Early Prediction of Overall Survival in Oncology Trials Using Tumor Dynamic Neural-ODE

Gengbo Liu[1]*, Victor Poon[1], Omid Bazgir[1], Phyllis Chan[1]
Kenta Yoshida[1], Pascal Chanu[1]
[1]Modeling & Simulation / Clinical Pharmacology, Genentech
{liu.gengbo, poon.victor, bazgir.omid, chan.phyllis,
yoshida.kenta, chanu.pascal}@gene.com

Abstract

Accurate early prediction of overall survival (OS) is crucial in oncology drug development, where late-stage trial failures remain common. Traditional tumor growth inhibition (TGI) models provide predictive biomarkers of OS but rely on restrictive assumptions that limit generalizability. We evaluate Tumor Dynamic Neural Ordinary Differential Equations (TDNODE), a pharmacology-informed neural network that learns patient-specific tumor kinetics from longitudinal tumor size data. Using 8,121 patients across 10 phase II/III atezolizumab trials in five tumor types, we show that TDNODE-derived kinetic metrics consistently outperform TGI metrics in OS prediction, achieving higher Concordance Indices and lower Integrated Brier Scores. Notably, accurate prediction is possible with as little as 16 weeks of tumor data, often surpassing TGI models on full trajectories. These results demonstrate that TDNODE generalizes across trials and tumor types while enabling early survival prediction, offering a promising framework to support oncology development and accelerate decision-making.

1 Introduction

Drug development is a high-stakes and resource-intensive process, with oncology trials facing particularly high failure rates Chirmule and Ghalsasi [2025]. Accurately predicting clinical trial outcomes—such as overall survival (OS), a widely used primary endpoint in oncology—early in the trial is critical for guiding strategic decisions and reducing the risk of late-stage failures.

Tumor dynamic models, which characterize longitudinal tumor size data, have been a cornerstone of predictive pharmacometrics in oncology Ribba et al. [2014]. These approaches often employ tumor growth inhibition (TGI) models that estimate metrics such as tumor shrinkage rates, regrowth rates, and time-to-regrowth. Such metrics have been shown to correlate with OS across multiple solid tumor types Claret et al. [2009], Stein et al. [2011]. However, TGI models rely on restrictive phenomenological assumptions, which can introduce inductive bias and limit their generalizability across trials, tumor types, and treatment settings Chan et al. [2021].

Neural ordinary differential equations (Neural ODEs) provide a more flexible paradigm for modeling continuous-time dynamics from longitudinal data Chen et al. [2018]. The Tumor Dynamic Neural Ordinary Differential Equation (TDNODE) framework Laurie and Lu [2023] integrates pharmacological priors with neural ODEs to learn patient-specific tumor kinetics directly from longitudinal tumor size data. TDNODE produces individualized kinetic parameters that more accurately capture tumor progression and have demonstrated improved OS prediction compared to traditional TGI metrics.

*Corresponding author: liu.gengbo@gene.com

The NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development (AI4D3 2025), San Diego, California, USA, 2025.

Nonetheless, prior evaluations have been limited to a single trial, leaving open questions regarding cross-trial generalizability, robustness across tumor types, and the feasibility of early OS prediction from truncated data.

In this work, we extend the evaluation of TDNODE to a large-scale, pan-indication dataset of 8,121 patients across 10 phase II/III atezolizumab trials spanning five solid tumor types. Specifically, we:

- **Cross-trial and multi-tumor generalizability:** Demonstrate that TDNODE-derived metrics are robust across independent clinical trials and diverse oncology indications.
- **Early prediction capability:** Evaluate the feasibility of accurate OS prediction using truncated observation windows (16, 24, and 32 weeks).

Our results show that TDNODE-derived kinetic metrics consistently outperform traditional TGI metrics, enabling accurate OS prediction even with limited data. This highlights TDNODE as a promising framework for de-risking oncology development and accelerating early clinical decision-making.

2 Methods

2.1 Data Summary

We assemble a large, pan-indication dataset of 8,121 patients from 10 atezolizumab clinical trials. The dataset includes patients with five cancer types: non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC), renal cell carcinoma (RCC), triple-negative breast cancer (TNBC), and metastatic urothelial carcinoma (mUC). Tumor lesions were assessed by computed tomography (CT) or magnetic resonance imaging (MRI) at baseline and at regular intervals (approximately every 6–9 weeks for the first year and then every 9–12 weeks thereafter, until disease progression, death, or loss of follow-up). Longitudinal tumor size data was defined as the sum of the longest diameters (SLD) of target lesions per RECIST 1.1 guidelines Eisenhauer et al. [2009] at each visit and was used to estimate TGI metrics. The SLD data per patient were split into baseline (prior to therapy initiation) and post-treatment (after therapy initiation) components. Post-treatment data were further split into seen and unseen measurements. To evaluate early prediction capabilities, we constructed observation windows at 16, 24, and 32 weeks and assessed whether tumor metrics (TGI and TDNODE-derived) could predict OS at the end of the clinical trial.

In addition, we incorporated 10 baseline patient-level covariates, including inflammatory markers—baseline albumin (ALBU), neutrophil count (NEU), neutrophil-to-lymphocyte ratio (NLR), , lactate dehydrogenase (LDH), and total protein (TPRO)—along with common prognostic factors: ECOG performance status, hemoglobin (HGB), presence of liver metastasis status, number of metastatic sites, and years since initial diagnosis (YSD). OS outcomes were represented as time-to-event data with censoring. For evaluation, we adopted a cross-trial validation design: in each iteration, nine trials were used for training and one held out for testing, repeated across all 10 trials. This cross validation approach was applied for both TDNODE based kinetic metrics generation and OS prediction.

2.2 Tumor Growth Inhibition (TGI) Metrics

To derive TGI metrics, we applied a biexponential TGI model to the longitudinal SLD data, as described by Stein et al. Stein et al. [2011]. This nonlinear mixed-effects model, implemented in NONlinear Mixed Effects Modeling (NONMEM) Bauer [2019], was fitted separately to each trial to characterize individual tumor dynamics.

The model is given by:

$$SLD(t) = TS_0 \cdot e^{-KG \cdot t}, \quad t \leq 0$$
$$SLD(t) = TS_0 \cdot \left[e^{KG \cdot t} + e^{-KS \cdot t} - 1\right], \quad t > 0$$

where TS_0 is baseline tumor size, KS the tumor shrinkage rate constant, KG the tumor regrowth rate constant, and t time since treatment initiation. This formulation captures the competing processes of tumor shrinkage due to treatment and regrowth driven by natural progression during treatment.

From this model, we extracted four key patient-level TGI metrics:

- **Tumor growth rate constant (KG)**: the estimated rate at which the tumor grows before treatment or regrows during treatment. Lower values indicate more effective growth suppression.
- **Tumor shrinkage rate constant (KS)**: the estimated rate at which the tumor decreases in response to treatment. Higher values reflect more effective tumor reduction.
- **Time to tumor regrowth (TTG)**: the predicted time from treatment initiation until tumor regrowth resumes. Longer TTG suggests a more durable treatment effect.
- **Tumor ratio at week 24 (TR24)**: the predicted ratio of tumor size at week 24 relative to baseline.

Together, these metrics summarize patient-specific tumor dynamics and have been shown to be predictive of OS across multiple solid tumor indications Chan et al. [2021] Velasquez et al. [2024].

2.3 TDNODE-Derived Kinetic Metrics

The Tumor Dynamic Neural-ODE (TDNODE) framework Laurie and Lu [2023], is designed to learn patient-specific tumor kinetics directly from longitudinal SLD data. Here, we briefly summarize the architecture for completeness. Each patient's measurement times are normalized by their last observed timepoint, and SLD values are Z-score standardized using training-set statistics. Pre-treatment tumor profiles are processed through the *Initial Condition Encoder* (ICE) to generate the patient-specific initial tumor state $z_i(0)$. Truncated post-treatment profiles are processed through the *Parameter Encoder* (PE), which outputs a 4-dimensional vector of kinetic metrics, $p_i \in \mathbb{R}^4$, scaled by each patient's last observed time to yield interpretable units of inverse time. These parameters capture individualized tumor shrinkage and regrowth dynamics, analogous to traditional TGI metrics but learned in a data-driven manner.

The Neural-ODE module integrates the initial state and kinetic metrics via:

$$\frac{dz(t)}{dt} = f_\theta(z(t), p_i), \quad t \in [0, T],$$

where f_θ is a neural network shared across patients. The solution $z(t)$ produces a continuous latent trajectory, which is then reduced to scalar tumor size predictions. Training is performed end-to-end to minimize the root mean square error (RMSE) between predicted and observed SLD measurements. The key output for downstream survival modeling is the patient-specific kinetic metric vector p_i. These learned metrics provide a compact representation of individual tumor dynamics, enabling their direct use as predictors in survival models.

2.4 Overall Survival Prediction with XGBoost

In total, we derived four TGI metrics and four TDNODE-derived kinetic metrics from the SLD data at observation windows of 16 weeks, 24 weeks, 32 weeks, and the full follow-up. Each set of tumor dynamic metrics (TGI or TDNODE) was combined with baseline covariates and used as input for OS prediction. For modeling, we employed the XGBoost Survival Embeddings (XGBSE) framework Vieira et al. [2021], which adapts gradient-boosted decision trees for survival analysis. To assess generalizability, we used a cross-trial validation design: in each iteration, nine trials were used for training and one held out for testing, repeated across all 10 trials. Performance was evaluated using the concordance index (C-index), which quantifies the model's ability to correctly rank patients by survival time, and the Integrated Brier Score (IBS) Graf et al. [1999], which measures the overall accuracy of predicted survival probabilities across time.

3 Results

We first assessed the ability of TDNODE to predict unseen SLD trajectories under a cross-trial validation setting, given different amounts of observed (seen) SLD data. As shown in Figure 1, the predictive performance improved with longer observation windows across all trials, as reflected by increasing R^2 values. Notably, R^2 exceeded 0.9 in all cases, indicating consistently good fit. Importantly, TDNODE was able to accurately predict tumor dynamics even when the test trial indication was not represented in the training set, highlighting the TDNODE model's strong generalizability.

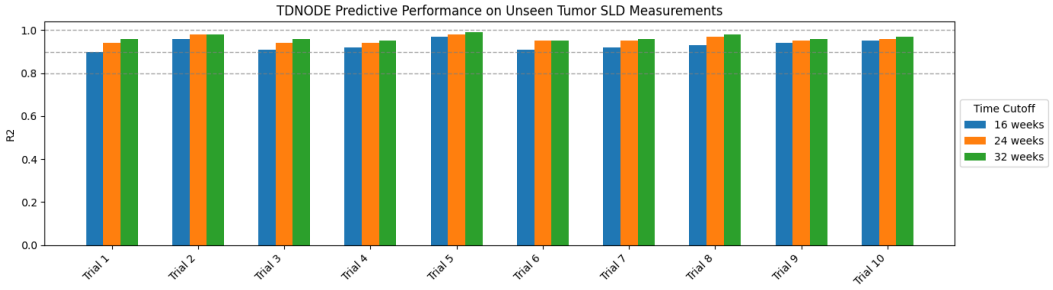

Figure 1: TDNODE-predicted tumor dynamics under cross-trial validation. Accuracy improves with longer observation windows.

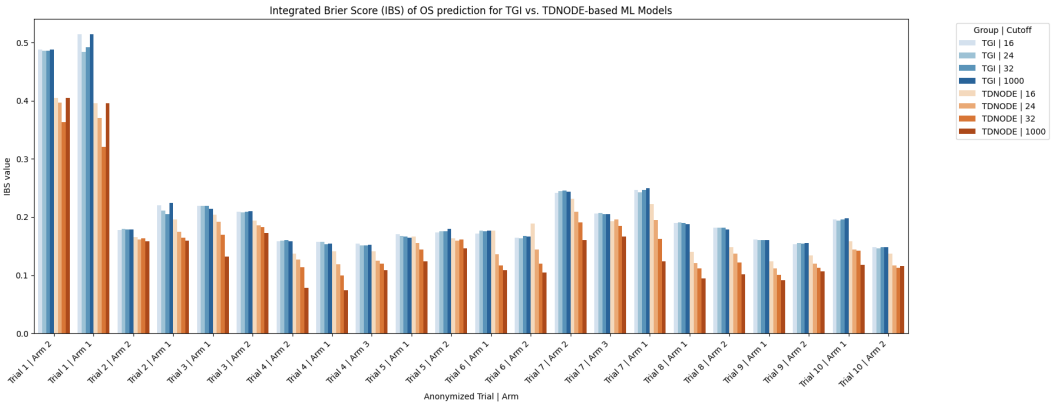

Figure 2: Comparison of OS prediction using TDNODE-derived vs. TGI-derived metrics across trials. TDNODE consistently achieves lower Brier scores, even with truncated observation windows.

Overall, these results demonstrate that TDNODE provides robust predictive performance for tumor dynamics in unseen trial data.

We next evaluated the predictive power of TDNODE-derived kinetic metrics compared to traditional TGI metrics for overall survival. Figure 2 and Figure 3 in Appendix summarizes performance across all trials, tumor types, and observation windows. Using XGBSE models, TGI-based metrics achieved median concordance indices (C-index) above 0.60 and integrated Brier scores (IBS) below 0.51 across all trials. In contrast, TDNODE-based metrics consistently outperformed TGI, yielding C-index values above 0.64 and IBS values below 0.40. Among the two approaches, TDNODE-derived kinetic metrics consistently outperformed TGI-based metrics, yielding higher C-index values and lower IBS across all trials, treatment arms, and observation cutoffs. For Trial 1, both approaches showed weaker predictive performance compared to other trials in IBS, yet the TDNODE-based model still achieved better results than the TGI-based model. We also observed a clear trend that predictive accuracy improved with longer observation windows. Importantly, in most settings, models trained on TDNODE metrics using only 16 weeks of observed data surpassed TGI-based models trained on the entire SLD trajectory. Together, these results demonstrate that TDNODE-derived kinetic metrics not only generalize well across diverse trial settings but also enable accurate and early OS prediction, making them a promising framework for informing future clinical trials and accelerating oncology drug development.

4 Conclusion

We presented the first large-scale, multi-trial evaluation of TDNODE for overall survival prediction in oncology. Across diverse tumor types and trials, TDNODE-derived metrics consistently outperformed traditional TGI metrics and enabled accurate early prediction with limited data. These results establish TDNODE as a promising tool for de-risking oncology trials and guiding early drug development

decisions. Our future work will focus on validating predictions with Kaplan–Meier and hazard ratio analyses.

References

Robert J Bauer. Nonmem tutorial part i: description of commands and options, with simple examples of population analysis. *CPT: pharmacometrics & systems pharmacology*, 8(8):525–537, 2019.

Phyllis Chan, Mathilde Marchand, Kenta Yoshida, Shweta Vadhavkar, Nina Wang, Alyse Lin, Benjamin Wu, Marcus Ballinger, Nitzan Sternheim, Jin Y Jin, and René Bruno. Prediction of overall survival in patients across solid tumors following atezolizumab treatments: A tumor growth inhibition–overall survival modeling framework. *CPT: pharmacometrics & systems pharmacology*, 10(10):1171–1182, 2021.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Narendra Chirmule and Vivek V Ghalsasi. The life cycle of drug development. In *Approved: The Life Cycle of Drug Development*, pages 1–34. Springer Nature, 2025.

Laurent Claret, Pascal Girard, Paulo M Hoff, Eric Van Cutsem, Klaas P Zuideveld, Karin Jorga, Jan Fagerberg, and René Bruno. Model-based prediction of phase iii overall survival in colorectal cancer on the basis of phase ii tumor dynamics. *Journal of Clinical Oncology*, 27(25):4103–4108, 2009.

Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, Daniel Sargent, Richard Ford, Janet Dancey, Steven Arbuck, Samuel Gwyther, Margaret Mooney, Larry Rubinstein, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

Mark Laurie and James Lu. Explainable deep learning for tumor dynamic modeling and overall survival prediction using neural-ode. *NPJ Systems Biology and Applications*, 9(1):58, 2023.

Benjamin Ribba, Nick H Holford, Paolo Magni, I Trocóniz, I Gueorguieva, Pascal Girard, C Sarr, M Elishmereni, Charlotte Kloft, and Lena E Friberg. A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis. *CPT: pharmacometrics & systems pharmacology*, 3(5):1–10, 2014.

Wilfred D Stein, James L Gulley, Jeff Schlom, Ravi A Madan, William Dahut, William D Figg, Yang-min Ning, Philip M Arlen, Douglas Price, Susan E Bates, and Tito Fojo. Tumor regression and growth rates determined in five intramural nci prostate cancer trials: the growth rate constant as an indicator of therapeutic efficacy. *Clinical Cancer Research*, 17(4):907–917, 2011.

Erick Velasquez, Nastya Kassir, Sravanthi Cheeti, Denison Kuruvilla, Rucha Sane, Steve Dang, Dale Miles, and James Lu. Predicting overall survival from tumor dynamics metrics using parametric statistical and machine learning models: application to patients with ret-altered solid tumors. *Frontiers in Artificial Intelligence*, 7:1412865, 2024.

Diego Vieira, Guilherme Gimenez, Gabriel Marmorela, and Victor Estima. Xgboost survival embeddings: Improving statistical properties of xgboost survival analysis implementation. *Loft Python*, 2021.

A Appendix

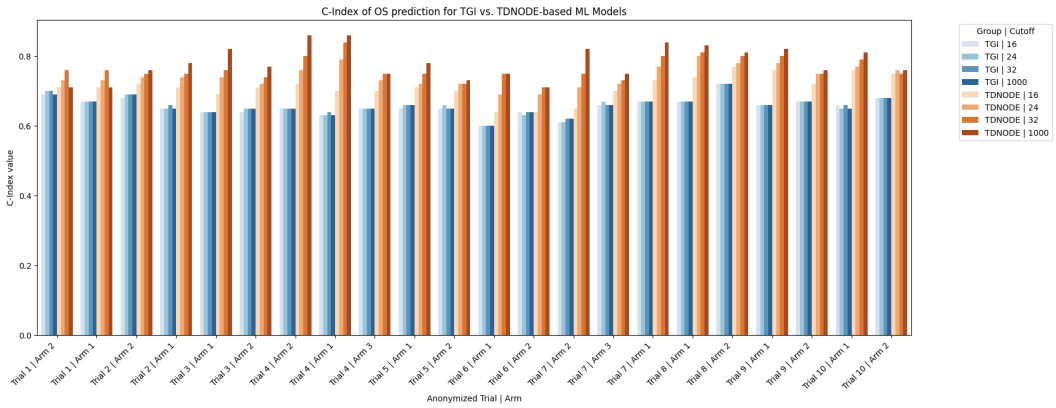

Figure 3: Comparison of OS prediction using TDNODE-derived vs. TGI-derived metrics across trials. TDNODE consistently achieves higher concordance indices, even with truncated observation windows.