From Real-World Data (RWD) to Digital Twins: Building Models for Patient-Level Counterfactual Prediction in Oncology

Sandra Griffith, PhD[1]; Joe Manfredonia, ME[1]; Marcello Ricottone, MA[1]; Richard Knoche, PhD[1]; Aaron B Cohen, MD, MSCE[1,2]; Jacqueline Law, PhD[1]; Melissa Estevez, MS[1]

[1]Flatiron Health, New York, NY; [2]Department of Medicine, Grossman School of Medicine, New York University, New York

MSR219

Scan for abstract and supplemental data

Background

- Digital twins (DT), models capable of generating patient-level counterfactual predictions, may improve oncology drug development success rates by informing trial design, contextualizing results, and enhancing statistical power
- Training DT models require large, diverse, and longitudinally-rich data, and an understanding of methodological approaches
- This study leverages the depth and scale of real-world data (RWD) to evaluate the feasibility of DT models, compare performance across approaches to understand strengths and limitations of different models, and evaluate the performance of a DT-powered composite risk score

Methods

- **Data source:** The US-based, longitudinal Flatiron Health Research Database—an electronic health record-derived (EHR), deidentified database, comprising patient-level data originated from ~220 US practices (~1600 sites of care; primarily community oncology settings) and curated via technology-enabled abstraction.[1] Features were generated using demographics, structured, and ML/LLM-extracted variables (eg, Charlson comorbidity index [CCI] and sites of metastases [SOM])
- **Setting:** The study included patients diagnosed with Stage IV non–small cell lung cancer initiating first-line (1L) platinum chemotherapy between 2011 and 2016, contemporaneous to chemotherapy use in trials
- **Main outcome measures:** Real-world overall survival (rwOS); patients were censored at the date of last activity in the EHR or the end of the study period
- **Statistical analysis:**
  - We trained four models (penalized pooled logistic regression [LR], XGBoost [XGB], Multi-layer Perceptron [MLP], and Graph Attention Network [GAT]) to predict rwOS using discrete time survival methods
  - Missing data were handled via multiple imputation with chained equations (LR and XGB) or a denoising autoencoder trained to reconstruct missing values from observed features (MLP, GAT)
  - Feature selection and hyperparamter optimization were performed iteratively on a held-out validation set
  - We used an outcome model-based approach to compute standardized survival curves in the held-out test set using patient-level predictions based on baseline patient characteristics
  - Model performance was assessed using C-index, area under the curve (AUC(t)), integrated Brier score (IBS), mean absolute difference (MAD) between predicted and observed survival curves, and median rwOS
  - Individual DT-powered composite risk scores were generated using the cumulative hazard based on the model predictions. The association between risk strata and rwOS was assessed with Kaplain-Meier methods, log-rank test, and Cox Proportional Hazards models

Results

- Participants**:** A total of 19,988 pts were randomly split 60/20/20 for train/validation/test
- All four methods performed well, comparable to published results using different methods[2] (**Figure 1**)
- LR/XGB demonstrated better discrimination (C-index, AUC[t]), while GAT exhibited superior calibration (MAD)
- Predicted median rwOS (9–10 months) aligned with observed rwOS (10 months) (**Figure 1**)
- Top features varied by method and included LLM-extracted variables (CCI, SOM), in addition to established prognostic variables (eg, ECOG) and laboratory results (eg, albumin) (LR: **Figure 2**; feature importance plots for additional models available via **QR code**)
- Risk groups derived from DT-powered composite risk scores successfully stratified rwOS, with the high-risk population demonstrating over 3-fold increase in the hazard of death compared to the low-risk group (**Figure 3**)

Figure 1. Observed vs predicted average real-world overall survival in Test set

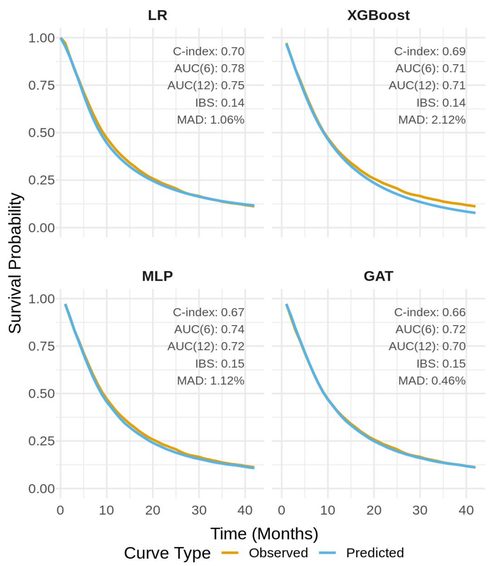


Figure 2. Penalized odds ratios (LR model)

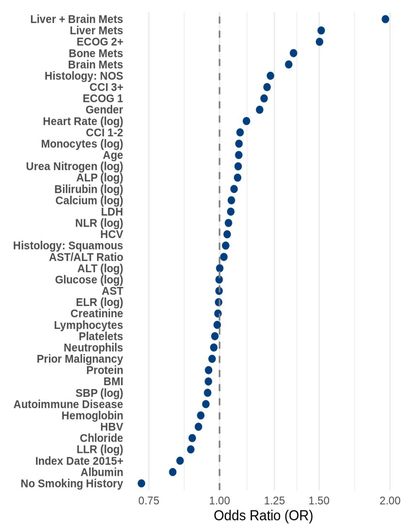
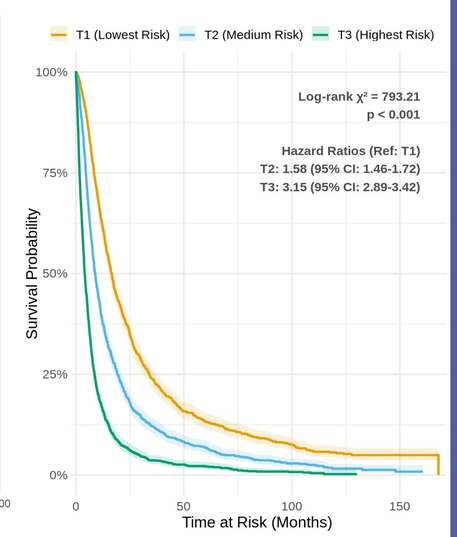


Figure 3. Kaplan-Meier curves stratified by risk score in Test set (LR model)



Abbreviations/definitions: ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; CCI, Charlson Comorbidity Index; ELR, eosinophils to leukocytes ratio; HBV, hepatitis B virus; HCV, hepatitis C virus; LDH, lactate dehydrogenase; LLR, lymphocytes to leukocytes ratio; mets, metastases; NLR, neutrophils to lymphocytes ratio; NOS, not otherwise specified; SBP, systolic blood pressure. Sites of Metastases (SOM) variables follow a hierarchical definition: Liver + Brain mets = both liver and brain mets; Liver mets = liver mets w/o brain mets; Brain mets = brain w/o liver mets; Bone mets = bone w/o liver or brain mets; reference category = other mets w/o liver, brain, or bone mets. Reference categories for other categorical variables: ECOG 0, Non-squamous histology, CCI 0, female gender, history of smoking.

RWD-powered **Digital Twin** models generate highly prognostic composite risk scores for **patient stratification**, demonstrating promising utility for clinical trial applications.

References

1. Flatiron Health. Database Characterization Guide. Flatiron.com. Published March 18, 2025. Accessed April 3, 2026. https://flatiron.com/database-characterization
2. Becker T, Weberpals J, Jegg AM, et al. An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. Annals of Oncology. 2020;31(11):1561-1568. doi:10.1016/j.annonc.2020.07.013

Disclosures: This study was sponsored by Flatiron Health, Inc.—an independent member of the Roche Group. SG, JM, MA, RK, AC, JL, and ME reported employment with Flatiron Health, Inc. and stock ownership in Roche. SG also reported employment at Griffith Scientific Consulting LLC and consulting fees from Flatiron Health. Data first presented at ISPOR 2026 in Philadelphia, PA on May 20, 2026.
Contact information: Jacqueline Law, jacqueline.law@flatiron.com

Discussion

- While simpler models (LR, XGB) achieved stronger discrimination, all four approaches produced well-calibrated survival curves (MAD < 2.2%), suggesting that for standard tabular clinical data, increased model complexity does not necessarily improve predictive accuracy
- Investing in richer LLM-based feature extraction from unstructured EHR data can augment existing feature sets to better align with RCT inclusion/exclusion criteria and meaningfully contribute to model explainability beyond what is available in standard structured datasets
- For DT applications that require transporting predictions to new populations (eg, from RWD to RCT cohorts), calibration performance may be as important as discrimination
- A DT-powered composite risk score can have multiple applications for drug development, including:
  - Patient stratification into risk groups that produce well-discriminated survival curves, which could be used to support prognostic enrichment or stratified randomization during trial design
  - Covariate adjustment, which has the potential to increase power for a clinical trial by mitigating bias and reducing variability in the treatment estimate
- External validation of DT models in clinical trial datasets is ongoing