CPT: Pharmacometrics & Systems Pharmacology

ARTICLE OPEN ACCESS

Follow-Up Bias in Tumor Dynamic Modeling: A Comparison of Classical and Neural-ODE Approaches

David C. Turner[1] | Mark Laurie[1] | Shuhui Li[1,2] | Huy X. Ngo[1] | Chunze Li[1] | Jin Y. Jin[1] | James Lu[1] | Logan Brooks[1]

[1]Department of Clinical Pharmacology, Genentech, South San Francisco, California, USA | [2]Department of Pharmaceutical Sciences, University of Maryland, Baltimore, Maryland, USA

Correspondence: David C. Turner (turner.david_@gene.com)

Received: 26 October 2025 | **Revised:** 2 February 2026 | **Accepted:** 26 February 2026

Keywords: artificial intelligence | machine learning | MIDD | oncology | tumor dynamic modeling

ABSTRACT

Tumor dynamic models are vital for evaluating oncology treatments and guiding clinical drug development decisions. However, few studies rigorously assess their predictive capabilities, especially when forecasting tumor trajectories from clinical trials with short or inconsistent follow-up across treatment arms. Poor predictive performance or biases related to follow-up time could potentially limit the general utility of tumor growth inhibition (TGI) models. This study quantitatively evaluates prediction bias across several established tumor dynamic models, comparing five classical pharmacometric TGI models with the deep learning-based Tumor Dynamic Neural-ODE (TDNODE) framework. Using time-truncated clinical trial data from 3106 patients with non-small cell lung cancer (NSCLC) across four completed atezolizumab phase III studies, we consistently observed moderate-to-high positive bias in the predictions from pharmacometric models, particularly with more limited follow-up. By examining the structures of these models and comparing them to observed data, we highlight how the assumed kinetic patterns potentially lead to biased parameter estimation and systemic overestimation of tumor size when applied to immature datasets. In contrast, the TDNODE framework, using deep learning, demonstrated promising early results, exhibiting improved predictive performance in the same evaluations. These findings underscore the critical need to address prediction bias in tumor dynamic modeling with immature data and to consider alternative approaches to established paradigms for certain drug development applications. This study also generally demonstrates the potential of novel methods, such as deep learning, to potentially enhance the reliability of tumor dynamics modeling, especially in challenging early-phase clinical decision-making scenarios.

1 | Introduction

Evaluating tumor response to treatment is crucial in developing novel oncology drugs and holds significant influence on the interpretations of treatment efficacy, early Go-/No-go development decisions, and exposure-response evaluations for dose selection. Traditionally, tumor response assessment has relied on criteria such as the Response Evaluation Criteria in Solid Tumors (RECIST), which involves categorizing tumor size data into response strata and further dichotomizing them into responders and non-responders [1]. However, this approach fails to fully capture the rich, longitudinal information contained within the on-study tumor measurements, such as the rates of tumor shrinkage and regrowth, the specific depth of response (nadir), and the overall trajectory over time, and may oversimplify these complex response patterns. Many of these inherent limitations are well recognized and often critiqued [2–4]. The US FDA's Project Optimus, which aims to improve dose-finding and optimization strategies in oncology, has highlighted the need for increased application of model-informed drug development (MIDD) and integration of more advanced quantitative approaches to better capture treatment response for dose

Study Highlights

- What is the current knowledge on the topic?
  - Tumor dynamic models (TGI) are vital for evaluating oncology treatments and guiding drug development decisions. However, few studies rigorously assess their predictive capabilities when forecasting tumor trajectories from clinical trials with short or inconsistent follow-up, which can lead to biases. Addressing follow-up bias and thoroughly assessing model performance are recognized needs in the field.
- What question did this study address?
  - This study addressed how prediction bias varies across five classical TGI models and the deep learning-based Tumor Dynamic Neural-ODE (TDNODE) framework when used to predict future tumor trajectories from time-truncated clinical trial data in patients with NSCLC.
- What does this study add to our knowledge?
  - The study demonstrates that classical TGI pharmacometric models consistently exhibit moderate-to-high positive prediction bias, significantly overestimating tumor size when applied to immature datasets. This bias is linked to model structural assumptions and increasing errors in key parameter estimation (e.g., the growth parameter KG). In contrast, the TDNODE framework consistently demonstrated improved predictive performance and substantially lower prediction bias across all scenarios evaluated.
- How might this change drug discovery, development, and/or therapeutics?
  - These findings underscore the critical need to address prediction bias in tumor dynamic modeling, especially when using immature data prevalent in early-phase clinical trials. Using alternative, more flexible methods like TDNODE could potentially enhance the reliability of tumor dynamics modeling, thereby supporting early clinical decision-making scenarios and dose optimization efforts.

optimization purposes in early clinical trials. In addition to the already robust applications toward internal drug development decision-making, Project Optimus has created new opportunities for application of tumor dynamic modeling more broadly as a regulatory tool in early-stage development.

The fundamental purpose of tumor dynamics or tumor growth inhibition (TGI) models is to quantify tumor growth dynamics and the effects of therapeutic agents, describing the change in tumor size over time, often quantified as the sum of the longest diameters (SLD) of target lesions. Various TGI models have been developed and widely applied in oncology drug development, demonstrating their usefulness in decision-making [5–7], particularly in randomized phase Ib/II clinical settings [8]. These models were originally developed and validated during an era dominated by cytotoxic chemotherapy and targeted therapies and typically incorporate mathematical expressions that reflect the anticipated biological mechanisms of these treatments—notably, an initial tumor shrinkage phase followed by eventual resistance and regrowth. Hence, the mathematical structure of these models often includes terms for exponential growth and drug-induced decay [9–13], which align with observed clinical patterns of these traditional oncology treatments, where acquired resistance through secondary mutations or compensatory pathway activations is common.

The advent of immune checkpoint inhibitors (ICIs) and various immune-oncology (IO) agents, among other emerging therapeutic modalities, introduced novel patterns of tumor response that sometimes challenge the assumptions of classical parametric models [4, 14]. Unlike the predictable decline-then-regrowth pattern often seen with chemotherapy, ICIs can often produce durable responses and long-term stable disease states that sometimes persist, even after treatment cessation [4]. These sustained responses conflict with some models' assumption of inevitable progression. Furthermore, phenomena such as pseudoprogression and dissociated responses—where an initial growth can occur prior to shrinkage or different lesions within the same patient respond differently to treatment—can present general challenges for modeling [4, 15]. These distinct response patterns imply that the mathematical structures developed for chemotherapy-era treatments lack the required degree of generalizability to accurately describe outcomes for patients receiving IO treatments. Despite these challenges, classical tumor dynamics models have been widely used and many times considered sufficient to describe patient response to IO treatments in different situations [13, 16].

Aside from novel response patterns, another obstacle for routine application of tumor dynamic modeling, especially in early-phase drug development, is the need to analyze incomplete or truncated clinical trial datasets. Due to various factors, early development decisions, such as an initial Phase II dose selection or Go-/No-go choices, are often required before mature trial data become available, and sparse tumor measurements with limited follow-up may impact the accuracy of estimated tumor metrics [17]. The task of estimating and accounting for follow-up bias, particularly when current models include unbounded exponential growth terms, is a recognized concern that may amplify bias in long-term predictions if the model's parameters are not appropriately estimated using early data [18].

Therefore, evaluating the assumptions of tumor dynamics model structures and potential follow-up bias for different treatments are aspects that have gained recognition as important components of model verification, though these aspects remain less frequently characterized and somewhat underreported in the literature. This highlights a future need for more thorough assessments of model performance and increased publications detailing the relationships between follow-up duration and derived tumor parameter estimates across different models, treatments, and trial settings. Such information, that is, defining these models' operational characteristics under more diverse clinical contexts, would be invaluable for guiding industry standards and refining tumor dynamics modeling best practices, especially when using data from early clinical trials.

A range of tumor dynamics models have been proposed [9–13], each with different functional forms, also leading to some

potential for inductive bias inherent to their specific mathematical equations [19]. Consequently, predictive capabilities and interpretation of these models, for example, in estimating tumor doubling times, may vary significantly depending on the chosen structure of the model [20]. In response to these challenges and acknowledging recent advances in the field, non-parametric deep learning (DL) approaches have been gaining increasing interest as alternatives to classical models and their parametric assumptions. Notably, within translational oncology and clinical tumor size modeling, a DL approach based on the neural-ordinary differential equation (neural-ODE) formalism has demonstrated recent success in enabling efficient data-driven model discovery [3, 21].

Hence, motivated by the overall predictive potential of tumor dynamic modeling in guiding early clinical trial decision-making and untapped applications in early development dose optimization support, particularly in single-arm cohorts with immature follow-up [8], this study aims to evaluate different modeling approaches for predicting tumor size trajectories based on hypothetical early trial data. Specifically, the focus is on characterizing the predictive capability of various published models using a bias assessment approach applicable to both classical pharmacometric [9–13] and DL-based models [21] of tumor dynamics. Leveraging a large dataset from four phase III clinical trials in patients with NSCLC, encompassing treatment arms spanning IO agents, chemotherapy, targeted therapies, and their combinations, each model's ability to predict future 'unseen' tumor sizes was assessed using truncated datasets ranging from 25% to 75% of the total trial duration. This data truncation was implemented to reflect hypothetical decision-making scenarios at regular milestone intervals of the trial duration, such as a potential pre-planned interim analysis with less-than-fully mature trial follow-up. Through objective quantitative comparison, we aim to highlight the predictive capabilities of different tumor dynamics modeling approaches and encourage the use of similar measures of evaluating tumor dynamic models in future publications and model applications prior to decision-support.

2 | Methods

2.1 | Modeling Development Equations and Architectures

This study is intended to evaluate the ability of Tumor dynamic neural-ODE (TDNODE) [21] and 5 classical and analytical TGI models (Stein, Claret, Bonate, Moore, and Wang [9–13]) to predict longitudinal tumor dynamics under different clinical trial settings based on collected clinical trial data and enrollment information; please refer to the Supporting Information for complete details. This study compared five published TGI models (Stein, Claret, Bonate, Moore, Wang) implemented as nonlinear mixed-effects models in NONMEM, with random effects on key kinetic parameters (e.g., growth and drug-effect rates), against the Tumor Dynamic Neural-ODE (TDNODE) framework. TDNODE uses an encoder–decoder design: a recurrent encoder summarizes each patient's baseline and on-treatment SLD history into a latent state, which a Neural-ODE decoder integrates to generate a continuous tumor-size trajectory. To improve generalization, TDNODE training used L2 weight decay (10−3) and a masking scheme that prevents any use of post-cutoff ("unseen") observations during fitting (details in Supporting Information).

2.2 | Experimental Design

Figure 1 summarizes the evaluation workflow. For each trial, we defined total study duration as the interval between first dose in the first enrolled patient and the last available SLD measurement in the last enrolled patient. We then created three "immature" data cuts at 25%, 50%, and 75% of that duration. For a given cutoff date, only SLD observations collected on/before the cutoff were used to fit/train each model; all later observations were held out as "unseen" for evaluation. Because enrollment was staggered, patients enrolling after an early cutoff contribute few (or no) scans at 25%, so the number of evaluable patients increases at later cuts, preserving the real-world imbalance in follow-up seen at interim readouts (illustrated in Figure S1).

For each patient, models were calibrated using that patient's "seen" data (estimating individual parameters or latent state), and forecasts for the "unseen" period were generated conditional on that same patient history (i.e., patient-specific rather than population-mean predictions). The identical train/test split was used for all models.

2.3 | Data Summary

The tumor dynamics models were assessed using a compiled dataset consisting of four randomized, phase III multicenter clinical trials that enrolled patients with NSCLC: IMPower130, IMPower131, IMPower132, and IMPower150 [20–22]. SLD data from both the control arm and atezolizumab-containing arm were included in the current analysis. For complete trial details, please refer to prior publications. Brief details are provided in the Supporting Information.

Patients in the atezolizumab-containing treatment arms received the therapy intravenously at 1200 mg every 3 weeks. In each trial, the SLD scans for specified target lesions were collected at baseline as well as at regular intervals on-study ranging from 6 to 9 weeks. All trials obtained institutional review board approval, and all patients provided written informed consent prior to enrollment.

2.4 | Model Development

Refer to Supporting Information for details.

2.5 | Model Evaluation

After fitting/training on the "seen" data, we evaluated predictions against observed SLD at held-out ("unseen") time points using mean/median relative percent error (RPE; $\varepsilon = 1$ mm for zero SLD), root mean squared error (RMSE), and R^2 (formulas in S1).

To probe sensitivity to informative dropout (patients who progress/discontinue early and therefore have fewer observations), we

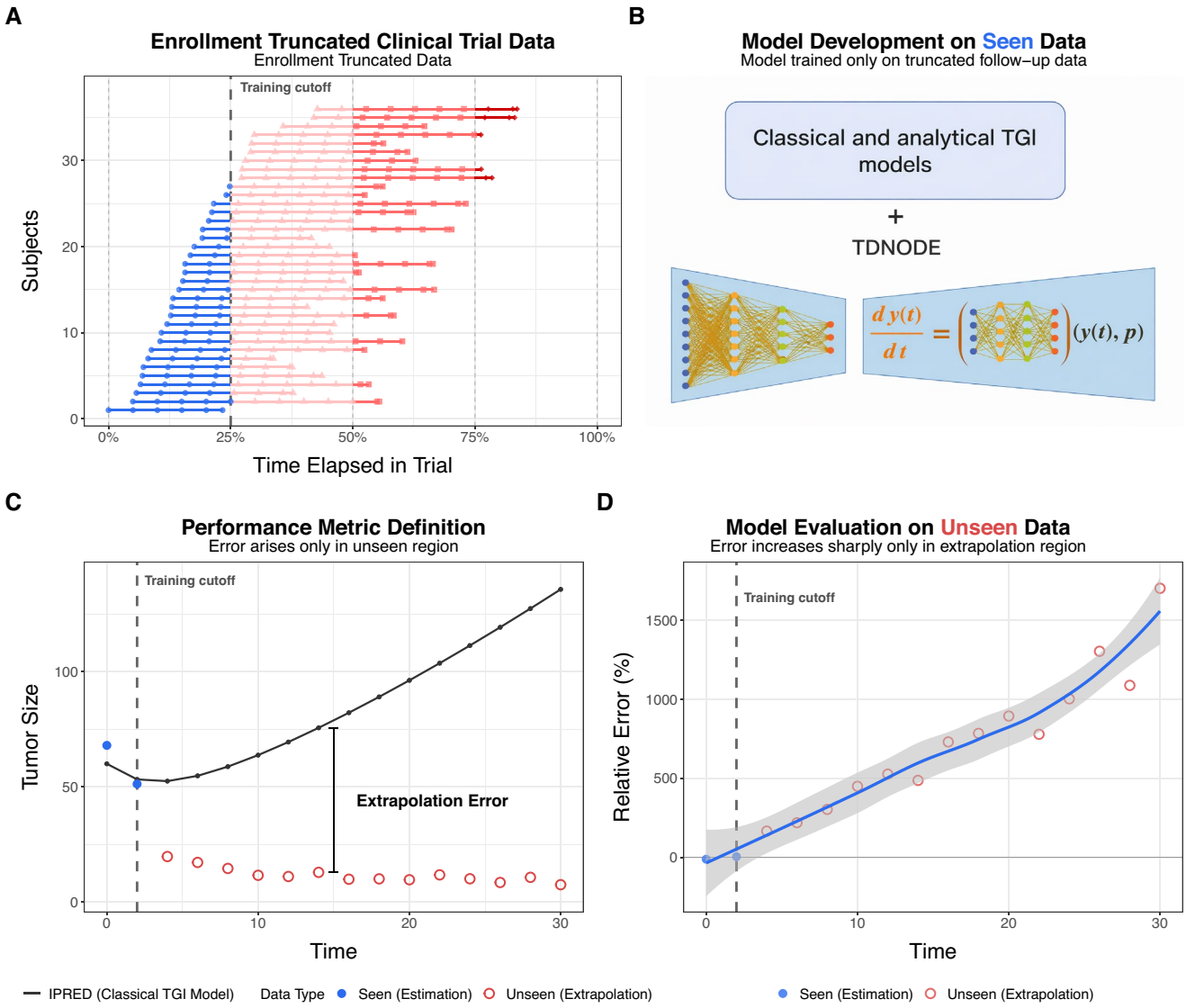

FIGURE 1 | Schematic of the workflow in the comparison of Classical TGI models and TDNODE for tumor dynamic prediction. (A) Taking data from a clinical trial and based on the enrollment times splitting into a "seen" and "unseen" set. (B) Classical TGI models and TDNODE were applied to model the "seen" data. (C) Comparison of individual prediction against the unseen values is used in the performance metric evaluation. (D) Evaluation of model predictive bias (defined as the percent difference between the tumor dynamic prediction and observation) on the unseen data. DV, observed value of dependent variable (SLD); IPRED, individual predictions; SLD, sum-of-longest-diameter; TDNODE, tumor dynamic neural-ODE; TGI, tumor growth inhibition.

performed a weighted bootstrap (500 iterations) that up-weighted patients with fewer scans when resampling prediction errors (conceptually similar to inverse-probability weighting). In each iteration, we resampled patients with replacement and then sampled a fixed number of observations per selected patient. This resampling was used as a post hoc bias analysis tool in our study, not as a routine modeling step. In prospective applications, one would ideally handle informative dropout analytically (e.g., via joint modeling of tumor size and survival) rather than via resampling.

2.6 | Tumor Dynamic Pattern Categorization

To examine whether prediction error differed by response pattern, we categorized each patient's full SLD trajectory into seven groups (IPR, SDP, SDM, PRP, PRM, CRP, CRM) using RECIST-like thresholds; full definitions are provided in the Supporting Information.

3 | Results

A total of 4949 patients with non-small cell lung cancer (NSCLC) were enrolled in the studies included in this analysis, forming the basis of our tumor dynamics modeling dataset. From this cohort, 3106 patients had two or more on-study scans, qualifying them as evaluable for this assessment of prediction bias. These 3106 patients, spanning across eight treatment arms, contributed a total of 20,232 SLD observations, with a median of six scans per patient. Summary

statistics detailing the number of patients, seen and unseen SLD measurements in each truncation scenario, and the median number of measurements are presented in Table 1, while additional visualizations of seen versus unseen data across different truncation scenarios are provided in Figures S1 and S2; Table S1. Importantly, because of staggered patient enrollment, the number of patients contributing 'seen' versus 'unseen' data differs by truncation level (Table 1). Patients enrolled later have no scans in the early cut (25% follow-up), so the evaluable sample grows for the 50% and 75% scenarios. Thus, differences in model performance across follow-up times are considered indicative of performance at a particular trial follow-up, reflecting both the data maturity and total sample size or number of scans.

3.1 | Consistent Alignment of Models With Observed Tumor Size Data in Training Set

Comparisons between observed data and the individual predicted values within the seen data provide insights into each model's performance and interpolation capability within a given follow-up time window. Generally, most observed tumor size data points are aligned closely with the identity line (slope of 1) in plots of observed versus model-estimated SLD values across different truncation scenarios (Figure 2). Model diagnostics based on residuals, such as conditional weighted residuals (CWRES) versus predicted values and time, showed some bias particularly in the truncated data but overall captured the observed data reasonably, suggesting tumor dynamics models generally provide adequate description of SLD data within the 'seen' training datasets, although stability and over parameterization issues still occur.

TABLE 1 | Overview of patient and observation information at 25%, 50% and 75% follow-up time.

	Number of subjects	Number of observations	Number observations per subject[a]
25% trial follow-up			
Estimation	349	985	3 (2–6)
Extrapolation	265	1600	5 (1–17)
50% trial follow-up			
Estimation	1771	7100	4 (2–12)
Extrapolation	1037	5265	4 (1–12)
75% trial follow-up			
Estimation	3111	16,579	5 (2–15)
Extrapolation	1121	4165	4 (1–7)

Note: For each choice of follow-up time, the 5 classical TGI models and TDNODE were fitted or trained using seen scans and then evaluated at the time values corresponding to the set of unseen tumor size measurements. Patients with less than two seen scans.
Abbreviations: TDNODE, tumor dynamic neural-ODE; TGI, tumor growth inhibition.
[a]Median (min-max).

3.2 | Model Extrapolation Performance and Prediction Error

When the models were used to make predictions beyond the truncation point to 'unseen' time points, the agreement between observed and predicted values deteriorated compared to the fits of the training data. As demonstrated in Figure 2 and other error metrics, the five classical TGI models showed systematically increasing positive bias over longer extrapolation horizons. In particular, these models displayed a tendency to consistently overpredict tumor sizes at later follow-up times, resulting in higher relative percent error (RPE) and root mean square error (RMSE). This consistent growth in prediction error over time is clearly visualized in Figure 3. In contrast, the TDNODE model consistently maintained a lower prediction error. For instance, when trained on only 25% of the follow-up data, the classical Stein model exhibited a mean RPE of 334.06%, while the TDNODE model's was substantially lower at 61.18% (Table 2). This preliminary finding suggests that a DL-based, neural-ODE approach, less constrained by assumptions of a particular mathematical functional form, may better capture tumor trajectories from these truncated datasets.

3.3 | Contextual Considerations and Model Limitations

Classical TGI models, developed largely for cytotoxic settings where tumors often shrink and later regrow, may not fully capture IO-associated patterns such as delayed responses or prolonged stable disease. Across both IO+ and IO− arms, classical models tended to overpredict late tumor rebound, with overprediction more pronounced in the IO+ arms (Table S3). Consistent with this, SLD-based progression after achieving response or stable disease was relatively infrequent, particularly in IO+ arms (Table S6 and Figure S3). All approaches—including TDNODE—had difficulty forecasting CR trajectories, but TDNODE generally showed improved performance across most trajectory categorizations (Tables S4 and S5).

3.4 | Illustrative Example of Extrapolation Bias and Model Comparison

Figure 4 illustrates a subset of patients who achieved a partial response (PR) and maintained that response without progression throughout the trial follow-up period. Solid lines and shaded regions represent model predictions (with confidence intervals) up to data truncation, while dashed lines indicate the observed SLD trajectory post-truncation. TGI models often displayed upward drift in SLD predictions over time, suggesting a loss of treatment effect or impending regrowth, thereby overestimating tumor size where this did not occur. This overprediction bias became more pronounced with extended extrapolation beyond the training window. Conversely, the TDNODE framework more accurately tracked sustained tumor shrinkage or stable disease, with predictions remaining near the actual observed trajectory, even with up to 75% follow-up truncation. This consistency may imply that TDNODE

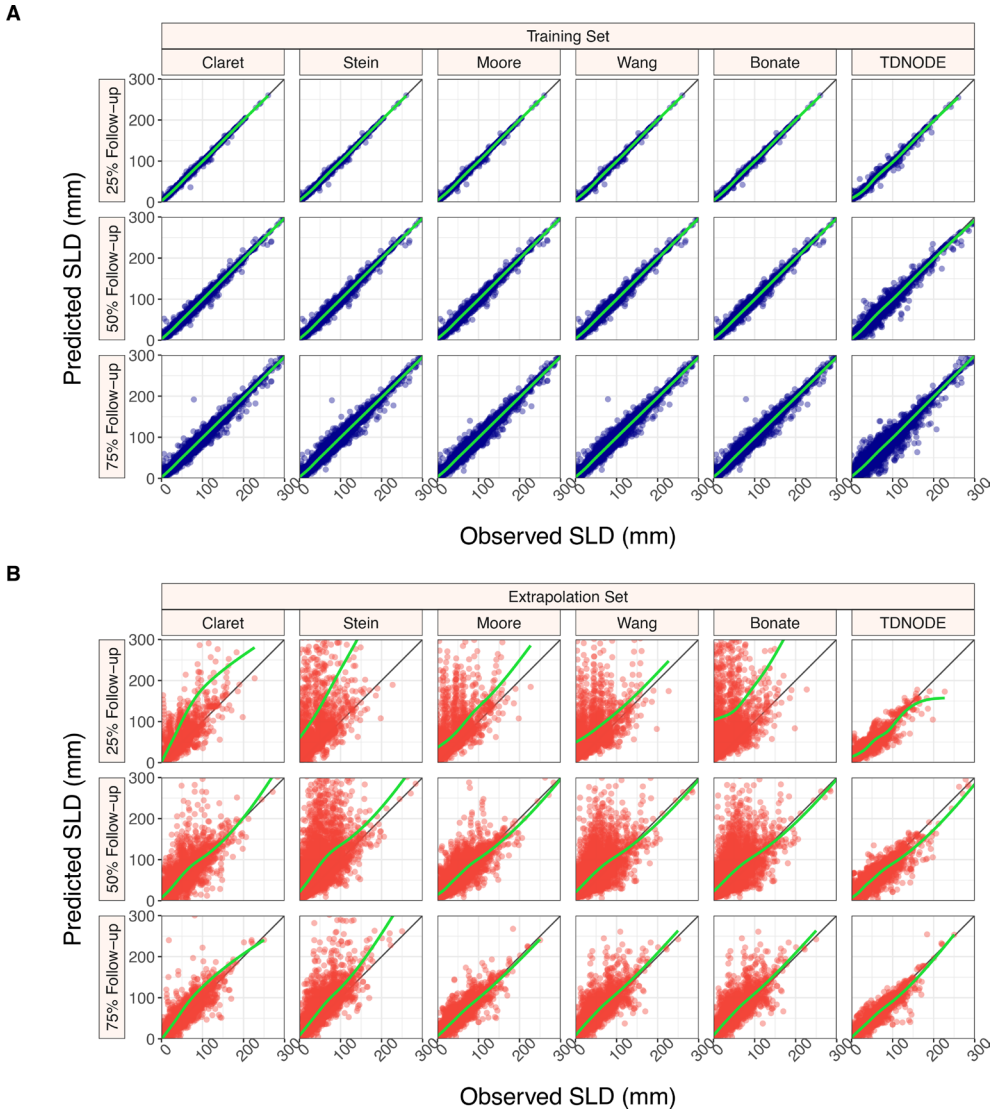

FIGURE 2 | Comparison of Classical TGI models and TDNODE predictions versus observations used in model estimation (goodness-of-fit on 'seen' training data) (A); Comparison of Classical TGI models and TDNODE predictions versus extrapolated observations not used in estimation ('unseen' extrapolation data) (B). SLD, sum of longest diameter.

is more flexibly guided by the observed data patterns learned despite a limited training interval.

3.5 | Connection Between Parameter Estimation Error and Extrapolation Bias

Lastly, we examined how truncated data may influence parameter estimations used in decision-making. As an illustrative example, for the Stein TGI model, the growth parameter (KG) estimated from truncated datasets showed increasing deviation from its "ground truth" value estimated on the complete dataset, and these deviations correlated strongly with greater extrapolation biases (Figure 5). Relatively speaking, the tumor shrinkage parameter showed no such correlation. This pattern underscores how limited early trial information can potentially mislead growth/shrinkage parameter estimation in TGI models, especially if true biological dynamics diverge from model assumptions.

4 | Discussion

This study provides a comprehensive evaluation of how follow-up duration in clinical trial data can impact the extrapolation performance of tumor dynamics models in patients with NSCLC treated with both atezolizumab and chemotherapy. Our systematic analysis, employing date-based truncation of target lesion SLD datasets to simulate immature trial readouts, reveals a consistent trend: published tumor dynamics models exhibit increasingly positive prediction bias with shorter follow-up, frequently overestimating future tumor sizes and suggesting a premature return of tumor growth. This finding potentially underscores a critical consideration for the application of these models in early drug development, where decisions are often made with incomplete data, which could lead to an overly conservative interpretation of long-term efficacy.

Varying degrees of prediction bias were observed among the tested TGI models. Such bias can stem from errors in individual

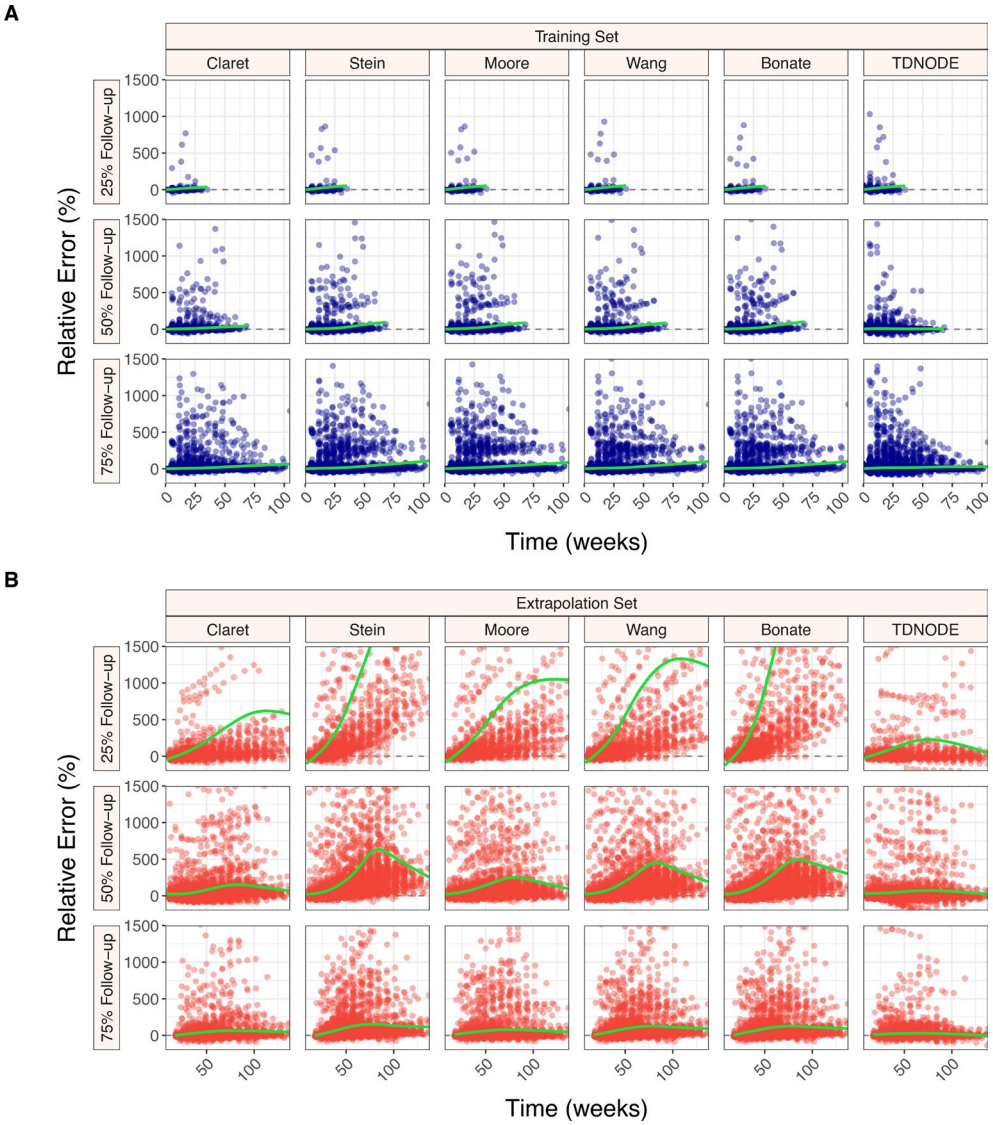

FIGURE 3 | Comparison of Classical TGI models and TDNODE's relative error on seen (A) and unseen (B) clinical data. All models were fitted or trained using available estimation data (shown as blue points) with follow-up time indicated as the plot label, and each model was used to make individual predictions at time values beyond them (i.e., "extrapolation", shown as red points). While the TGI models (from left to right columns: Claret, Stein, Moore, Wang, and Bonate, respectively) exhibited a clear tendency to over predict actual tumor size values and thereby resulting in a positive bias which grows with time, TDNODE (last column to the right) exhibits significantly less bias. Relative Percent Error (RPE) calculated as defined in the methods section. *x*-axis 'time' is the elapsed time since baseline for each individual patient. TDNODE, tumor dynamic neural-ODE; TGI, tumor growth inhibition.

patient parameter estimation, which may be particularly susceptible to shrinkage or misestimation when models are fitted to incomplete datasets. In essence, with limited data, a TGI model might estimate a parameter set that adequately describes the 'seen' observations in a given patient, yet this set may not accurately represent the true underlying tumor kinetics that would be revealed with more complete follow-up. When applied, potentially biased parameter estimates determine an entire predicted tumor trajectory beyond the measurement window. Therefore, while a model might effectively describe the seen data, the inferred underlying dynamics (e.g., growth rates or secondary metrics) derived from its parameters can still be biased. This implies that follow-up bias in tumor dynamics models has potential implications not just explicitly for long-term predictions, but also potential impact on the accuracy of model-derived parameters and descriptive metrics, even in situations where no overt extrapolation is performed.

Classical TGI models, including those developed by Stein, Claret, Bonate, Moore, and Wang, predominantly rely on mathematical expressions incorporating somewhat simple exponential growth and decay terms. Development of these empirical model structures was largely informed by the typical response patterns observed with cytotoxic chemotherapies, that is, characterized by initial tumor shrinkage followed by eventual resistance and regrowth. However, the diverse response patterns associated with immune checkpoint inhibitors like atezolizumab, including delayed responses and prolonged periods of stable disease, frequently deviate from these assumed kinetic profiles. The exponential growth terms in these classical models predispose

TABLE 2 | Quantification of bias and error for TGI and TDNODE models assessed at various values of follow-up times.

	Mean RPE (%)[a,b]	Median RPE (%)[a,b]	R^2[a,c]	RMSE[a,d]
25% enrollment				
Claret	97.11 (26.61)	3.93 (1.42)	0.20 (0.26)	264.47 (221.56)
Stein	334.06 (68.35)	18.32 (3.68)	0.07 (0.03)	141.82 (30.29)
Moore	193.46 (39.60)	9.47 (1.98)	0.34 (0.05)	46.89 (5.00)
Wang	237.81 (41.79)	17.69 (3.43)	0.25 (0.05)	48.22 (4.64)
Bonate	461.75 (93.37)	28.06 (4.74)	0.04 (0.02)	133.76 (31.35)
TDNODE	61.18 (14.72)	−3.09 (1.09)	0.81 (0.02)	16.83 (0.81)
50% enrollment				
Claret	44.64 (5.50)	3.95 (0.59)	0.54 (0.14)	37.31 (10.89)
Stein	137.77 (13.97)	15.71 (1.22)	0.33 (0.04)	57.89 (5.50)
Moore	77.23 (8.19)	6.78 (0.61)	0.75 (0.02)	20.83 (0.93)
Wang	115.55 (10.83)	14.47 (1.22)	0.52 (0.04)	32.97 (1.85)
Bonate	123.25 (11.85)	15.03 (1.17)	0.51 (0.03)	33.61 (1.82)
TDNODE	32.20 (5.15)	−0.77 (0.48)	0.86 (0.01)	14.95 (0.49)
75% enrollment				
Claret	33.08 (7.03)	5.66 (0.44)	0.31 (0.32)	173.52 (154.76)
Stein	79.30 (5.57)	22.11 (0.87)	0.46 (0.09)	47.65 (10.47)
Moore	46.84 (4.27)	7.22 (0.59)	0.85 (0.01)	14.13 (0.45)
Wang	68.28 (4.68)	20.94 (1.13)	0.68 (0.03)	24.95 (2.13)
Bonate	68.32 (4.80)	20.96 (1.19)	0.68 (0.03)	25.01 (2.14)
TDNODE	19.49 (2.45)	1.87 (0.36)	0.90 (0.01)	10.99 (0.41)

Note: For each value of follow-up time (25%, 50% and 75% of total trial duration), all models were fitted or trained with the corresponding "seen" data and subsequently evaluated on the remaining unseen data. The predictive performance was evaluated using several evaluation metrics based on the unseen data. To derive the predictive performance statistics, 500 bootstraps were performed from which the respective metric was computed.
[a]Values represent the bootstrapped mean (SD) unless indicated otherwise; SD, standard deviation.
[b]Relative % Error (RPE) = 100 × ((Predicted − Observed)/Observed + epsilon).
[c]R^2, *R*-squared.
[d]RMSE, root mean square error.

them to predict eventual tumor rebound, a tendency that becomes problematic when applied to truncated datasets from IO trials where long-term durable responses or stable disease by SLD is prevalent. Our analysis suggests that a fundamental mismatch between these models' structural assumptions and the observed tumor dynamics, especially the relative infrequency of SLD-based progression after initial response or stable disease appears to contribute to this prediction bias. For instance, the Stein model, with its assumption of unconstrained exponential growth, exhibited the most pronounced extrapolation errors with bootstrapped mean RPE ranging from 79% at the 75% truncation scenario to 334% at 25% truncation. While the Claret model, incorporating a resistance parameter that moderates the tumor growth term over time, showed relatively less bias (RPE of 33%, 45%, and 97% for 75%, 50%, and 25% truncation, respectively), the underlying assumption of inevitable regrowth still contributed to overpredictions in many cases of sustained response. Tumor regrowth patterns assumed by all of these models appear to be a less common occurrence in the observed data while sustained responses or stable disease of SLD target lesions comprise approximately 60% of the cases observed in the current dataset of patients treated with atezolizumab. It should be noted that more complex 'immune-response' or non-resistance tumor models exist (e.g., models allowing prolonged tumor control or oscillating tumor-immune dynamics), but these were not included in our analysis. We chose to evaluate the standard pharmacometric models that are most commonly used in practice for early decision-making, even in IO trials, to illustrate the extent of bias under current modeling paradigms.

The predominance of stable post-response trajectories in this evaluable cohort also highlights measurement and selection considerations. We classified progression using target-lesion SLD only, whereas RECIST progression also includes non-target/new lesions and clinical deterioration; therefore, SLD-based "no progression" does not necessarily imply absence of clinical progression. Additionally, we required ≥2 on-study SLD measurements to evaluate prediction, which excludes patients who

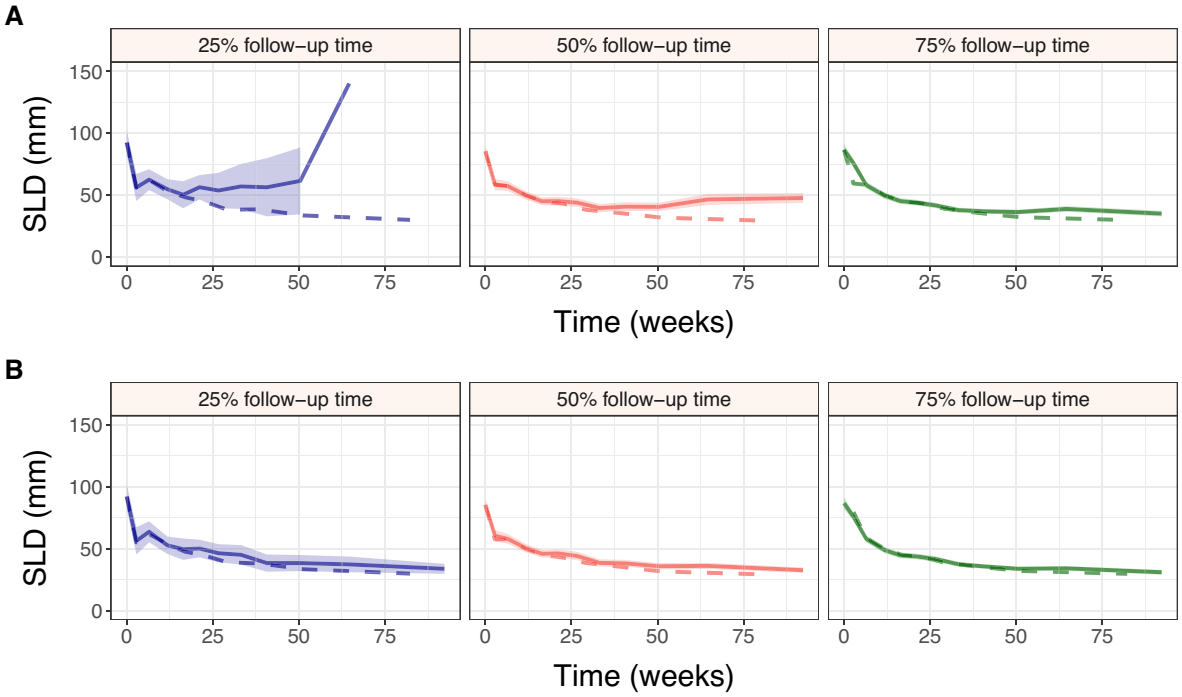

FIGURE 4 | Comparison of the observed versus predicted SLD between Claret TGI model (A) and TNDODE (B) across the subset of patients who had a partial response (PR) with no progression based on the modified RECIST categories outlined in the methods. Solid line represents the mean predicted tumor dynamic profile across time, shaded region represents the 95% confidence interval around the prediction, and dashed line represents the mean observed tumor dynamic profile across time.

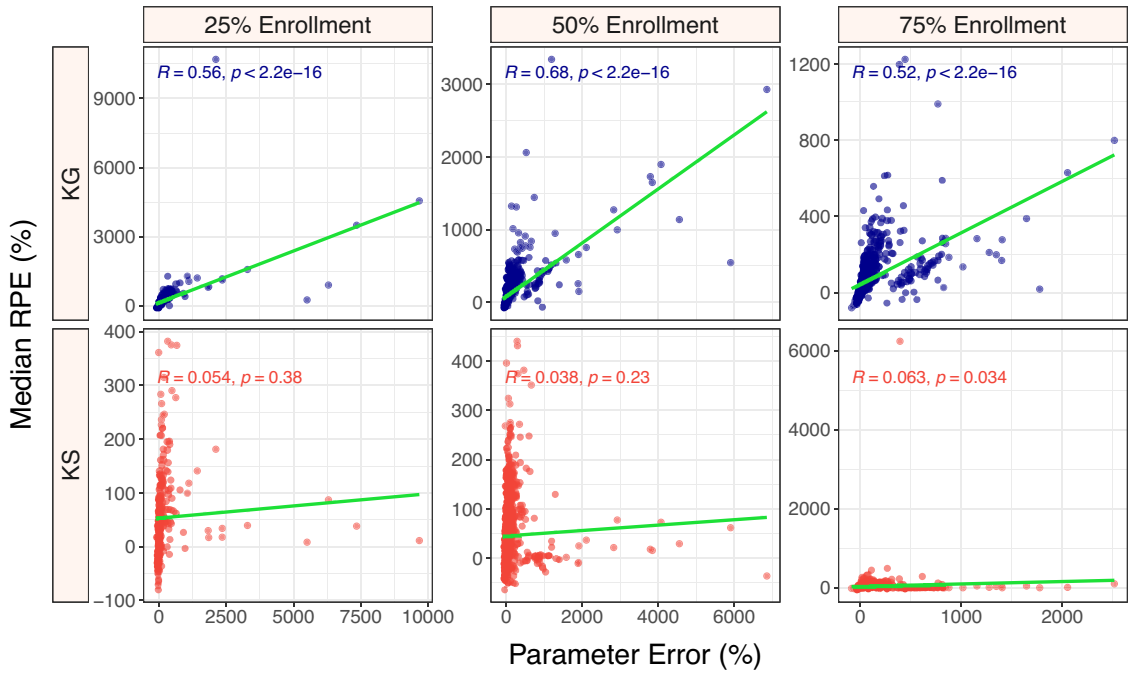

FIGURE 5 | Relationship between extrapolation error and parameter estimate error for the Stein model across enrollment cuts. Extrapolation error defined as the median relative percent error (RPE) for unseen data. Parameter error was defined as the relative percent error between the truncated individual parameter estimates and the ground truth parameter estimates based on the complete dataset. The blue line represents a linear fit and associated standard error. R = Pearson's correlation coefficient and its associated p-value.

progressed at the first follow-up and discontinued without further scans. This necessary design choice likely underrepresents early progressors and should be considered when interpreting both the observed trajectory frequencies and models' apparent ability to forecast progression.

The observed link between trial duration and predictive bias suggests that derived tumor dynamics metrics are potentially similarly affected. This potential connection between parameter estimates and trial duration has possible implications for the use of these metrics in decision-making contexts with limited follow-up, such as early Phase Ib/II studies [8]. As our illustrative analysis of the Stein model demonstrated, the growth parameter (Kg) estimated from truncated data showed increasing deviation from the "ground truth" value obtained from the full dataset, and this parameter bias was strongly correlated with increased extrapolation error (Figure 5; $p < 2.2 \times 10^{-16}$). This finding highlights the potential for early, limited data to potentially yield misleading tumor model parameter estimates, particularly for tumor growth-related parameters, when the assumed model structure does not align with the true underlying tumor dynamics and there is insufficient data to establish the true dynamics. In practical terms, an early-trial model that overestimates the tumor growth rate (Kg)—as we saw with truncated data—could lead teams to falsely conclude a lack of drug efficacy, potentially triggering a premature No-go decision for an agent that might actually show sustained benefit with longer follow-up. Conversely, underestimating growth due to model assumptions could mask a failing therapy. This underscores that biased model parameters can directly misinform critical Go/No-go and dose optimization decisions. Our findings suggest that deploying more flexible data-driven models (like TDNODE) in interim analyses could potentially reduce the risk of these decision errors by providing more accurate patient-specific forecasts when data are immature. The DL-based TDNODE framework, unconstrained by pre-defined mathematical assumptions, consistently exhibited the lowest prediction bias across all truncation scenarios outlined here. Its ability to learn complex patterns directly from the data appears to allow for more accurate capture of new types of tumor trajectories, even with similarly limited information, suggesting a potential advantage of such flexible approaches in mitigating some of the inductive biases inherent in classical pharmacometric models. While these early results for TDNODE are promising, further research is warranted to validate the robustness of TDNODE or other similar approaches across more diverse clinical settings and tumor types to further define the operational characteristics of this framework.

More broadly, our results support prior recommendations to evaluate tumor dynamic models under time-truncated conditions to understand operational performance before applying them for decisions [18]. It is already often acknowledged in survival analyses of time-to-event data that follow-up bias and time-dependent factors can often result in misleading associations [22, 23] and have, for example, motivated the use of joint models of tumor kinetics and overall survival (OS) and some prior evaluations of follow-up duration [24]. Potential mitigations against tumor modeling follow-up bias highlighted here include ensuring comparable follow-up across arms, restricting use to settings with sufficient follow-up for the endpoint of interest, or considering adopting alternative model structures that better match modern IO response patterns. Historically, many successful TGI applications have relied on the derivation of relative metrics (e.g., comparing TGI parameters or hazard ratios between concurrently enrolling randomized groups), which can be more robust to shared follow-up limitations. In contrast, absolute SLD forecasting in single-arm or early-readout settings places greater weight on the structural assumptions and patient-level parameter identifiability, making systematic follow-up bias potentially more consequential. Insights from this work suggest that directly tackling the fundamental causes of tumor dynamics model-related follow-up biases and validating model structural assumptions, rather than solely relying on relative comparisons, could refine these approaches in such a way that it would enhance the reliability of future results in a broader range of clinical decision-making scenarios.

This study has limitations. First, analyses were restricted to NSCLC and treatments within the atezolizumab program; generalizability to other tumor types, modalities, and earlier-phase trials remains to be established. Second, the inverse-weighted bootstrap was used as a retrospective sensitivity analysis to probe informative dropout; prospective applications should ideally address dropout analytically (e.g., joint tumor kinetics and time-to-event modeling). Third, although TDNODE showed improved predictive performance here, its mechanistic interpretability is limited and warrants future work (e.g., explainable AI or hybrid models). Finally, the "ground truth" parameters estimated on the full dataset remain model-dependent, and understanding how follow-up-driven bias propagates into downstream frameworks (e.g., TGI–OS models) is an important next step.

In summary, our analysis underscores the importance of addressing follow-up bias to further expand successful applications of tumor dynamics modeling in drug development. The observed positive bias in classical TGI model predictions emphasizes the need for careful consideration when applying these models to immature datasets, which are particularly prevalent in early-phase clinical trials. This study also demonstrates the potential of alternative DL-based approaches, such as TDNODE, to improve the reliability of longitudinal SLD predictions in early-phase clinical decision-making. Continued research focused on understanding and mitigating bias, validating novel tumor dynamics modeling frameworks with modern oncology treatments, and defining operational characteristics to ensure the applicability of these tools across diverse clinical settings is essential for advancing model-informed drug development.

Author Contributions

All authors wrote the manuscript; D.C.T. and J.L. designed the research; M.L., S.L., H.X.N., and L.B. performed the research; M.L., S.L., H.X.N., and L.B. analyzed the data.

Acknowledgments

We would like to express our gratitude to the following for stimulating discussions as well as supporting this work: Matts Kågedal, René Bruno, Pascal Chanu, and Amita Joshi. Large language model tools

were used for language editing and refinement during the preparation of this manuscript.

Funding

This work was supported by F. Hoffmann-La Roche Ltd.

Conflicts of Interest

All authors are current or former employees of Genentech Inc., and hold stocks and options in F. Hoffmann-La Roche Ltd. As an Associate Editor for *CPT: Pharmacometrics & Systems Pharmacology*, James Lu was not involved in the review or decision process for this paper.

References

1. E. A. Eisenhauer, P. Therasse, J. Bogaerts, et al., "New Response Evaluation Criteria in Solid Tumours: Revised RECIST Guideline (Version 1.1)," *European Journal of Cancer Oxford, England* 45, no. 2 (1990): 228–247, https://doi.org/10.1016/j.ejca.2008.10.026.

2. D. F. Heitjan, A. Manni, and R. J. Santen, "Statistical Analysis of In Vivo Tumor Growth Experiments," *Cancer Research* 53, no. 24 (1993): 6042–6050.

3. W. F. Forrest, B. Alicke, O. Mayba, et al., "Generalized Additive Mixed Modeling of Longitudinal Tumor Growth Reduces Bias and Improves Decision Making in Translational Oncology," *Cancer Research* 80, no. 22 (2020): 5089–5097, https://doi.org/10.1158/0008-5472.CAN-20-0342.

4. N. Ghaffari Laleh, C. M. L. Loeffler, J. Grajek, et al., "Classical Mathematical Models for Prediction of Response to Chemotherapy and Immunotherapy," *PLoS Computational Biology* 18, no. 2 (2022): e1009822, https://doi.org/10.1371/journal.pcbi.1009822.

5. A. Yin, D. J. A. R. Moes, J. G. C. van Hasselt, J. J. Swen, and H.-J. Guchelaar, "A Review of Mathematical Models for Tumor Dynamics and Treatment Resistance Evolution of Solid Tumors," *CPT: Pharmacometrics & Systems Pharmacology* 8, no. 10 (2019): 720–737, https://doi.org/10.1002/psp4.12450.

6. N. Al-Huniti, Y. Feng, J. J. Yu, et al., "Tumor Growth Dynamic Modeling in Oncology Drug Development and Regulatory Approval: Past, Present, and Future Opportunities," *CPT: Pharmacometrics & Systems Pharmacology* 9, no. 8 (2020): 419–427, https://doi.org/10.1002/psp4.12542.

7. R. Bruno, D. Bottino, D. P. de Alwis, et al., "Progress and Opportunities to Advance Clinical Cancer Therapeutics Using Tumor Dynamic Models," *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 26, no. 8 (2020): 1787–1795, https://doi.org/10.1158/1078-0432.CCR-19-0287.

8. R. Bruno, M. Marchand, K. Yoshida, et al., "Correction: Tumor Dynamic Model-Based Decision Support for Phase Ib/II Combination Studies: A Retrospective Assessment Based on Resampling of the Phase III Study IMpower150," *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 29, no. 20 (2023): 4314, https://doi.org/10.1158/1078-0432.CCR-23-2525.

9. W. D. Stein, J. L. Gulley, J. Schlom, et al., "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, no. 4 (2011): 907–917, https://doi.org/10.1158/1078-0432.CCR-10-1762.

10. L. Claret, P. Girard, P. M. Hoff, et al., "Model-Based Prediction of Phase III Overall Survival in Colorectal Cancer on the Basis of Phase II Tumor Dynamics," *Journal of Clinical Oncology* 27, no. 25 (2009): 4103–4108, https://doi.org/10.1200/JCO.2008.21.0807.

11. Y. Wang, C. Sung, C. Dartois, et al., "Elucidation of Relationship Between Tumor Size and Survival in Non-Small-Cell Lung Cancer Patients Can Aid Early Decision Making in Clinical Drug Development," *Clinical Pharmacology and Therapeutics* 86, no. 2 (2009): 167–174, https://doi.org/10.1038/clpt.2009.64.

12. P. L. Bonate and A. B. Suttle, "Modeling Tumor Growth Kinetics After Treatment With Pazopanib or Placebo in Patients With Renal Cell Carcinoma," *Cancer Chemotherapy and Pharmacology* 72, no. 1 (2013): 231–240, https://doi.org/10.1007/s00280-013-2191-0.

13. I. González-García, V. Pierre, V. F. S. Dubois, et al., "Early Predictions of Response and Survival From a Tumor Dynamics Model in Patients With Recurrent, Metastatic Head and Neck Squamous Cell Carcinoma Treated With Immunotherapy," *CPT: Pharmacometrics & Systems Pharmacology* 10, no. 3 (2021): 230–240, https://doi.org/10.1002/psp4.12594.

14. E. Borcoman, Y. Kanjanapan, S. Champiat, et al., "Novel Patterns of Response Under Immunotherapy," *Annals of Oncology: Official Journal of the European Society for Medical Oncology* 30, no. 3 (2019): 385–396, https://doi.org/10.1093/annonc/mdz003.

15. M. Tazdait, L. Mezquita, J. Lahmar, et al., "Patterns of Responses in Metastatic NSCLC During PD-1 or PDL-1 Inhibitor Therapy: Comparison of RECIST 1.1, irRECIST and iRECIST Criteria," *European Journal of Cancer (Oxford, England: 1990)* 88 (2018): 38–47, https://doi.org/10.1016/j.ejca.2017.10.017.

16. L. Claret, J. Y. Jin, C. Ferté, et al., "A Model of Overall Survival Predicts Treatment Outcomes With Atezolizumab Versus Chemotherapy in Non-Small Cell Lung Cancer Based on Early Tumor Kinetics," *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 24, no. 14 (2018): 3292–3298, https://doi.org/10.1158/1078-0432.CCR-17-3662.

17. S. M. Krishnan and L. E. Friberg, "Bayesian Forecasting of Tumor Size Metrics and Overall Survival," *CPT: Pharmacometrics & Systems Pharmacology* 11, no. 12 (2022): 1604–1613, https://doi.org/10.1002/psp4.12869.

18. A. Ruiz-Garcia, P. Baverel, D. Bottino, et al., "A Comprehensive Regulatory and Industry Review of Modeling and Simulation Practices in Oncology Clinical Drug Development," *Journal of Pharmacokinetics and Pharmacodynamics* 50, no. 3 (2023): 147–172, https://doi.org/10.1007/s10928-023-09850-2.

19. W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, eds., *Encyclopedia of Systems Biology* (Springer New York, 2013), https://doi.org/10.1007/978-1-4419-9863-7.

20. H. Murphy, H. Jaafari, and H. M. Dobrovolny, "Differences in Predictions of ODE Models of Tumor Growth: A Cautionary Example," *BMC Cancer* 16 (2016): 163, https://doi.org/10.1186/s12885-016-2164-x.

21. M. Laurie and J. Lu, "Explainable Deep Learning for Tumor Dynamic Modeling and Overall Survival Prediction Using Neural-ODE," *NPJ Systems Biology and Applications* 9, no. 1 (2023): 58, https://doi.org/10.1038/s41540-023-00317-1.

22. C. van Walraven, D. Davis, A. J. Forster, and G. A. Wells, "Time-Dependent Bias Was Common in Survival Analyses Published in Leading Clinical Journals," *Journal of Clinical Epidemiology* 57, no. 7 (2004): 672–682, https://doi.org/10.1016/j.jclinepi.2003.12.008.

23. J. Beyersmann, P. Gastmeier, M. Wolkewitz, and M. Schumacher, "An Easy Mathematical Proof Showed That Time-Dependent Bias Inevitably Leads to Biased Effect Estimation," *Journal of Clinical Epidemiology* 61, no. 12 (2008): 1216–1221, https://doi.org/10.1016/j.jclinepi.2008.02.008.

24. C. Tardivon, S. Desmée, M. Kerioui, et al., "Association Between Tumor Size Kinetics and Survival in Patients With Urothelial Carcinoma Treated With Atezolizumab: Implication for Patient Follow-Up," *Clinical Pharmacology and Therapeutics* 106, no. 4 (2019): 810–820, https://doi.org/10.1002/cpt.1450.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** psp470239-sup-0001-Supinfo1.txt. **Table S1:** Summary of trial data cutoff dates and determination of data truncation by study. **Figure S1:** Illustrative samples of tumor measurement times observed across patients at different follow-up times within the respective clinical trials. For each of the four NSCLC trials (IMPower130, IMPower131, IMPower132, IMPower150), the trial duration corresponds to the difference between the time of the last SLD measurement and the time of first dose for the first patient. Times are shown as a percentage of the trial duration from the first dose given. The line colors denote data that fall within the following time intervals: blue denotes between (0, 25%) time, red denotes between (25%, 50%) time, green denotes between (50%, 75%) time, orange denotes between (75%, 100%) time; note that not all subjects in the trial are shown, only a representative subset for illustrative purposes. SLD: sum-of-longest-diameter. **Figure S2:** Distributions of seen (observed) and unseen data for each study at 25%, 50%, and 75% enrollment cuts. (A) IMPower130 (B) IMPower131 (C) IMPower132 dataset (D) IMPower150 dataset. **Table S2:** Summary of TGI Models. **Figure S3:** Longitudinal Tumor Size Reduction by Immunotherapy Status. The solid lines represent the mean tumor size for the observations across time with the shaded region representing the 95% confidence interval. To visualize the longitudinal trend, continuous time points were grouped into bins based on a quantile method which divides the data into time bins with approximately equal number of points per time bin. **Figure S4:** Sample of longitudinal patient trajectories by tumor dynamic cluster. Random Sample of 500 subjects demonstrating the qualitative trends captured in each response category. **Table S3:** Quantification of error for TGI and TDNODE models on unseen data assessed at various values of follow-up times based on inclusion of an immune checkpoint inhibitor (IO+/IO−). For each value of follow-up time (25%, 50% and 75% of total trial duration), all models were fitted or trained with the corresponding "seen" data, and subsequently evaluated on the remaining unseen data. The predictive performance was evaluated using several evaluation metrics based on the unseen data. To derive the predictive performance statistics, 500 bootstraps were performed from which the respective metric was computed. **Table S4:** Quantification of error for TGI and TDNODE models on unseen data assessed at various values of follow-up times based on tumor dynamic cluster. For each value of follow-up time (25%, 50% and 75% of total trial duration), all models were fitted or trained with the corresponding "seen" data, and subsequently evaluated on the remaining unseen data. The predictive performance was evaluated using several evaluation metrics based on the unseen data. To derive the predictive performance statistics, 500 bootstraps were performed from which the respective metric was computed. **Table S5:** Quantification of error for TGI and TDNODE models on unseen data assessed at various values of follow-up times by best overall response. For each value of follow-up time (25%, 50% and 75% of total trial duration), all models were fitted or trained with the corresponding "seen" data, and subsequently evaluated on the remaining unseen data. The predictive performance was evaluated using several evaluation metrics based on the unseen data. To derive the predictive performance statistics, 500 bootstraps were performed from which the respective metric was computed. Patients with progressive disease (PD) were excluded due to very limited unseen scans. **Table S6:** Percentage of subjects by cluster and IO treatment status.