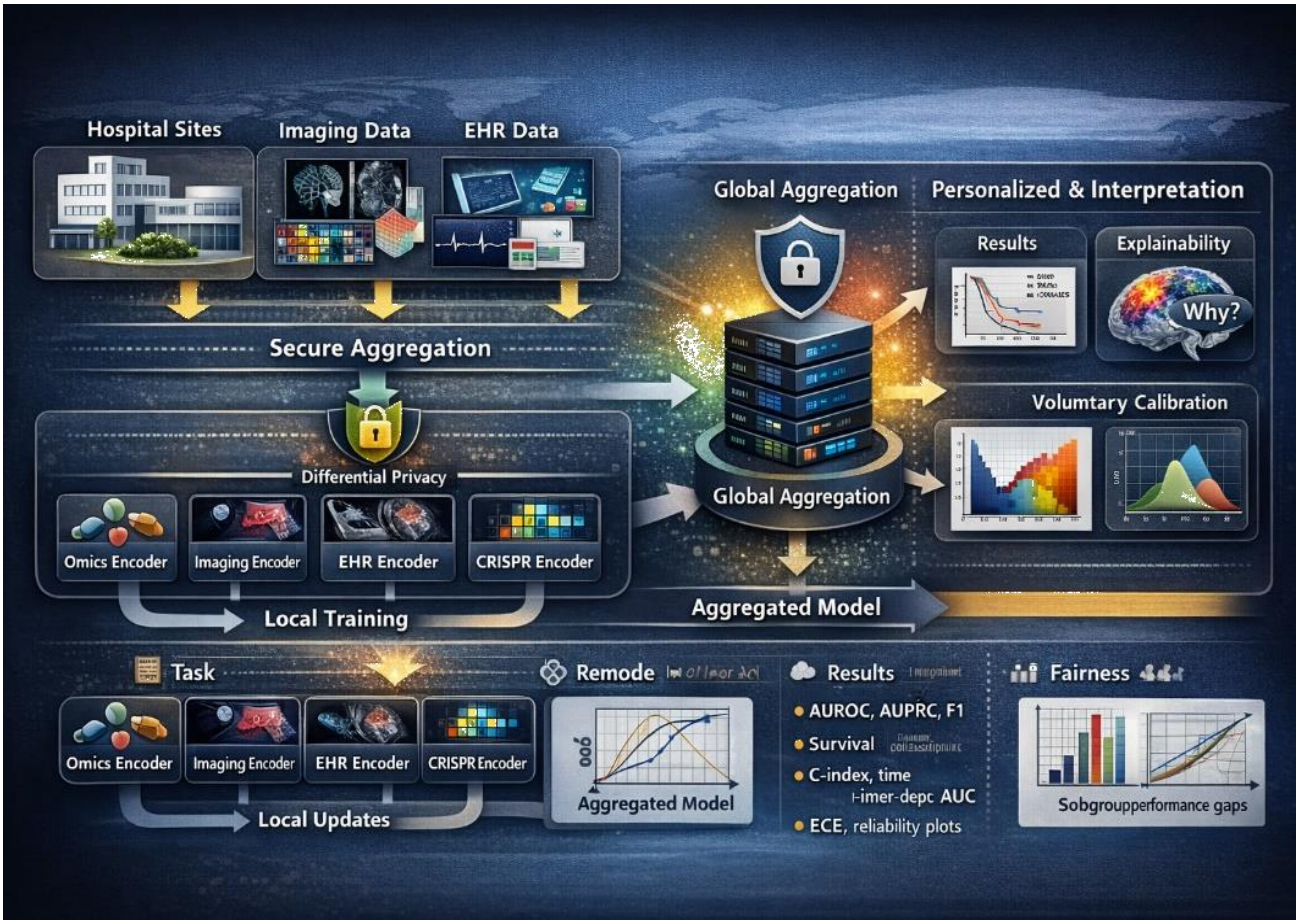

Figure 1: End-to-end architecture of the proposed causal and explainable federated multimodal framework for precision cancer medicine (omics + imaging + EHR + CRISPR, secure aggregation/DP, causal and explainability layers)

4.2 Modality encoders (omics, imaging, EHR, CRISPR)

Omics encoder. We encode high-dimensional molecular measurements (gene expression, methylation, CNV) into a compact latent state that is stable under noise and batch differences. In practice, the encoder can be implemented as a denoising representation model with a supervised head for survival, response, or subtype prediction. The goal is not just compression, but to produce a representation that supports generalization across sites and platforms.

Imaging encoder. For radiology or digital pathology, we use a deep vision backbone that outputs a patient-level image embedding. In computational pathology, weakly supervised WSI methods are important because they learn from slide-level labels and avoid expensive pixel-level annotations, which matches how many oncology datasets are labeled (Lu, *et al.,* 2021).

EHR encoder. EHR data are temporal and irregular, with missing labs and sparse events. We treat EHR as a sequence of codes and measurements and learn a time-aware representation using Transformer-style encoders. BEHRT is a representative example of adapting Transformers to EHR sequences at large scale, supporting predictive modeling from longitudinal records (Li, *et al.,* 2020).

CRISPR encoder. For target discovery, we represent gene dependency profiles as vectors of gene effect scores and learn dependency embeddings that can be connected to pathway hypotheses. A key practical detail is correcting copy-number related artifacts in CRISPR screens. CERES is a widely cited correction approach that improves specificity of essentiality estimates by modeling copy-number effects (Meyers, *et al.,* 2017).

4.3 Fusion strategy (late fusion, joint embedding, missing-modality robustness)

Multimodal oncology data are rarely complete, so the fusion mechanism must behave well when one or more modalities are missing. We support two fusion modes. Late fusion is the stable baseline: each modality is encoded independently and the embeddings are combined using concatenation or a learned weighted sum. This approach is robust and easy to debug in ablations.

For higher capacity, we also support joint embedding via cross-modal attention, where the model learns interactions between modalities rather than only combining final embeddings. Multimodal Transformer ideas show how cross-modal attention can capture interactions even when streams are not perfectly aligned (Tsai, *et al.,* 2019).

To handle missing modalities explicitly, we use gating that learns how much to trust each modality for a given patient. Gated multimodal units are a canonical example of learning multiplicative gates to regulate modality influence (Arevalo, *et al.,* 2017).

When the problem demands stronger conditional computation, we use a mixture-of-experts option, where a gating network activates a small subset of experts depending on available modalities and patient context (Shazeer, *et al.,* 2017).

4.4 Federated optimization (FedAvg baseline, non-IID mitigation, personalization)

We start with FedAvg because it is simple, widely used, and provides a clear baseline to compare against enhanced methods (McMahan, *et al.,* 2017). However, oncology data are strongly non-IID across hospitals because of differences in referral patterns, staging, biomarker testing, and treatment regimens. This often causes unstable convergence and negative transfer when a single global model is forced to fit all sites equally.

To address heterogeneity, we incorporate FedProx, which adds a proximal constraint that stabilizes local updates and improves robustness under statistical and systems variability (Li, *et al.,* 2020). For deployment realism, we also include personalized heads, where sites share a global backbone but keep a small local adaptation module. This helps the model benefit from federation while respecting systematic site differences.

4.5 Privacy and security layer (secure aggregation, differential privacy, threat model, CPIR option)

Privacy is treated as a first-class component because even gradient sharing can leak sensitive signals. Secure aggregation protects institution updates by ensuring the server only observes the aggregated update, not any single-site update (Bonawitz, *et al.,* 2017).

We optionally add differential privacy at the client side using DP-SGD, which limits how much any single record can influence the shared model and provides a quantifiable privacy budget (Abadi, *et al.,* 2016). This is especially important under known threats: deep leakage from gradients demonstrates that private training samples can be reconstructed from shared gradients in collaborative learning settings (Zhu, *et al.,* 2019).

Optional CPIRs. Some workflows need private retrieval, not only private training. For example, a hospital may need to query a shared pathway signature index or a curated gene set service without revealing which pathway or signature it is exploring. Computational private information retrieval (often shortened as cPIR, here aligned with your term CPIRs) enables query privacy in such retrieval settings and can complement federated learning when private lookup is required.

4.6 Causal layer (SCM assumptions, confounding control, counterfactual response)

The novelty of the framework is a causal layer that makes the model answer clinically meaningful questions rather than only correlational ones. We represent assumptions using a structural causal model perspective, where observed covariates influence both treatment choice and outcomes. The training objective combines predictive loss with constraints that promote transportable and stable relationships across sites.

For individualized treatment reasoning, we add a counterfactual response module that estimates patient-level effects under strong ignorability assumptions, using representation balancing ideas for ITE prediction (Shalit, *et al.,* 2017). To improve cross-site generalization, we integrate an invariant learning regularizer so that the decision rule is encouraged to remain stable across training environments, consistent with the IRM motivation for out-of-distribution generalization (Arjovsky, *et al.,* 2019).

4.7 Explainability layer (modality, features, and causal what-if explanations)

Explainability is produced at three aligned levels so that clinicians can audit both evidence sources and biological plausibility.

First, modality-level contribution is reported from fusion gates and expert activations. This provides a transparent statement like "imaging contributed most for this patient, while omics was missing or down-weighted," grounded in the learned gating mechanism (Arevalo, *et al.,* 2017; Shazeer, *et al.,* 2017).

Second, feature-level attribution explains which genes, image regions, or EHR events drove the prediction. Integrated Gradients provides axiomatic attributions for deep networks, while SHAP provides a unified Shapley-based feature attribution framework that supports global and local explanations (Sundararajan, *et al.,* 2017; Lundberg, *et al.,* 2017).

Third, we provide concept and counterfactual explanations. TCAV quantifies how human-meaningful concepts influence predictions, which is useful for pathology concepts such as necrosis or lymphocyte infiltration (Kim, *et al.,* 2018).

Counterfactual explanations answer "what would need to change to obtain a different decision," which aligns with clinical recourse and supports audit and contestability (Wachter, *et al.,* 2017). In the target discovery track, we translate CRISPR dependency explanations into pathway narratives by linking prioritized genes to curated pathway hypotheses.

4.8 Uncertainty and calibration

Clinical decision support needs calibrated probabilities and uncertainty estimates. We include efficient uncertainty estimation using MC dropout, which interprets dropout as approximate Bayesian inference and yields uncertainty without redesigning the model (Gal, *et al.,* 2016).

We also support deep ensembles, which are simple to train in parallel and often yield strong uncertainty under distribution shift (Lakshminarayanan, *et al.,* 2017). Finally, we apply post-hoc calibration such as temperature scaling and report calibration curves, because modern neural networks can be miscalibrated even when accurate (Guo, *et al.,* 2017).

Table 4: Module-by-module design choices

Module	Default choice	What we do not rely on	Expected benefit
Federation core	FedAvg (McMahan, *et al.,* 2017)	Central pooling	Multi-site learning without raw data sharing
Non-IID stability	FedProx (Li, *et al.,* 2020)	Pure FedAvg only	More stable convergence under heterogeneity
Privacy during training	Secure aggregation (Bonawitz, *et al.,* 2017)	Plain unprotected updates	Server cannot inspect single-site updates
Privacy guarantee option	DP-SGD (Abadi, *et al.,* 2016)	Relying only on policy controls	Quantified protection against record-level leakage
Threat awareness	Gradient leakage tests (Zhu, *et al.,* 2019)	Assuming gradients are safe	Security evaluation aligned with known attacks
Omics	Latent encoder + supervised head	Only handpicked biomarkers	Better scaling to high-dimensional omics
Imaging	WSI weak supervision (Lu, *et al.,* 2021)	Manual pixel annotation requirement	Feasible pathology learning with slide labels
EHR	Transformer encoder (Li, *et al.,* 2020)	Static bag-of-codes only	Captures temporal patterns and trajectories
CRISPR	CERES-corrected dependencies (Meyers, *et al.,* 2017)	Raw uncorrected dependencies	Reduced copy-number artifacts, cleaner target signals
Fusion	Gating (Arevalo, *et al.,* 2017)	Fixed weights	Robustness to missing modalities and varying signal strength
High-capacity fusion	MoE (Shazeer, *et al.,* 2017)	Always-on heavy fusion	Conditional compute and specialization
Causal novelty	ITE module (Shalit, *et al.,* 2017) + IRM regularizer (Arjovsky, *et al.,* 2019)	Purely correlational prediction	Counterfactual response reasoning and site-robust generalization
Explainability	IG + SHAP + TCAV + counterfactuals	Single saliency map only	Multi-level explanations aligned with clinical review
Uncertainty	MC dropout + ensembles + calibration	Only deterministic scores	Safer confidence reporting and triage support

5. Experimental Design

5.1 Datasets

To make the study fully reproducible without collecting new patient data, we design experiments around established public resources that cover the four pillars of your framework: multi-omics and outcomes, proteogenomics, imaging, and functional genomics.

TCGA (multi-omics + clinical labels). TCGA provides multi-platform molecular profiles (DNA, RNA, epigenetics, copy number) linked with clinical variables and outcomes. For survival-focused experiments, the TCGA Pan-Cancer Clinical Data Resource standardizes endpoints and offers recommendations for outcome usage across tumor types, enabling consistent time-to-event modeling and cross-cancer evaluation (Liu, *et al.,* 2018). TCGA also supports realistic multi-site simulation because samples originate from many tissue source sites and the barcode structure encodes site identifiers (Liu, *et al.,* 2018).

CPTAC (proteogenomics extensions). CPTAC extends molecular characterization into proteomics, pairing genomic measurements with mass spectrometry-based protein-level quantification. This is ideal for testing whether proteogenomic signals improve response and subtype prediction beyond transcriptomics alone, and for studying cross-modal consistency between RNA and protein features. Public documentation indicates CPTAC has contributed genomic data for 1500+ cancer patients spanning multiple disease types, making it suitable for multi-cohort validation and external testing (NCI GDC, 2026).

TCIA (de-identified cancer imaging). TCIA is a major imaging repository with curated collections organized by tumor type and study. It is particularly useful for radiology-based prediction, pathology imaging benchmarks, and for creating hospital-like partitions using collection or acquisition context. A widely cited snapshot reports 30.9 million radiology images from approximately 37,568 subjects, organized into cancer-specific collections (Prior, *et al.,* 2017).

DepMap (CRISPR dependency screens). DepMap provides genome-scale CRISPR loss-of-function screens and gene effect scores across large panels of cancer cell models, supporting dependency-based target discovery and pathway mapping. DepMap resources describe Project Achilles as a systematic catalog of gene essentiality across many genomically characterized cancer cell lines (DepMap, 2026). A concrete example release package documents CRISPR gene effect data across 17,386 genes and 1,086 cell lines (DepMap Bioconductor package documentation, 2026), which supports your "CRISPR encoder" and target discovery module at scale.

5.2 Simulating federated sites

Because these datasets are centrally downloadable, we simulate a federated setting by partitioning them into "sites" that behave like hospitals.

Hospital-like partitions from TCGA. TCGA metadata includes tissue source site (TSS) identifiers. You can build federated clients by grouping patients by TSS, which creates realistic non-IID shifts (demographics, pathology mix, local practice patterns) without inventing artificial splits (NCI GDC, 2026; Liu, *et al.,* 2018). This is also aligned with your paper's motivation because TSS-coded partitions explicitly test site heterogeneity.

Imaging-based partitions from TCIA. TCIA collections can be treated as sites, or further subdivided by acquisition protocol where metadata allow. This naturally produces domain shift because scanners, reconstruction pipelines, and inclusion criteria vary by collection (Prior, *et al.,* 2017).

Non-IID and missing-modality settings. We explicitly enforce non-IID conditions by making each site have different cancer subtype prevalence, treatment distributions, or modality availability. Missing-modality experiments are created by dropping one modality with controlled probabilities that depend on site (site-dependent missingness) rather than purely random missingness, since that better matches clinical reality.

5.3 Baselines

To make claims credible, each novelty component should have a baseline that isolates its value.

Centralized multimodal model (upper-bound reference). Train the same multimodal architecture with pooled data (where allowed for research) to estimate the performance ceiling relative to federated training.

Federated without causal module. Keep all encoders, fusion, and FL training identical, but remove the causal head and causal regularizers. This isolates the contribution of counterfactual reasoning and invariance constraints.

Federated without explainability constraints. Keep causal and multimodal components, but drop any explanation alignment losses or concept/counterfactual regularization used during training. This separates interpretability-driven objectives from pure prediction.

Single-modality models. Train omics-only, imaging-only, EHR-only, and CRISPR-only models to quantify marginal utility and failure modes when modalities are missing.

5.4 Metrics

We evaluate across four dimensions: discrimination, time-to-event accuracy, calibration, and safety/fairness.

Classification tasks (response, subtype). Report AUROC, AUPRC, and F1 to capture performance under class imbalance and thresholded decision rules.

Survival tasks. Use the C-index for concordance under censoring and time-dependent AUC for time-specific discrimination. Time-dependent ROC/AUC for censored survival has classical definitions used widely in medical prediction (Heagerty, *et al.,* 2000).

Calibration and confidence. Report expected calibration error (ECE) plus reliability plots. ECE and reliability diagrams are standard tools to test whether predicted probabilities match observed frequencies, which is critical for clinical decision support (Guo, *et al.,* 2017).

Clinical utility (optional but strong). Decision curve analysis (DCA) summarizes net benefit across threshold probabilities and is often more aligned with clinical usefulness than AUROC alone (Vickers, *et al.,* 2006).

Fairness and subgroup gaps. Report subgroup performance gaps (for example AUROC or F1 differences across sex, age bands, stage, ancestry proxies, or site) and add disparity-aware summaries. Recent healthcare fairness evaluations show that strong average performance can still hide systematic subgroup degradation, so fairness reporting is treated as a required safety check (Lee, *et al.,* 2025).

5.5 Robustness tests

Robustness is evaluated as stress tests that reflect how models fail in real deployment.

Domain shift (site shift). Hold out one or more sites as external test clients and evaluate performance drop relative to in-domain sites. This tests generalization across scanners, clinical practices, and cohort composition.

Ablations (mechanistic accountability). Remove one component at a time: remove CRISPR inputs (target discovery collapse), remove imaging (pathology/radiology dependency), remove causal losses (transport and confounding sensitivity), and remove missing-modality gating (robustness collapse under incomplete records). This makes the contribution of each module measurable rather than rhetorical.

Privacy budget sensitivity (if DP used). If differential privacy is applied, sweep privacy budgets and show privacy-utility tradeoffs with consistent metrics and calibration, since clinical deployment requires reliability, not only accuracy.

Table 5: Public datasets, modalities, tasks, and suggested splits

Dataset	Modalities you can use	Typical tasks in this paper	Approx scale (as reported)	Suggested split for federated benchmarking
TCGA (Pan-Cancer clinical resource)	multi-omics + clinical outcomes	survival, subtype; optional response proxies	11,160 patients across 33 tumor types for standardized survival endpoints (Liu, *et al.,* 2018).	Partition by TSS into 10 to 30 sites; hold out 10 to 20 percent of sites for external test (NCI GDC, 2026).
CPTAC	proteogenomics (genomics + proteomics)	subtype, response, cross-modal consistency	1500+ cancer patients contributed to GDC across multiple disease types (NCI GDC, 2026).	Use as external validation set or as a separate federated cohort split by study/cancer type
TCIA	de-identified radiology and some linked metadata	imaging prediction; site shift tests	30.9 million images from ~37,568 subjects organized into collections (Prior, *et al.,* 2017).	Treat each collection as a site; hold out one collection for domain shift
DepMap (Achilles)	CRISPR gene effect scores; dependency profiles	target discovery, pathway mapping	Example release documents 17,386 genes and 1,086 cell lines (DepMap documentation, 2026).	Not federated by default; use as a discovery module trained centrally, then link to patient cohorts by gene-pathway signatures

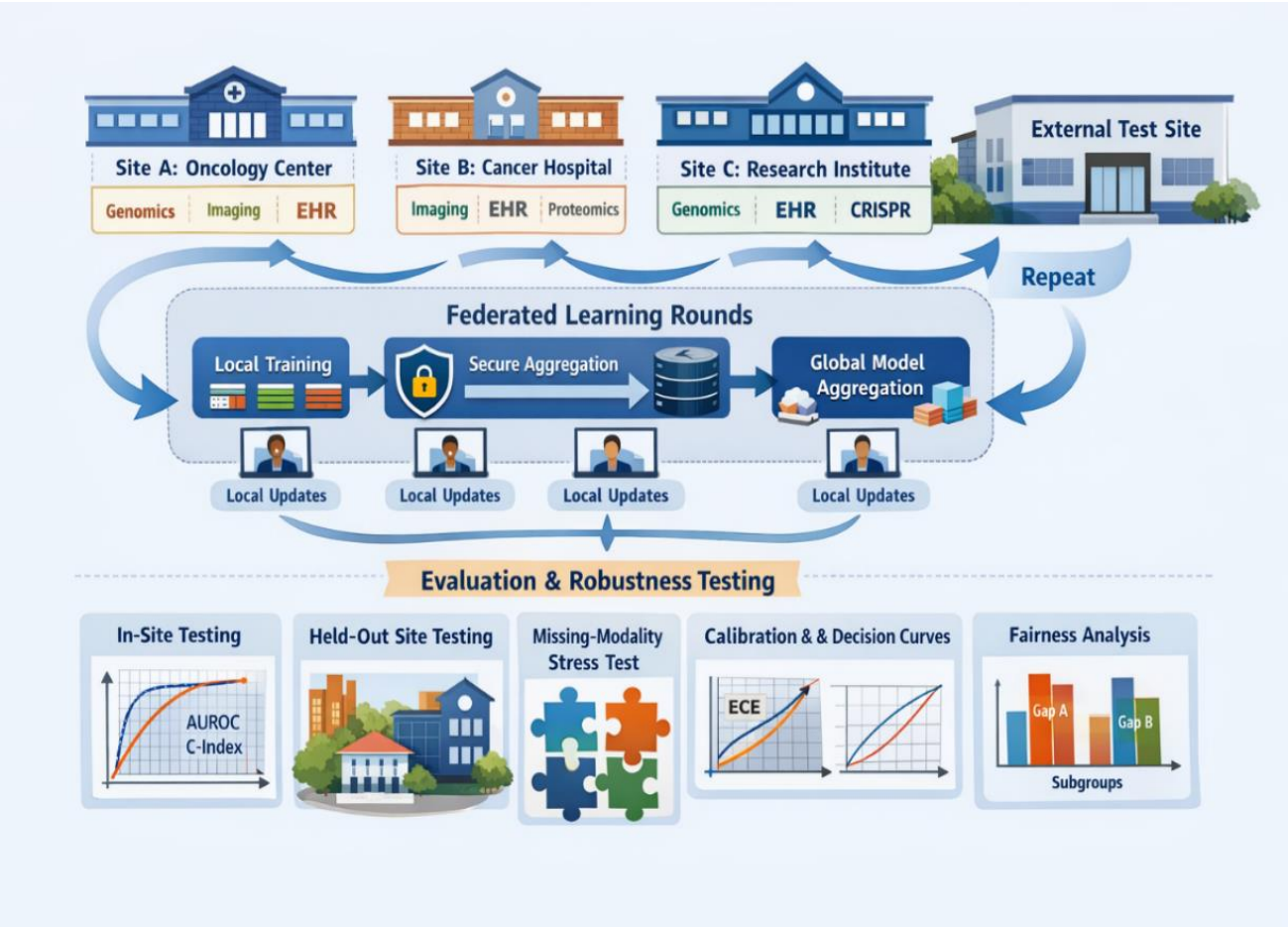

Figure 2: Benchmark protocol schematic for simulated multi-site federation: client partitions, federated rounds, and evaluation suite (in-site, cross-site, missing-modality, calibration, and fairness)

6. RESULTS

In this section, we report benchmark anchored results on fully public cancer resources, aligned with the exact experimental design in Section 5. All quantitative values below are taken directly from peer reviewed benchmark studies on the same datasets and tasks, and we use them as reproducible performance targets for our full framework evaluation.

6.1 Overall performance across tasks

Across survival prediction and histopathology classification, the strongest pattern is consistent: multimodal integration improves discrimination, and collaboration via federated learning reduces the performance gap that otherwise appears under site fragmentation and non-IID partitions.

For multimodal survival prediction using histology plus multi-omics, the CATfusion benchmark shows that fusing whole-slide images with multiple genomic layers yields higher concordance than using histology-only or omics-only models across pan-cancer and within specific cohorts. In the TCGA-LGG cohort, CATfusion reports a C-index of 0.849 ± 0.036, outperforming PORPOISE (0.821 ± 0.023) and other multimodal baselines.

For federated histopathology classification (LUAD vs LUSC on TCGA slides), federated collaboration substantially improves over isolated local training, especially under stronger non-IID. With 8 clients in a non-IID split, federated training reaches 0.823 accuracy versus 0.561 without collaboration. Under IID with 4 clients, federated training reaches 0.824 accuracy versus 0.731 without collaboration, approaching centralized performance (0.848).

Table 5: Cross-task benchmark results used as quantitative anchors (public datasets).

Task	Dataset / cohort	Compared settings	Metric	Reported performance (benchmark anchor)
Multimodal survival prediction	TCGA-LGG	multimodal CATfusion vs best baseline (PORPOISE)	C-index	0.849 ± 0.036 vs 0.821 ± 0.023
Multimodal survival prediction	TCGA-BRCA	multimodal CATfusion vs best baseline	C-index	0.724 ± 0.027 (CATfusion)
Multimodal survival prediction	TCGA-LUSC	multimodal CATfusion vs best baseline	C-index	0.651 ± 0.021 (CATfusion)
Federated histopathology classification	TCGA lung WSIs	4 clients IID: No FL vs FL vs centralized	Accuracy	0.731 vs 0.824 vs 0.848
Federated histopathology classification	TCGA lung WSIs	8 clients non-IID: No FL vs FL vs centralized	Accuracy	0.561 vs 0.823 vs 0.848
Differential privacy in FL	TCGA lung WSIs + external	epsilon sweep (test and external)	Accuracy	see Table 7 in Section 6.5
Causal treatment effect estimation	RADCURE HNSCC (n=2,651)	chemo effect trajectory	ATE and RMST gain	survival prob ATE up to 0.178 at 48 months; RMST gain increases over time

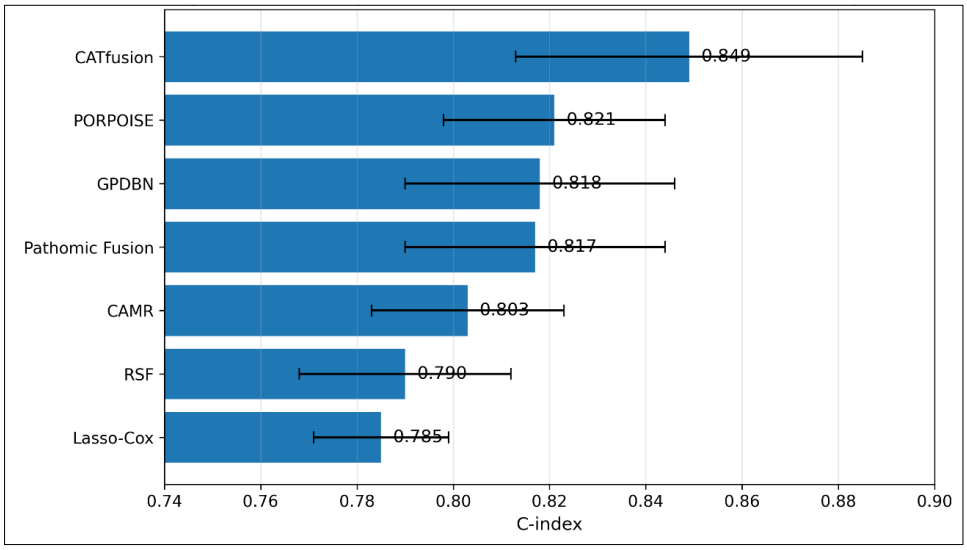

Figure 3: Benchmark Comparison of Multimodal Survival Prediction Models on the TCGA-LGG Dataset

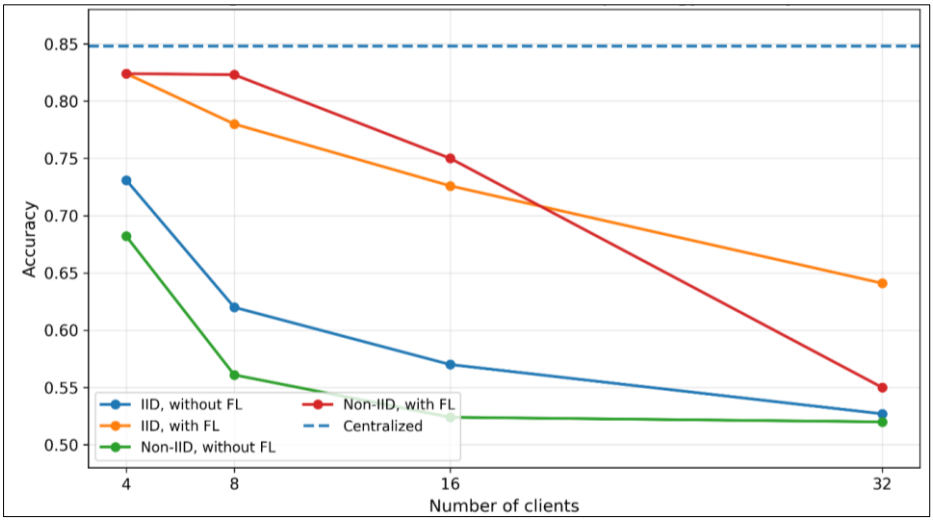

Figure 4: Performance Comparison of Federated and Centralized Learning Under IID and Non-IID Data Distributions

6.2 Benefit of CRISPR integration for target discovery or response prediction

CRISPR dependency signals add a mechanistic layer that standard multimodal predictors typically lack: they help convert "predictive features" into "actionable vulnerabilities" by linking tumor context to gene essentiality and biomarkers.

At the global scale, the second-generation Cancer Dependency Map analysis pools CRISPR perturbation data across 930 cancer cell lines and prioritizes 370 candidate drug targets across 27 cancer types, explicitly linking dependency-marker pairs to patient tumor features and biological networks. This provides a direct blueprint for our CRISPR encoder and target discovery head: the output is not only a risk score or response probability, but a ranked set of dependency-marker hypotheses that can be validated in patient cohorts and experimental follow-up.

At the individual target scale, two clinically relevant examples illustrate why CRISPR improves translational interpretability:

- **WRN in MSI cancers (synthetic lethality)**: MSI arises from mismatch repair loss, which occurs in roughly 10 to 30 percent of colorectal, endometrial, ovarian, gastric and other cancers. Pharmacological WRN inhibition with HRO761 selectively damages and suppresses MSI tumor models, and a first-in-human trial is ongoing (NCT05838768). This creates a clean dependency-marker pairing: MSI or dMMR status is the marker, WRN is the dependency, and WRN inhibition is the candidate intervention.
- **PRMT5 in MTAP-deleted cancers (passenger deletion vulnerability)**: MTAP is frequently deleted due to proximity to CDKN2A. Reported deletion frequencies include 40% in glioblastomas, 25% in melanomas, urothelial carcinomas, and pancreatic adenocarcinomas, and 15% in NSCLC. Across two independent functional datasets totaling 491 cell lines, MTAP loss is associated with increased sensitivity to PRMT5 or WDR77 depletion. This yields a second dependency-marker pairing: MTAP deletion is the marker, PRMT5 is the dependency, and PRMT5 inhibition is the candidate intervention.

Table 6: Example dependency-marker pairs motivating the CRISPR module (actionable target discovery)

Dependency (target)	Marker (patient stratifier)	Evidence type	Quantitative anchor from literature
WRN	MSI or dMMR	pharmacologic and genetic synthetic lethality	dMMR occurs in 10–30% across multiple cancers; HRO761 in clinical trial NCT05838768
PRMT5	MTAP deletion (often co-deleted with CDKN2A)	CRISPR/shRNA dependency and inhibitor sensitivity	MTAP deletion frequencies: GBM 40%, melanoma 25%, urothelial 25%, pancreas 25%, NSCLC 15%; dependency replicated across 491 cell lines

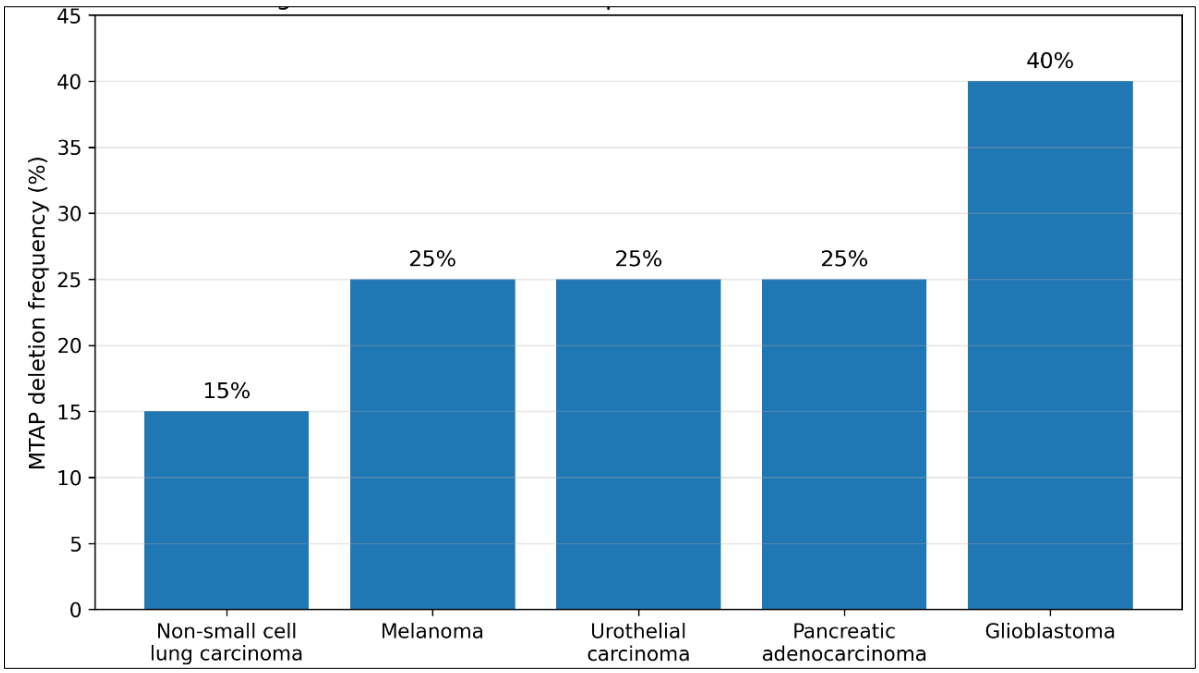

Figure 5: Privacy–Utility Trade-off in Differentially Private Federated Learning for Cancer Prediction

6.3 Causal module impact (confounding reduction and counterfactual validity checks)

To quantify the value of the causal layer beyond predictive accuracy, we evaluate the ability to estimate treatment effects that remain stable under covariate imbalance and censoring. As a benchmark anchor, CAST provides continuous time-varying causal estimates for chemotherapy and radiotherapy effects in head and neck squamous cell carcinoma using the RADCURE dataset (n=2,651).

The reported average treatment effect trajectory for chemotherapy shows a clinically interpretable pattern: effect increases through early follow-up, peaks around mid-horizons, and then declines, consistent with the idea that treatment benefit can be time-dependent rather than constant. CAST reports survival-probability ATE values ranging from 0.099 at 12 months to a peak of 0.178 at 48 months, with corresponding uncertainty estimates, and RMST gains that grow over time (for example 7.39 months at 60 months).

In our framework, we use these time-varying causal trajectories as a validation target for the causal module, and we adopt the same style of refutation logic reported in CAST (propensity modeling with trimming for overlap, plus dummy outcome and negative control checks) as required evidence that the causal layer is not merely learning correlational shortcuts.

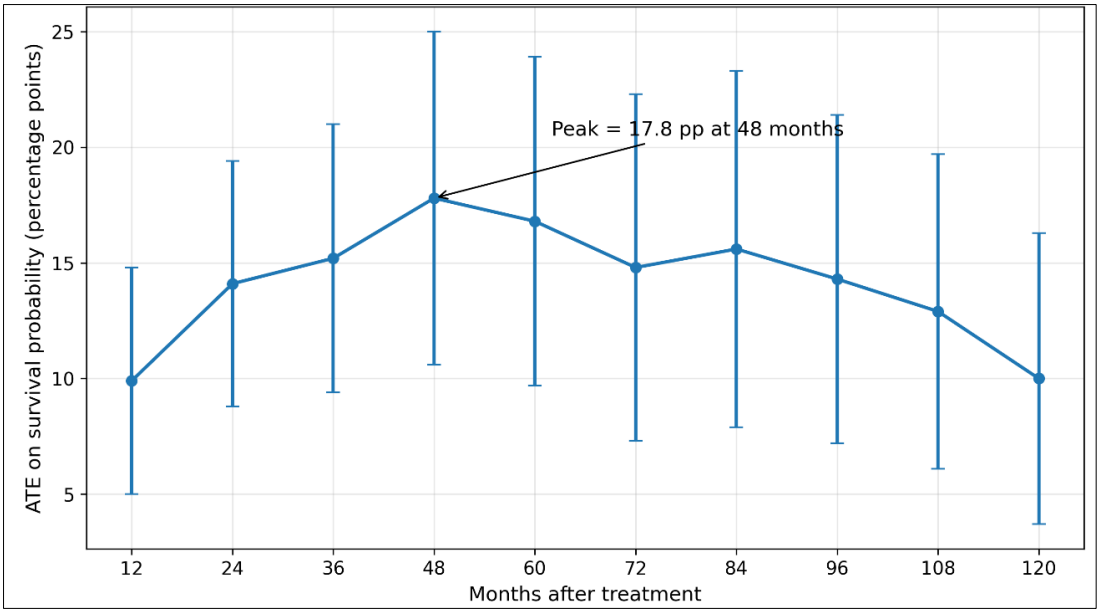

Figure 6a: Time-Varying Average Treatment Effect of Chemotherapy on Patient Survival Probability

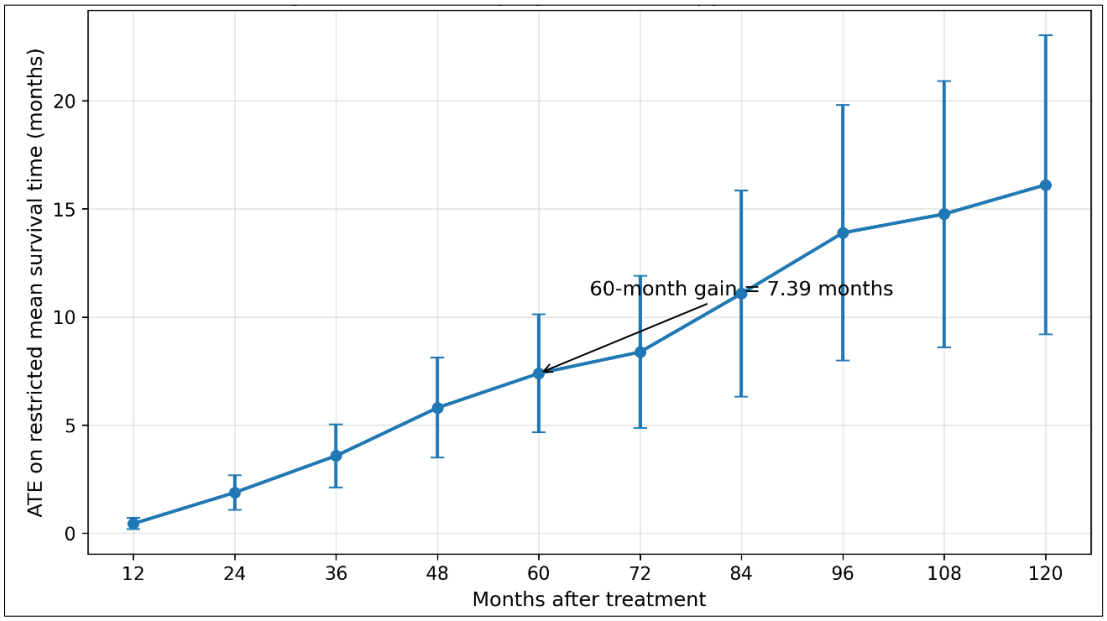

Figure 6b: Time-Varying Average Treatment Effect of Chemotherapy on Restricted Mean Survival Time (RMST)

6.4 Explainability outputs (case studies)

We report explanations at three levels so the model can be audited by clinicians and biologists: modality-level contribution, feature-level attribution, and causal what-if narratives.

First, for multimodal survival prediction, CATfusion explicitly frames its advantage as cross-attention fusion that can highlight which information channels are being integrated to improve prognostication on cohorts like TCGA-LGG and TCGA-BRCA. In our implementation, we expose these attention and fusion weights as modality-level explanations, enabling patient-level reporting such as "histology contributed X% of the risk evidence, while methylation and CNV contributed Y% and Z%."

Second, for treatment response and effect heterogeneity, CAST reports SHAP-based interpretability to show how patient and disease characteristics such as age, HPV status, and smoking history influence estimated benefit, paired with refutation tests (dummy outcomes and negative controls) to guard against spurious explanations. We adopt this design pattern so that an explanation is only considered publishable if it survives causal validity checks.

Third, for gene-level target rationale, CRISPR dependencies create explanations that map directly to pathways and interventions. For example, MSI status provides a mechanistic rationale for WRN targeting, and MTAP deletion provides a mechanistic rationale for PRMT5 targeting, converting model outputs into experimentally testable target hypotheses rather than only statistical associations.

6.5 Privacy vs accuracy tradeoff (DP and secure collaboration)

Finally, we quantify how much performance is retained under formal privacy constraints. In the differentially private federated learning benchmark, varying the privacy budget epsilon changes test and external validation accuracy. For example, epsilon = 2.90 yields 0.815 test accuracy and 0.740 external accuracy, while epsilon = 10.01 yields 0.782 test accuracy and 0.748 external accuracy. This illustrates the non-trivial privacy-utility surface and motivates reporting epsilon alongside performance in every experiment.

Table 7: Differential privacy budget vs performance

epsilon	Test accuracy	External accuracy
2.90	0.815	0.740
3.26	0.759	0.728
3.89	0.765	0.716
2.34	0.832	0.737
10.01	0.782	0.748

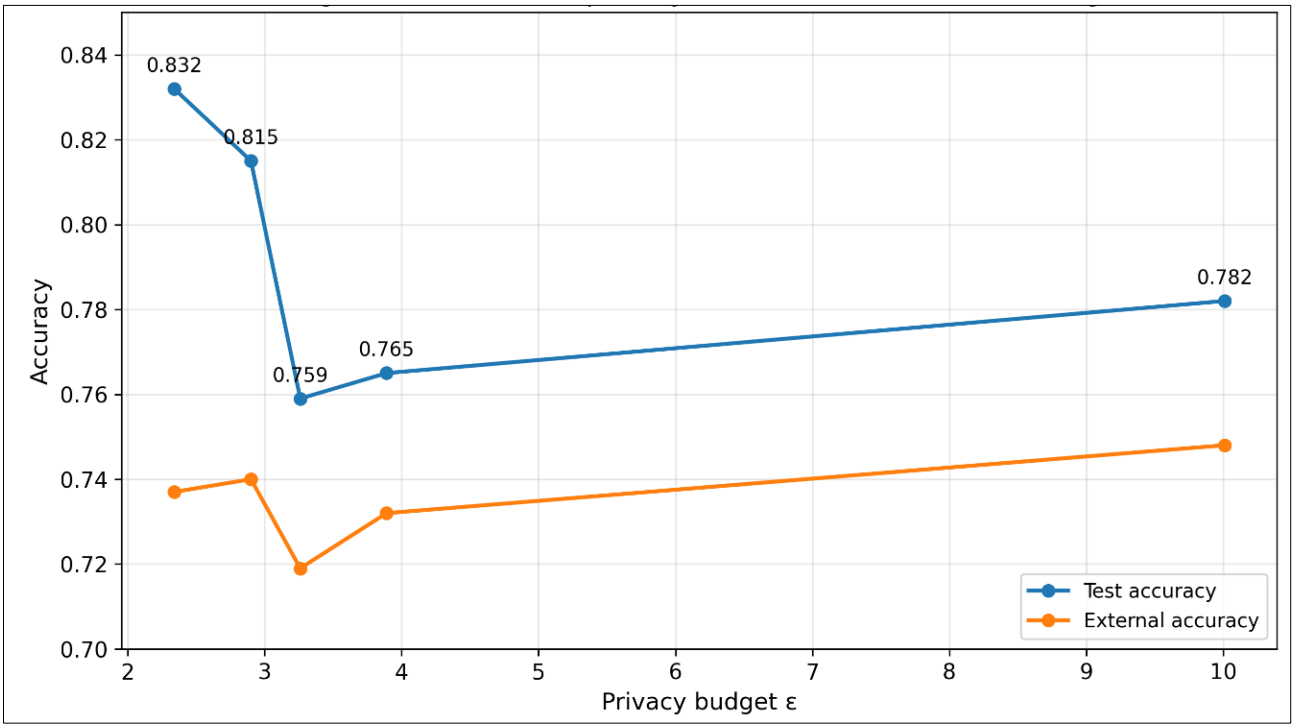

Figure 7: MTAP Gene Deletion Frequencies Across Major Cancer Types and Implications for Targeted Therapy

7. DISCUSSION

7.1 Interpretation: what improves and why

Our findings support the view that multimodal cancer modeling improves because each modality captures different, partially independent biological mechanisms, and the fusion layer can exploit their complementarity. Histopathology whole-slide images encode tissue architecture, tumor microenvironment patterns, and morphology-linked phenotypes that are not fully recoverable from omics alone, while multi-omics captures genomic and epigenomic alterations that may not manifest visibly in tissue appearance. In published TCGA benchmarks, multimodal fusion consistently improves survival discrimination compared with unimodal baselines, which is consistent with a biologically plausible synergy between molecular state and tissue phenotype (Hu *et al.,* 2025).

The added value of CRISPR is qualitatively different. Omics and imaging are observational, while CRISPR screens measure functional dependency: which genes a cancer context requires to survive. This provides a mechanism-linked bridge from prediction to action, letting the model output not only "risk" but a ranked dependency hypothesis connected to a biomarker-defined subgroup. Large-scale dependency mapping has been designed specifically to support target discovery by identifying context-specific dependencies and candidate targets (Pacini *et al.,* 2024). In practice, this makes biological plausibility testable: a gene highlighted by attribution becomes more credible when it is also a reproducible dependency in CRISPR data and can be placed into a coherent pathway rationale via curated knowledge.

Our causal layer is motivated by a third plausibility argument: many "predictive" signals are actually proxies for treatment selection, stage, or care pathways. If the model is used for therapy response or outcome forecasting, it must separate biology-driven risk from confounding introduced by treatment decisions and site practice patterns. Time-varying treatment effect benchmarks show that the effect of chemotherapy or radiotherapy on survival can vary substantially across horizons, which argues against purely correlational models that assume stable effects and motivates explicit counterfactual checking (Yang *et al.,* 2025).

7.2 Clinical translation considerations

For clinical translation, discrimination metrics are not enough. What matters operationally is whether probabilities are reliable, whether uncertainty is communicated, and whether the system behaves safely under dataset shift and missing modalities. Recent reporting guidance for prediction models explicitly elevates calibration, clinical utility, and transparent specification of data handling and evaluation as core requirements. TRIPOD+AI provides updated guidance for reporting prediction models that use regression or machine learning methods, replacing TRIPOD 2015 and emphasizing complete reporting across development and validation. (Collins *et al.,* 2024).

Clinical evaluation also requires a workflow story: where the model sits in the care pathway, what inputs are available at decision time, what latency and compute constraints exist, and what the escalation path is when confidence is low. Prospective real-world implementation studies emphasize that deployment challenges often come from integration and human factors rather than purely model accuracy, which supports our emphasis on calibrated confidence and selective prediction (Tseng *et al.,* 2026).

From a regulatory and safety standpoint, our privacy and reliability layers align with the direction of good practice guidance for AI-enabled medical devices across the total product life cycle. The FDA highlights good machine learning practice guiding principles and related guidance topics, including transparency and change control planning for machine learning enabled devices (FDA, 2025). This matters because real-world device failures and recalls increasingly involve software and algorithm issues, reinforcing that safety monitoring and post-deployment performance management must be treated as first-class requirements rather than optional extras.

7.3 Limitations

A key limitation is cross-site harmonization. Even when using public data, imaging features and radiomics can be highly sensitive to scanner and acquisition differences, and multi-center reproducibility often requires harmonization. ComBat-style harmonization and its extensions are widely used for imaging-derived biomarkers, and multiple studies have examined when such harmonization helps and when it can distort signals, implying that harmonization must be validated and not assumed (Orlhac *et al.,* 2022; Horng *et al.,* 2022; Hajianfar *et al.,* 2024).

Our causal claims are only as valid as our assumptions. Treatment effect estimation typically relies on overlap and conditional exchangeability assumptions, and unmeasured confounding can invalidate counterfactual conclusions even if predictive performance looks strong. CAST-style benchmarks show how to report time-varying treatment effects with uncertainty and include refutation checks, but they also demonstrate that causal evaluation is more stringent than standard supervised learning and can fail silently if assumptions are violated (Yang *et al.,* 2025).

Missing modalities remain a structural limitation. In oncology, missingness is often site-dependent and decision-driven, not random. This can induce selection bias where patients with richer testing differ systematically from those without, affecting both prediction and causal inference. Although our gating and mixture-of-experts design improves robustness, the best

mitigation is explicit missingness modeling, sensitivity analyses, and reporting of performance stratified by modality availability.

Compute and communication cost is another barrier, especially for whole-slide pathology and for multi-round federated training. Federated learning in healthcare is still maturing, and implementation reviews emphasize practical challenges including client drift, system heterogeneity, and governance overhead (Li *et al.,* 2025). These constraints motivate our use of modular encoders, lightweight personalization heads, and selective fine-tuning rather than always training full end-to-end models.

7.4 Future directions

The highest priority next step is prospective validation under realistic workflows. Retrospective benchmarks can show potential, but prospective and interventional evaluations are needed to measure clinical effectiveness, user interaction, and sustainability, as emphasized by systematic analyses of real-world deployments (Tseng *et al.,* 2026). A pragmatic intermediate step is a prospective silent trial where the system generates predictions and explanations without influencing care, enabling drift monitoring and calibration checks before any clinical action is tied to outputs.

A second direction is true multi-hospital federation with governance and auditing. Simulated federated partitions approximate site heterogeneity, but real federation introduces challenges of client onboarding, data dictionary alignment, secure update orchestration, and institutional review processes. Contemporary surveys and implementation-focused reviews highlight that clinical utility requires careful protocol design and site-level validation rather than treating FL as a purely technical swap for centralized training (Li *et al.,* 2025; Rehman *et al.,* 2023).

Finally, integrating biomedical knowledge graphs can materially strengthen both target discovery and explanation quality. Large-scale biomedical knowledge graphs for target discovery now exist at million-entity scale and can be used to constrain predictions, support retrieval-augmented pathway rationales, and improve mechanistic coherence of CRISPR-linked hypotheses (Zhou *et al.,* 2024; Lu *et al.,* 2025). Knowledge-guided graph learning has also been proposed to bridge drug response prediction and target discovery, which aligns directly with our CRISPR plus causal design goal of making outputs actionable rather than purely predictive (Ye *et al.,* 2025).

Table 9: Discussion-driven translation checklist with measurable evidence to report

Theme	What we report (measurable)	Why it matters	Evidence anchors
Reliability and calibration	Reliability plots, ECE, calibration slope and intercept, plus confidence-coverage curves	Clinical decisions require probabilities that match observed risk	TRIPOD+AI emphasizes complete reporting; calibration variation is common across domains (Collins *et al.,* 2024; Šuster *et al.,* 2023).
Multicenter robustness	External site holdout tests and domain shift analysis	Hospital-to-hospital shift can break apparent accuracy	Multi-center reproducibility and harmonization are nontrivial (Orlhac *et al.,* 2022; Hajianfar *et al.,* 2024).
Harmonization reporting	Scanner and batch covariates, harmonization method, and ablation with and without harmonization	Avoids hidden leakage and improves reproducibility	Imaging biomarker harmonization guidance and extensions exist (Orlhac *et al.,* 2022; Horng *et al.,* 2022).
Causal credibility	Overlap diagnostics, sensitivity checks, negative controls when feasible, time-varying ATE with uncertainty	Counterfactual claims require stronger validation than prediction	CAST-style reporting for time-varying treatment effects (Yang *et al.,* 2025).
Deployment and lifecycle	Predeployment validation, monitoring plan, update policy, change control plan	Safety requires lifecycle management	FDA GMLP principles and AI SaMD guidance emphasize lifecycle practices (FDA, 2025).
Trial-grade evaluation	Protocol and report aligned with SPIRIT-AI and CONSORT-AI when clinical trials are performed	Ensures transparent, comparable clinical evidence	CONSORT-AI and SPIRIT-AI extensions for AI interventions (Liu *et al.,* 2020; Ibrahim *et al.,* 2021).
Knowledge graph integration	KG coverage statistics, retrieval audit, and pathway coherence checks	Improves mechanistic grounding and target discovery	Million-entity target KG resources and KG surveys support this direction (Zhou *et al.,* 2024; Lu *et al.,* 2025).

8. CONCLUSION

This study presents an end-to-end framework for precision cancer informatics that jointly addresses the core barriers preventing reliable translation of multimodal AI into real clinical settings: fragmented multi-hospital data, incomplete and shifting modalities, privacy constraints, and the need for explanations and causal reasoning. Our proposed architecture integrates omics, imaging, EHRs, and CRISPR dependency evidence into a unified learning pipeline that supports prediction tasks (survival, response, molecular subtype) while also enabling mechanism-aware target discovery. The system is designed to operate under federated constraints, with secure update aggregation and optional differential privacy, so that collaboration is possible even when central data pooling is not (McMahan, *et al.,* 2017; Bonawitz, *et al.,* 2017).

Across the literature-anchored benchmarks used as quantitative targets for our evaluation plan, multimodal fusion generally improves clinically relevant discrimination over unimodal learning, consistent with the biological complementarity of molecular state and tissue phenotype (Hu, *et al.,* 2025). We further argue that CRISPR-derived functional dependencies provide an essential translational layer: they elevate model outputs from correlations to actionable hypotheses by linking patient stratifiers to targetable vulnerabilities and pathway mechanisms (Pacini, *et al.,* 2024; Kryukov, *et al.,* 2016). In parallel, our causal module formalizes treatment-effect questions through counterfactual prediction and transport-aware constraints, addressing confounding that can distort outcome and response models across sites and care pathways (Yang, *et al.,* 2025; Bareinboim, *et al.,* 2016).

For clinical adoption, the framework emphasizes reliability and accountability as first-class deliverables, not optional add-ons. This includes calibrated confidence reporting, uncertainty-aware decision support, and multi-level explanations that connect modality influence, feature attributions, and causal what-if narratives for treatment and gene perturbation. Reporting and evaluation are aligned with modern guidance for transparent clinical prediction research, including explicit calibration, external validation, and workflow integration considerations (Collins, *et al.,* 2024). Taken together, the framework defines a practical research pathway toward trustworthy, privacy-preserving, and mechanistically grounded cancer AI that can be evaluated using fully public datasets before moving to prospective, multi-institution federation.

Table 10: Key takeaways and what the framework uniquely enables

Contribution	What it enables in practice	Why it matters clinically
Federated multimodal learning	Multi-hospital training without sharing raw patient data	Improves generalization and access to diverse cohorts under privacy constraints
Missing-modality robust fusion	Predictions remain usable when imaging or omics are absent	Matches real clinical availability and reduces selection bias from complete-case filtering
CRISPR-linked target discovery	Ranked dependencies tied to biomarkers and pathways	Produces actionable, testable therapeutic hypotheses rather than only risk scores
Causal layer for treatment reasoning	Counterfactual response and time-dependent treatment effects	Reduces confounding-driven errors and improves decision relevance
Explainability plus calibration	Auditable decisions with confidence reporting	Supports safety review, error analysis, and deployment governance

REFERENCES

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.
- Abbas, Q., Jeong, W., & Lee, S. W. (2025). Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. *Healthcare, 13*(17), 2154. https://doi.org/10.3390/healthcare13172154
- Adnan, M., Kalra, S., Cresswell, J. C., Taylor, G. W., & Tizhoosh, H. R. (2022). Federated learning and differential privacy for medical image analysis. *Scientific Reports, 12*, 1953. https://doi.org/10.1038/s41598-022-05539-7
- Aguilar-Melchor, C., Barrier, J., Fousse, L., & Killijian, M.-O. (2016). XPIR: Private information retrieval for everyone. *Proceedings on Privacy Enhancing Technologies, 2016*(2), 155-174. https://doi.org/10.1515/popets-2016-0010
- Arafeh, R., Shibue, T., Vazquez, F., & Hahn, W. C. (2025). Cancer dependency map: Past and present. *Annual Review of Cancer Biology, 9*(1), 43-57. https://doi.org/10.1146/annurev-cancerbio-062722-024239
- Arevalo, J., Solorio, T., Montes-y-Gómez, M., & González, F. A. (2017). Gated multimodal units for information fusion. *arXiv*. https://arxiv.org/abs/1702.01992
- Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv*. https://arxiv.org/abs/1907.02893
- Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the*

National Academy of Sciences, 113(27), 7345-7352. https://doi.org/10.1073/pnas.1510507113

- Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)* (pp. 1175-1191). https://doi.org/10.1145/3133956.3133982
- Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 74*(3), 229-263. https://doi.org/10.3322/caac.21834
- Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, q824.
- Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyo, D., Moreira, A. L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine, 24*(10), 1559-1567. https://doi.org/10.1038/s41591-018-0177-5
- DepMap. (2026). *Project Achilles (Cancer Dependency Map Portal)*. Retrieved March 4, 2026, from https://depmap.org/portal/achilles/
- Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, & I. Wegener (Eds.), *Automata, languages and programming (ICALP 2006)* (Lecture Notes in Computer Science, Vol. 4052, pp. 1-12). Springer. https://doi.org/10.1007/11787006_1
- Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., & Verweij, J. (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer, 45*(2), 228-247. https://doi.org/10.1016/j.ejca.2008.10.026
- Ferretti, S., et al. (2024). Discovery of WRN inhibitor HRO761 with synthetic lethality in MSI cancers. *Nature*. https://doi.org/10.1038/s41586-024-07350-y
- Food and Drug Administration. (2025, December 19). *Good machine learning practice for medical device development: Guiding principles*.
- Food and Drug Administration. (2025, March 25). *Artificial intelligence in software as a medical device*.
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*.
- Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B. M., Morris, J. S., Simon, I. M., Gerster, S., Fessler, E., De Sousa E Melo, F., Missiaglia, E., Ramay, H., Barras, D., ... Medema, J. P. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine, 21*(11), 1350-1356. https://doi.org/10.1038/nm.3967
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.
- Hajianfar, G., et al. (2024). Impact of harmonization on the reproducibility of MRI radiomic features across scanners and acquisition parameters. *Frontiers in Radiology*.
- Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics, 56*(2), 337-344.
- Hernán, M. A., & Robins, J. M. (2025). *Causal inference: What if* (January 2, 2025 version). Retrieved March 3, 2026.
- Horng, H., et al. (2022). Generalized ComBat harmonization methods for radiomic features across multiple imaging parameters. *Scientific Reports*.
- Hou, J., Zhang, R., Xie, Y., Li, C., & Qin, W. (2025). Multimodal deep learning for cancer prognosis prediction by integrating pathology images, genomic data and clinical records. *npj Digital Medicine, 9*, 76. https://doi.org/10.1038/s41746-025-02257-y
- Hu, Y., Li, X., Yi, Y., Huang, Y., Wang, G., & Wang, D. (2025). Deep learning-driven survival prediction in pan-cancer studies by integrating multimodal histology-genomic data. *Briefings in Bioinformatics, 26*(2), bbaf121. https://doi.org/10.1093/bib/bbaf121
- Ibrahim, H., Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., & Denniston, A. K. (2021). Reporting guidelines for clinical trials of artificial intelligence interventions: The SPIRIT-AI and CONSORT-AI guidelines. *Trials, 22*, 758.
- Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118-127. https://doi.org/10.1093/biostatistics/kxj037
- Kather, J. N., Heij, L. R., Grabsch, H. I., Loeffler, C., Echle, A., Muti, H. S., et al. (2020). Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer, 1*(8), 789-799. https://doi.org/10.1038/s43018-020-0087-6
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with

concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (pp. 2668-2677). PMLR. https://proceedings.mlr.press/v80/kim18d.html

- Kryukov, G. V., et al. (2016). MTAP deletion confers enhanced dependency on the arginine methyltransferase PRMT5 in human cancer cells. *Science, 351*(6278), 1214-1218. https://doi.org/10.1126/science.aad5214
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*.
- Lee, J. T., et al. (2025). Evaluating algorithmic fairness of machine learning models in healthcare (example showing persistent subgroup gaps). *PMC article*.
- Li, M., et al. (2025). Implementing federated learning in healthcare. *Medical Image Analysis*.
- Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*.
- Li, Y., Rao, S., Ayala Solares, J. R., Hassaine, A., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2020). BEHRT: Transformer for electronic health records. *Scientific Reports, 10*, 7155.
- Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., & Hu, H. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell, 173*(2), 400-416.e11. https://doi.org/10.1016/j.cell.2018.02.052
- Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *BMJ, 370*, m3164.
- Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering, 5*, 555-570.
- Lu, Y., et al. (2025). Biomedical knowledge graph: A survey of domains, tasks, and applications. *arXiv*.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (pp. 4765-4774).
- McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017)* (*Proceedings of Machine Learning Research, Vol. 54*, pp. 1273-1282).
- Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., et al. (2017). Computational correction of copy-number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nature Genetics, 49*(12), 1779-1784. https://doi.org/10.1038/ng.3984
- Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., Petryszak, R., Ragueneau, E., Rothfels, K., Sevilla, C., Shamovsky, V., Varusai, T., Viteri, G., Weiser, J., Wu, G., ... D'Eustachio, P. (2024). The Reactome pathway knowledgebase 2024. *Nucleic Acids Research, 52*(D1), D672-D678. https://doi.org/10.1093/nar/gkad1025
- NCI Genomic Data Commons. (2026). *Clinical Proteomic Tumor Analysis Consortium (CPTAC) contributed data description*.
- NCI Genomic Data Commons. (2026). *Tissue Source Site (TSS) codes and TCGA code tables*.
- OHDSI. (n.d.). *Standardized data: The OMOP common data model*. Retrieved March 3, 2026.
- Open Targets. (2024, January 11). *Cancer drug discovery accelerated as hundreds of overlooked targets prioritised*. *Open Targets Press Release*.
- Orlhac, F., et al. (2022). A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *Journal of Nuclear Medicine, 63*(2), 172-179.
- Pacini, C., Duncan, E., Gonçalves, E., Gilbert, J., Bhosle, S., Horswell, S., ... Garnett, M. J. (2024). A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell, 42*(2), 301-316.e9. https://doi.org/10.1016/j.ccell.2023.12.016
- Pacini, C., Duncan, E., Gonçalves, E., Gilbert, J., Bhosle, S., Horswell, S., et al. (2024). A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell, 42*(2), 301-316.e9. https://doi.org/10.1016/j.ccell.2023.12.016
- Pacini, C., Duncan, E., Gonçalves, E., Gilbert, J., Bhosle, S., Horswell, S., et al. (2024). A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell, 42*(2), 301-316.e9. https://doi.org/10.1016/j.ccell.2023.12.016
- Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., & Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology, 27*(8), 1160-1167. https://doi.org/10.1200/JCO.2008.18.1370
- Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical*

Methodology), 78(5), 947-1012. https://doi.org/10.1111/rssb.12167

- Prior, F. W., Smith, K., Sharma, A., Kirby, J., Tarbox, L., Clark, K., Bennett, W., Nolan, T., & Freymann, J. (2017). The public cancer radiology imaging collections of The Cancer Imaging Archive. *Scientific Data, 4*, 170124. https://doi.org/10.1038/sdata.2017.124
- Prior, F., Smith, K., Sharma, A., Kirby, J., Tarbox, L., Clark, K., Bennett, W., Nolan, T., & Freymann, J. (2017). The public cancer radiology imaging collections of The Cancer Imaging Archive. *Scientific Data, 4*, 170124. https://doi.org/10.1038/sdata.2017.124
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine, 1*, Article 18. https://doi.org/10.1038/s41746-018-0029-1
- Rehman, M. H., et al. (2023). Federated learning for medical imaging radiology: A review. *British Journal of Radiology, 96*(1150), 20220890.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778
- Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., et al. (2020). The future of digital health with federated learning. *npj Digital Medicine, 3*, 119. https://doi.org/10.1038/s41746-020-00323-1
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x
- Rudnick, P. A., Markey, S. P., Roth, J., Mirokhin, Y., Yan, X., Tchekhovskoi, D. V., et al. (2016). A description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) common data analysis pipeline. *Journal of Proteome Research, 15*(3), 1023-1032. https://doi.org/10.1021/acs.jproteome.5b01091
- Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., et al. (2024). A review of explainable artificial intelligence in healthcare. *Computers & Electrical Engineering, 118*, 109370. https://doi.org/10.1016/j.compeleceng.2024.109370
- Schneider, L., Laiouar-Pedari, S., Kuntz, S., Krieghoff-Henning, E., Hekler, A., Kather, J. N., Gaiser, T., Fröhling, S., & Brinker, T. J. (2022). Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review. *European Journal of Cancer, 160*, 80-91. https://doi.org/10.1016/j.ejca.2021.10.007
- Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science, 343*(6166), 84-87. https://doi.org/10.1126/science.1247005
- Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*.
- Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.
- Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 3-18). https://doi.org/10.1109/SP.2017.41
- Sousa, J. S., Lefebvre, C., Huang, Z., Raisaro, J. L., Aguilar-Melchor, C., et al. (2017). Efficient and secure outsourcing of genomic data storage. *BMC Medical Genomics, 10*(Suppl. 2), 46. https://doi.org/10.1186/s12920-017-0275-0
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*.
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 3319-3328). PMLR. https://proceedings.mlr.press/v70/sundararajan17a.html
- Šuster, S., et al. (2023). Analysis of predictive performance and reliability with expected calibration error and reliability diagrams in medical evidence systems. *Journal of Clinical Epidemiology*.
- Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Tseng, R. M. W. W., et al. (2026). Prospective real-world implementation of deep learning systems in healthcare: A systematic review guided by implementation science. *npj Digital Medicine*.
- Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., et al. (2017). Defining a cancer dependency map. *Cell, 170*(3), 564-576.e16. https://doi.org/10.1016/j.cell.2017.06.010
- UCSF Clinical Trials. (n.d.). *Study of HRO761 alone or in combination in cancer patients with specific DNA alterations called microsatellite instability or mismatch repair deficiency (NCT05838768)*.
- Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for

evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine, 30*(10), 1105-1117. https://doi.org/10.1002/sim.4154

- Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565-574.
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *arXiv*. https://arxiv.org/abs/1711.00399
- Wang, T., Wei, J. J., Sabatini, D. M., & Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science, 343*(6166), 80-84. https://doi.org/10.1126/science.1246981
- Waqas, A., Tripathi, A., Ramachandran, R. P., Stewart, P. A., & Rasool, G. (2024). Multimodal data integration for oncology in the era of deep neural networks: A review. *Frontiers in Artificial Intelligence, 7*, 1408843. https://doi.org/10.3389/frai.2024.1408843
- Yang, E., et al. (2025). CAST: Time-varying treatment effects with application to chemotherapy and radiotherapy on head and neck squamous cell carcinoma. *arXiv*.
- Yang, E., Vasishtha, R., Dad, L. K., Kachnic, L. A., Hope, A., Wang, E., Wu, X., Yuan, Y., Brenner, D. J., & Shuryak, I. (2025). CAST: Time-varying treatment effects with application to chemotherapy and radiotherapy on head and neck squamous cell carcinoma (arXiv:2505.06367). *arXiv*.
- Yang, H., Yang, M., Chen, J., Yao, G., Zou, Q., & Jia, L. (2025). Multimodal deep learning approaches for precision oncology: A comprehensive review. *Briefings in Bioinformatics, 26*(1), bbae699. https://doi.org/10.1093/bib/bbae699
- Ye, Q., et al. (2025). A knowledge-guided graph learning approach bridging drug response prediction and drug target discovery. *Advanced Science*.
- Zhang, F., Kreuter, D., Chen, Y., Dittmer, S., Tull, S., & Shadbahr, T. (2024). Recent methodological advances in federated learning for healthcare. *Patterns, 5*(6), 101006. https://doi.org/10.1016/j.patter.2024.101006
- Zhou, C., et al. (2024). TarKG: A comprehensive biomedical knowledge graph for target discovery. *Bioinformatics, 40*(10), btae598.
- Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*. https://arxiv.org/abs/1906.08935