Comparing Clinical Trajectories in Wilms Tumor: Multi-State Additive Hazards Models from NWTS-3 and NWTS-4

Qikai Jiang

Introduction

Wilms tumor (WT) is the most common pediatric kidney cancer, originating from embryonal renal precursors known as nephrogenic rests (Yang et al., 2021). This malignancy is classified into five stages based on the tumor's anatomical extent and resectability (Coppes et al., 1991): Stage I (tumor confined to the kidney, completely excised), Stage II (tumor extends locally but remains fully resected), Stage III (residual abdominal disease or lymph node involvement), Stage IV (distant metastases), and Stage V (bilateral disease). Histological analysis further stratifies prognosis (Szychot et al., 2014), with favorable histology (found in 90% of cases) associated with excellent outcomes, while anaplastic histology, either focal or diffuse, is associated with chemoresistance and an increased risk of relapse. It is important to distinguish between institutional and central pathology histology: the latter, Central pathology as conducted by expert panels under the Children's Oncology Group (COG) Wilms Tumor Study (Ehrlich, 2017), ensures uniform histological classification and minimizes inter-institutional discrepancies.

Age at diagnosis and tumor burden are both critical prognostic factors in Wilms tumor outcomes, as consistently demonstrated in the National Wilms Tumor Study Group (NWTSG) trials. Under NWTSG protocols, 5-year overall survival has improved significantly, reaching approximately 90% in the United States (Green et al., 1995; Yang et al., 2021). However, prognosis still varies markedly by age: children under 5 years generally experience the most favorable outcomes (>90% 5-year survival), likely due to early detection and more favorable tumor biology, whereas survival drops to around 80% in those aged 5–10 years, and is poorer in adolescents, who often present with more advanced disease and a higher incidence of anaplastic histology (Breslow, 2006). Even among infants under 12 months, survival remains

high, thoug this group is more commonly associated with genetic syndromes such as WAGR or Beckwith-Wiedemann. Tumor diameter at diagnosis has emerged as a clinically relevant prognostic indicator in Wilms tumor, with larger lesions often signaling more advanced disease and increased metastatic potential. Prior studies, including those from the National Wilms Tumor Study Group (NWTSG), have demonstrated that tumors exceeding 10 centimeters in maximal diameter are more frequently associated with higher stage at presentation and adverse histologic features, necessitating intensified therapeutic regimens (Pshak et al., 2014). Moreover, tumor size frequently co-occurs with other high-risk features such as older age at diagnosis and the presence of lymph node involvement, reinforcing its role in risk stratification algorithms and treatment planning.

The NWTS-3 trial (1979–1986) established a risk-adapted, multimodal treatment strategy for Wilms tumor, combining surgery, chemotherapy, and radiation therapy to tailor therapy by stage (D'Angio et al., 1989). Stage I patients typically received nephrectomy followed by single-agent chemotherapy (vincristine or actinomycin D), while Stages II and III were treated with flank radiation (10.8 Gy) and intensified chemotherapy regimens, often adding doxorubicin. For Stage IV or anaplastic tumors, more aggressive approaches were used, including cyclophosphamide and whole-lung irradiation. Although building on gains from NWTS-2 (Neville et al., 2002), NWTS-3 did not yield significant further improvements in overall survival and raised concerns about overtreatment and long-term toxicities such as anthracycline-induced cardiotoxicity and increased risk of secondary malignancies. In response, NWTS-4 (1986–1994) focused on maintaining survival while reducing treatment intensity and duration (Shamberger et al., 1999). Stage I patients received just 10 weeks of chemotherapy, compared to 15 months in NWTS-3, and Stage II patients received 6 months, with no changes in radiation protocols for advanced stages. NWTS-4 demonstrated that survival outcomes could be preserved with less intensive therapy, reducing the long-term risks of cardiovascular complications and secondary cancers.

In the event of relapse, post-relapse management differed in intensity depending on the initial treatment protocol under NWTS-3 or NWTS-4. Patients treated under NWTS-3 typically received substantially

escalated therapy at relapse, often involving higher doses of agents such as doxorubicin, cyclophosphamide, and etoposide, some of which were not included in their frontline treatment (Malogolowkin et al., 2008). Radiation was frequently intensified and directed at metastatic sites, and repeat nephrectomy was performed when the relapse was localized and resectable. For those with early recurrence, extensive metastases, or treatment-resistant disease, high-dose chemotherapy supported by autologous stem cell transplantation was occasionally used. In contrast, relapse treatment following NWTS-4, which emphasized reduced initial intensity, focused on adding or escalating chemotherapy (e.g., cyclophosphamide, etoposide, doxorubicin) while preserving the efficacy of the initial regimen. Radiation remained targeted, and stem cell transplantation was considered in high-risk cases, with increased attention to long-term toxicity and marrow recovery (Malogolowkin et al., 2008).

Despite the availability of extensive time-to-event data from the National Wilms Tumor Study (NWTS) cohorts, most survival analyses in Wilms tumor research have centered on overall survival, typically modeled using the Cox proportional hazards framework (Honeyman et al., 2012; Mullen et al., 2018). These studies often fail to incorporate intermediate clinical events, particularly relapse and rarely model transitions between distinct disease states such as from disease-free to relapse and from relapse to death. As a result, the full clinical trajectory of patients is not adequately captured.

Moreover, few analyses rigorously assess the proportional hazards assumption, and diagnostic checks for violations are rarely presented in a transparent or reproducible manner. The reliance on the Cox model persists even in settings where time-varying effects or non-proportional hazards are plausible. While some studies acknowledge differences across NWTS cohorts, particularly the marked changes in treatment protocols between NWTS-3 and NWTS-4, these variations are infrequently accounted for in the modeling strategy. Stratification by study era is occasionally implemented (e.g., Cotton et al., 2007), but formal modeling of between-study heterogeneity remains limited.

Notably, the seminal work of Kulich and Lin (2004), which introduced an efficient additive hazards model for case-cohort designs, applied their method to the NWTS dataset to illustrate violations of the proportional hazards assumption. However, their analysis did not incorporate the dynamic transitions between disease states, nor did it differentiate between NWTS-3 and NWTS-4 or address the impact of evolving treatment protocols. Furthermore, modern data visualization tools that could illustrate time-varying effects and multistate transitions have not yet been widely adopted in this literature.

Overall, existing approaches fall short in capturing the dynamic and multifaceted nature of disease progression in Wilms tumor, especially in the presence of relapse and shifting risk profiles over time. A more flexible modeling framework, one that incorporates intermediate events, time-dependent effects, and multistate transitions, is needed to better reflect the clinical reality experienced by patients across NWTS cohorts.

In this study, we propose a comprehensive analytical framework for Wilms tumor that accounts for differences across treatment eras (NWTS-3 and NWTS-4), time-varying effects, and disease state transitions. We use Kaplan–Meier estimators for initial exploratory analyses, followed by Cox and Accelerated Failure Time models with diagnostic checks. To address non-proportional hazards, we apply additive hazards models. Finally, we implement multi-state modeling to capture transitions from diagnosis to relapse and death, and simulate state-occupancy probabilities for dynamic prognostic insights. This framework offers a more granular, clinically relevant view of patient trajectories than conventional survival approaches.

Methods

We analyzed individual-level data from the NWTSCO dataset, derived from the National Wilms Tumor Study Group (NWTSG) clinical trials, specifically, the NWTS-3 and NWTS-4 studies, which enrolled children diagnosed with Wilms tumor between 1979 and 1994 (Green et al., 1998; Green et al., 2001). These multi-center studies were later integrated into the Children's Oncology Group (COG), which

standardized trial procedures and ensured central pathology review for consistent histological classification (Ehrlich, 2017). The NWTSCO dataset, publicly available via the addhazard package in R, contains detailed time-to-event data on patient demographics, tumor characteristics (including stage at diagnosis, tumor weight, tumor diameter, and histology as determined by central review), relapse status, and survival outcomes.

As part of our exploratory analysis, we summarized both categorical and continuous patient characteristics separately for NWTS-3, NWTS-4, and the combined cohort to assess cohort comparability and inform model specification (Table 1). Continuous variables such as tumor diameter and age at diagnosis were examined in their original form and later categorized using clinically informed cut points to facilitate interpretation and align with model assumptions (Table 3). To characterize the temporal relationship between relapse and subsequent outcomes, we cross-tabulated relapse and vital status, capturing expected event ordering within the follow-up window (Table 2).

Finally, we evaluated pairwise associations among candidate predictors using three complementary measures: Spearman correlation for continuous and ordinal variables, chi-square statistics for categorical pairs, and point-biserial (Pearson) correlation for continuous–binary combinations. These ranked association metrics (Figure 1) informed variable selection and assessed potential redundancy prior to multivariable and multi-state modeling.

We defined three clinically relevant time-to-event outcomes for survival analysis. Time to relapse was measured from the date of initial diagnosis to the first documented recurrence of Wilms tumor. Time to death captured Wilms tumor–specific mortality, defined as the interval from diagnosis to death attributed to Wilms tumor. Time from relapse to death quantified post-relapse survival and was defined for the subset of patients who experienced a relapse (i.e., relapse status = 1); this interval measured the duration from the date of relapse to Wilms tumor–related death or censoring. Patients who had not experienced the event of interest by the end of follow-up were administratively censored at the date of last known contact.

These outcome definitions reflect distinct stages of disease progression and allow for separate modeling of initial risk, recurrence, and survival after relapse.

To evaluate differences in time-to-event distributions across key clinical subgroups, we applied the log-rank test for each of the three survival outcomes: time to relapse, time to death, and time from relapse to death. The log-rank test assesses whether survival functions differ significantly across groups by comparing the observed and expected number of events at each event time, under the null hypothesis of no difference between strata. For each covariate, we report the chi-square test statistic, degrees of freedom, and corresponding p-values (Table 4), using a significance threshold of $\alpha = 0.05$. For clinical variables with statistically significant log-rank test results, we generated Kaplan–Meier survival curves to visualize differences in survival probabilities over time. These stratified survival estimates provided preliminary insight into heterogeneity in prognosis across subgroups and informed covariate selection for multivariable survival and multi-state modeling. Detailed interpretation of the Kaplan–Meier curves is provided in the results section.

To inform the specification of the survival modeling, we began by evaluating the treatment of study membership using the Akaike Information Criterion (AIC) across three model formulations: including study as a covariate, stratifying by study, and stratifying by study with covariate-by-study interaction terms. Models were fit for each of the three time-to-event outcomes of time to relapse, time to death, and time from relapse to death, using the set of candidate predictors retained after correlation screening (Figure 1), where continuous variables were included in their original form. AIC values were compared to assess relative model fit (Table 5). Substantially lower AIC values for the stratified or interaction-stratified models provided evidence in favor of accounting for study-level heterogeneity through stratification in Cox models or fitting separate models in alternative survival frameworks such as the accelerated failure time or additive hazards models.

To assess the appropriateness of modeling age at diagnosis and tumor diameter as continuous covariates, we examined martingale residual plots for each outcome (Figure 4). Departures from linearity in these residuals indicated the need to categorize continuous variables using clinically meaningful cut points. Based on these diagnostics, final model specifications used the categorical forms of age and tumor diameter. We then evaluated the proportional hazards assumption for each Cox model using the Grambsch–Therneau test based on scaled Schoenfeld residuals (Table 6). The null hypothesis of this test posits that covariate effects are constant over time; p-values were reported globally and for each covariate, using a significance threshold of $\alpha = 0.05$. Rejection of the null globally or for the majority of covariates indicated meaningful violations of the proportional hazards assumption, thereby motivating the use of alternative modeling strategies such as additive hazards models, or accelerated failure time models that do not rely on the proportionality assumption.

Given that the proportional hazards assumption was found to be violated, we adopted accelerated failure time (AFT) models as an alternative framework and used Bayesian Information Criterion (BIC) to guide distributional selection. Unlike AIC, BIC is preferred here because our focus lies in estimation and inference rather than prediction, and BIC imposes a stronger penalty for model complexity, favoring more parsimonious models, particularly important when comparing nested parametric families. For each clinical endpoint (time to relapse, time to death, and time from relapse to death), we fitted AFT models assuming Weibull, exponential, log-normal, log-logistic, normal, and gamma distributions. Table 7 presents BIC values either separately for Study 3 and Study 4 or as a single model, depending on Table 5; if Table 5 supports modeling study as a covariate, a single model is shown, whereas evidence for study-specific effects leads to separate models.

Following model selection based on BIC, Figure 5 presents diagnostics for the best-fitting AFT models, including standardized residual plots, residual histograms, and normal Q-Q plots. These diagnostics allow us to assess key AFT model assumptions: (i) the appropriateness of the specified error distribution (e.g., normality for log-time), (ii) homoscedasticity, and (iii) model fit to the log-transformed survival time. We

examined residual behavior across all candidate distributions to ensure that the selected model provides a reasonable fit. Even in cases where diagnostics in Figure 5 indicated severe violations of AFT model assumptions, all models are retained and reported to establish a consistent baseline for later comparison with additive hazards modeling. Table 8 presents the corresponding coefficient estimates, time ratios, standard errors, 95% confidence intervals, and p-values ($\alpha = 0.05$) for the best-fitting AFT models while detailed interpretations will be provided. These models were selected based on diagnostic results from Figure 5 when an alternative distribution showed substantially better model diagnostics than the BIC-selected model in Table 7; otherwise, when diagnostics were comparable across distributions, the model with the lowest BIC in Table 7 was used.

While in case of diagnostics in Figure 5 revealed evident violations of AFT model assumptions across all candidate distributions, we turned to the additive hazards modeling framework proposed by Lin and Kulich (2004), which accommodates time-varying covariate effects. To formally assess the presence of non-constant effects over time, we applied both the supremum test and the Kolmogorov–Smirnov (K–S) test, with results reported in Table 9 for each clinical endpoint (Martinussen and Scheike, 2006). In this context, the null hypothesis for both tests is that the covariate effects are constant over time. The supremum test evaluates the largest deviation of the cumulative coefficient process from a constant line, while the K–S test examines the maximum distance between the observed cumulative process and its expected trajectory under the null. Rejection of the null at $\alpha = 0.05$ suggests significant time-varying effects, justifying the use of the additive hazards model.

To visualize these effects, Figure 6 presents the estimated cumulative regression functions from Aalen's additive hazards model for the three time-to-event endpoints in NWTS Studies 3 and 4, with 95% pointwise confidence bands shown for each covariate. In the survival context, these cumulative functions represent the integrated effect of a covariate on the hazard rate over time. A flat trajectory indicates a constant effect, while upward or downward trends reflect time-varying influence. Nonlinear or non-monotonic patterns suggest changes in the direction or magnitude of covariate effects during follow-up.

The inclusion of confidence bands allows assessment of statistical uncertainty, helping to determine whether observed variation departs meaningfully from zero. Together, these plots offer detailed insight into the temporal dynamics of risk factors, information that would be masked under proportional hazards or fully parametric AFT assumptions.

In survival analysis, multi-state modeling provides a versatile framework for analyzing time-to-event data where individuals may progress through a series of clinically relevant states, including both intermediate and terminal events (Putter, Fiocco, and Geskus, 2007). This methodology is particularly appropriate in cancer research settings such as the National Wilms Tumor Studies (NWTS), where disease progression is not limited to a single endpoint but may follow multiple distinct pathways. For instance, patients can transition from initial diagnosis and treatment (state 1, or "healthy") to relapse (state 2), from relapse to death (state 3), or from the initial state directly to death without prior relapse, a clinically important pathway that is not captured in simpler models. Unlike conventional survival models, which estimate a single hazard for the time until a terminal event, multi-state models decompose the disease course into transition-specific components. This allows the modeling of separate hazard functions for each transition (e.g., 1→2, 1→3, 2→3), accommodating differences in timing, risk factors, and treatment effects associated with each phase of disease. By doing so, multi-state models not only enhance interpretability but also enable dynamic characterization of patient trajectories over time, which is critical for both clinical decision-making and long-term risk assessment.

To evaluate the appropriateness of the additive hazards framework for each transition, we first examined time-varying effects using supremum and Kolmogorov–Smirnov (K–S) tests within Aalen's additive model. These test results, reported in Table 10, assess the null hypothesis that covariate effects remain constant over time for each transition (1→2, 1→3, 2→3), separately for Study 3 and Study 4. Rejection of the null at the 5% significance level indicates significant deviation from time-constancy, supporting the use of nonparametric, time-varying effect models. Figure 7 displays the estimated cumulative regression functions for each covariate in each transition, stratified by study, using Aalen's additive hazards model.

These functions represent the cumulative effect of covariates on the transition-specific hazard rates over time; flat curves suggest constant effects, while non-linear shapes indicate time-varying influence. As with earlier models, confidence bands are included to assess statistical uncertainty. This flexible modeling approach allows us to capture the dynamic and potentially non-proportional influence of covariates on each stage of disease progression.

Table 11 provides essential context for the three modeled transitions, relapse (1→2), direct death without relapse (1→3), and death after relapse (2→3), by summarizing the number of individuals at risk and the cumulative number of observed events at key time points, separately for NWTS Studies 3 and 4. This information reflects how the population at risk evolves over time within each transition pathway, offering insight into the data structure that supports hazard estimation. By aligning the timing of events with the corresponding risk sets, the table helps clarify the temporal distribution and empirical basis of transition-specific models, ensuring transparency in how observed data inform the estimated hazard functions.

To operationalize the multi-state additive hazards model in a clinically meaningful way, we define a set of covariate profiles that reflect distinct patterns of disease severity and patient characteristics. Table 12 lists ten representative profiles, each defined by a unique combination of four categorical variables: age at diagnosis (younger vs. older), tumor diameter (small vs. large), disease stage (early vs. advanced), and histology (favorable vs. unfavorable). These variables were chosen based on their established prognostic relevance in Wilms tumor and their inclusion in the fitted models. The profiles are designed to capture a range from low-risk to high-risk clinical scenarios, enabling clear contrasts in disease progression across patient types. These fixed covariate combinations form the basis for interpreting model outputs and for generating simulated outcomes and model-based functions in subsequent figures and tables. In particular, they provide a structured framework for evaluating how different patient profiles experience transitions through the illness course, such as relapse or death, under the multi-state model, and how those trajectories differ by treatment era (i.e., NWTS 3 vs. NWTS 4).

Figures 8 and 9 display the estimated state occupancy probabilities over time for the ten covariate profiles defined in Table 12, stratified by study (NWTS-3 and NWTS-4). These probabilities, denoted $P_{ik}(t)$, quantify the likelihood that an individual with covariate profile i occupies state $k \in \{1, 2, 3\}$ (corresponding to Healthy, Relapse, or Dead) at time t following initial treatment. The estimates are obtained through Monte Carlo integration using 5,000 simulated individual trajectories per profile, based on the cumulative baseline hazards and estimated time-varying regression coefficients from the fitted additive hazards models for each transition (1→2, 1→3, and 2→3). These simulations solve the Kolmogorov forward equations for the multi-state process under the Aalen additive framework, accounting for covariate effects that vary with time (Norris, 1997). As such, the figures provide a full probabilistic characterization of each profile's projected disease course over 25 years, capturing the dynamic interplay between relapse risk and mortality across clinical subgroups. This approach aligns with the general principles of non-Markovian multi-state modeling in continuous time (Andersen and Keiding, 2002; Putter, Fiocco, and Geskus, 2007), and it illustrates how model-based estimates can be translated into clinically interpretable trajectories for risk stratification and outcome evaluation.

Table 13 presents estimated transition probabilities from the three-state illness-death model for the ten predefined covariate profiles, stratified by study cohort (NWTS-3 and NWTS-4) and evaluated at selected time points. Each profile combines clinically relevant attributes, age at diagnosis, tumor size, stage, and histology, that reflect the heterogeneity of disease severity observed across the NWTS population. In this framework, the transition probability $P_{ik}^{(i)}(t)$ represents the probability that a patient with covariate profile i, initially in the Healthy state (state 1), will occupy state $k \in \{1, 2, 3\}$ (Relapse, Dead, or remain Healthy) at time t. These probabilities are derived by numerically solving the Kolmogorov forward equations for the continuous-time Markov process, using the estimated cumulative hazards from additive hazard models fitted to each transition: Healthy → Relapse (1→2), Healthy → Direct Death (1→3), and Relapse → Death (2→3). While the dataset does not include direct information on treatment regimens or therapeutic protocols, the stratification by study allows indirect inference about differences between

NWTS-3 and NWTS-4, which are known to differ in clinical management strategies and trial design. By holding covariate profiles constant across studies, the transition probabilities isolate the impact of unmeasured temporal or systemic factors including likely treatment evolution on disease dynamics. Therefore, Table 13 not only facilitates covariate-specific interpretation of disease progression but also enables contextual comparison between studies, offering insight into how structural or unobserved changes between NWTS-3 and NWTS-4 may influence relapse and mortality risks in Wilms tumor.

Figures 10 and 11 present a transition-specific view of disease progression using cumulative incidence functions (CIFs), $F_{l_1 l_2}(t|Z_i)$, estimated for each of the ten covariate profiles in the NWTSCO dataset and stratified by study (NWTS-3 and NWTS-4). These functions quantify the probability that an individual initially in state $l_1 \in \{1, 2\}$ (Healthy or Relapse) will transition to state $l_2 \in \{2, 3\}$ (Relapse or Death) by time t, given covariate values Z_i. Crucially, the CIF framework accounts for competing risks, such that the probability of relapse from a healthy state, for example, is appropriately adjusted for the risk of dying before relapse. The functions are derived analytically from the estimated cumulative baseline hazards and additive covariate effects for each transition, without simulation, ensuring interpretability and precision. In contrast to Figures 8 and 9, which display overall state occupancy probabilities by integrating over all possible transition paths, Figures 10 and 11 focus on the timing and likelihood of specific events, providing a more granular understanding of individual transition risks. This separation enhances clinical insight into the direct risk of relapse or death, including mortality following relapse, and reveals how these risks evolve over time across different patient subgroups. The NWTSCO dataset, with its detailed patient-level information across treatment eras, enables a data-driven assessment of how baseline characteristics influence disease trajectories, even in the absence of direct treatment variables. Together, these CIF plots underscore the value of multi-state modeling in capturing heterogeneous and time-dependent risks in pediatric oncology (Putter, Fiocco, & Geskus, 2007).

In summary, our methodological approach integrates conventional survival analysis with flexible multi-state modeling to characterize disease progression in Wilms tumors across distinct clinical phases. We

began with Kaplan–Meier estimators and log-rank tests to explore unadjusted differences in time-to-event outcomes across clinical subgroups. To address limitations of proportional hazards assumptions and linear covariate effects, we adopted the additive hazards framework proposed by Aalen, which models the hazard function as a sum of baseline and covariate-specific time-varying components. This approach allows the effect of each covariate on the hazard rate to evolve over time and is particularly well-suited for detecting departures from proportionality. Time-varying effects were formally assessed using supremum and Kolmogorov–Smirnov tests and visualized through cumulative regression function plots with pointwise confidence bands. Extending this framework, we implemented a nonparametric three-state illness-death model using transition-specific additive hazards models for diagnosis-to-relapse (1→2), diagnosis-to-death without relapse (1→3), and relapse-to-death (2→3) transitions. This multi-state structure accommodates the sequential and competing nature of events and allows each transition to have its own time-varying covariate effects. Transition probabilities and state occupancy probabilities were derived by numerically solving the Kolmogorov forward equations, incorporating both the cumulative baseline hazards and the estimated cumulative covariate effects. Simulations were conducted across predefined covariate profiles to capture heterogeneity in disease progression across clinical subgroups and study eras. In the following sections, we present the results of these analyses, highlighting covariate-specific differences in risk patterns and interpreting their implications for long-term outcome evaluation and clinical risk stratification.

Results

Patient Characteristics and Outcome Measures

Table 1 summarizes the baseline characteristics and outcome distributions for patients enrolled in Study 3 (N = 1,671), Study 4 (N = 2,244), and the combined cohort (N = 3,915). The overall mortality rate was 11.3%, with 13.2% of patients dying in Study 3 and 10.0% in Study 4. Relapses occurred in 17.1% of patients overall, with a slightly higher rate in Study 3 (18.4%) compared to Study 4 (16.1%).

Histology distributions were similar across studies. Based on central pathology review, 88.8% of patients had favorable histology and 11.2% had unfavorable histology; this proportion was nearly identical across studies. Institutional histology review classified 90.2% of tumors as favorable overall, with a higher proportion in Study 4 (91.5%) than Study 3 (88.6%).

Stage at diagnosis varied across studies. Stage I was the most common overall, comprising 39.4% of the total cohort (41.8% in Study 3 and 37.6% in Study 4). Stage II occurred more frequently in Study 4 (28.2%) than in Study 3 (21.6%). Notably, Stage III and Stage IV were more prevalent in Study 3, accounting for 24.1% and 12.5%, respectively, compared to 22.4% and 11.8% in Study 4. These differences suggest a slightly higher burden of advanced disease in the Study 3 cohort.

Key baseline characteristics were similar across Study 3 and Study 4 cohorts in the NWTS dataset, with only minor differences observed in tumor size measured in centimeters, specimen weight measured in grams, and age at diagnosis and surgical removal of tumors being performed in years. The mean diameter of the primary tumor was 11.45 cm (standard deviation [SD] of 3.9) in Study 3 and 11.03 cm (SD of 3.8) in Study 4, yielding a combined mean of 11.21 cm (SD of 3.8). Specimen weight distributions were also comparable between studies, with an average weight of 607.8 grams (SD of 408.2) in Study 3 and 602.2 grams (SD of 394.9) in Study 4; the overall mean weight was 604.6 grams (SD of 400.6). Age at diagnosis showed minimal variation, with Study 3 patients diagnosed at a mean age of 3.50 years (SD of 2.6), compared to 3.56 years (SD of 2.5) in Study 4, resulting in a pooled average of 3.53 years (SD of 2.6). These continuous variables demonstrate broad consistency across the two cohorts, suggesting that any structural or outcome-related differences between Study 3 and Study 4 are unlikely to stem from large imbalances in these fundamental clinical features.

Follow-up duration differs substantially between studies. The mean time to death or last known vital status was longer in Study 3 (13.24 years) than in Study 4 (8.16 years), yielding a combined mean of 10.33 years (SD 5.5). In Study 3, the median follow-up for death was 15.23 years, with the third quartile

reaching 17.87 years and a maximum of 22.50 years. In contrast, Study 4 exhibited a shorter median follow-up of 8.59 years, a third quartile of 11.08 years, and a maximum of 15.18 years.

A similar trend was observed in time to relapse or last relapse-free follow-up. Study 3 participants had a mean time to relapse of 12.39 years, compared to 7.60 years in Study 4 (overall mean: 9.65 years). The median, third quartile, and maximum time to relapse were 14.87, 17.78, and 22.50 years, respectively, in Study 3, versus 8.16, 10.94, and 15.18 years in Study 4.

Among individuals who experienced a relapse, event times were notably shorter. In Study 3, the mean time to relapse among relapsed patients was 1.83 years (median: 0.84; third quartile: 1.60; maximum: 17.52), compared to 1.41 years in Study 4 (median: 0.90; third quartile: 1.67; maximum: 12.91). Similarly, for individuals who died, the mean time to death was 2.85 years in Study 3 (median: 1.53; third quartile: 3.12; maximum: 18.50) and 2.53 years in Study 4 (median: 1.71; third quartile: 3.34; maximum: 12.91).

These findings underscore important temporal and structural differences between Study 3 and Study 4 cohorts of the NWTS dataset. The substantially longer follow-up times observed in Study 3 reflect both its earlier enrollment period and likely differences in study design, clinical monitoring protocols, and data collection infrastructure relative to Study 4. Although follow-up durations varied markedly, the timing of relapse and death among those who experienced these events, was concentrated early in the disease course across both cohorts, suggesting that short-term risk dynamics may be comparably structured. Nonetheless, the observed disparities in follow-up time, event rates, and baseline characteristics (e.g., tumor stage distribution) raise concerns about potential heterogeneity in risk processes across studies. These differences support the analytic need to evaluate cohort effects formally such as through stratified Cox models or statistical tests for proportional hazards violations to assess whether pooled analysis is appropriate. The empirical contrasts described here thus motivate further investigation into whether

study-specific modeling is warranted to ensure valid inference and accurate estimation of covariate effects.

Event Sequencing and Temporal Ordering of Relapse and Mortality

To assess the internal consistency and expected temporal ordering of clinical events, we examined the joint distribution of relapses and vital status at last follow-up (Table 2). Across both studies, no patients were observed to be alive at last follow-up with a relapse indicator of zero and time to relapse equal to time to last contact, confirming that relapse and censoring were appropriately distinguished in the dataset. Among the 669 patients who experienced a relapse, 225 were deceased at last follow-up, while the remaining 444 were alive. Importantly, only 4 individuals (1.8%) were recorded with identical times for relapse and death, suggesting that virtually all deaths occurred after a recorded relapse event or in the absence of relapse (i.e., direct death). This supports the validity of using relapses as a distinct intermediate event in the multi-state modeling framework. The proportion of patients whose relapse preceded death was comparable across studies: 69% (151 of 220) in Study 3 and 73% (163 of 224) in Study 4. These patterns are consistent with the clinical trajectory of Wilms tumor, where death typically follows disease recurrence. Overall, the observed temporal progression of relapses and mortality supports our modeling assumptions regarding the ordering of intermediate and terminal events.

Assessment of Correlation and Justification for Covariate Selection

To identify potential redundancy and inform covariate selection for multivariable and multi-state models, we assessed pairwise associations among candidate predictors. These associations are summarized in Figure 1, which presents three panels based on the type of variable comparison.

The panel on the left shows ranked Spearman correlations among continuous and ordinal variables. The strongest relationship was between specimen weight and tumor diameter ($\rho = 0.76$), reflecting their shared dependence on tumor size. Moderate correlations were also observed between age at diagnosis and

specimen weight ($\rho = 0.35$), as well as between age and tumor stage ($\rho = 0.28$), suggesting that older children tended to present with larger or more advanced tumors. Other correlations, such as tumor diameter with tumor stage ($\rho = 0.25$), specimen weight with tumor stage ($\rho = 0.24$), and age with tumor diameter ($\rho = 0.22$), were weaker but still notable. None of the associations indicated problematic collinearity, but they helped clarify which variables contributed unique prognostic information.

The panel in the middle presents chi-squared statistics for associations among categorical variables. As expected, the most substantial association was between institutional and centrally reviewed histology ($\chi^2 = 2303.4$), reflecting a high level of agreement between the two classification systems. Moderate associations were also found between institutional histology and tumor stage ($\chi^2 = 80.7$), and between central histology and stage ($\chi^2 = 44.1$), both consistent with known relationships between histologic subtype and disease severity. Associations involving the study group were smaller: study group versus stage ($\chi^2 = 22.1$), study group versus institutional histology ($\chi^2 = 9.3$), and only negligible association with central histology ($\chi^2 = 0.22$), suggesting that histologic classification remained relatively consistent across studies.

The panel on the right displays point-biserial (Pearson) correlations for binary–continuous variable pairs. All observed values were small ($|r| < 0.06$), indicating weak linear relationships between binary indicators (e.g., study cohort or histology group) and continuous measures such as age or tumor size. The largest effect, though still weak, was a negative correlation between study group and tumor diameter ($r = -0.054$), hinting at slightly smaller tumors in more recent cohorts. All other correlations, including those involving age and specimen weight, were close to zero.

Overall, these analyses supported retaining *age at diagnosis*, *tumor diameter*, *central pathology histology*, *tumor stage*, and *study group* as key covariates. Correlations among these variables were modest enough to avoid concerns about multicollinearity, and their established clinical relevance justified their inclusion in the subsequent modeling framework.

Covariate Categorization: Tumor Diameter and Age at Diagnosis

To enhance clinical interpretability and ensure methodological rigor across our analyses, we categorized tumor diameter and age at diagnosis using thresholds that are both empirically informed by the observed distributions in our cohort and supported by prior Wilms tumor research. These categorizations are essential for conducting Kaplan–Meier survival estimation and log-rank tests, which rely on discrete group comparisons, and they provide a consistent framework for evaluating time-to-event outcomes. Furthermore, the defined categories support subsequent modeling decisions, including whether to stratify or separate cohorts in Cox proportional hazards, accelerated failure time (AFT), and additive hazards models. This approach allows us to evaluate the effect of heterogeneity across trials while maintaining alignment with clinically meaningful cutoffs used in existing literature.

Tumor diameter was classified into three categories: <10 cm, 10–15 cm, and >15 cm. These thresholds align with surgical and risk-stratification considerations observed in earlier studies, where tumor size has been shown to correlate with resectability, treatment complexity, and prognosis (Green et al., 1998; Shamberger et al., 1999; Dome et al., 2015). For example, tumors ≥10 cm are often associated with higher stage at presentation and may influence decisions around preoperative chemotherapy or surgical planning.

Age at diagnosis was categorized as <2 years, 2–4 years, and >4 years, reflecting well-established age-dependent variations in biological behavior and treatment response. Several studies have demonstrated that children diagnosed before the age of 2 have better survival outcomes, potentially due to differences in tumor biology and lower stage at presentation (Breslow et al., 1993; Green et al., 2001; Dome et al., 2006). Furthermore, treatment protocols in the NWTS and subsequent Children's Oncology Group (COG) trials often stratify therapeutic regimens using similar age cutoffs.

Observed distributions across studies supported these groupings. For tumor diameter, roughly one-third of patients in each study had tumors <10 cm (32.9% in Study 3 vs. 34.4% in Study 4), and over half fell into

the 10–15 cm range (54.2% vs. 54.7%). Slightly more patients in Study 3 had tumors >15 cm (12.9% vs. 10.9%). Age distributions were also similar: children <2 years accounted for 32.0% (Study 3) and 31.9% (Study 4); ages 2–4 years comprised 34.2% and 35.5%, respectively; and those >4 years were slightly more common in Study 4 (35.1% vs. 33.8%).

Together, these categorization strategies maintain clinical relevance, harmonize with historical and ongoing trial structures, and allow for meaningful stratification of covariates in both univariable and multivariable analyses.

Kaplan–Meier Survival Estimates and Log-Rank Tests by Clinical Covariates

To preliminarily assess whether time-to-event distributions differ across levels of categorical covariates, we employed the log-rank test, a standard nonparametric tool in survival analysis. The log-rank test evaluates the null hypothesis that the hazard functions are equal across all groups defined by a given covariate. Operationally, it compares the observed number of events within each group to the number expected under the assumption of identical hazard rates over time. These comparisons are accumulated across all observed event times to yield a test statistic that, under the null hypothesis, follows an approximate chi-squared distribution with degrees of freedom equal to the number of groups minus one. The test is particularly sensitive to proportional differences in hazards and is commonly used as an initial screen for covariate associations prior to more complex multivariable modeling.

We applied the log-rank test to three distinct time-to-event outcomes: time to relapse, time to death, and time from relapse to death. The third outcome was restricted to individuals who experienced a relapse. All hypothesis tests were evaluated at a significance level of $\alpha = 0.05$. For each covariate and endpoint, we report the chi-squared test statistics, corresponding degrees of freedom, and the p-value in Table 4, interpreting significance relative to the null hypothesis of equal survival functions across categories.

When comparing across study cohorts, which reflect enrollment in NWTS-3, NWTS-4, or other combined sub-cohorts (df = 2), the p-values were 0.255 for time to relapse, 0.057 for time to death, and 0.012 for time from relapse to death. While the first two endpoints did not meet the conventional significance threshold, the result for post-relapse survival was statistically significant. This suggests that, although overall relapse and mortality rates may not differ substantially between studies, there may be important differences in survival after relapse. This could reflect differences in post-relapse treatment protocols or follow-up practices between the study cohorts.

Tumor stage at diagnosis showed highly significant differences in time-to-event distributions across all outcomes. With degrees of freedom equal to three, the test statistics were 147.22 for time to relapse, 182.64 for time to death, and 91.70 for time from relapse to death, all with p-values less than 0.001. These findings confirm the well-established clinical importance of staging in Wilms tumor prognosis, with progressively higher stage associated with poorer outcomes. The large test statistics reflect substantial divergence in survival profiles across the four stage groups, particularly for time to death.

Histology, categorized according to central pathology review, exhibited the most evident differences in survival outcomes across all covariates examined. With two degrees of freedom, the test statistics were exceptionally large: 410.3 for time to relapse, 498.1 for time to death, and 107.3 for time from relapse to death, all with p-values well below 0.001. These results underscore the dominant prognostic role of histologic subtype, as centrally reviewed pathology appears to distinguish survival trajectories more sharply than other baseline factors.

Age at diagnosis and tumor diameter were also evaluated, each with three-level categorizations yielding two degrees of freedom. Age was a more informative prognostic factor than tumor size. The log-rank test for age yielded p-values below 0.001 for both time to relapse and time to death, with test statistics of 23.81 and 37.93, respectively. However, the test for time from relapse to death was not significant (p = 0.120), indicating that age-related differences in prognosis may diminish once relapse occurs. In contrast,

tumor diameter showed weaker and less consistent associations. The p-values were 0.076 for time to relapse, 0.002 for time to death, and 0.355 for time from relapse to death, indicating some evidence of a relationship with overall survival but limited prognostic separation otherwise.

Taken together, these univariate log-rank tests provide an initial assessment of how key clinical covariates relate to relapse, death, and post-relapse survival. The results highlight tumor stage, central histology, and age at diagnosis as the most evidently informative variables for subsequent multivariable modeling. All analyses were conducted on the combined study population without stratification by study cohort, to preserve sample size and provide a global view of prognostic associations in this population.

Figure 2 displays Kaplan–Meier survival curves for three key time-to-event outcomes, time to relapse, time to death, and time from relapse to death, stratified by two important clinical covariates: tumor stage at diagnosis (top row) and central pathology histology (bottom row). These nonparametric estimates visualize the unadjusted survival experience across strata, without imposing assumptions about the shape or proportionality of the underlying hazard functions. The stepwise form of the curves reflects updates at each observed event time, and although right-censoring is not explicitly marked, the width of the confidence bands provides insight into follow-up uncertainty and sample size over time.

In addition to displaying survival probabilities, these visualizations offer preliminary diagnostic insight into the suitability of common survival models. Patterns such as clear vertical separation between curves (e.g., between extreme stages or histology types), non-overlapping confidence intervals, and consistent ordering of curves over time support the potential adequacy of proportional hazards models, such as the Cox model. However, in settings where curves are non-parallel, converge or diverge over time, or show time-varying gaps in survival, particularly for more aggressive subgroups, such behavior may indicate non-proportional hazards or time-varying covariate effects. These features suggest that more flexible approaches, such as additive hazards models or accelerated failure time (AFT) models, may be necessary to appropriately capture covariate-outcome relationships. Thus, Figure 2 and later Figure 3 serve not only

as a descriptive summary but also as an initial guide for selecting an appropriate class of survival models in subsequent analyses.

The top left panel presents time to relapse by tumor stage. Relapse-free survival varies substantially across stages, with Stage I patients showing the most favorable outcomes, survival probabilities remain above 90% at 2 years and above 80% at 10 years post-diagnosis. In contrast, Stage IV exhibits the steepest decline, with relapse-free survival falling below 70% within the first 2 years and continuing to decrease steadily, reflecting the most aggressive disease progression. Stages II and III lie between these extremes, but their curves are closely aligned, particularly beyond 15 years, where they nearly overlap. The confidence intervals for Stages II and III also overlap extensively throughout the follow-up, indicating limited evidence of a meaningful difference in relapse risk between these two groups. At 5 years, relapse-free survival is approximately 88–90% for Stage II and 85–88% for Stage III; by 10 years, both range around 80–85%, further underscoring their similar trajectories. Overall, these patterns highlight a clear inverse relationship between tumor stage and relapse-free survival, with the strongest contrast between Stage I and Stage IV, and less distinct separation between the intermediate stages.

The top middle is time to death by tumor stage. Survival probabilities differ clearly by tumor stage. Survival probabilities differ clearly by tumor stage. At 5 years, patients with Stage I disease show the most favorable survival, exceeding 95%, and remaining above 90% at 10 years. Stage II and Stage III demonstrate intermediate outcomes, Stage II survival hovers around 90% at 5 years, while Stage III is slightly lower at approximately 85–88%. However, the survival curves for Stages II and III are more closely aligned and exhibit overlapping 95% confidence intervals throughout much of the follow-up, suggesting more modest differentiation between these two groups. In contrast, Stage IV patients have the poorest prognosis, with survival dropping to roughly 80% at 5 years and approaching 70% by 10 years, with their curve distinctly separated from earlier stages. Confidence intervals widen noticeably after year 15 across all groups, particularly for the more advanced stages, reflecting reduced precision due to smaller numbers of patients remaining at risk.

The top right panel illustrates survival following relapses, stratified by tumor stage. Among patients who experienced a relapse, post-relapse survival differs markedly by initial stage. Stage I patients exhibit the most favorable outcomes, with survival remaining above 60% at 2 years and just above 50% at 5 years post-relapse. Stage II patients follow closely, with survival near 55% at 2 years and around 45–50% at 5 years, indicating a smaller gap from Stage I than seen in the time-to-relapse and time-to-death curves. In contrast, survival declines are steeper for Stages III and IV. Stage III patients show survival dropping below 50% within 2 years and approaching 35–40% at 5 years. Stage IV patients face the poorest outcomes, with survival falling rapidly below 30% by year 2 and nearing 20% by year 5.

Compared to the more gradual trends observed in the time-to-relapse and overall survival panels, the curves here demonstrate more abrupt declines across all stages, underscoring the high mortality risk following relapse. Confidence intervals are wider in this setting, particularly for Stages III and IV, reflecting increased uncertainty due to smaller sample sizes after relapse. While Stage I and II curves are more closely aligned and their confidence intervals frequently overlap, the curves for Stage III and IV are distinctly separated and show minimal overlaps with lower-stage groups, reinforcing the pronounced prognostic effect of advanced stage on post-relapse survival. Overall, this figure highlights that not only is the likelihood of relapsing stage-dependent, but the ability to survive a relapse is as well, those diagnosed at higher stages face a compounded disadvantage.

The bottom left panel presents time to relapse by central pathology histology. Patients with favorable histology (coded 0) show substantially better relapse-free survival than those with unfavorable histology (coded 1). For the favorable group, the survival curve remains high and stable, relapse-free survival stays around 90% at 2 years, above 85% at 5 years, and remains close to 85% even at 10 years, reflecting long-term disease control. In contrast, patients with unfavorable histology experience a much steeper decline, particularly within the first year, where relapse-free survival drops sharply below 60%, and falls further to approximately 50% at 5 years. This early divergence between the curves demonstrates that relapse occurs much sooner and more frequently in the unfavorable histology group. The separation between the two

survival curves is pronounced and sustained throughout the entire follow-up period, indicating a clear prognostic distinction. Confidence intervals further support this contrast, intervals for favorable histology are narrow and stable over time, indicating precise estimation due to a larger number of events and consistent outcomes, whereas the unfavorable group shows wider confidence intervals, especially beyond 3–5 years, reflecting increased uncertainty due to fewer patients remaining at risk. These patterns confirm central pathology histology as a powerful predictor of relapse-free survival, with unfavorable histology associated with significantly earlier and more frequent relapses.

The bottom middle panel presents time to death by central pathology histology, revealing clear but somewhat less pronounced differences in overall survival compared to relapse-free survival. Patients with favorable histology maintain high survival probabilities across time, approximately 95% at 2 years, around 92–93% at 5 years, and near 90% at 10 years, indicating excellent long-term outcomes. In contrast, patients with unfavorable histology exhibit a steeper decline in survival, though the decrease is more gradual than for relapse-free survival: survival drops to about 80% by 2 years, around 70–75% at 5 years, and approximately 60–65% at 10 years. The survival gap between the two histology groups remains wide and persistent throughout the follow-up period, though the rate of decline is slower than in the relapse analysis. The 95% confidence intervals show minimal overlap between the groups over most of the follow-up time, particularly in the earlier years, supporting statistically and clinically meaningful differences in mortality risk. Confidence bands for favorable histology are consistently narrower, reflecting greater precision, while those for unfavorable histology widen notably in later years, indicating more uncertainty due to fewer patients at risk. Overall, histology remains a strong prognostic factor for survival, though the contrast in death rates is less abrupt than for relapse.

The bottom right panel illustrates time from relapse to death by central pathology histology, clearly demonstrating that histologic subtype remains a powerful prognostic factor even after relapse. Patients with unfavorable histology experience a very steep and early decline in survival: within just 6 months post-relapse, survival probability falls to around 50%, and by 1 year it drops below 40%. This trend

continues rapidly which by 18 months, survival declines to nearly 25–30%, and falls further below 20% by year 2. The short horizontal span of the unfavorable histology curve reflects the limited post-relapse survival for this group, with very few patients surviving beyond 3 years. In contrast, patients with favorable histology who relapse show substantially better survival: their probability of surviving 1 year post-relapse remains above 80%, drops to around 70% by 2 years, and remains above 60% through year 3. The confidence intervals for unfavorable histology are wide, particularly beyond 1 year, indicating sparse data and increasing uncertainty, while favorable histology confidence bands are narrower in the early years and gradually widen with time. The persistent and wide gap between the curves, especially in the first two years, reflects the profound survival disadvantage faced by patients with unfavorable histology following relapse. These findings emphasize that unfavorable histology not only increases the likelihood of relapse but is also strongly associated with shortened survival thereafter.

Figure 3 displays Kaplan–Meier survival curves stratified by NWTS study group, age at diagnosis, and tumor diameter category across three clinical endpoints: time to relapse, time to death, and time from relapse to death. These variables were selected for visualization based on their statistical significance or borderline significance in preceding log-rank tests. The figure complements earlier univariate results by illustrating time-to-event distributions and highlighting the degree of separation (or lack thereof) among subgroups. While these covariates show less pronounced differences than stage or histology, the figure provides valuable context for interpreting their prognostic relevance and informs subsequent decisions regarding model specification and covariate inclusion.

The top left panel of Figure 3 displays Kaplan–Meier estimates for time to death stratified by NWTS Study 3 and Study 4 cohorts. The survival curves for the two groups remain closely aligned over the entire follow-up period, showing minimal divergence and no points of crossover. At 5 years, both cohorts exhibit high survival probabilities exceeding 90%, with gradual declines reaching approximately 85%–87% by 20 years. The shaded 95% confidence intervals around each curve overlap substantially at all time points, reflecting a lack of statistically meaningful separation in mortality risk between the two study

groups. This close alignment reinforces the earlier log-rank result indicating no significant difference in time to death across studies and suggests that differences in follow-up duration between cohorts do not translate into substantial differences in long-term survival outcomes.

The bottom left panel of Figure 3 presents Kaplan–Meier estimates for time from relapse to death, stratified by NWTS Study 3 and Study 4. Both study groups demonstrate steep declines in post-relapse survival, particularly within the first year, where survival probabilities drop below 60%. By 2 years post-relapses, survival falls below 50% in both groups, and by 5 years, Study 3 shows survival probabilities approaching or falling just below 30%, while Study 4 remains slightly higher at around 35%. Despite this modest difference, the 95% confidence intervals for the two curves are wide and extensively overlap throughout the entire follow-up, indicating no statistically meaningful difference in survival outcomes after relapse. The minimal separation between curves and broad confidence bands suggest that variation in post-relapse mortality between Study 3 and Study 4 is negligible, consistent with the weak significance observed in the log-rank test for this endpoint. These findings reinforce that post-relapse prognosis is comparably poor across cohorts, and study group alone is not a strong discriminator of survival in the relapsed population.

The top middle panel of Figure 3 presents relapse-free survival by age at diagnosis (<2 years, 2–4 years, >4 years). The curves for the <2 and 2–4 age groups are very closely aligned throughout follow-up, even showing potential crossing near 5–10 years, with both groups maintaining relapse-free survival above 90% at 5 years and around 85%–88% at 10 years. In contrast, the >4 group consistently shows lower relapse-free survival across the entire period, dropping to about 85% by 5 years and down to roughly below 80% by 10 years. The confidence intervals for all three groups broadly overlap, especially between <2 and 2–4, indicating little evidence of meaningful separation between these two. The >4 group, while not dramatically different, maintains a consistently lower survival probability and does not overlap as tightly with the other two, suggesting a modest trend toward worse relapse-free survival for older children at diagnosis.

The bottom middle panel of Figure 3 displays overall survival by age at diagnosis. All three age groups exhibit very similar survival trajectories, with survival probabilities exceeding 90% at 5 years and gradually declining to approximately 85% by 15–20 years. The curves are closely aligned, and their 95% confidence intervals overlap substantially throughout, indicating minimal differences in mortality across age groups. There is some suggestion of potential curve crossing between the <2 and 2–4 groups during follow-up, reflecting the absence of a consistent ordering. Notably, unlike in the time-to-relapse panel, the >4 group shows survival probabilities that more closely resemble the other two groups, remaining near 88% at 10 years and just under 85% by year 20. This consistent proximity and overlapping intervals suggest that age at diagnosis is not a strong determinant of overall survival in this population.

The top right panel of Figure 3, displaying time to relapse by tumor diameter category, shows that patients with tumors >15 cm and those with tumors between 10–15 cm have nearly identical relapse-free survival in the first year, with both curves starting above 95% and overlapping almost completely. This early overlap suggests that tumor size does not strongly differentiate relapse risk immediately after diagnosis. However, after the first year, the survival curves begin to diverge: the >15 cm group experiences a steeper decline, with relapse-free survival dropping to roughly 85% at 5 years and below 80% by 10 years. The 10–15 cm group shows a more gradual decline, reaching about 88% at 5 years and 82% at 10 years. Patients with tumors <10 cm consistently have the most favorable outcomes, maintaining around 90% survival at 5 years and 85% at 10 years. Although confidence intervals for all three groups overlap to some extent across the follow-up period, particularly between the >15 cm and 10–15 cm groups, the overall pattern suggests a delayed but progressive effect of tumor size on relapse risk, with larger tumors eventually associated with worse outcomes.

The bottom right panel of Figure 3, showing time to death by tumor diameter category, presents the most overlapping set of survival curves among all panels. Across the entire follow-up period, the curves for the three groups are nearly indistinguishable, frequently crossing one another both early and late in the timeline. All groups maintain survival probabilities above 90% at 5 years and decline gradually to around

85%–87% by year 15–20, with no consistent ordering by tumor size. The 95% confidence intervals overlap substantially throughout, offering little evidence of distinct survival patterns by diameter category. This convergence of curves and confidence bands suggests that, in contrast to other covariates like histology or stage, tumor diameter has a minimal influence on overall mortality in this cohort.

The Kaplan–Meier survival curves and corresponding log-rank tests provided a comprehensive, univariate assessment of the prognostic relevance of key clinical covariates across three event endpoints: time to relapse, time to death, and time from relapse to death. Among all variables examined, tumor stage and central pathology histology consistently demonstrated the most pronounced and persistent survival differences across all outcomes, with well-separated curves and large log-rank test statistics ($p < 0.001$), highlighting their critical prognostic value. Age at diagnosis showed moderate differentiation, particularly for time to relapse and death, while tumor diameter and study cohort revealed weaker or more delayed effects, with minimal curve separation and overlapping confidence intervals, particularly for overall survival.

Several Kaplan–Meier plots suggested non-proportional hazards, particularly when survival curves crossed or diverged at non-constant rates. For example, age-specific relapse-free survival curves showed potential crossing between <2 and 2–4 year groups, and tumor diameter curves diverged only after the first year, indicating potential time-varying effects. Similarly, the abrupt early declines in post-relapse survival for patients with unfavorable histology or advanced-stage disease suggest hazard rates that are not proportional over time. These graphical patterns indicate that the proportional hazards assumption may not hold uniformly, particularly for variables like age, tumor diameter, and histology, depending on the outcome considered.

Taken together, these findings reinforce the importance of tumor stage, histology, and age as strong candidates for inclusion in subsequent multivariable models. Furthermore, the visual and statistical evidence of potential violations of the PH assumption motivates consideration of alternative modeling

strategies, such as additive hazards models or multi-state modeling, to more accurately capture complex prognostic dynamics in this pediatric cancer population.

Model Specification and Assessment of Proportional Hazards Assumptions

To determine the appropriate handling of study effects in our Cox regression models, we evaluated three modeling strategies for each of three clinically meaningful endpoints: time to relapse, time to death, and time from relapse to death. All models included age at diagnosis (in years) and tumor diameter (in centimeters) as continuous covariates, and Wilms tumor stage and histology as categorical covariates. In the first approach, we included study (NWTS-3 vs. NWTS-4) as a categorical covariate. In the second approach, we stratified the Cox model by study, allowing the baseline hazard to vary across studies but constraining covariate effects to be the same. In the third approach, we stratified by study and additionally included interaction terms between study and each covariate, thereby allowing covariate effects to differ between studies. This third model approximates a fully study-specific analysis and serves as a diagnostic tool to assess whether separate models should be fit by study in subsequent analyses, particularly when using methods where stratification is not straightforward (e.g., AFT or additive hazards models).

To compare these models, we computed the Akaike Information Criterion (AIC), a widely used measure of model fit that balances model complexity and goodness-of-fit. AIC is defined as $AIC = 2k - 2log(L)$, where kk is the number of estimated parameters and LL is the maximized likelihood. Lower AIC values indicate a better tradeoff between goodness-of-fit and model parsimony. Since the models being compared are not nested, there is no relationship of null and alternative models here, classical likelihood ratio tests are not applicable, and AIC provides an appropriate and rigorous basis for model comparison.

Table 5 presents the AIC values for each modeling approach. For time to relapse, the AICs were 10,557 (study as covariate), 9,646 (stratified), and 9,635 (stratified + interactions). For time to death, AICs were 6,767, 6,160, and 6,166, respectively. For time from relapse to death, AICs were 5,233, 4,619, and 4,625, respectively. These results consistently show a substantial reduction in AIC when moving from a model

that treats study as a covariate to one that stratifies by study, suggesting a markedly improved model fit when allowing for study-specific baseline hazards. However, incorporating interaction terms yields minimal additional improvement in AIC, indicating little evidence that covariate effects differ meaningfully across studies. Based on these results, we stratify by study in subsequent Cox model analyses to account for heterogeneous baseline hazards. For subsequent analyses using accelerated failure time (AFT) and additive hazards models, we fit models separately within each study, since stratification, as implemented in Cox models, is not inherently compatible with the modeling structure of AFT and additive hazards frameworks. This approach ensures appropriate accommodation of study-specific baseline differences across modeling strategies.

To evaluate whether age at diagnosis and tumor diameter should be modeled as continuous or categorical variables, we examined the functional form of these covariates using Martingale residual plots from Cox proportional hazards models for each of the three clinical endpoints: time to relapse, time to death, and time from relapse to death. Martingale residuals are commonly used for assessing the adequacy of functional form in survival models; systematic departures from a horizontal line at zero in the residual plots may indicate nonlinearity and suggest that a transformation or categorization of the covariate may be appropriate.

Figure 4 displays Martingale residuals plotted against age at diagnosis (top row) and tumor diameter (bottom row), with a Loess smooth superimposed to assess potential non-linear relationships across three clinical endpoints: time to relapse, time to death, and time from relapse to death.

For age at diagnosis, the residual plot for time to relapse (top left) shows a generally flat smoothed curve throughout most of the age range, with a slight upward curvature emerging after approximately age 10. This suggests that while the linearity assumption may hold for younger children, it may be violated for older patients. In the plot for time to death (top center), a similar pattern appears, with the curve remaining flat until around age 10 before gradually trending upward, indicating a potential

underestimation of mortality risk among older patients if modeled linearly. The pattern becomes more pronounced in the time from relapse to death plot (top right), where the smoothed residual curve initially dips and then rises sharply beginning around age 6 or 7. This more substantial nonlinearity implies that the association between age and post-relapse mortality may vary meaningfully across the age spectrum.

For tumor diameter, the residual plot for time to relapse (bottom left) reveals a relatively flat trend initially, with a steady upward slope developing after approximately 10 cm. This suggests that larger tumors may carry increasing relapse risk in a nonlinear fashion. The plot for time to death (bottom center) shows a concave shape, with the residual curve dipping slightly in the midrange of tumor sizes (approximately 10–20 cm) before trending upward again, implying a potentially non-monotonic association between tumor size and mortality. In the time from relapse to death plot (bottom right), the residual curve is mostly flat up to about 15 cm but rises steeply thereafter, once again indicating that very large tumors may confer elevated post-relapse mortality. Notably, the extreme right of this plot features a persistent outlier with a markedly high residual value. This point may unduly influence the smoothed curve in the upper range, and its clinical plausibility or data quality should be carefully evaluated.

Taken together, these plots suggest that the assumption of linearity may not hold for either age at diagnosis or tumor diameter when modeled as continuous covariates. Given these observed patterns, and the presence of clinically meaningful categorical thresholds for both variables as mentioned previously, we chose to model age and tumor diameter categorically in subsequent analyses. This approach improves interpretability, addresses potential nonlinearity, and ensures consistency across modeling frameworks including Cox, AFT, and additive hazards models, where the ability to flexibly model nonlinear effects is limited.

To assess the validity of the proportional hazards assumption in Cox regression, we applied the Grambsch–Therneau test based on scaled Schoenfeld residuals for each covariate and globally across all covariates. The null hypothesis of this test is that the log hazard ratios are constant over time, that is, the

effect of each covariate on the hazard is proportional and does not vary with time. A small p-value leads to rejection of this null, indicating time-varying covariate effects and a violation of the proportional hazards assumption. We performed this test for each covariate as well as a global test, using stratified Cox models that account for the true stratification by study (NWTS-3 and NWTS-4), and included age at diagnosis, tumor diameter, stage of disease, and histology as categorical variables with clinically defined levels. The results are reported in details of Table 6.

For the time to relapse endpoint, the Grambsch–Therneau test yielded p-values of 0.43 for age group and 0.11 for tumor diameter, suggesting no strong evidence of time-varying effects for these covariates. In contrast, the p-values for histology and stage of disease were both <0.001, indicating clear violations of the proportional hazards assumption. The global test also returned a p-value <0.001, reinforcing the conclusion that the proportional hazards assumption does not hold in this model.

For time to death, the age group had a p-value of 0.015, and tumor diameter was 0.078. While tumor diameter showed no strong evidence against proportionality, the p-value for age group suggests potential time-dependent effects. As with the relapse model, histology and stage again produced highly significant p-values (<0.001), and the global test p-value was also <0.001. This again indicates strong evidence of non-proportionality across multiple covariates.

For time from relapse to death, the age group and tumor diameter had p-values of 0.066 and 0.37, respectively, suggesting no meaningful departure from proportionality for these variables. However, histology had a p-value of 0.01 and stage of disease had a highly significant p-value (<0.001), leading to a global test p-value <0.001.

Overall, these results consistently point to violations of the proportional hazards assumption across all three clinical endpoints. In particular, the effects of histology and stage of disease appear to vary significantly over time, regardless of the endpoint. Even when some covariates individually appear proportional, the global tests uniformly reject the assumption, indicating that the Cox proportional

hazards model may not be adequate for these data. Therefore, alternative modeling strategies that do not rely on the proportional hazards assumption, such as accelerated failure time models or additive hazards models, are warranted for further analysis.

Accelerated Failure Time Models: Distributional Selection, Model Diagnostics, and Parameter Estimation

Given the clear violations of the proportional hazards assumption for all three clinical endpoints, time to relapse, time to death, and time from relapse to death, we adopted the accelerated failure time (AFT) framework as a semiparametric alternative that directly models the (log-transformed) survival time. The AFT model assumes that covariates act multiplicatively on the survival time scale, offering a natural and interpretable structure when proportional hazards do not hold.

In these models, we used clinically meaningful categorical representations for each covariate: age at diagnosis (<2 years, 2–4 years, >4 years), tumor diameter (<10 cm, 10–15 cm, >15 cm), histology (favorable vs. unfavorable), and disease stage (I–IV). Based on prior evidence from stratified Cox models, we fit separate AFT models for Study 3 and Study 4 to accommodate study-specific differences in baseline survival distributions.

To determine the most appropriate error distribution for each model, we considered six commonly used parametric families within the AFT framework: Weibull, exponential, log-normal, log-logistic, normal, and gamma. Model selection was guided by the Bayesian Information Criterion (BIC). Unlike the Akaike Information Criterion (AIC), which emphasizes predictive accuracy, BIC incorporates a stronger penalty for model complexity and is more appropriate in settings where the goal is estimation and inference. This makes BIC particularly suitable for comparing nested or non-nested parametric families, where overfitting is a concern and parsimony is desirable. All results are being recorded in Table 6 across the three chosen endpoints.

For time to relapse, the gamma distribution provided the best fit for both Study 3 and Study 4, with BIC values of 2519 and 2999, respectively. These values were substantially lower than those from alternative distributions, indicating a clearly superior fit. Log-normal and log-logistic models were next best, with log-normal yielding BICs of 2670 (Study 3) and 3079 (Study 4), but these values were not competitive with gamma, reinforcing the robustness of the gamma distribution for modeling relapse timing.

For time to death, the log-normal distribution provided the best fit for Study 4, with a BIC of 1839. In Study 3, the BIC values for log-normal (1858), gamma (1855), and log-logistic (1870) were relatively close. Given the stronger evidence in favor of log-normal for Study 4 and the nearly equivalent performance in Study 3, we selected the log-normal distribution as the optimal choice for modeling time to death across studies. This selection reflects a balance between goodness-of-fit and consistency across endpoints, while also acknowledging that slight differences in BIC within 2–6 points are not typically meaningful in isolation.

For time from relapse to death, the optimal distribution was less definitive. In Study 3, the gamma distribution had the lowest BIC (867), while in Study 4, the log-normal distribution performed best with a BIC of 880. Log-logistic models were again competitive in both studies but did not outperform either gamma or log-normal. Given the comparability of BIC values for gamma and log-normal, we selected the log-normal distribution for consistency with the time to death endpoint and for interpretability of time ratios, particularly under symmetric error assumptions. The log-normal model also produced more stable diagnostic behavior across residual-based assessments, as discussed in the following section.

Together, these distributional choices reflect a careful balance of statistical fit (via BIC), model interpretability, and diagnostic robustness. All candidate models were retained and evaluated to establish a consistent comparative foundation, but the log-normal and gamma distributions emerged as the most appropriate for inference across endpoints and studies within the AFT framework.

Following distributional selection based on BIC (presented in Table 7), we evaluated the fit of the best-performing accelerated failure time (AFT) models using standardized residual diagnostics across all three clinical endpoints: time to relapse (gamma distribution), time to death (log-normal distribution), and time from relapse to death (log-normal distribution). Each panel in Figure 5 corresponds to a diagnostic plot for one endpoint, stratified by Study 3 and Study 4 (left and right panels, respectively). The rows display standardized residuals versus linear predictors (top), histograms of standardized residuals (middle), and normal Q-Q plots (bottom), allowing assessment of distributional assumptions and overall model adequacy.

For time to death, modeled with a log-normal distribution, the standardized residual plots (top-left) reveal a pronounced downward trend, particularly in Study 3, indicating a systematic relationship between residuals and the linear predictor that suggests heteroskedasticity and potential model misspecification. The corresponding histograms (middle-left) display residuals skewed to the left in both studies, deviating from the symmetry expected under the log-normal assumption and implying potential underestimation of survival times for a significant number of patients. Normal Q-Q plots (bottom-left) further highlight this concern, showing strong downward curvature in the lower tail, especially in Study 3, consistent with heavy-tailed deviations from normality. These diagnostic features indicate that the log-normal model does not fully capture the error structure for this endpoint.

However, upon examining diagnostics across all candidate distributions, no single distribution emerged as consistently superior across both studies. Similar degrees of deviation from model assumptions were observed for the Weibull, log-logistic, gamma, and normal models, making it difficult to identify a clearly better alternative. Therefore, the log-normal model, selected based on its favorable BIC values, was retained for inference and reporting. This choice reflects a pragmatic balance between statistical fit, parsimony, and the need for a consistent modeling framework across endpoints and studies.

For time to relapse, modeled using a gamma distribution, the standardized residual plots (top-center) show a relatively clear pattern with reasonably consistent spread across the linear predictor in both studies, especially in Study 4, suggesting no severe heteroskedasticity. However, the residual histograms (middle-center) reveal strong departures from symmetry, with pronounced left skewness and sharp peaks, indicating that the gamma model does not adequately capture the distributional shape of the residuals. The normal Q-Q plots (bottom-center) further underscore this misfit: residuals in both studies deviate substantially from the theoretical normal line, with systematic crossing of the diagonal and no sustained overlap, particularly in the left tails. These diagnostic results suggest that, despite being selected based on BIC, the gamma distribution may poorly approximate the underlying error structure, especially in Study 3. However, diagnostics for alternative candidate distributions (e.g., log-normal, log-logistic) showed similarly poor performance, and no model clearly outperformed others across both studies. Thus, the gamma AFT model was retained for time to relapse to maintain consistency with the BIC-based model selection strategy and enable coherent comparison across endpoints, while acknowledging its limitations in capturing residual behavior.

For time from relapse to death, modeled using a log-normal distribution, the standardized residual plots (top-right) display a mild downward trend, suggesting some potential deviation from model assumptions, though the violation is less pronounced compared to the other two endpoints. The residuals are also more sparsely distributed, likely reflecting the smaller number of patients experiencing both relapse and death. This sparser yet more balanced event distribution, between those who die post-relapses and those censored, may contribute to better model performance. Supporting this, the histograms (middle-right) appear reasonably symmetric and less skewed, while the Q-Q plots (bottom-right) show strong alignment with the diagonal reference line, especially in Study 4, with only modest deviations in the tails. Taken together, these diagnostics suggest that the log-normal AFT model provides a comparatively better fit for this endpoint than for the others and supports its continued use as the BIC-optimal model for time from relapse to death.

Building on the residual diagnostics in Figure 5, Table 8 presents the estimated time ratios (TRs), standard errors, 95% confidence intervals, and p-values from the BIC-selected AFT models for each of the three clinical endpoints. Although the residual diagnostics revealed clear deviations from model assumptions for time to relapse and time to death, particularly in terms of skewness, heteroskedasticity, and Q-Q plot curvature, these models were nonetheless retained for two reasons. First, across all candidate distributions considered, no alternative consistently outperformed the selected models, and the gamma and log-normal distributions remained optimal under the BIC criterion. Second, despite their limitations, these AFT models provide a useful baseline for comparison in subsequent additive hazard analyses, enabling a consistent and interpretable point of reference. The model for time from relapse to death, which showed comparatively better fit across diagnostics, lends additional support to retaining this framework for presentation. We thus present Table 8 not as definitive inference, but as a structured summary of effect estimates under the AFT paradigm, with acknowledgment of the models' limitations.

Based on Table 8, for time to relapse, the estimated TRs for unfavorable vs. favorable histology were 0.31 (95% CI: 0.17–0.57, $p < 0.001$) in NWTS-3 and 0.38 (95% CI: 0.24–0.61, $p < 0.001$) in NWTS-4. These indicate that, adjusting for other covariates, children with unfavorable histology relapsed in approximately one-third the time of those with favorable histology, a 69% and 62% reduction in median time to relapse, respectively. This underscores the aggressive early disease course associated with unfavorable histology.

The adverse prognostic impact of histology was even more pronounced for overall survival. In NWTS-3, the TR for unfavorable histology was an extraordinary 0.02 (95% CI: 0.01–0.05, $p < 0.001$), implying a 98% reduction in survival time relative to favorable histology. A similarly stark effect was observed in NWTS-4 (TR = 0.05, 95% CI: 0.03–0.09, $p < 0.001$). These findings demonstrate the overwhelming influence of histological subtype on mortality risk in Wilms tumor.

Among children who relapsed, histological subtype continued to be a key determinant of outcome. For time from relapse to death, TRs were 0.15 (95% CI: 0.07–0.31, $p < 0.001$) in NWTS-3 and 0.18 (95% CI:

0.11–0.30, $p < 0.001$) in NWTS-4. These results reflect an 82–85% reduction in post-relapse survival time for patients with unfavorable histology. Taken together, these findings confirm that unfavorable histology is a potent and persistent risk factor across the disease trajectory, with critical implications for clinical risk stratification and treatment intensification.

Tumor stage, with Stage 1 as the reference, was a strong predictor of time to relapse, especially for Stage 4 disease. In NWTS-3, only Stage 4 was significantly associated with shorter relapse-free survival (TR = 0.24, 95% CI: 0.13–0.45, $p < 0.001$), suggesting a 76% faster time to relapse. In NWTS-4, both Stage 2 (TR = 0.57, 95% CI: 0.39–0.84, $p = 0.004$) and Stage 4 (TR = 0.42, 95% CI: 0.26–0.68, $p < 0.001$) were significantly associated with accelerated relapse. Stage 3 was not statistically significant in either study but consistently trended toward shorter relapse times. These results support a general pattern of progressively poorer relapse outcomes with increasing disease stage, though statistical significance varied by cohort.

For time to death, tumor stage demonstrated a consistent, stepwise inverse association across both studies. In NWTS-3, TRs were 0.34 (Stage 2, $p = 0.01$), 0.11 (Stage 3, $p < 0.001$), and 0.03 (Stage 4, $p < 0.001$), indicating 66%, 89%, and 97% reductions in survival times, respectively. NWTS-4 yielded similar gradients: TRs were 0.35 (Stage 2), 0.29 (Stage 3), and 0.07 (Stage 4), all with $p < 0.001$. Notably, the effect of Stage 3 was more severe in NWTS-3 (TR = 0.11) compared to NWTS-4 (TR = 0.29), suggesting potential differences in cohort characteristics or treatment efficacy over time. Overall, tumor stage at diagnosis is a strong, independent, and consistent predictor of mortality in Wilms tumor.

When examining time from relapse to death, NWTS-3 showed similar TRs for Stages 2–4 (range: 0.26–0.27), all statistically significant, indicating a uniformly poor post-relapse prognosis for advanced stages. In NWTS-4, however, a clearer gradient was observed: Stage 2 patients had a TR of 0.56 (95% CI: 0.30–1.03, $p = 0.06$), while Stage 3 and Stage 4 had significantly lower TRs of 0.23 and 0.13, respectively (both $p < 0.001$). This suggests that in NWTS-4, more advanced stage was associated with incrementally

worse outcomes after relapse, highlighting the prognostic relevance of initial stage even in late disease phases.

Age effects varied across endpoints and studies. For time to relapse, age was not a statistically significant factor in either cohort. NWTS-3 showed nearly identical TRs close to 1.0 across age groups (all $p > 0.75$), while NWTS-4 suggested a potential delay in relapse among children aged 2–4 years (TR = 1.43, $p = 0.06$), though not statistically significant.

In contrast, time to death showed a stronger age-related pattern. In NWTS-3, children aged 2–4 had significantly longer survival than those under 2 (TR = 2.43, 95% CI: 1.17–5.00, $p = 0.02$). NWTS-4 also showed prolonged survival for this age group (TR = 1.70, $p = 0.049$), albeit with borderline significance. Older children (>4 years) showed mixed effects: NWTS-3 showed a non-significant survival benefit (TR = 1.53, $p = 0.24$), while NWTS-4 showed a non-significant reduction in survival (TR = 0.89, $p = 0.64$). The opposing directions for the oldest group suggest possible differences in treatment response or disease biology across studies.

For time from relapse to death, age effects were not statistically significant in either study. However, both NWTS-3 and NWTS-4 indicated longer post-relapse survival among children aged 2–4 years (TR = 1.60 and 1.47, respectively), with modest improvements also seen in the >4 group in NWTS-4 (TR = 1.46). Though not conclusive, these trends suggest a potential survival advantage for children diagnosed after infancy, particularly in the 2–4-year group.

The impact of tumor diameter on time to relapse differed between studies and was not statistically significant in either. NWTS-3 suggested shorter time to relapse for larger tumors (TR = 0.86 for 10–15 cm, TR = 0.75 for >15 cm), while NWTS-4 showed a non-significant trend toward delayed relapse with increasing size (TR = 1.17 and 1.40, respectively). The opposing directions and non-significant p-values suggest that tumor diameter is not a reliable predictor of relapse timing.

For time to death, the associations with tumor size remained non-significant and inconsistent. In NWTS-3, intermediate-sized tumors were associated with longer survival (TR = 1.29), while very large tumors suggested shorter survival (TR = 0.72). NWTS-4 showed little difference for 10–15 cm tumors (TR = 0.98) and a non-significant survival benefit for tumors >15 cm (TR = 1.46). These mixed patterns indicate a lack of consistent prognostic utility for tumor size in predicting mortality.

The most unexpected finding appeared in the model for time from relapse to death in NWTS-3, where larger tumors were associated with significantly longer post-relapse survival. TRs were 2.32 (10–15 cm) and 3.05 (>15 cm), both statistically significant ($p = 0.03$), suggesting 2–3 fold longer survival after relapse for patients with larger tumors. This counterintuitive result contrasts with clinical expectations and may reflect unmeasured confounding, treatment selection bias, or heterogeneity in relapse biology. In NWTS-4, no such effect was observed (TRs near 1.0, $p > 0.80$), reinforcing the notion that the NWTS-3 finding may be study-specific and warrants cautious interpretation.

To conclude, Table 8 provides a structured summary of covariate effects estimated from BIC-selected AFT models for time to relapse, time to death, and time from relapse to death. Across all three endpoints and both studies, histology emerged as the most consistently significant and clinically consequential predictor: unfavorable histology was associated with drastically shorter times to relapse and death, and significantly reduced survival even after relapse. Tumor stage also showed a strong and monotonic association with prognosis. Stage 4 consistently predicted earlier relapse and markedly reduced overall and post-relapse survival, with intermediate stages demonstrating a graded pattern of worsening outcomes. Age at diagnosis was a less consistent predictor, with only the 2–4 year age group demonstrating potential survival advantages in select endpoints, and mixed results observed for older children across studies. Tumor diameter did not show significant or consistent effects for relapse or death in either study, though NWTS-3 revealed an unexpected survival advantage post-relapse for patients with larger tumors, a finding not replicated in NWTS-4 and interpreted cautiously.

While the AFT models in Table 8 provide interpretable summaries of covariate effects across the three clinical endpoints, their inferential reliability is limited by clear violations of model assumptions, most notably for time to relapse and time to death. Residual diagnostics revealed substantial skewness, heteroskedasticity, and curvature in Q-Q plots, raising concerns about the adequacy of the log-normal and gamma distributions, despite their selection based on BIC. These limitations are further underscored by the unexpected and counterintuitive result in NWTS-3, where larger tumor diameter was associated with significantly longer survival following relapses. This paradoxical finding, inconsistent with clinical expectations and not replicated in NWTS-4, raises additional doubts about model validity and suggests potential unmeasured confounding or structural misspecification. Accordingly, we do not treat these AFT results as definitive but rather as preliminary benchmarks for comparison. To address these concerns and accommodate time-varying effects and non-proportional hazards, we proceed with additive hazards modeling. This semiparametric framework provides greater flexibility in modeling dynamic covariate effects and offers a more robust foundation for inference considering the diagnostic shortcomings observed here.

Assessment and Visualization of Time-Varying Covariate Effects Using Additive Hazards Models

To further investigate the time-varying nature of covariate effects suggested by the AFT model diagnostics, we fit separate Aalen additive hazards models for NWTS-3 and NWTS-4. This semiparametric approach allows covariate effects to evolve flexibly over time, offering a more robust alternative to proportional hazards or parametric survival models when such assumptions are violated. For each of the three clinical endpoints, time to relapse, time to death, and time from relapse to death, we specified models that included age at diagnosis, tumor diameter, histologic classification, and disease stage as categorical predictors. This structure ensures interpretability and consistency across covariates, while leveraging the additive model's capacity to estimate cumulative regression functions that capture both the magnitude and temporal dynamics of covariate effects.

Table 9 presents formal hypothesis tests for detecting time-varying covariate effects using two nonparametric statistics, the supremum test and the Kolmogorov–Smirnov (K–S) test, applied to the cumulative coefficient functions estimated from Aalen's additive hazards model. These tests evaluate, for each covariate and clinical endpoint (time to relapse, time to death, and time from relapse to death), whether the effect of that covariate remains constant over follow-up or varies with time.

The supremum test examines the maximum absolute deviation of the estimated cumulative coefficient process $\hat{B}(t)$ from a horizontal line at zero, under the null hypothesis $H_0: B(t) = 0$. In other words, it tests whether there is any significant effect of the covariate over time, regardless of whether that effect is constant or time-varying. A p-value below 0.05 for the supremum test indicates that the covariate has a statistically significant effect on the hazard function at some point during follow-up, rejecting the null of no effect.

In contrast, the K–S test focuses on the constancy of the covariate effect over time. It evaluates the null hypothesis $H_0: B(t) = \beta t$, i.e., that the cumulative coefficient grows linearly over time, implying a time-constant hazard contribution (analogous to a proportional effect). The K–S test measures the maximum distance between the observed cumulative coefficient trajectory and its expected linear form under the null. A p-value below 0.05 here indicates that the covariate's effect varies over time, rejecting the assumption of a constant effect.

Together, the supremum and Kolmogorov–Smirnov (K–S) tests provide complementary information for assessing the nature of covariate effects in the additive hazards framework. A supremum test p-value below 0.05 indicates that a covariate has a statistically significant effect at some point during follow-up, regardless of whether that effect is constant or time-varying. In contrast, a K–S test p-value below 0.05 suggests that the effect of the covariate is not constant over time, indicating the presence of time-varying influence. When both p-values fall below 0.05, the covariate is interpreted as having a significant and time-varying effect. If only the supremum test is significant, the covariate likely has a time-constant effect

that is nonetheless non-zero. Conversely, when only the K–S test is significant, the covariate may exhibit transient or oscillating effects that vary over time but average to zero, a scenario that is less common. Finally, if both p-values exceed 0.05, there is no statistically detectable effect or variation in that covariate. In this analysis, these tests were conducted separately for NWTS-3 and NWTS-4, allowing for a comparison of temporal patterns across studies, an important consideration given differences in treatment regimens, enrollment periods, and patient characteristics between the two cohorts.

Based on Table 9 of NWTS-3 for time to relapse, the intercept term exhibits strong evidence of both non-zero and time-varying effects (Sup < 0.001, K–S < 0.001), indicating substantial changes in baseline relapse risk over time. Age at diagnosis emerges as an important prognostic factor: both age 2–4 years and age >4 years show statistically significant non-zero effects (Sup < 0.001 and 0.007, respectively), but only the younger age group (2–4) demonstrates evidence of time-varying influence (K–S = 0.001), whereas age >4 years likely exerts a constant but non-zero effect (K–S = 0.104). Tumor diameter, in contrast, shows no significant association with relapse risk across time (Sup and K–S > 0.05 for both diameter categories), suggesting that size alone may not influence relapse dynamics in this cohort. Histologic classification, particularly unfavorable histology, displays strong and time-varying effects (Sup < 0.001, K–S < 0.001), highlighting the adverse influence of histologic subtype on early and changing relapse hazards. Among stage groups, Stage 2 appears unrelated to relapse (Sup = 0.507, K–S = 0.180), while Stage 3 and Stage 4 both exhibit robust and time-varying effects (Sup and K–S < 0.001), underscoring increasing relapse risk over time for patients with more advanced disease.

In NWTS-4, a similar pattern emerges in several areas, yet notable distinctions arise. The intercept again suggests a non-constant baseline hazard for relapse (Sup = 0.002, K–S = 0.024). Age effects are less pronounced compared to NWTS-3: the 2–4 year group has marginal evidence of non-zero effect (Sup = 0.073) and no time-variation (K–S = 0.315), whereas the age >4 group is significant in both Supremum and K–S tests (0.041 and 0.023), indicating a dynamically varying risk for older children in this cohort, reversing the pattern observed in NWTS-3. Consistent with NWTS-3, tumor diameter groups do not show

significant associations with relapse (Sup and K–S > 0.05), reinforcing the limited prognostic relevance of size. Unfavorable histology is again time-varying and significant (Sup and K–S < 0.001), confirming its consistent adverse role. For stage, Stage 2 also becomes significant and time-varying (Sup and K–S < 0.001), in contrast to NWTS-3 where it was negligible. Stage 3 shows borderline significance in both tests (Sup = 0.031, K–S = 0.056), and Stage 4 remains consistently time-varying and significant (Sup and K–S < 0.001). Comparing across studies, the time-varying nature of stage-related relapse risks is evident in both NWTS-3 and NWTS-4, though NWTS-3 shows stronger and broader age-related effects. These contrasts may reflect evolving treatment protocols, temporal cohort differences, or changes in clinical staging practices between studies.

In NWTS-3 for time to death, nearly all covariates exhibit evidence of both significant and time-varying effects on the hazard of death. The intercept term is strongly significant (Sup <0.001, K–S <0.001), indicating that the baseline hazard changes considerably over time. Both age groups, 2–4 years and >4 years, show strong time-varying behavior (Sup and K–S <0.001), implying that the prognostic impact of age is not only substantial but also dynamic throughout the course of follow-up. Tumor diameter >15 cm shows a borderline effect in the supremum test (Sup = 0.050), but the K–S test is non-significant (K–S = 0.533), suggesting that any influence of large tumor size may be constant and weak. Unfavorable histology continues to play a prominent role, with strong evidence of both significance and time-variation (Sup and K–S < 0.001). For disease stage, both Stage 3 and Stage 4 are associated with time-varying and highly significant effects (all p < 0.001), while Stage 2 shows weaker evidence of association (Sup = 0.075, K–S = 0.039), potentially indicating a mild and possibly time-constant effect.

In comparison, NWTS-4 shows a more limited pattern of time-varying covariate behavior. The intercept is non-significant (Sup = 0.119, K–S = 0.500), suggesting a relatively stable baseline hazard across time, which may reflect more standardized care protocols or overall lower baseline mortality risk. For age, only the 2–4-year group shows a significant and time-varying effect (Sup and K–S < 0.001), while the >4-year group is marginally significant in the supremum test (Sup = 0.051) but not time-varying (K–S = 0.382).

Tumor diameter remains non-significant for both size categories, mirroring results from NWTS-3. Unfavorable histology again demonstrates a robust, time-varying effect (Sup and K–S <0.001). All three stage groups are significant in NWTS-4, with Stage 3 and Stage 4 showing clear time-varying behavior (Sup and K–S <0.001), while Stage 2 shows weaker yet present variation (Sup = 0.002, K–S = 0.018). In summary, NWTS-3 presents broader time-varying behavior across baseline risk, age, and stage, whereas in NWTS-4, time-varying effects are more concentrated in histology and higher disease stages, reflecting possible shifts in treatment efficacy or cohort characteristics between the studies.

In NWTS-3, the additive hazards model reveals that the intercept term is again highly significant and time-varying (Sup < 0.001, K–S < 0.001), reflecting substantial changes in baseline hazard following relapse. In contrast to the earlier endpoints, age appears to have no statistically significant impact on post-relapse survival: neither the 2–4-year group nor the > 4-year group shows significant results in either the supremum or K–S tests (all p > 0.05), suggesting that age at diagnosis does not strongly influence outcomes after recurrence in this cohort. Tumor diameter categories (10–15 cm and >15 cm) are similarly non-significant, indicating no detectable effect on hazard after relapse. Unfavorable histology, however, continues to exert a strong and time-varying effect (Sup and K–S < 0.001), underscoring its critical role even in the late disease course. Among the staging variables, Stage 2 and Stage 3 show only marginal or non-significant effects (Sup = 0.091 and 0.082, respectively), while Stage 4 demonstrates a clear and dynamic impact on hazard (Sup and K–S <0.001), emphasizing the severity of advanced-stage disease post-relapse.

In NWTS-4, the post-relapse hazard structure differs notably from NWTS-3, particularly in the role of age. The intercept again shows strong significance (Sup < 0.001, K–S = 0.001), indicating a shifting baseline hazard. Unlike NWTS-3, both age groups, 2–4 years and >4 years, exhibit moderate yet statistically significant time-varying effects (Age 2–4: Sup = 0.004, K–S = 0.007; Age > 4: Sup = 0.004, K–S = 0.023), suggesting that older age at diagnosis may be associated with evolving risks after relapse in this later cohort. Tumor diameter remains non-significant (all p > 0.2), mirroring the earlier pattern of

limited prognostic value. Unfavorable histology and Stage 4 both retain their strong and time-varying effects (all $p < 0.001$), consistently confirming their prognostic importance across both studies. Stage 3 also emerges as statistically significant in NWTS-4 (Sup = 0.010), with some evidence of time-varying influence (K–S = 0.003), in contrast to its more muted role in NWTS-3. Taken together, these findings highlight how the prognostic landscape following relapse may have shifted over time: while histology and advanced staging remain dominant in both cohorts, age at diagnosis emerges as a more dynamic and relevant factor in NWTS-4, potentially reflecting improved salvage therapies or cohort-specific differences in disease biology or treatment response.

Taken together, the additive hazards analyses from NWTS-3 and NWTS-4 illustrate both shared and divergent patterns in the temporal dynamics of risk factors across the disease trajectory of Wilms tumor. Across all three endpoints, time to relapse, time to death, and time from relapse to death, unfavorable histology and advanced disease stage (especially Stage 4) consistently emerge as significant and time-varying prognostic indicators, underscoring their persistent impact on patient outcomes. Age at diagnosis shows more nuanced behavior: it plays a pronounced role in NWTS-3 for earlier endpoints, particularly time to relapse and time to death, whereas in NWTS-4, its influence intensifies during the post-relapse period. This shift may reflect changing treatment protocols, improved relapse management, or biological differences in later-era patient cohorts. Tumor diameter, by contrast, exhibits limited prognostic utility across all endpoints and both studies, suggesting that size alone is not a robust indicator of temporal hazard in Wilms tumor.

These results are visually reinforced by Figure 6, which plots the cumulative regression functions estimated from Aalen's additive hazards model. The trajectories in these plots mirror the patterns uncovered through supremum and Kolmogorov–Smirnov tests: covariates with strong and time-varying effects display steep or evolving slopes, indicating shifting contributions to hazard over time. In contrast, flat or slowly changing lines correspond to covariates with negligible or time-constant effects. Comparing NWTS-3 and NWTS-4 side-by-side, the plots in Figure 6 highlight how baseline hazard dynamics, as

well as the prognostic roles of age and stage, have evolved, supporting the hypothesis that treatment era, staging practices, or patient mix influenced the underlying survival processes. In the broader context of Wilms tumor survivorship, these findings emphasize the importance of modeling time-varying effects to capture the complex, evolving nature of relapse and mortality risk, which may inform both clinical risk stratification and long-term management strategies.

The cumulative regression functions for time to relapse in Figure 6 reveal both consistent and evolving patterns of covariate influence across NWTS-3 (left) and NWTS-4 (right), with interpretation enriched by the inclusion of 95% pointwise confidence bands. In both studies, the intercept function rises sharply in the early months, then levels off, capturing elevated baseline relapse risk shortly after diagnosis. The relatively narrow confidence bands during early follow-up indicate precise estimation in this critical window, while widening bands over time reflect increasing uncertainty as at-risk populations diminish. Unfavorable histology shows the strongest and most persistent cumulative effect in both studies, with a steeper trajectory in NWTS-3 (~0.45 vs. ~0.35 in NWTS-4), and consistently tight confidence bands, affirming their robust, adverse, and time-varying impact on relapse hazard. Stage 4 similarly contributes substantial cumulative hazard, particularly in NWTS-3, where its upward trend and narrow bands indicate a reliably elevated risk. Stage 3 follows with a moderate slope, though the confidence bands occasionally touch zero in NWTS-4, suggesting diminished or less certain effects in that cohort. Stage 2 remains near flat and statistically indistinguishable from zero in both studies, confirming minimal influence on relapse.

The age-related trajectories in NWTS-3 highlight important differential risk dynamics. The cumulative regression function for age 2–4 years shows a marked and consistently negative slope, with the curve trending downward well below zero and its confidence band remaining non-overlapping with the horizontal axis for much of the follow-up period. This indicates a significant protective effect against relapse that intensifies over time. In contrast, the age >4 years group shows a more modest downward trend, with its cumulative effect closer to zero and broader uncertainty, suggesting a weaker, possibly constant, and less definitively protective influence on relapse risk. In NWTS-4, however, the pattern

shows difference: age >4 years has a higher cumulative effect, with divergence from zero beyond year one, while age 2–4 years shows a flatter trend and wider, often overlapping confidence bands, signaling reduced or less reliable age-related differentiation in more recent patients. As for tumor diameters in Study 3, the >15 cm diameter group shows a slight upward trend in its cumulative regression function for time to relapse, suggesting a weakly increasing effect over time, whereas the 10–15 cm group remains flat throughout. In Study 4, both diameter categories exhibit stable trajectories close to zero, with no discernible time-varying effect. These patterns indicate a potential minor influence of very large tumors in NWTS-3 but consistently limited prognostic value for tumor size overall.

The cumulative regression functions for time to death in NWTS-3 (left panel) and NWTS-4 (right panel) illustrate distinct temporal patterns in covariate effects, consistent with earlier test-based findings and model diagnostics. In NWTS-3, several covariates exhibit pronounced time-varying effects, particularly histology and stage, with unfavorable histology showing the steepest early increase and the highest cumulative impact on the hazard function. The cumulative effect of Stage 4 similarly rises rapidly and remains distinctly above that of other stage groups, highlighting the persistent excess mortality associated with advanced disease. In Study 3, both age groups (2–4 years and >4 years) exhibit clear negative cumulative regression trends, indicating a protective effect on mortality risk following diagnosis, with more pronounced downward slopes observed for children aged 2–4 years. The intercept function shows a steep rise within the first 5 years, then plateaus, reflecting substantial early baseline hazard that diminishes with time. In contrast, tumor diameter groups (10–15 cm and >15 cm) display flat, near-zero trajectories, reinforcing their limited prognostic impact on time to death in this cohort.

In NWTS-4, the cumulative regression functions for time to death reveal a broadly consistent yet more tempered prognostic structure compared to NWTS-3, with several covariate-specific differences clearly illustrated through the additive hazards framework. Most notably, the intercept trajectory remains relatively flat throughout the follow-up period, and its 95% confidence band consistently overlaps zero, graphically reinforcing its non-significance in both the supremum and Kolmogorov–Smirnov tests and

indicating a stable baseline hazard across time. The cumulative effect of unfavorable histology remains the most prominent, showing a steep upward trajectory with tight confidence bands, confirming a strong, time-varying prognostic role for this factor. In contrast, age-related effects are notably attenuated in this cohort; the curve for Age 2–4 years stays close to zero with wide, overlapping intervals, indicating minimal and unstable influence comparing to Study 3. Tumor diameter categories (10–15 cm and >15 cm) remain flat and tightly clustered around zero, with broad confidence intervals throughout, supporting the statistical findings of no significant association with mortality. Disease stage continues to stratify risk effectively, with Stage 3 and especially Stage 4 exhibiting clear, positive, and upward cumulative functions, though these appear somewhat less steep than in NWTS-3 especially Stage 3, suggesting a possible attenuation of late-stage mortality through improved interventions. Collectively, the shape, separation, and uncertainty of these trajectories not only validate the formal hypothesis tests but also reflect shifts in temporal hazard structures and therapeutic responsiveness between the two NWTS cohorts.

In NWTS-3, the cumulative regression functions for time from relapse to death reveal distinct temporal patterns among covariates, with unfavorable histology and Stage 4 disease showing the steepest and highest cumulative effects, strongly indicating time-varying, adverse impacts on post-relapse survival. These curves rise rapidly within the first two years and remain elevated throughout follow-up, with tight confidence intervals that confirm the robustness of these associations. The intercept term also shows a substantial increase over time, nearly matching the trajectories of unfavorable histology and Stage 4, implying considerable changes in the baseline hazard post-relapse. Stage 3 begins with a steeper slope than Stage 2 in the early phase; however, by year 3, the cumulative hazard for Stage 2 overtakes that of Stage 3 and maintains a consistently higher trajectory thereafter. This crossover suggests that although Stage 3 may be associated with more immediate post-relapse risk, Stage 2 may confer a more delayed but eventually greater hazard, perhaps reflecting heterogeneous subgroups within Stage 2. In contrast, covariates such as age groups and tumor diameter remain flat or even decline slightly, with cumulative

regression curves tightly clustered around zero and enveloped by wide confidence bands, indicating non-significant and unstable effects.

In NWTS-4, a broadly similar structure is evident, but with notable refinements. Unfavorable histology and Stage 4 continue to display the most prominent time-varying influences, with steep cumulative regression curves and narrow confidence intervals that nearly replicate their NWTS-3 counterparts. Stage 3 similarly rises and levels off at a high cumulative risk, though in this cohort, Stage 2 shows virtually no increase over time and stays close to zero, contrasting sharply with its upward trajectory in NWTS-3. This lack of effect in Stage 2 may reflect more effective management of intermediate-stage patients in the later trial era. Age effects in NWTS-4 remain minimal, mirroring NWTS-3, with trajectories for both age groups flat and hovering near zero, supported by wide and overlapping confidence bands. Interestingly, the intercept term increases early but reaches a lower maximum than in NWTS-3, suggesting a more stable or reduced baseline hazard following relapse, potentially a result of improved supportive care or earlier salvage interventions in the NWTS-4 treatment protocols. Together, these patterns emphasize the consistent and dominant roles of histology and advanced stage (particularly Stage 4) across both studies, while also highlighting subtle cohort differences in baseline hazard and the long-term impact of intermediate-stage disease.

Across both NWTS-3 and NWTS-4, cumulative regression function analyses reinforce the dominant and time-varying prognostic roles of unfavorable histology and advanced disease stage, particularly Stage 4, across all three endpoints. These covariates consistently exhibit steep, positive cumulative effects with early and sustained influence on relapse, mortality, and post-relapse survival. Age at diagnosis shows a protective effect in NWTS-3 for time to death, with both younger age groups demonstrating declining hazard trends; however, such patterns are notably attenuated or absent in NWTS-4. Tumor diameter, by contrast, demonstrates flat and near-zero trajectories across both studies and endpoints, suggesting limited clinical utility as a prognostic factor in this setting. Importantly, NWTS-4 reveals a more stable baseline hazard, particularly post-relapses, suggesting improvements in supportive care or therapeutic protocols

between study eras. Overall, while key prognostic factors remain consistent across studies, the evolution of treatment and risk over time is captured in the nuanced temporal dynamics of intermediate-stage disease and baseline hazard structures, underscoring the added value of the additive hazards framework in evaluating changing risk landscapes in pediatric oncology.

Assessment of Time-Varying Covariate Effects in Multi-State Additive Hazards Models

In earlier sections, we evaluated time-to-event outcomes using separate additive hazards models for three endpoints: time to relapse, time to death, and time from relapse to death. While these models provide useful marginal insights into the impact of covariates on each outcome, they do not fully reflect the sequential, clinically dependent structure of disease progression in Wilms tumor. To capture the full complexity of the patient journey, we now transition to a multi-state modeling framework, which offers a more comprehensive and structured approach to event history analysis (Andersen and Keiding, 2002). In this framework, each individual is considered to occupy one of several clinically defined states over time: state 1 (initial remission or "healthy"), state 2 (relapse), and state 3 (death). Patients may follow one of several progression pathways, such as directly transitioning from state 1 to death (1→3) or progressing through relapse before death (1→2→3).

Traditional single-endpoint models, particularly those analyzing time to death, combine all fatal events, regardless of whether relapse occurred beforehand. This aggregation obscures important clinical distinctions and limits interpretability, particularly when relapse status modifies the hazard of mortality. The multi-state model overcomes this by separately modeling each transition-specific hazard: 1→2 (relapse), 1→3 (direct death without relapse), and 2→3 (death after relapse). This decomposition allows for dynamic, stage-specific estimation of covariate effects, accommodates time-varying risks across different phases of illness, and correctly aligns risk sets and event times with the clinical pathways they

represent. Moreover, multi-state models naturally incorporate intermediate events like relapse, which are both clinically meaningful and statistically informative for downstream outcomes.

As emphasized by Andersen and Keiding (2002), multi-state models provide a principled statistical framework for longitudinal and time-to-event disease processes, enabling joint modeling of multiple outcomes with proper handling of competing risks, censoring, and transition dependencies. This is particularly advantageous in pediatric oncology, where treatment response and subsequent outcomes are tightly linked. In this study, adopting a multi-state perspective allows us to move beyond endpoint-specific hazard estimation toward a more integrated, path-based understanding of disease evolution. It also facilitates the estimation of transition probabilities and state occupancy over time, enhancing both prognostic accuracy and clinical utility.

Table 11 presents evidence of time-varying effects for the transition from state 1 (disease-free) to state 2 (relapse). In Study 3, significant violations of the proportional hazards (PH) assumption are observed for histology (sup < 0.001, KS < 0.001), age 2–4 years (sup = 0.006, KS = 0.029), and stage 3 (sup = 0.001, KS = 0.002), suggesting that relapse risk associated with these covariates changes notably over time. Specifically, the strong non-proportionality for histology aligns with its central prognostic role and indicates a dynamic hazard that may be elevated in early follow-up. Stage 3's early relapse risk is also pronounced, while age 2–4 years shows a temporal trend not shared by age >4 years, which exhibits more stable behavior (sup = 0.085, KS = 0.621). Tumor diameter groups show no evidence of time-varying effects ($p > 0.05$ across tests), implying constant risk over time. In Study 4, the PH assumption is again violated for histology (sup < 0.001, KS < 0.001), and now for age >4 years (sup < 0.001, KS < 0.001), marking a shift in age-related relapse dynamics. Additionally, stage 2 displays time-varying effects (sup = 0.001, KS = 0.004), while stage 3 (sup = 0.089, KS = 0.083) and stage 4 (sup = 0.075, KS = 0.047) exhibit more modest, marginal deviations. Age 2–4 years and tumor diameter again appear time-invariant. Comparing studies, Study 3 highlights temporal effects for age 2–4 years and stage 3, whereas Study 4

emphasizes age >4 years and stage 2. In both, unfavorable histology shows persistent non-proportionality, underscoring its evolving influence on relapse risk across time and study settings.

As shown in Table 11 for the direct transition from state 1 (disease-free) to state 3 (death without relapse), both studies provide strong evidence of time-varying effects across several covariates, indicating violations of the proportional hazards assumption. In Study 3, statistically significant non-proportionality is observed for age 2–4 years (sup = 0.025, KS = 0.035), age >4 years (sup = 0.022, KS = 0.004), stage 3 (sup = 0.047, KS = 0.002), stage 4 (sup < 0.001, KS < 0.001), and unfavorable histology (sup < 0.001, KS < 0.001). These results suggest that the hazard of direct mortality from diagnosis varies substantially over time, particularly for older children, advanced-stage disease, and non-favorable histologic subtypes. By contrast, tumor diameter and stage 2 show no significant time-dependent behavior (all $p > 0.05$), indicating stable effects. In Study 4, the extent of non-proportionality is even more pronounced. Both age groups, 2–4 years (sup = 0.001, KS = 0.001) and >4 years (sup = 0.021, KS = 0.007), display clear time-varying effects, as does the intercept itself (sup = 0.004), suggesting baseline hazard instability. Consistent with Study 3, unfavorable histology (sup < 0.001, KS < 0.001), stage 3 (sup = 0.050, KS = 0.003), and stage 4 (sup < 0.001, KS = 0.025) again demonstrate significant deviations from the PH assumption. Tumor diameter remains time-invariant, reinforcing its limited role in predicting early versus late mortality from the initial state. In comparison, Study 4 reveals broader and more statistically robust patterns of non-proportionality, particularly across age and stage groups, possibly reflecting evolving treatment strategies or age-specific vulnerabilities over time. Nonetheless, unfavorable histology and stage 4 consistently emerge as dominant time-varying prognostic factors in both cohorts, underscoring their critical role in shaping mortality risk directly from diagnosis.

Additionally, as detailed in Table 11, the transition from state 2 (relapse) to state 3 (death) reveals markedly different patterns of non-proportionality between Study 3 and Study 4. In Study 3, nearly all covariates demonstrate strong violations of the proportional hazards assumption. Both age 2–4 years (sup = 0.029, KS = 0.001) and age >4 years (sup = 0.015, KS = 0.008) exhibit significant time-varying effects,

suggesting that mortality risk after relapse evolves with age, possibly reflecting age-specific treatment responses or biological behavior post-relapse. Similarly, both tumor diameter categories, ≤10 cm (sup = 0.001, KS < 0.001) and >10 cm (sup = 0.008, KS < 0.001), show time-dependent hazards, indicating that tumor burden at diagnosis continues to impact post-relapse survival in a non-constant manner. Histology (sup < 0.001, KS < 0.001) and all stage categories, stage 2 (sup < 0.001, KS = 0.003), stage 3 (sup < 0.001, KS = 0.001), and stage 4 (sup < 0.001, KS = 0.001), also violate the PH assumption, underscoring the complex and evolving mortality risk after relapse, potentially modulated by disease aggressiveness and initial extent of spread. In contrast, Study 4 shows a more selective pattern of non-proportionality. Only unfavorable histology (sup < 0.001, KS < 0.001), stage 3 (sup = 0.002, KS = 0.005), and stage 4 (sup = 0.001, KS = 0.005) display significant time-varying effects, while age groups, tumor diameter, and stage 2 appear proportional (all p > 0.05), suggesting more stable hazard functions post-relapse. This suggests that in the later study era, mortality following relapse may have become more consistent across demographic and clinical subgroups, possibly due to advances in salvage therapies or more standardized post-relapse management. Comparatively, Study 3 reveals broad, systemic violations of the PH assumption post-relapses, indicating a heterogeneous risk landscape influenced by nearly all patient and tumor characteristics. Study 4, on the other hand, reflects a more concentrated time-varying pattern driven primarily by histologic subtype and advanced stage, highlighting evolving care practices and possibly reduced variability in post-relapse survival trajectories in more recent cohorts.

In Study 3 (top-left panel), the cumulative regression functions for the transition from state 1 (initial diagnosis) to state 2 (relapse), shown in Figure 7 (left panel), reveal clear time-varying patterns that reinforce the hypothesis testing results in Table 11. The intercept term rises sharply within the first two years and then plateaus, indicating a high baseline risk of relapse early in follow-up. This early hazard is tightly bounded, with narrow confidence intervals, highlighting the stability of this estimated effect. Unfavorable histology exhibits the most pronounced time-varying effect: its curve climbs steadily over time with relatively wide but non-overlapping confidence bands, demonstrating both the strength and

uncertainty of its cumulative impact. Stage 3 also displays a substantial positive slope in its regression function, with consistently tight confidence intervals, suggesting a robust and increasing risk of relapse over time. In contrast, Stage 2 remains flat throughout, with a cumulative function centered around zero and narrow bands, indicating no effect. A particularly notable observation is that age 2–4 years exhibits a consistently negative cumulative regression curve, suggesting a protective effect against relapse, with confidence bands that do not cross zero until after year 5. This underscores a meaningful early benefit for children in this age group. Age >4 years, tumor diameter 10–15 cm, and diameter >15 cm all show flat trajectories with wide and overlapping confidence intervals, providing no strong evidence of either constant or time-varying effects on relapse hazard.

In Study 4 (Figure 7, top-right panel), the cumulative regression patterns diverge in meaningful ways from those observed in Study 3, suggesting cohort or treatment-era differences in relapse dynamics. The intercept again rises early but does so more gradually and levels off quickly, indicating a lower and more stable baseline hazard over time, confirmed by wide but symmetric confidence intervals. Unfavorable histology retains a steadily increasing effect with narrow and consistently positive confidence bands, underscoring its persistent influence on relapse risk. A striking change is observed for age >4 years, whose cumulative regression function rises steadily and surpasses zero early in follow-up, accompanied by non-overlapping confidence intervals, confirming a significant and time-varying risk effect not present in Study 3. In contrast, age 2–4 years shows a flat curve near zero with wide confidence bands, consistent with its statistical non-significance. Another key difference is the emerging role of Stage 2, which now displays a weakly positive trajectory that trends upward with confidence bands occasionally drifting away from zero, suggesting a marginally time-varying effect. Stages 3 and 4 continue to exhibit rising cumulative effects, though the slope for Stage 3 is more modest than in Study 3, and Stage 4's trajectory is slightly flatter but still significant, both supported by tight confidence intervals. Tumor diameter groups again remain flat with overlapping confidence intervals. In summary, Figure 7 complements Table 11 by illustrating how time-varying effects evolve distinctly across studies, particularly for age groups and

staging variables, reaffirming the necessity of flexible multi-state additive models to accurately capture evolving relapse risks over time.

In Study 3 (middle-left panel of Figure 7), the cumulative regression functions for the transition from state 1 (initial diagnosis) to state 3 (direct death without prior relapse) illustrate several time-varying covariate effects that align closely with the formal test results in Table 11. Unfavorable histology stands out with a steeply increasing trajectory and narrow confidence intervals, underscoring its strong and persistent additive contribution to the hazard of direct death. Stage 4 similarly displays a consistently rising cumulative effect, beginning early in follow-up and maintaining an upward trend beyond 15 years, with confidence intervals well above zero, indicating a significant and evolving risk associated with advanced disease. Stage 3 also shows a modest but positive slope, with its confidence band largely above the zero line, suggesting a time-varying impact of intermediate-stage disease. Age >4 years exhibits a distinct pattern: the cumulative function rises quickly early on but begins to flatten, indicating a strong initial effect that diminishes over time. In contrast, age 2–4 years demonstrates a mostly flat or slightly negative trajectory, with its curve hovering near or below zero, suggesting a protective influence for this age group. These age-specific effects mirror the test results, which showed time-varying significance for both age categories. Meanwhile, both tumor diameter categories (10–15 cm and >15 cm) remain close to zero across time, with wide and overlapping confidence intervals, providing no indication of either a significant or time-dependent role in direct mortality. The intercept term rises modestly before stabilizing, indicating a moderate but relatively constant baseline hazard. Together, these patterns reinforce the importance of histology, stage, and age in shaping dynamic mortality risk from diagnosis, as captured through the additive hazards framework.

In Study 4 (Figure 7, middle-right panel), the cumulative regression functions for transition 1→3 (direct death without prior relapse) show widespread but more moderate time-varying effects across several covariates compared to Study 3, reinforcing the supremum and K–S test findings in Table 11. Unfavorable histology again emerges as the most dominant covariate, with a steeply increasing curve and

tight confidence intervals, reflecting a consistently elevated risk over time. Stage 4 also shows a sustained upward trajectory that accelerates in later follow-up, though widening confidence intervals suggest increasing uncertainty at longer times. Stage 3 presents a more modest, yet consistent, positive trend. The age covariates both show declining cumulative effects, with age >4 years decreasing more sharply than age 2–4 years early on, though both trajectories are flatter than in Study 3. This suggests that while age-related mortality risk remains time-varying, the magnitude of its effect has lessened in the later cohort. Notably, the tumor diameter covariates, which were flat in Study 3, now exhibit slight downward trends, indicating weak but potentially time-varying protective effects, though confidence intervals remain wide and overlapping. The intercept function rises a bit in Study 4 than in Study 3 during early follow-up, indicating a more dynamic baseline hazard. Compared to Study 3, where the age covariate effects declined more sharply and diameters remained completely flat, Study 4 reveals a more distributed pattern of moderate time-varying influence across clinical covariates, suggesting that temporal risk dynamics in NWTS-4 may reflect changes in mortality patterns, possibly due to improvements in early treatment or shifts in cohort characteristics.

In Study 3 (Figure 7, bottom-left panel), the cumulative regression functions for the transition from relapse to death (2→3) reveal strong time-varying effects across nearly all covariates. The intercept exhibits a sharp and persistent rise, reaching above 2.0, indicating an increasing baseline hazard over follow-up, with narrow confidence intervals that reflect high precision. Among the clinical variables, unfavorable histology, Stage 4, and Stage 3 show steep, sustained increases, suggesting their dominant additive contributions to post-relapse mortality. Notably, Stage 2, often assumed to carry lower risk, shows a similarly strong and increasing cumulative effect, nearly overlapping with the curves for Stages 3 and 4, implying a convergence in risk once relapse occurs. In contrast, age groups and tumor diameter groups demonstrate clear negative cumulative regression functions. Specifically, both age 2–4 years and age >4 years show declining trajectories, with age >4 years decreasing more steeply toward the negative axis, though to a lesser extent than diameter groups. The most striking decline is seen in Diameter >15

cm, which sharply drops below –2.5, while Diameter 10–15 cm follows a similar, though slightly less extreme, downward pattern. These suggest potential protective effects or time-varying interactions with treatment efficacy or disease trajectory post-relapse. All of these trends are supported by the statistically significant supremum and Kolmogorov–Smirnov test results for nearly every covariate in Table 11, highlighting pervasive non-proportionality in Study 3.

In Study 4 (Figure 7, bottom-right panel), the covariate effects in the 2→3 transition are notably more stable and attenuated compared to Study 3. The intercept remains essentially flat around zero with wide, overlapping confidence bands, indicating no significant time-varying baseline hazard. While Stage 4, Stage 3, and unfavorable histology retain positive and increasing cumulative functions, their trajectories are gentler, with broader confidence intervals than in Study 3, reflecting both weaker effects and increased uncertainty. Stage 2, in particular, diverges significantly from its pattern in Study 3, showing only a slight upward slope and remaining well below Stage 3 and 4 curves, reaffirming its relatively lower risk in NWTS-4. Importantly, both age groups now remain nearly flat around zero, with minimal time-dependent patterns and wide confidence bands, suggesting a lack of dynamic age-related effect post-relapses. Similarly, tumor diameter groups, which previously showed pronounced downward slopes, now also stay flat with overlapping confidence intervals, reinforcing their limited role in NWTS-4 mortality dynamics. This visual contrast with Study 3 highlights substantial differences in post-relapse risk architecture between the cohorts. The relatively stable effects in Study 4 may reflect improvements in salvage therapy, refined risk stratification, or changes in supportive care, and these findings further corroborate the covariate-specific PH test results reported in Table 11. Taken together, the comparative analysis of Figure 7 underscores how additive hazards modeling captures the evolution and attenuation of covariate effects in the shifting therapeutic landscape across NWTS-3 and NWTS-4.

The additive hazards multi-state modeling framework offers a powerful and interpretable approach to evaluating the evolving impact of prognostic factors across clinically meaningful transitions in Wilms tumor progression. Compared to traditional methods that assess covariate effects at a small number of

fixed time points, the cumulative regression function plots (Figure 7) provide a richer, more granular depiction of time-varying effects, allowing us to trace how covariate influence accumulates or wanes throughout follow-up. Formal supremum and Kolmogorov–Smirnov test statistics (Table 11) validate these patterns and confirm the presence of temporal heterogeneity in several key covariates, particularly in NWTS-3.

In the transition from diagnosis to relapse (state 1→2), NWTS-3 displays dynamic and often nonlinear effects, with clear increases for unfavorable histology and stage 3, and a marked protective signal for age 2–4 years. In contrast, NWTS-4 exhibits subtler dynamics, with moderate increases for age >4 years and stage 2. For direct death without relapse (1→3), histology and stage 4 emerge as dominant, time-varying risks in both studies, though NWTS-3 shows sharper increases for age >4 and a steeper intercept, while NWTS-4 reflects more stable covariate effects over time. The transition from relapse to death (2→3) further reveals NWTS-3's high temporal heterogeneity: stage 2 aligns with stages 3, 4, and the intercept in a sharp upward trend, while age and tumor diameter decline steeply which indicates differential hazard accumulation and possible treatment response. By contrast, NWTS-4 presents a flatter hazard landscape, with nearly all covariate effects remaining stable and close to zero.

Altogether, these results reinforce the value of continuous-time modeling via additive hazards in capturing nuanced and time-varying prognostic behavior, which would be obscured under conventional three-timepoint comparisons. The contrast between NWTS-3 and NWTS-4 also highlights how treatment era and protocol evolution can shape not only absolute risks but also the dynamics by which those risks accumulate, underscoring the importance of flexible survival modeling in longitudinal oncology research.

Dynamic Predictions and Covariate-Driven Heterogeneity in Multi-State Progression: Transition Probabilities, Occupancy, and Incidence Under Additive Hazards Modeling

Table 11 presents risk set sizes and cumulative event counts at clinically informative time points for each of the three transitions, 1→2 (relapse), 1→3 (direct death without relapse), and 2→3 (death after relapse),

under the multi-state framework, stratified by Study 3 and Study 4. These empirical counts provide the structural basis for transition probability estimation, state occupancy modeling, and cumulative incidence curves in subsequent analyses, and they reflect the temporal unfolding of events within each study cohort.

In Study 3, the risk set for transition 1→2 (relapse) begins at 1,484 individuals at year 1 and decreases steadily: 1,395 at year 2, 1,354 at year 3, 1,301 at year 5, and 1,229 by year 8. By year 10, 1,158 remain at risk, dropping to 817 at year 15 and reaching 0 by year 25. Correspondingly, relapse events accumulate rapidly in early follow-up, from 133 at year 1 to 199 (year 2), 219 (year 3), and 230 (year 5). The curve flattens thereafter, reaching 234 (year 8), 237 (year 15), and ultimately 238 total relapses. This suggests that the majority of relapses occur within the first 5–8 years. For transition 1→3, the number at risk declines from 1,577 at year 1 to 1,461 (year 3), 1,395 (year 5), and 1,241 (year 10), with a steep decline to 873 by year 15 and full depletion by year 25. Direct deaths increase from 43 (year 1) to 54 (year 5), reaching 58 by year 10 and plateauing at 69 by year 25, indicating that direct mortality without prior relapses is rare and unfolds slowly over extended follow-up. For transition 2→3, the risk set grows slightly before declining, starting at 93 (year 1), peaking at 115 (year 2), then dropping to 107 (year 3), 94 (year 5), and just 83 by year 10, reflecting the entry of new relapses. Cumulative deaths rose sharply from 40 at year 1 to 84 (year 2), 111 (year 3), 134 (year 5), and 148 by year 15, ultimately reaching 151. These patterns highlight a high fatality rate following relapse, with most deaths occurring within 10 years post-relapses.

In Study 4, the transition dynamics follow similar shapes but at a higher volume and faster pace. For transition 1→2, the initial risk set is larger (2,024 at year 1), consistent with broader enrollment. It declines to 1,891 (year 2), 1,800 (year 3), 1,623 (year 5), and only 746 remains at risk by year 10, plummeting to 6 at year 15 and reaching 0 by year 20. Events accrue more rapidly than in Study 3, with 164 relapses already by year 1, 250 by year 2, 288 by year 3, and a maximum of 301 by year 10, indicating that nearly all relapses occur within the first decade. For transition 1→3, the risk set decreases from 2,158 (year 1) to 2,049 (year 2), 1,918 (year 3), 1,755 (year 5), and 800 by year 10. Deaths accrue

slowly, 32 (year 1), 44 (year 3), 53 (year 5), and 57 (year 10), reaching 61 by year 25, similar to Study 3. For transition 2→3, the risk set begins at 134 (year 1), expands to 158 (years 2–3), and then declines to 132 (year 5), 83 (year 10), and 0 by year 25. Post-relapse deaths show a sharp rise: 28 (year 1), 86 (year 2), 121 (year 3), 142 (year 5), and 163 by year 15, suggesting an even more concentrated period of mortality post-relapse than in Study 3.

Comparing Studies 3 and 4, several distinctions emerge. First, Study 4 consistently has larger risk sets due to greater sample size, but events also occur earlier and accumulate more rapidly. Relapse (1→2) events are nearly completed by year 10 in Study 4 (301 vs. 234 in Study 3 at year 10), suggesting more aggressive disease detection or earlier failure of therapy. Despite this, direct deaths (1→3) are similarly rare in both studies and show slow accumulation. The 2→3 transition (death after relapse) is more compressed in Study 4, with 163 deaths by year 15 (vs. 148 in Study 3), but most occur before year 10, indicating a tighter window for post-relapse mortality. In contrast, Study 3 exhibits a more gradual event distribution, especially for 1→2 and 2→3 transitions, suggesting longer follow-up and slower event processes. These differences shape the foundation for transition probabilities, occupancy estimates, and cumulative incidence curves in Figures 8–11 and reflect temporal shifts in treatment protocols, relapse management, and follow-up practices between NWTS-3 and NWTS-4.

The selected covariate profiles summarized in Table 12 serve as the core clinical scenarios for downstream estimation of transition probabilities, state occupancy curves, and cumulative incidence functions, as visualized in Figures 10 and 11 and reported numerically in Table 13. These profiles represent carefully constructed combinations of patient-level characteristics, age group, tumor diameter, disease stage, and histologic subtype, designed to capture a clinically meaningful spectrum of risk severity in the context of Wilms tumor progression.

A total of 10 covariate profiles are ranked by increasing prognostic severity. The first profile, serving as the reference group, includes patients under 2 years old, with tumor diameter <10 cm, stage 1 disease, and

favorable histology, reflecting the lowest-risk configuration. Subsequent profiles introduce stepwise increases in clinical burden across multiple dimensions. For example, Profile 2 represents a mild elevation in risk, age 2–4 years, diameter 10–15 cm, stage 2, with favorable histology. Profile 3 considers patients older than 4 years with large tumors (>15 cm) but otherwise favorable histological features, while Profile 5 introduces both stage 4 disease and large tumor size, though histology remains favorable.

More aggressive risk configurations are represented in Profiles 6 through 10, which all involve unfavorable histology. For instance, Profile 6 includes a patient with older age (>4 years) and unfavorable histology, but low-stage and mid-size tumor, reflecting a mixed-risk profile. Profile 8 balances small tumor size (<10 cm) with stage 3 disease and unfavorable histology, illustrating the complex interplay of anatomical and histopathologic risk. The most severe case, Profile 10, combines oldest age group, largest tumor diameter, highest disease stage (4), and unfavorable histology, representing the poorest prognosis across all modeled transitions.

These profiles were chosen not only for their clinical relevance but also for their statistical representativeness across the NWTS cohort, and they allow us to quantify risk across realistic patient archetypes. They provide the structure for Figures 10–11 (cumulative incidence), Figures 8–9 (state occupancy), and Table 13 (transition probabilities), enabling direct comparison of time-dependent outcome probabilities across varying levels of clinical severity in both Study 3 and Study 4.

State occupancy probabilities quantify the time-varying likelihood that an individual resides in a particular clinical state, typically “event-free,” “relapsed,” or “dead”, conditional on having entered the process at diagnosis. Within the theoretical framework of multi-state models, these probabilities are computed by integrating the transition hazards over all possible paths leading into each state, capturing both direct and indirect transitions across a finite state space. This construction yields a probabilistic characterization of patient histories that goes beyond standard survival or cumulative incidence functions, which only marginalize over time-to-first event. Instead, state occupancy probabilities offer a complete

joint description of the stochastic process, accounting for both the occurrence and temporal ordering of intermediate events. From a modeling standpoint, they reflect solutions to Kolmogorov forward equations or their additive hazard analogs, depending on the assumed semi-Markovian or non-Markovian structure. Clinically, they provide interpretable quantities, such as the probability that a patient is alive and relapse-free 10 years post-diagnosis, enabling communication of long-term risk trajectories in ways that are directly meaningful for patients, clinicians, and policymakers.

Figures 8 and 9 present estimated state occupancy probabilities over a 25-year time horizon for ten representative covariate profiles (defined in Table 12), separately for NWTS Study 3 and Study 4. These figures illustrate the temporal dynamics of disease progression under varied combinations of baseline prognostic factors. Crucially, they reveal how the accumulation of adverse covariate levels, through their impact on transition intensities, reshapes the long-term distribution of patients across clinical states. For instance, favorable profiles maintain high event-free occupancy for decades, while unfavorable profiles experience rapid early transitions into relapse and death. By stratifying these curves across discrete covariate scenarios and study cohorts, the figures enable detailed comparative insight into how clinical and biological risk factors alter the likelihood of cure, the window of vulnerability to relapse, and the residual risk of late mortality. Furthermore, they allow for temporal disaggregation of competing events, permitting identification of critical periods during which patients are most at risk of clinical deterioration. As such, the state occupancy framework provides a statistically rigorous yet clinically interpretable mechanism for translating multi-state hazard estimates into longitudinal prognostic expectations.

Across Figure 8, representing Study 3, state occupancy trajectories clearly differentiate prognostic profiles in a logical and clinically interpretable progression. In Profiles 1 through 5, all defined by favorable histology, the healthy state probability dominates throughout follow-up but degrades progressively with increasing disease burden. Profile 1 exhibits excellent outcomes, with healthy occupancy remaining above 90% through year 25, relapse probability peaking below 5%, and virtually no mortality, consistent with early-stage, low-volume, biologically indolent disease. Profile 2 maintains a

high healthy probability, modestly declining to ~85% by year 25, while relapse increases to ~10% and death remains below 5%. In Profile 3, healthy state stays above 80% by year 25, relapse rises to ~15%, and death reaches ~10%, reflecting greater risk associated with older age and larger tumor size. Profile 4 shows faster attrition from the healthy state falling below 75% while relapse and death both reach ~20%. Profile 5 marks the highest stage among favorable histology patients, with healthy occupancy declining sharply to ~60% by year 25; relapse peaks at 20% and death approaches 22%, underscoring the cumulative adverse effects of increasing stage even in favorable histology settings.

Profiles 6 through 10 in Study 3 include patients with unfavorable histology and show consistently poorer outcomes, with steep early declines in healthy state occupancy and high death probabilities. In Profile 6, the healthy probability plunges below 70% by year 10, relapse rises to ~20%, and death exceeds 25% by year 25, highlighting the dominant effect of histology, even in low-stage disease. Profile 7 shows a further decline: despite small tumor size, healthy probability falls below 60% by year 10, and death exceeds relapse within 5 years, ultimately reaching over 20% similarly, lower because the age is younger. In Profile 8, advanced stage drives aggressive dynamics: relapse reaches ~25% and death approaches 30% by year 25, with healthy state occupancy falling under 40% by year 10. Profile 9 demonstrates rapid deterioration despite patient age under 2, with healthy probability dropping to ~30% by year 5 and death surpassing 35% by year 25, emphasizing how advanced stage and tumor size outweigh age as prognostic factors. Finally, Profile 10 reflects the worst-case scenario, with healthy occupancy collapsing below 20% by year 5, relapse peaking early, and death surpassing 40% well before year 10. This panel vividly illustrates the compounded risk from all adverse covariate levels.

In sum, the state occupancy probabilities displayed in Figure 8 for Study 3 delineate a clear, clinically interpretable stratification of long-term outcomes across covariate profiles, capturing both the pace and magnitude of relapse and mortality. Profiles with favorable histology (Profiles 1–5) maintain dominant event-free probabilities over time, with stage, tumor size, and age incrementally eroding prognosis in a graded manner. In contrast, Profiles 6–10, defined by unfavorable histology, show steep early declines in

remission and sharply rising death probabilities, even in cases with otherwise limited disease burden, underscoring the dominant prognostic role of tumor biology. Importantly, older age emerges as a potent modifier of risk in this high-risk group: despite shared unfavorable histology, older patients (e.g., Profile 8) exhibit markedly faster deterioration and higher mortality than their younger counterparts (e.g., Profiles 6 and 7), even when tumor size or stage is modest. This pattern suggests an additive and potentially synergistic, interaction between age and poor histology, likely reflecting increased treatment resistance or biological aggressiveness with age. The early convergence of relapse and death curves in these higher-risk profiles further highlights limited salvageability. Together, these findings emphasize the heterogeneous and nonlinear risk landscape of Wilms tumor in Study 3 and demonstrate the value of state occupancy modeling in capturing clinically actionable prognostic dynamics across patient subgroups.

Study 4 (Figure 9) reveals broadly similar patterns of state occupancy probabilities across the ten covariate profiles but demonstrates consistent, clinically meaningful improvements in survival and relapse control compared to Study 3, particularly among lower-risk patients. In Profile 1, the event-free state remains above 90% across the full 25-year horizon, with relapse and death remaining under 5%, mirroring the trajectory in Study 3 but with slightly more prolonged stability. Profile 2 shows a healthy state plateauing around 85%, with relapse below 8% and death under 5%, modestly improving upon its Study 3 counterpart. In Profile 3, relapse rises to approximately 15–18% and death to around 10% by year 25, but transitions are delayed, particularly for mortality, suggesting extended survival following initial relapse. Profile 4 also shows a marginally more favorable trajectory, with healthy occupancy remaining around 80–85% and a slower increase in relapse and death, indicating more durable remission despite moderate risk. Profile 5 continues to exhibit early relapse, reaching 20% by year 5, but the mortality curve climbs more gradually, peaking at around 12–13%, suggesting potential gains in post-relapse survival.

For profiles with unfavorable histology (6–10), Study 4 offers limited but discernible improvements. In Profile 6, the healthy state remains over 50% at year 10, slightly higher than in Study 3, and death approaches 30% by year 25 with a flatter rise, pointing to potentially improved salvage rates. Profile 7

retains the steep early loss in healthy status and rising mortality seen in Study 3, indicating that unfavorable histology continues to dominate prognosis regardless of modest changes in treatment era. Profile 8, representing an older patient with poor histology but moderate stage and size, shows slightly earlier relapse and mortality curves than in Study 3, yet still reflects the survival benefit of younger age relative to higher-risk profiles. Profile 9 displays rapid early deterioration in healthy state (falling below 30% by year 5), with death exceeding 30% by year 25, comparable to Study 3 but with somewhat earlier transitions. Finally, Profile 10, the worst-risk configuration, exhibits rapid decline in the event-free state and death surpassing 30% before year 10, closely matching the trajectory in Study 3 and reaffirming the persistent lethality of combined high-risk factors. Overall, while the general structure of risk stratification remains stable between studies, Study 4 demonstrates modest survival gains across most profiles, underscoring therapeutic improvements over time, especially for patients with favorable histology and early-stage disease.

Covariate impacts in Study 4 largely mirror those observed in Study 3, but clear improvements in long-term outcomes are evident across all profiles. For favorable histology profiles (1–5), the healthy state is more persistently maintained, with death probabilities consistently lower and relapse curves slightly attenuated, reflecting gains likely attributable to refined risk-adapted therapies, enhanced treatment protocols, or improved supportive care. Among unfavorable histology profiles (6–10), although overall mortality remains substantial, several profiles exhibit delayed transitions to adverse states. Notably, Profile 6 shows higher healthy occupancy (approximately 50% at year 10) and a more gradual rise in death probability compared to Study 3, suggesting incremental benefit for patients with early-stage, poor histology disease. Similarly, Profile 8 demonstrates a modest delay in death trajectory, reinforcing the protective effect of younger age even in biologically aggressive settings. While Profiles 9 and 10 continue to show steep deterioration and early mortality, their trajectories in Study 4 are slightly less abrupt, indicating marginal but consistent improvements even in the most challenging prognostic groups. Overall,

the across-the-board survival gains in Study 4 underscore meaningful clinical progress, particularly for patients with intermediate risk, and highlight the value of treatment refinements over time.

Across both Study 3 and Study 4, histology remains the single most influential determinant of long-term clinical trajectories, with unfavorable histology consistently associated with sharply reduced event-free survival, earlier relapse, and higher mortality across all covariate combinations. Stage at diagnosis also exerts a strong gradient of risk, particularly within the favorable histology group, where increasing stage leads to stepwise declines in healthy state occupancy and earlier onset of adverse events. Age at diagnosis further modifies these effects, especially when histology is unfavorable, as older age amplifies mortality risk even in the presence of limited tumor burden or early stage. By contrast, tumor diameter shows minimal independent prognostic effect once histology and stage are accounted for, suggesting its contribution to risk stratification is likely mediated through these stronger covariates. Comparing the two studies, Study 4 demonstrates modest but consistent improvements in survival outcomes across nearly all profiles, particularly in earlier-stage disease and among patients with favorable histology. These improvements likely reflect protocol evolution, more precise risk stratification, and better supportive care. Together, the findings highlight both the enduring dominance of biological aggressiveness in shaping Wilms tumor outcomes and the capacity for incremental therapeutic refinements to shift population-level survival curves, even among high-risk groups. This reinforces the clinical importance of early and accurate risk classification and underscores the need for continued innovation in treating the most biologically aggressive disease subtypes.

Table 13 reports the estimated transition probabilities under the three-state illness-death model, specifically, the probabilities of remaining event-free (State 1), experiencing relapse but surviving (State 2), or having died (State 3), for a set of ten covariate profiles of clinical interest. Estimates are presented at multiple landmark time points (1, 2, 3, 5, 8, 10, 15, and 20 years post-diagnosis) and stratified by study cohort (NWTS-3 vs. NWTS-4). These time-specific probabilities represent marginal state occupancy

values, computed from the fitted multi-state additive hazards model by integrating transition intensities over time across all admissible paths.

This table complements the state occupancy probability plots (Figures 8 and 9) by translating continuous state probability trajectories into discrete, interpretable numerical summaries at clinically relevant intervals. While the plots offer a dynamic, time-continuous visualization of patient state evolution, Table 13 provides concrete benchmarks that clinicians and trialists may more readily apply, such as estimating the proportion of patients in remission, relapsed, or deceased at 5, 10, or 15 years for a given prognostic profile. This is particularly useful in pediatric oncology, where risk communication, long-term survivorship planning, and stratified follow-up protocols often hinge on absolute risk levels at fixed horizons.

From a theoretical standpoint, these transition probabilities provide an alternative but equally informative lens into the structure of the underlying multi-state process. Unlike traditional survival analysis outputs that focus on hazard ratios or marginal survival probabilities, multi-state transition probabilities explicitly differentiate *which failure pathway* a patient is likely to follow. They quantify the cumulative influence of sequential transitions (e.g., from event-free to relapse, and then to death) as well as direct transitions from remission to death, offering more nuanced clinical insight, particularly into the timing and burden of intermediate events like relapse.

By presenting these estimates across a range of covariate profiles, Table 13 also facilitates transparent comparison of prognostic trajectories and study-level differences. The explicit timing allows us to detect both early- and late-phase divergences in survival outcomes, and to assess the effectiveness of protocol modifications over time. For instance, improvements in long-term survival for patients with unfavorable histology or advanced stage disease can be precisely quantified in terms of delayed transitions to relapse or death. Thus, this table serves as both a numerical distillation of the multi-state modeling results and a

practical tool for informing long-term care strategies, patient counseling, and evidence-based updates to clinical guidelines.

In Study 3, estimated transition probabilities for selected covariate profiles provide a granular, time-specific view of disease progression in children diagnosed with Wilms tumor, decomposing survival into remission, relapse, and post-relapse death states. Among the five profiles defined by favorable histology, long-term event-free survival remains high, relapses relatively rare, and direct mortality minimal. For instance, Profile 1 (Stage I, age ≤ 4 years, tumor diameter <10 cm) demonstrates exemplary remission durability: the probability of remaining in remission (1→1) is 88.8% at year 1, 84.6% at year 3, and 71.4% at year 10, with low cumulative relapse risk (1→2: 7.0% at year 1, 13.4% at year 10) and minimal direct mortality (1→3: under 4.2% by year 10). Profile 2 (Stage II, age ≤ 4, tumor 10–15 cm) follows a similar pattern, maintaining 82.8% in remission at year 3 and 69.6% at year 10, with a 1→2 risk of 13.6% and a 1→3 risk of only 2.3% by year 10. Profile 3 (Stage I, age > 4, tumor 10–15 cm) reflects the modest adverse effect of older age and larger tumors: remission remains favorable (1→1: 74.3% at year 5), though relapse increases slightly (1→2: 19.8% at year 5). Profile 4 (Stage III, age > 4, tumor 10–15 cm) shows more relapse burden, 1→1 declines to 59.9% by year 5, while 1→2 increases to 26.7% and 1→3 reaches 11.6%. The most adverse of the favorable histology profiles, Profile 5 (Stage IV, age ≤ 4, tumor 10–15 cm), exhibits accelerated transition out of remission: 1→1 drops to 54.9% by year 5 and 35.3% by year 15, while 1→2 grows to 36.7%. Despite these differences, a unifying theme among all favorable histology profiles is that relapse is the critical tipping point: post-relapse survival is poor, with 2→3 transitions approaching 99% by year 20, and 2→2 (alive in relapse) probabilities rarely exceeding 0.3%.

In contrast, profiles with unfavorable histology (Profiles 6–10) demonstrate markedly more aggressive disease trajectories, with high early relapse rates and substantially elevated direct mortality from the initial remission state. Profile 6 (Stage I, age > 4, tumor <10 cm) highlights how unfavorable histology can override otherwise good clinical prognostic factors. By year 3, the remission probability drops to 69.6%, 1→2 rises to 25.3%, and 1→3 climbs to 4.9%, with 1→1 falling further to 42.5% by year 10.

Profiles 7 and 8 (Stage II–III, age > 4, tumor 10–15 cm) continue this unfavorable trajectory, with Profile 8 reaching 35.4% 1→2 transition and 13.9% direct mortality (1→3) by year 5. Profile 9 (Stage IV, age ≤ 4, tumor >10 cm) exhibits extreme early failure: remission drops to 64.6% by year 1 and direct death rises to 21.5%, with 2→3 exceeding 98% by year 5. Profile 10 (Stage IV, age > 4, tumor >10 cm) displays slower initial transition but rapidly escalating relapse and death: 1→2 climbs to 34.5% by year 5, and 1→3 reaches 23.6% by year 10. In all unfavorable profiles, post-relapse survival is virtually absent, with 2→3 nearly 100% beyond year 3 and 2→2 rarely exceeding 0.2%.

Taken together, the contrast between profiles reinforces the central role of histology as the dominant prognostic determinant. At 10 years, favorable histology profiles retain 53.7%–71.4% probability of remission (1→1), while unfavorable profiles often fall below 45%. Similarly, relapse probabilities (1→2) range from 13.4%–36.7% under favorable histology but exceed 30% in most unfavorable profiles, with direct death (1→3) reaching up to 23.6% in the latter group. The additive hazards framework offers significant theoretical advantages in parsing these dynamics, allowing covariate effects on transition hazards to evolve over time and uncovering subtleties in long-term risk, such as how small increases in 1→2 can dramatically affect overall survival due to near-certain post-relapse mortality. Notably, tumor size (within the 10–15 cm range) exerts only modest influence, particularly when histology is known, and age at diagnosis (>4 years) appears to exacerbate poor outcomes primarily in the presence of aggressive tumor biology. These results emphasize that therapeutic strategies must prioritize relapse prevention, especially for patients with unfavorable histology, where even early-stage or small tumors provide limited protection against long-term mortality.

In Study 4, transition probabilities across favorable histology profiles (Profiles 1–5) exhibit generally high long-term remission rates and low mortality, though with greater variation than in Study 3. Profile 1 (Stage I, age <2 years, tumor <10 cm) shows excellent outcomes: 1→1 remains at 94.5% at year 5 and 92.8% even at year 20, with minimal direct death and a relapse-to-death (2→3) rate below 44% by year 10. Profile 2 (Stage II, age 2–4, tumor 10–15 cm) reveals higher early relapse risk, 1→2 reaches 21.7% by

year 10, and gradual erosion in remission to 78.3%. Profile 3 (Stage II, age >4, tumor >15 cm) maintains high 1→1 early (88.8% at year 5), but relapse rises steadily (1→2: 14.2% at year 10), and post-relapse death accelerates, with 2→3 reaching 36.2% by year 10. Profile 4 (Stage III, age <2, tumor >15 cm) sees faster transitions to relapse (1→2: 7.1% by year 5), and unlike prior favorable profiles, a sustained post-relapse population persists (2→2: ~21.6% at year 20), although mortality (2→3) remains high (~78%). Profile 5 (Stage IV, age 2–4, tumor >15 cm) has the most aggressive course among the favorable group, with remission dropping to 66.7% by year 15, relapse rising (1→2: 18.1%), and the highest long-term death risk among these profiles (1→3: 16.6% by year 20; 2→3: >82%).

Unfavorable histology profiles (Profiles 6–10) show rapid decline in remission and steep increases in both relapse and direct mortality. Profile 6 (Stage I, age >4, tumor 10–15 cm) starts with reasonable remission (82% at year 1), but by year 5, only 67.4% remain in remission, and 27.6% have relapsed. Profile 7 (Stage II, age >4, tumor <10 cm) deteriorates more quickly: 1→1 is only 60.5% at year 5, with relapse nearing 31% and 1→3 death exceeding 8%. Profile 8 (Stage III, age 2–4, tumor <10 cm) shows further weakening, by year 5, remission is just 57.8%, relapse 32.1%, and death 10.1%, with post-relapse mortality at 96.1%. Profile 9 (Stage IV, age <2, tumor 10–15 cm) shows the sharpest early failure: 1→1 declines to 53.5% by year 5, 1→3 rises to 19.1%, and relapse is nearly uniformly fatal (2→3: 97.9%). Profile 10 (Stage IV, age >4, tumor >15 cm) has similarly poor outlook: 1→1 drops below 55% by year 5, and cumulative risk of death (direct or post-relapse) exceeds 39% by year 10, with 2→3 exceeding 97%.

Comparing covariates across profiles reinforces key implications seen in Study 3. Favorable histology remains the dominant protective factor, preserving long-term remission across a range of ages, stages, and tumor sizes. Within favorable histology, tumor size >15 cm and advanced stage (III/IV) slightly reduce remission probability and increase relapse, but the effect is modest compared to histologic differences. Age <2 years does not uniformly confer benefit, some younger patients (e.g., Profile 4) still experience substantial relapse and death. In contrast, unfavorable histology overwhelms otherwise mild covariates: Profiles 6–10 include patients with early-stage disease or small tumors, yet exhibit high relapse and rapid

mortality, especially once relapse occurs. Notably, Study 4 suggests somewhat lower post-relapse mortality than Study 3, particularly for some favorable histology profiles where 2→2 occupancy remains non-negligible even at 10–20 years. This may reflect evolving treatment approaches, improved salvage therapies, or cohort differences, although the near inevitability of death following relapse in most unfavorable profiles remains a consistent and sobering finding across both studies.

Building directly upon the nuanced findings from state occupancy probabilities, the transition probability analysis offers a sharper, event-focused lens on when and how these differences manifest, particularly across Study 3 and Study 4. While occupancy curves describe the net accumulation of time in each state, transition probabilities decompose the risk dynamics at each decision point, highlighting the *rate* at which patients leave remission, relapse, or succumb to disease. These results reinforce and clarify the previous findings: in Study 4, favorable histology profiles exhibit consistently lower 1→2 (remission to relapse) and 1→3 (remission to death) probabilities across all time points, suggesting more effective suppression of early progression compared to Study 3. Importantly, transition intensities confirm that older age, particularly age >4 years, amplifies the likelihood of early transition to death, a trend not fully apparent from state occupancy results alone. Furthermore, while occupancy plots showed only modest variation across tumor sizes, the transition analysis more clearly isolates tumor diameter as a *non-significant contributor* once histology and stage are accounted for. In profiles with unfavorable histology, transition probabilities confirm the stark and nearly uniform rise in 2→3 (relapse to death) transitions within 1–3 years post-relapse, emphasizing that the post-relapse window remains a high-risk, low-survival phase regardless of other covariates. Interestingly, Study 4 shows slightly attenuated 2→3 transitions in certain favorable histology profiles, aligning with the modest survival gains observed in state occupancy but now pinpointed to delayed or reduced mortality after relapse. Thus, while state occupancy curves highlight broad improvements in remission duration and survival, transition probabilities illuminate where along the clinical trajectory these improvements originate, offering a mechanistic understanding of therapeutic impact and identifying covariate patterns most associated with favorable or refractory disease dynamics.

To complement the insights gained from state occupancy probabilities and transition probabilities, Figures 10 and 11 present cumulative incidence functions (CIFs) for key transitions in Studies 3 and 4, respectively. Estimated from the multi-state additive hazards model, these CIFs quantify the cumulative probability of each specific transition occurring over time, for example, from remission to relapse (1→2), remission to death without relapse (1→3), or relapse to death (2→3). Each panel corresponds to a selected covariate profile, reflecting combinations of age, tumor size, stage, and histology representative of the clinical heterogeneity in Wilms tumor.

While state occupancy probabilities summarize the proportion of patients in each state at a given time, and transition probabilities characterize the instantaneous risk of moving between states, CIFs address a distinct and clinically interpretable quantity: the absolute risk of experiencing a particular event by a given time, properly accounting for the presence of competing events. For example, a patient cannot relapse after dying directly from remission; the CIF for relapse must therefore be adjusted for the competing risk of early death. This cumulative view provides a time-integrated perspective on event risk, enriching the picture provided by more instantaneous or state-based summaries.

In this framework, the CIFs serve two main purposes. First, they sharpen the interpretation of how covariate effects such as histology or stage translate into event-specific risks over time, e.g., by showing that patients with unfavorable histology not only relapse more often but do so earlier. Second, CIFs offer a clinically grounded summary that aligns with common prognostic questions, such as the likelihood of relapse within five years or the cumulative probability of death after relapse. By examining these CIFs alongside the transition and occupancy probabilities, we can disentangle whether improved survival in Study 4, for example, reflects reduced relapse incidence, delayed timing of transitions, or diminished post-relapse mortality risk.

Together, Figures 10 and 11 allow for a more comprehensive understanding of disease progression and treatment impact across studies, anchoring theoretical insights in cumulative, interpretable quantities that directly inform prognosis and clinical decision-making.

Favorable histology profiles of Figure 10 of Study 3 (Profiles 1–5) collectively illustrate a consistent but graduated risk structure governed by patient age, tumor size, and disease stage. Profile 1, young age (<2 years), small tumor (<10 cm), and stage I, represents the most favorable constellation. Here, relapse (1→2, green) remains exceptionally rare, staying below 10% throughout 15 years of follow-up, and direct mortality without relapse (1→3, orange) is virtually absent until a negligible rise after year 15, likely attributable to rare late effects such as treatment-induced secondary malignancies, cardiotoxicity, or unrelated causes of death. Once relapse occurs, however, the 2→3 transition (purple) displays a sharp surge, reaching nearly 80% within a short time window. This rapid escalation, coupled with a brief mid-curve dip, suggests high fatality post-relapse and reflects potential numerical fluctuations due to sparse relapse events and censoring-related instabilities in the Aalen-Johansen estimator. In Profile 2 (age 2–4, 10–15 cm tumor, stage II), relapse incidence modestly increases to ~13%, and direct mortality remains low with a small late rise. The 2→3 curve again rises steeply after relapse and displays a minor non-monotonic feature around year 10, likely due to a few censored relapsed patients who were alive at last contact, creating localized estimator variability. In Profile 3 (age >4, tumor >15 cm, stage II), cumulative relapse risk increases to ~20%, while 1→3 direct mortality being stable around only 2%. The 2→3 curve retains its steepness, though with a slightly more gradual slope, potentially reflecting longer post-relapse survival among older children, or variation in salvage treatment responsiveness. A visible dip mid-curve persists, consistent with earlier profiles, again reflecting finite-sample behavior rather than clinical reversal. Profile 4 (age <2, tumor >15 cm, stage III) combines a favorable age with higher disease burden. Relapse incidence climbs to ~26%, while direct death reaches ~10%. The 2→3 curve shows a steep initial increase followed by a shallow plateau, possibly indicating transient success with salvage protocols, or reflective of staggered relapse timing and a sparsely populated risk set over time. Finally, Profile 5, age 2–

4, tumor >15 cm, and stage IV, marks the most severe among favorable histology groups, with relapse approaching 28% and direct mortality ~20%. The post-relapse fatality remains substantial, as seen in the steep 2→3 curve, but again a minor mid-curve decline emerges. Across all favorable histology profiles, these downward fluctuations in the 2→3 curves are minor and best interpreted as artifacts of the nonparametric estimation process under censoring and competing risks. They do not imply clinical improvement in survival after relapse but reflect limitations in observed data density within the relapsed strata. In summary, although patients with favorable histology demonstrate relatively good long-term outcomes, relapse, once it occurs, is uniformly associated with high lethality, and small estimator dips should be interpreted cautiously within the framework of finite-sample multi-state survival analysis.

Unfavorable histology profiles (Profiles 6–10) consistently reveal starkly elevated risks across all transitions, with histology dominating the prognosis even in early-stage or otherwise favorable clinical settings. In Profile 6 (age >4, tumor 10–15 cm, stage I), relapse (1→2) reaches ~35% within five years, and direct mortality (1→3) begins early, reaching ~12–14% by year 10. The 2→3 curve rises almost vertically post-relapse, surpassing 90%, highlighting the devastating lethality of relapse in unfavorable histology, even with stage I disease. Profile 7 (age >4, <10 cm, stage II) shows a similar relapse risk (~38%) and slightly higher direct mortality, demonstrating that increasing stage modestly worsens prognosis, but histology remains the primary driver. In Profile 8 (age 2–4, <10 cm, stage III), the cumulative incidence of relapse increases further (~40%), and direct death exceeds 15%, while the 2→3 transition occurs almost immediately after relapse with little delay, underscoring the rapid progression and limited efficacy of salvage therapy. Profile 9 (age <2, tumor 10–15 cm, stage IV) shows accelerated transitions: relapse surpasses 40% within five years, and direct death begins early. Despite young age, a favorable factor in other settings, the presence of metastasis and unfavorable histology leads to poor outcomes, and the 2→3 curve behaves like a step function, again approaching 1 rapidly. Finally, Profile 10 (age >4, >15 cm, stage IV) represents the worst-case scenario, with all adverse covariates compounding risk. Here, 1→2 exceeds 45% by year 10, 1→3 climbs above 25%, and the 2→3 transition

curve surges almost immediately to 1. The visual pattern of all three CIFs, each elevated, left-shifted, and steep, clearly reflects the cumulative and possibly synergistic burden of older age, larger tumor size, advanced stage, and unfavorable histology. Across these five profiles, the pattern is unequivocal: unfavorable histology leads to high relapse rates, substantial early direct mortality, and an overwhelmingly lethal course following relapse, with minimal temporal separation between relapse and death.

Across the cumulative incidence functions (CIFs) shown in Figure 10 of Study 3, each covariate, histology, stage, tumor size, and age, exerts a distinct and interpretable effect on relapse (1→2), direct death (1→3), and post-relapse mortality (2→3). Most prominently, histology acts as the dominant determinant of prognosis: profiles with unfavorable histology (Profiles 6–10) exhibit markedly elevated CIFs for all transitions, regardless of stage or age. Even in early-stage disease (e.g., Profile 6), relapse exceeds 35% and direct death surpasses 12% within ten years, while the 2→3 transition approaches 100% shortly after relapse, confirming poor salvage outcomes. In contrast, favorable histology profiles (Profiles 1–5) maintain substantially lower risks, with 1→2 typically under 30% and 1→3 often negligible across 15 years, though 2→3 curves remain steep, indicating high lethality after relapse across histology types. Stage contributes incrementally: as stage advances from I to IV within a histological stratum, both 1→2 and 1→3 CIFs rise consistently. For example, among favorable histology profiles, relapse rises from ~10% (Profile 1, Stage I) to ~28% (Profile 5, Stage IV), and direct death climbs from nearly 0% to ~20%, confirming that higher stage amplifies both relapse risk and competing mortality. Tumor size shows a similar additive effect; larger tumors correlate with higher cumulative incidence of relapse and direct death. Within-stage comparisons (e.g., Profiles 2 vs. 3 or 6 vs. 10), typically for relapses, demonstrate that tumor diameter >15 cm worsens prognosis, especially when paired with advanced stage or unfavorable histology. Finally, age exhibits nuanced effects: in favorable histology, older age often predicts higher relapse and mortality (Profile 3), though in some cases younger age appears protective. However, under unfavorable histology, the protective effect of young age disappears, as in Profile 9, where infants with

stage IV disease and intermediate tumor size face >40% relapse and high direct death within five years. Overall, the CIF curves illustrate how histology exerts the strongest prognostic influence, but each additional adverse factor, higher stage, larger tumor, or older age, exerts cumulative or synergistic effects, particularly under poor histologic conditions.

In Study 4, favorable histology profiles (Profiles 1–5) exhibit a gradation of risk shaped by age, tumor size, and disease stage. Profile 1 (Age <2, Diameter <10 cm, Stage I) demonstrates a low-risk trajectory: relapse incidence (1→2) climbs gradually to about 5% by year 10 and then flattens, while direct death without relapse (1→3) remains below 4%, with only a slight rise near the end of follow-up. Death after relapse (2→3) rises steeply, reaching approximately 38% by year 11, confirming that post-relapse survival remains compromised even in this low-risk group. Profile 2 (Age 2–4, Diameter 10–15 cm, Stage II) reflects higher baseline risk, with 1→2 reaching 12–13% by year 10, and 1→3 increasing slowly to 5%. The 2→3 curve escalates rapidly to around 55–60%, consistent with poor salvage outcomes following relapse. In Profile 3 (Age >4, Diameter >15 cm, Stage II), 1→2 climbs further to ~18%, while 1→3 stays under 3% but displays a non-monotonic decline, which is a statistical artifact due to censoring and sparse events, not an actual reversal in risk. The 2→3 cumulative incidence grows more gradually than in prior profiles, reaching ~35% by year 10, suggesting relatively longer post-relapse survival in this older cohort. Profile 4 (Age <2, Diameter >15 cm, Stage III) shows a sharp early rise in 1→2, plateauing near 26–27% by year 3. The 1→3 curve rises to ~6% early and then flattens. Notably, the 2→3 transition climbs steeply to ~60% by year 3 and approaches 70% by year 10, highlighting the severe risk after relapse. Finally, Profile 5 (Age 2–4, Diameter >15 cm, Stage IV) presents the most severe risk pattern among favorable histology groups. Relapse (1→2) approaches 30%, and direct death (1→3) shows a steady and steeper rise to ~15%. The 2→3 transition surges early, reaching ~70% by year 3 and nearly 80% by year 10, highest among favorable histology profiles, emphasizing that even favorable tumor biology cannot offset the prognostic burden of advanced disease stage and large tumor size.

In Study 4, unfavorable histology profiles (Profiles 6–10) consistently demonstrate markedly elevated risks across all transitions, emphasizing the dominant prognostic role of histology even in early-stage disease. Profile 6 (Age >4, Diameter 10–15 cm, Stage I) shows relapse cumulative incidence (1→2) exceeding 35%, with early direct death (1→3) climbing to ~10%. Once the relapse occurs, the 2→3 transition rises steeply, surpassing 85% by year 5, indicating poor salvage despite low-stage disease. Profile 7 (Age >4, Diameter <10 cm, Stage II) displays a similar relapse rate of ~38%, with the 1→3 curve rising above 10%. The 2→3 function again climbs rapidly, approaching 90%, affirming high lethality post-relapses. In Profile 8 (Age 2–4, Diameter <10 cm, Stage III), relapse probability reaches ~30%, and direct mortality (1→3) approaches 13%. The 2→3 curve surges past 90% almost immediately after relapse, underscoring the aggressive progression and limited effectiveness of salvage therapy in this high-risk group. Profile 9 (Age <2, Diameter 10–15 cm, Stage IV) exhibits early relapse reaching ~40%, with direct mortality rising sharply beyond 20% by year 15. The 2→3 curve climbs rapidly to ~95%, indicating near-certain mortality following relapses, despite the young age of the patient. Finally, Profile 10 (Age >4, Diameter >15 cm, Stage IV) represents the most extreme risk configuration. Here, relapse (1→2) exceeds 40%, direct death (1→3) rises above 23%, and the 2→3 cumulative incidence approaches 95% within just a few years, reflecting the additive, and likely synergistic, impact of all adverse prognostic factors. These profiles collectively reveal that unfavorable histology overwhelmingly determines relapse risk, direct mortality, and poor post-relapse survival, regardless of other covariate levels.

Across both NWTS-3 and NWTS-4, histology remains the most influential prognostic factor, with unfavorable histology profiles uniformly exhibiting markedly higher cumulative incidence of relapse (1→2), direct death without relapse (1→3), and especially post-relapse mortality (2→3), often exceeding 90% regardless of other covariates. However, NWTS-3 consistently shows greater absolute risks across all transitions and profiles than NWTS-4, highlighting temporal improvements in disease management. For favorable histology, relapse risks (1→2) are generally below 20% in NWTS-4 but often rise above

25% in NWTS-3 for higher-stage or older patients. Direct death (1→3) under favorable histology rarely exceeds 10% in NWTS-4 but is 5–10 percentage points higher in NWTS-3 across matching profiles. Notably, stage amplifies risk gradients in both studies: transitions 1→2 and 1→3 increase substantially from stage I to IV, with NWTS-3 showing steeper increments. For example, in NWTS-3, advancing from stage I to IV under favorable histology increases relapse from ~8% to ~28%, and direct death from <5% to ~20%; NWTS-4 shows a similar directional trend but with attenuated magnitudes (~5% to ~30% for relapse, ~3% to ~15% for death). Interestingly, tumor diameter shows a non-intuitive but consistent protective pattern in both studies: smaller tumors (<5 cm) are associated with higher risks of relapse and death than larger ones (>15 cm), particularly visible in profiles such as 8 vs. 9. This suggests that tumor size alone, when adjusted for stage and histology, may not linearly increase risk and could reflect underlying biological or treatment-related confounding. Age effects are more nuanced: in favorable histology, younger age (<2) is protective in both studies, with the lowest relapse and death rates seen in Profile 1, while older age (>4) correlates with higher risks; in unfavorable histology, however, the protective effect of young age disappears, and all age groups show high 2→3 mortality exceeding 85–95%, especially in NWTS-3. Additionally, NWTS-3 often exhibits more abrupt rises and mid-curve irregularities, especially in 2→3 transitions, compared to smoother, more stable CIFs in NWTS-4, likely due to differences in sample size, censoring, or estimation precision. In summary, although both studies validate the same broad risk hierarchy, histology > stage > age, with more complex, potentially protective diameter effects, NWTS-3 consistently presents higher cumulative risks, sharper CIF gradients, and less favorable outlooks across all covariate patterns, underscoring advances in treatment efficacy and study design reflected in NWTS-4.

In conclusion, while state occupancy and transition probability summaries provide valuable insights into the dynamic evolution of patient status over time, the cumulative incidence functions (CIFs) uniquely reveal the absolute, transition-specific risks critical for clinical interpretation. The CIFs expose nuanced patterns not readily captured by state probabilities alone, most notably, the sharply elevated post-relapse

mortality (2→3) in unfavorable histology across both studies, the attenuated relapse and direct death risks in NWTS-4 relative to NWTS-3, and the non-monotonic, potentially protective effect of tumor diameter. Unlike occupancy curves, which aggregate over transitions, or transition matrices, which reflect conditional movement at fixed time points, CIFs directly quantify the cumulative burden of each competing risk process, providing interpretable probabilities essential for prognosis. Thus, the CIF analysis not only confirms the overarching risk hierarchy across histology, stage, age, and diameter but also illustrates temporal improvements in survival outcomes from NWTS-3 to NWTS-4, underscoring its indispensable role in multi-state modeling of pediatric oncology cohorts.

Discussion

Summary of Results

Initial descriptive analysis of the NWTS-3 and NWTS-4 cohorts, as well as the combined sample, revealed notable temporal and clinical distinctions that shaped our subsequent modeling strategy. NWTS-3 exhibited a higher frequency of relapse and substantially worse post-relapse survival compared to NWTS-4, patterns that align with historical improvements in treatment protocols and risk stratification implemented in later studies. While unfavorable histology was less common overall, it was disproportionately represented among patients experiencing adverse outcomes in both cohorts, underscoring its persistent prognostic relevance. Tumor diameter and age at diagnosis displayed strong nonlinear associations with survival outcomes, prompting categorization into clinically meaningful strata: tumor diameter was split at <10 cm, 10–15 cm, and >15 cm, and age at diagnosis was grouped into <2 years, 2–4 years, and >4 years. Stage and centrally reviewed histology were retained as categorical predictors due to their clinical interpretability and robust statistical association with outcomes.

Model selection decisions were guided by a combination of theoretical considerations and empirical fit, including AIC comparisons and assessment of model diagnostics across endpoints. Consistently lower AIC values for stratified models suggested that separate analyses by cohort (rather than a pooled

approach) were warranted. This allowed for more accurate estimation of covariate effects and better accounted for evolving clinical practices between NWTS-3 and NWTS-4. The choice to model tumor diameter and age categorically was further supported by residual-based diagnostics, which revealed significant departures from linearity on the continuous scale. These foundational choices ensured that all subsequent models, whether Cox, AFT, or additive hazards, reflected both the biological complexity and clinical heterogeneity of Wilms tumor across eras of treatment.

Subsequent global tests based on Schoenfeld residuals revealed systematic violations of the proportional hazards (PH) assumption in all three endpoints, with especially severe departures for histology and stage. Visual diagnostics, including log(-log) Kaplan–Meier plots and time-varying coefficient estimates, supported these findings and further indicated that covariate effects varied substantially over time especially for time to relapse and time to death. In light of these violations, we adopted an accelerated failure time (AFT) framework, which directly models the log survival time and allows covariates to scale survival multiplicatively without requiring proportional hazards. Model selection proceeded via AIC and residual analysis tailored to each endpoint. For time to relapse, the log-normal distribution provided the best overall fit, outperforming both Weibull and gamma alternatives. For time to death, a gamma distribution yielded the most favorable diagnostics, while time from relapse to death was best modeled using a log-normal AFT model. Despite violations of AFT model assumptions, key prognostic signals were consistently identified. Unfavorable histology emerged as the strongest predictor across all endpoints and both cohorts, with children experiencing markedly shortened times to relapse, death, and post-relapse survival, underscoring the aggressive nature of this subtype. Tumor stage also exhibited a consistent gradient effect, particularly for Stage 4 disease, which was associated with significantly worse outcomes across all clinical phases. Age at diagnosis showed more nuanced effects: children aged 2–4 years demonstrated improved survival in NWTS-3 and modest benefit in NWTS-4, though patterns were less stable for older age groups. Tumor size yielded inconsistent associations, with an apparent survival advantage post-relapses in NWTS-3 not reproduced in NWTS-4. These findings, though informative,

remain constrained by distributional misfit and nonproportional effects, motivating the transition to additive hazards models that more flexibly accommodate time-varying risks and cohort-specific trajectories.

To more flexibly account for the time-varying nature of covariate effects observed in AFT model diagnostics and proportional hazards violations, we subsequently fit Aalen's additive hazards models separately for NWTS-3 and NWTS-4. This semiparametric framework permits covariate effects to evolve continuously over time, offering a robust alternative for modeling hazard structures when both AFT and PH assumptions are violated. Formal inference on the temporal behavior of covariates was conducted using two nonparametric tests tailored to the additive hazards setting: the supremum test, which assesses whether a covariate has any non-zero effect over time, and the Kolmogorov–Smirnov (K–S) test, which evaluates whether that effect is constant or varies with time. Covariates yielding p-values below 0.05 in both tests are interpreted as having statistically significant and time-varying impacts on the hazard function.

Across all three clinical endpoints, time to relapse, time to death, and time from relapse to death, unfavorable histology and advanced stage (particularly Stage 4) consistently emerged as dominant and dynamically evolving risk factors in both studies. Their effects were reflected in steep cumulative regression functions and supported by significant results from both the supremum and K–S tests. Age at diagnosis exhibited shifting prognostic roles across cohorts: in NWTS-3, children aged 2–4 years showed time-varying protective effects on early outcomes, whereas in NWTS-4, older age (>4 years) was more predictive of post-relapse mortality. Tumor diameter remained statistically non-significant throughout, with flat cumulative functions and consistently non-rejected null hypotheses, underscoring its limited prognostic relevance. Notably, baseline hazard functions (intercepts) were more volatile in NWTS-3, especially following relapses, while NWTS-4 displayed more stable hazard patterns, possibly reflecting evolving treatment protocols or improved supportive care. Together, these additive hazards results confirm and extend the AFT findings, revealing both persistent and evolving risk structures over time and

highlighting the clinical value of modeling non-proportional, time-varying covariate effects in long-term Wilms tumor outcomes.

While the additive hazards models fitted separately to time to relapse, time to death, and time from relapse to death allowed for flexible modeling of time-varying covariate effects, they inherently treated each endpoint in isolation, ignoring the dependent structure of clinical disease progression. These models estimated marginal hazards without explicitly accounting for the order or interdependence of events, particularly the distinction between deaths preceded by relapse and those occurring without relapse. As such, the model for time to death could not differentiate whether mortality arose from aggressive disease post-relapse or from early failure unrelated to relapse, and the time from relapse to death model could only be evaluated conditional on having relapsed, excluding patients who died first. By contrast, the multi-state additive hazards framework directly incorporated the illness-death structure by modeling transition-specific hazards for relapse (1→2), direct death without relapse (1→3), and death following relapse (2→3). This approach allowed for a decomposition of mortality pathways and offered a more clinically interpretable and statistically coherent picture of how covariate effects evolve dynamically across disease stages.

The differences in interpretation were most striking for histology, stage, age, and tumor size. While unfavorable histology showed strong and time-varying effects across all three single-outcome models, the multi-state approach revealed heterogeneity by transition: its effect was modest for relapse (1→2), markedly stronger for direct death (1→3), and most severe for post-relapse death (2→3), where the cumulative regression function increased steeply over time in both NWTS-3 and NWTS-4. Tumor stage likewise exhibited divergent behavior: endpoint models suggested a broad stage gradient for mortality, but multi-state analysis clarified that this was driven almost entirely by increased direct death (1→3) in Stage 4 patients, while post-relapse mortality (2→3) showed relatively attenuated stage differences. Age effects, which appeared equivocal in marginal models, were localized in the multi-state setting to reduced relapse risk (1→2) in younger children, especially in NWTS-3, whereas post-relapse mortality (2→3) was less

favorable or even worse in this group. Tumor size showed weak or inconsistent effects across separate endpoints but emerged in the multi-state model as a strong predictor of early relapse in NWTS-4 and a paradoxical survival benefit after relapse in NWTS-3, possibly reflecting differential treatment response in patients with larger localized tumors. Together, these findings underscore how single-endpoint analyses, even when accounting for time-varying effects, obscure important dependencies between failure types. Multi-state additive hazards modeling, by capturing the full event process structure, provided a refined and clinically actionable understanding of covariate influence across the disease trajectory.

A key advantage of the multi-state modeling framework over traditional endpoint-specific survival analysis is its ability to capture the full temporal dynamics of disease progression, accounting not only for the occurrence of individual events but also for their ordering and interdependence over time. By jointly estimating transition probabilities, cumulative incidence functions (CIFs), and state occupancy probabilities, this approach yields a comprehensive probabilistic characterization of patient trajectories across multiple clinically relevant states, event-free, relapsed, and deceased. In contrast to models focused solely on time to first event, multi-state methods incorporate intermediate events such as relapsing and modeling their downstream consequences, offering nuanced insight into long-term outcomes. To operationalize this framework, we defined ten clinically representative covariate profiles spanning a broad spectrum of prognostic severity based on combinations of histology, disease stage, age at diagnosis, and tumor diameter, from the most favorable to the most adverse risk configurations. These profiles serve as the foundation for Figures 8–11 and Table 13, enabling direct comparison of dynamic outcome probabilities across both NWTS-3 and NWTS-4. The state occupancy curves reveal how patient risk accumulates and redistributes across disease states over time, while transition probabilities and CIFs offer complementary perspectives on the likelihood and timing of specific events. Together, these tools provide a flexible, interpretable, and clinically meaningful framework for assessing long-term risk, highlighting the value of multi-state modeling in characterizing complex disease processes like Wilms tumor.

The state occupancy probability curves in Figures 8 and 9 reveal consistent time-dependent effects of histology, stage, age at diagnosis, and tumor diameter on disease progression, with meaningful contrasts between NWTS-3 and NWTS-4. Across both studies, histology is the most influential factor, with unfavorable histology leading to markedly earlier relapse and higher long-term mortality. However, in NWTS-4, the adverse impact of unfavorable histology appears slightly attenuated, suggesting modest therapeutic improvements over time. Stage exhibits a strong risk gradient independent of histology, especially among favorable histology patients; yet, transitions occur more gradually in NWTS-3, whereas NWTS-4 shows earlier clustering of relapse and post-relapse deaths, indicating tighter clinical follow-up and potentially earlier detection. Age at diagnosis also differentiates risk: older children tend to exit the event-free state more rapidly in both studies, but the temporal spread of these transitions is broader in NWTS-3, consistent with slower disease kinetics and longer post-relapse survival. Tumor diameter shows a comparatively weaker and less consistent effect across both studies, with limited prognostic influence after adjusting for histology, stage, and age. Together, these results demonstrate how multi-state modeling surpasses endpoint-specific approaches by capturing the full trajectory of disease progression and enabling time-resolved comparisons of how prognostic factors operate across treatment eras.

The contrast between Study 3 and Study 4 reveals both enduring and evolving effects of histology, stage, age at diagnosis, and tumor diameter on Wilms tumor progression, with transition probabilities providing new clinical insights beyond those offered by state occupancy curves. While both studies reaffirm histology as the primary determinant of prognosis, with unfavorable histology profiles showing consistently higher relapse (1→2) and death (1→3, 2→3) rates, the transition probability framework in Study 4 reveals important shifts in the timing and magnitude of these risks. In particular, favorable histology patients in Study 4 exhibit delayed transitions out of remission and reduced post-relapse mortality (2→3), suggesting incremental gains in long-term control and salvage efficacy not fully captured by occupancy-based summaries. For example, while occupancy curves showed durable remission in both studies, transition probabilities clarify that the reduction in 2→3 transitions is where

survival gains are concentrated, an effect especially pronounced in younger patients with advanced-stage, favorable histology disease. Conversely, for older children with unfavorable histology, Study 4 confirms persistently high early transitions to relapse and death, mirroring Study 3 and reinforcing their status as a high-risk subgroup, despite protocol evolution. Notably, tumor diameter, which appeared weakly prognostic in state occupancy results, shows minimal incremental contribution in transition probabilities once histology and stage are controlled, underscoring the dominance of tumor biology over anatomical burden. Finally, transition estimates more clearly demonstrate how age >4 years accelerates progression, particularly through higher 1→3 and 2→3 hazards, a nuance less apparent in cumulative state occupancy. Altogether, the transition probability analysis not only refines our understanding of where survival improvements have occurred between studies but also clarifies which covariate effects remain stable over time versus those that may be modifiable, thereby offering a richer, time-localized interpretation of risk that directly informs clinical monitoring and treatment adaptation.

Cumulative incidence functions (CIFs) provide a critical extension to multi-state survival analysis by directly quantifying absolute, event-specific risks over time in the presence of competing risks, offering clinical interpretability that state occupancy probabilities and transition intensities cannot. Whereas occupancy probabilities describe the distribution of patients across disease states and transition probabilities reflect instantaneous hazards, CIFs answer the clinically relevant question of how likely a specific event (e.g., relapse, death) is to occur by a given time, accounting for competing outcomes such as dying before relapse. Compared to NWTS-3, CIFs from NWTS-4 consistently show lower cumulative risks across transitions, particularly for relapse (1→2) and direct death without relapse (1→3) in favorable histology profiles, highlighting therapeutic advances over time; for instance, relapse rarely exceeds 20% in NWTS-4 but often surpasses 25% in NWTS-3, and direct death is generally under 10% in NWTS-4 versus 15–20% in NWTS-3 for similar profiles. Post-relapse mortality (2→3) remains steep in both studies, often exceeding 85–90%, but CIFs in NWTS-4 show slower early escalation, suggesting modest gains in salvage therapy. CIFs also sharpen interpretation of covariate effects: histology is the

dominant factor, with unfavorable histology profiles showing high relapse (>35%), elevated direct death (20–25%), and near-certain post-relapse mortality (>90%) regardless of other characteristics. Stage further stratifies risk within histologic groups, relapse and direct death increase from Stage I to IV, with more attenuated gradients in NWTS-4. Tumor diameter effects are complex: although large tumors are generally riskier, some profiles (e.g., with smaller tumors but high-stage disease) show higher CIFs, implying potential confounding. Age appears protective in favorable histology (especially <2 years) but loses prognostic value under unfavorable histology, where high post-relapse mortality prevails across all age groups. Importantly, CIFs also reveal estimation irregularities, such as mid-curve dips in 2→3 transitions, driven by censoring and small relapse counts in NWTS-3, underscoring limitations of nonparametric estimation in sparse strata. Ultimately, CIFs enhance the interpretive depth of multi-state models by providing cumulative, interpretable, and event-specific probabilities that clarify the interaction of covariates with timing and burden of clinical events, reveal temporal improvements from NWTS-3 to NWTS-4, and support nuanced prognostication beyond what transition probabilities and state occupancy alone can offer.

Clinical Interpretation

Across both NWTS-3 and NWTS-4, histology remained the strongest prognostic determinant, with unfavorable histology associated with markedly increased risks of relapse, direct mortality, and death following relapse. Cumulative incidence functions (CIFs) revealed that patients with unfavorable histology in NWTS-3 had relapse probabilities often exceeding 35%, direct death probabilities reaching 20–25%, and near-certain post-relapse mortality (>90%) by 8–10 years. While NWTS-4 showed similar qualitative patterns, the cumulative risks were attenuated, relapse rarely exceeded 25%, and direct death was often under 10% for comparable profiles, suggesting therapeutic gains in the more recent cohort. Multi-state transition probabilities and additive hazards estimates confirmed this improvement, showing delayed and lower hazard rates for transitions out of remission in NWTS-4. However, even in NWTS-4, post-relapse mortality remained steep and time-dependent, indicating that despite advancements in front-

line therapies, salvage regimens for relapsed disease remain insufficient, particularly for unfavorable histology patients.

Tumor stage also exhibited a clear and persistent prognostic gradient, with Stage IV patients facing the highest cumulative incidence and transition probabilities for all event types. This effect was especially pronounced in the 1→3 (direct death) transition, where multi-state modeling revealed that the stage-related mortality risk was largely driven by early deaths rather than relapse-mediated pathways. CIF comparisons between cohorts showed that NWTS-4 patients with advanced-stage disease had modestly reduced direct death risks compared to NWTS-3, indicating improved supportive care or staging accuracy over time. Age at diagnosis showed cohort-specific differences: in NWTS-3, children aged 2–4 years experienced a protective effect against early relapse and death, while in NWTS-4, this protective pattern shifted toward children under 2 years. Conversely, older children (>4 years) exhibited elevated transition intensities, particularly for 1→3 and 2→3 in NWTS-4, reflecting either age-related biological aggression or treatment resistance. These shifting age patterns across cohorts underscore the need for age-tailored treatment refinement, especially for older children with high-risk histologic or stage profiles.

Tumor diameter displayed the weakest and most inconsistent prognostic influence, but its interpretation varied across cohorts and transitions. In NWTS-3, large tumors (>15 cm) were associated with an apparent survival advantage after relapse, a paradox not replicated in NWTS-4, where larger tumors predicted earlier relapse (1→2) but had limited downstream mortality effects. These findings may reflect cohort-specific interactions between tumor bulk and response to therapy, potentially confounded by differences in surgical or chemotherapy regimens. Importantly, the broader cohort comparison across all covariates revealed that NWTS-4 patients, overall, experienced longer durations in the remission state, lower transition probabilities to relapse or death, and flatter cumulative regression functions for most risk factors. This suggests a general pattern of improved disease control, especially for favorable histology and lower-stage tumors. However, high-risk subgroups, particularly older children with unfavorable histology

and advanced stage, continued to fare poorly in both studies, reinforcing the need for intensified risk-adapted therapy and closer post-treatment monitoring in future protocols.

Comparison to Prior Literature

Our findings reinforce the critical prognostic role of histology and stage in Wilms tumor, consistent with both cooperative group studies (e.g., Kulich & Lin, 2004; Green et al., 1998) and institutional analyses (Honeyman et al., 2012). Across NWTS-3 and NWTS-4, children with unfavorable histology exhibited markedly elevated cumulative incidence across all transitions, relapse (1→2), death without relapse (1→3), and especially post-relapse death (2→3),reflecting substantially worse survival trajectories. These effects were most severe in NWTS-3, where post-relapse mortality curves rose sharply, underscoring limited salvage effectiveness in earlier treatment eras. Higher disease stage (III–IV) similarly portended increased transition hazards, particularly in NWTS-4, where later-stage patients experienced both higher relapse probability and elevated death risk, paralleling findings by both Honeyman et al. and prior NWTS reports. Tumor diameter, while showing monotonic effects in univariate models, did not emerge as a consistent transition-specific risk factor in our multi-state analysis, diverging from Honeyman et al.'s institutional findings and suggesting that tumor size may carry variable prognostic weight in multicenter trial populations. Age at diagnosis showed nuanced patterns, with infants (<1 year) displaying distinct relapse and mortality curves from older children, reaffirming previously reported non-monotonic age effects in Wilms tumor. Finally, our additive hazards modeling revealed substantial heterogeneity in time-varying effects, particularly in NWTS-3, supporting the move away from proportional hazards frameworks and toward dynamic, transition-aware modeling strategies that better reflect individualized clinical course.

By employing a multi-state framework with additive hazards modeling, we were able to uncover dynamic and clinically informative transition patterns not readily apparent from traditional survival endpoints alone. State occupancy probabilities provided an integrative view of patient distribution across health

states (no relapse, relapse, death), showing that unfavorable profiles quickly transitioned out of remission, while favorable-risk patients remained relapse-free with high probability well beyond five years. Transition-specific cumulative incidence functions (CIFs) added further granularity, revealing competing risks and temporal shifts: for instance, in NWTS-4, the risk of direct death without relapse (1→3) remained consistently low, while the relapse-related mortality pathway (1→2→3) was the dominant terminal trajectory for high-risk patients. Compared to earlier analyses such as Kulich & Lin (2004), which focused solely on relapse-free survival using full cohort or case-cohort Cox models, our multi-state estimates delineate the sequence and timing of events, allowing for more actionable clinical interpretation. Whereas prior studies emphasized parameter estimation efficiency (e.g., via CDW estimators), our focus on covariate effects across multiple transitions and states reveal that histology, stage, age, and tumor diameter not only influence the likelihood of events but also their temporal spacing and cumulative burden. These layered insights offer a more precise foundation for individualized surveillance and treatment planning, moving beyond static survival summaries to clinically relevant prognostication across disease trajectories.

A central theme emerging from our multi-state modeling is the striking cohort difference between NWTS-3 and NWTS-4, which echoes historical improvements in risk stratification, treatment protocols, and supportive care reported by the National Wilms Tumor Study Group (D'Angio et al., 1989; Green et al., 1998). Patients enrolled in NWTS-3 experienced substantially higher relapse rates and notably worse survival following relapse, particularly those with high-risk features such as unfavorable histology or stage IV disease, than their NWTS-4 counterparts. These patterns are consistent with the longitudinal findings described by Breslow et al. (2006), who emphasized the clinical gains achieved through protocol refinements, including adjusted chemotherapy dosing, radiation targeting, and biologically informed risk grouping implemented after NWTS-3. In contrast, NWTS-4 patients exhibited delayed and reduced incidence of relapse, along with improved post-relapse survival, suggesting the emergence of more effective salvage strategies and intensified surveillance in later eras. While both cohorts reaffirmed the

prognostic significance of histology and stage, the attenuation of transition-specific risks in NWTS-4, particularly the shallower post-relapse mortality curves, underscores the benefits of therapeutic evolution. These cohort-specific dynamics lend quantitative support to the evolving standard-of-care hypothesis articulated in earlier NWTSG publications and also reinforce similar conclusions drawn by Honeyman et al. (2012), who reported that outcome disparities often reflect institutional and temporal variability in treatment access and clinical decision-making. Our findings therefore highlight the necessity of era-stratified modeling in pediatric oncology research and caution against the extrapolation of risk estimates across historical cohorts.

Limitations

While the multi-state additive hazards model used in this study offers substantial advantages, particularly its ability to capture time-varying effects and competing risks, it does not produce individualized, covariate-specific estimates of transition probabilities or state occupancy. This limits its utility for clinical decision-making, where precise risk predictions for patient profiles (e.g., a 4-year-old with stage II, favorable histology, and 8 cm tumor diameter) are often needed. In contrast, approaches like that of Kulich & Lin (2004), which utilize case-cohort sampling and the CDW estimator, offer more actionable, personalized estimates of cumulative incidence. Moreover, while the additive hazards framework enables clear decomposition of covariate effects over time, the interpretation of these effects, especially through cumulative hazard plots, is less intuitive than hazard ratios or survival probabilities, potentially limiting its accessibility for clinicians or non-statistical stakeholders.

The use of NWTS-3 and NWTS-4 data provides a robust basis for internal comparisons across eras, but the absence of external validation poses a clear limitation. Findings may not generalize to later cohorts, such as those treated in NWTS-5 (Green et al., 1998) or contemporary COG AREN trials (Fernandez et al., 2017), where molecular markers and more intensive risk stratification guide therapy. Additionally, while the NWTS data are of high quality, key variables such as histologic classification and disease stage

may still be affected by misclassification or inter-center variability. Importantly, unmeasured prognostic markers, such as loss of heterozygosity at 1p/16q or WT1 mutations, were not available in these historical datasets, despite their recognized importance in more recent risk models. Tumor diameter, although included as a continuous covariate, showed no strong or consistent effect across transitions. This may reflect residual measurement error, heterogeneity across centers, or possibly unmodeled nonlinear effects and interactions.

Some limitations are methodological in nature. The additive hazards model imposes linearity and additivity on covariate effects, which may oversimplify more complex or interactive relationships among prognostic factors (Lin and Kulich, 2004). Although the model accommodates time-varying effects, it does not capture threshold or non-monotonic patterns unless explicitly modeled. A further concern involves the analysis of transition 2→3 (death after relapse), which is inherently conditional on having experienced a relapse. This introduces a form of survivorship bias, as patients who survive longer without relapse may differ systematically from those who relapse early, leading to biased post-relapse mortality estimates. More sophisticated methods, such as joint modeling of transitions or landmark analyses, might better address this dependency structure in future work.

Finally, some limitations arise from sparse data and statistical uncertainty in estimating rare transitions. Particularly in NWTS-4, late events, such as death after long remission or after late relapse, were infrequent, leading to small risk sets and unstable estimates in later follow-up periods. These limitations were evident in the visual behavior of several cumulative incidence functions and transition probability curves, which sometimes exhibited non-monotonic jumps or irregularities. Such behavior, especially for low-frequency transitions like 1→3 or 2→3, likely reflects sampling variability and sparse events rather than true clinical heterogeneity. While the additive hazards model remains theoretically sound in these settings, the resulting plots can be difficult to interpret and may obscure meaningful contrasts across covariate profiles.

Our analysis revealed limited and inconsistent prognostic influence of tumor diameter across all three transitions, relapse, direct death without relapse, and death after relapse, in both NWTS-3 and NWTS-4, contrasting with findings from institutional series that have identified tumor size as a significant risk factor (e.g., Honeyman et al., 2012). Several factors may underlie this discrepancy. First, the NWTS cohorts represent large, multicenter trial populations, and the resulting heterogeneity may dilute localized associations that are more apparent in single-institution studies with uniform imaging and staging protocols. Second, our modeling framework treated tumor diameter as a linear or discretized covariate, which may have obscured potential nonlinear effects or interactions with other clinical variables, such as stage or histology. It is plausible that tumor size exerts a threshold or joint effect not well captured by the additive hazards model in its current form. Lastly, the conversion of a continuous measurement into a few broad categories, while practical for clinical interpretability, may have led to a loss of prognostic resolution, especially if relevant distinctions occur within rather than across categories. Future studies incorporating flexible modeling strategies, such as splines or machine learning approaches, may be better suited to detect nuanced size-related risk patterns in large, heterogeneous cohorts.

Future Work

A major avenue for future work is developing methods that yield individualized, covariate-adjusted estimates of transition probabilities and state occupancy. While the additive hazards model captures population-level hazard contributions flexibly, it does not directly return the probability that a patient with a specific covariate profile (e.g., a 3-year-old with Stage III favorable histology and 6 cm tumor) will experience relapse or death by a certain time. Adopting or extending case-cohort approaches such as those used by Kulich & Lin (2004), or employing pseudo-observation-based frameworks, could enable estimation of individualized cumulative incidence functions. Such estimates are crucial for clinical decision-making and risk communication.

Given the inconsistent findings regarding tumor size in our study, future work should explore more refined and biologically plausible modeling strategies. Tumor diameter was treated categorically, which may obscure important prognostic patterns. Modeling it as a continuous covariate using splines or fractional polynomials could reveal non-linear associations with relapses or mortality risk. Furthermore, interactions between tumor size and stage or histology (e.g., whether size influences risk differently in diffuse anaplastic vs. favorable histology tumors) should be formally tested. Machine learning methods such as survival forests (Ishwaran et al., 2008) or gradient-boosted survival models (Chen et al., 2013; Hothorn et al., 2006) may also help uncover complex relationships that additive or AFT models may miss. These approaches are capable of modeling high-order interactions and non-linear covariate effects, making them especially well-suited for exploratory risk stratification in heterogeneous pediatric oncology populations.

Our findings are based solely on NWTS-3 and NWTS-4, which, despite their value, reflect older therapeutic eras. It is critical to validate these results in more recent cohorts such as NWTS-5 and the Children's Oncology Group (COG) AREN trials (e.g., Grundy et al., 2005; Ehrlich et al., 2022), which incorporate molecular risk factors (e.g., LOH 1p/16q, gain of 1q), refined histologic classification, and risk-adapted therapy. Repeating multi-state survival analyses in these newer datasets would assess the durability of our findings under current clinical practices and support dynamic risk prediction models that evolve with changing treatment protocols.

Sparse events and small risk sets at later follow-up intervals, particularly for rare transitions like direct death (1→3) and death following relapse (2→3) in NWTS-4, led to abrupt jumps and visual irregularities in our cumulative incidence functions (CIFs) and transition probability curves. Such fluctuations reflect the natural variance introduced by low event counts rather than underlying clinical phenomena. To address these instabilities, future studies should consider employing smoothing techniques such as kernel-smoothed CIF estimators (Soltanian and Mahjub, 2012), or penalized additive hazard models which regularize hazard contributions to produce smoother hazard function estimates (Jackson, 2023).

Alternatively, fully Bayesian spline-based hazard models can incorporate prior distributions and automatically select smoothing parameters, thereby yielding credible bands around CIFs and hazards that reflect both uncertainty and regularized tendencies (Köhler et al., 2018). Importantly, figures depicting these functions should include confidence or credible intervals to properly convey uncertainty and prevent overinterpretation of volatile fluctuations due to sparse data.

Both NWTS-3 and NWTS-4 were multicenter studies spanning multiple years, during which treatment approaches evolved. However, our models did not explicitly adjust for time-dependent treatment effects or center-level variability. Future analyses should incorporate variables such as timing and type of relapse therapy, radiation exposure, chemotherapy intensity, and institutional practice patterns. Multi-level additive hazard models or joint frailty models could capture heterogeneity across centers and help distinguish patient-level risk from institutional or temporal factors. Additionally, incorporating dynamic treatment regimes into multi-state models could facilitate evaluation of how real-time therapy adjustments influence long-term outcomes.

References

Andersen, P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical methods in medical research*, *11*(2), 91-115.

Breslow, N. E., Beckwith, J. B., Perlman, E. J., & Reeve, A. E. (2006). Age distributions, birth weights, nephrogenic rests, and heterogeneity in the pathogenesis of Wilms tumor. *Pediatric blood & cancer*, *47*(3), 260-267.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Coppes, M. J., Wilson, P. C., & Weitzman, S. (1991). Extrarenal Wilms' tumor: staging, treatment, and prognosis. *Journal of clinical oncology*, *9*(1), 167-174.

Cotton, C. A., Peterson, S., Norkool, P. A., & Breslow, N. E. (2007). Mortality ascertainment of participants in the National Wilms Tumor Study using the National Death Index: comparison of active and passive follow-up results. *Epidemiologic Perspectives & Innovations*, *4*, 1-8.

D'angio, G. J., Breslow, N., Beckwith, J. B., Evans, A., Baum, E., Delorimier, A., ... & Thomas, P. R. (1989). Treatment of Wilms' tumor. Results of the third national Wilms' tumor study. *Cancer*, *64*(2), 349-360.

Dome, J. S., Cotton, C. A., Perlman, E. J., Breslow, N. E., Kalapurakal, J. A., Ritchey, M. L., ... & Green, D. M. (2006). Treatment of anaplastic histology Wilms' tumor: results from the fifth National Wilms' Tumor Study. *Journal of clinical oncology*, *24*(15), 2352-2358.

Dome, J. S., Graf, N., Geller, J. I., Fernandez, C. V., Mullen, E. A., Spreafico, F., ... & Pritchard-Jones, K. (2015). Advances in Wilms tumor treatment and biology: progress through international collaboration. *Journal of Clinical Oncology*, *33*(27), 2999-3007.

Ehrlich, P., Chi, Y. Y., Chintagumpala, M. M., Hoffer, F. A., Perlman, E. J., Kalapurakal, J. A., ... & Dome, J. S. (2017). Results of the first prospective multi-institutional treatment study in children with bilateral Wilms tumor (AREN0534): a report from the Children's Oncology Group. *Annals of surgery*, *266*(3), 470-478.

Ehrlich, P. F., Tornwall, B., Chintagumpala, M. M., Chi, Y. Y., Hoffer, F. A., Perlman, E. J., ... & Dome, J. S. (2022). Kidney preservation and Wilms tumor development in children with diffuse hyperplastic perilobar nephroblastomatosis: a report from the children's oncology group study AREN0534. *Annals of surgical oncology*, *29*(5), 3252-3261.

Fernandez, C. V., Perlman, E. J., Mullen, E. A., Chi, Y. Y., Hamilton, T. E., Gow, K. W., ... & Shamberger, R. C. (2017). Clinical outcome and biological predictors of relapse after nephrectomy only for very low-risk Wilms tumor: a report from Children's Oncology Group AREN0532. *Annals of surgery*, *265*(4), 835-840.

Green, D. M., Thomas, P. R., & Shochat, S. (1995). The treatment of Wilms tumor: results of the National Wilms Tumor Studies. *Hematology/Oncology Clinics*, *9*(6), 1267-1274.

Green, Daniel M., et al. "Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: a report from the National Wilms' Tumor Study Group." *Journal of clinical oncology* 16.1 (1998): 237-245.

Grundy, P. E., Breslow, N. E., Li, S., Perlman, E., Beckwith, J. B., Ritchey, M. L., ... & Green, D. M. (2005). Loss of heterozygosity for chromosomes 1p and 16q is an adverse prognostic factor in favorable-histology Wilms tumor: a report from the National Wilms Tumor Study Group. *Journal of clinical oncology*, *23*(29), 7312-7321.

Honeyman, J. N., Rich, B. S., McEvoy, M. P., Knowles, M. A., Heller, G., Riachy, E., ... & La Quaglia, M. P. (2012). Factors associated with relapse and survival in Wilms tumor: a multivariate analysis. *Journal of Pediatric Surgery*, *47*(6), 1228-1233.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355-373.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests.

Jackson, C. H. (2023). survextrap: a package for flexible and transparent survival extrapolation. *BMC Medical Research Methodology*, *23*(1), 282.

Köhler, M., Umlauf, N., & Greven, S. (2018). Nonlinear association structures in flexible Bayesian additive joint models. *Statistics in medicine*, *37*(30), 4771-4788.

Kulich, M., & Lin, D. Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, *99*(467), 832-844.

Malogolowkin, M., Cotton, C. A., Green, D. M., Breslow, N. E., Perlman, E., Miser, J., ... & Weeks, D. (2008). Treatment of Wilms tumor relapsing after initial treatment with vincristine, actinomycin D, and doxorubicin. A report from the National Wilms Tumor Study Group. *Pediatric blood & cancer*, *50*(2), 236-241.

Martinussen, T., & Scheike, T. H. (2006). *Dynamic regression models for survival data* (Vol. 1). New York: Springer.

Mullen, E. A., Chi, Y. Y., Hibbitts, E., Anderson, J. R., Steacy, K. J., Geller, J. I., ... & Dome, J. S. (2018). Impact of surveillance imaging modality on survival after recurrence in patients with favorable-histology Wilms tumor: a report from the Children's Oncology Group. *Journal of Clinical Oncology*, *36*(34), 3396-3403.

Neville, Holly L., and Michael L. Ritchey. "WILMS'TUMOR: Overview of National Wilms' Tumor Study Group Results." *Urologic Clinics of North America* 27.3 (2000): 435-442.

Norris, S. (1997). JR; Markov Chains, Cambridge Uni.

Pshak, T. J., Cho, D. S., Hayes, K. L., & Vemulakonda, V. M. (2014). Correlation between CT-estimated tumor volume, pathologic tumor volume, and final pathologic specimen weight in children with Wilms' tumor. *Journal of Pediatric Urology*, *10*(1), 148-154.

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, *26*(11), 2389-2430.

Shamberger, R. C., Guthrie, K. A., Ritchey, M. L., Haase, G. M., Takashima, J., Beckwith, J. B., ... & Breslow, N. E. (1999). Surgery-related factors and local recurrence of Wilms tumor in National Wilms Tumor Study 4. *Annals of surgery*, *229*(2), 292-297.

Soltanian, A. R., & Mahjub, H. (2012). A non-parametric method for hazard rate estimation in acute myocardial infarction patients: kernel smoothing approach.

Szychot, E., Apps, J., & Pritchard-Jones, K. (2014). Wilms' tumor: biology, diagnosis and treatment. *Translational pediatrics*, *3*(1), 12.

Yang, S., Wallach, M., Krishna, A., Kurmasheva, R., & Sridhar, S. (2021). Recent developments in nanomedicine for pediatric cancer. *Journal of Clinical Medicine*, *10*(7), 1437.

Table 1. Patient Characteristics and Outcome Measures for Study 3, Study 4, and Combined Cohorts

Variables	Category	Study 3 (N=1671) n (%)	Study 4 (N=2244) n (%)	Total (N=3915) n (%)
Death	**Alive when last seen**	1451 (86.8)	2020 (90.0)	3471 (88.7)
	Died	220 (13.2)	224 (10.0)	444 (11.3)
Relapse	**Alive no relapses when last seen**	1364 (81.6)	1882 (83.9)	3246 (82.9)
	Relapsed	307 (18.4)	362 (16.1)	669 (17.1)
Central Path Histology	**Favorable**	1479 (88.5)	1997 (89.0)	3476 (88.8)
	Unfavorable	192 (11.5)	247 (11.0)	439 (11.2)
Institutional Histology	**Favorable**	1480 (88.6)	2053 (91.5)	3533 (90.2)
	Unfavorable	191 (11.4)	191 (8.5)	382 (9.8)
Stage of Disease	**I**	699 (41.8)	844 (37.6)	1543 (39.4)
	II	361 (21.6)	632 (28.2)	993 (25.4)
	III	402 (24.1)	504 (22.4)	906 (23.1)
	IV	209 (12.5)	264 (11.8)	473 (12.1)
Time to Death or Last Date Seen (yr), Mean (SD)		13.24 (6.2)	8.16 (3.8)	10.33 (5.5)
Time to Relapse or Last Date Seen (yr), Mean (SD)		12.39 (6.8)	7.60 (4.2)	9.65 (5.9)
Diameter of Tumor (cm), Mean (SD)		11.45 (3.9)	11.03 (3.8)	11.21 (3.8)
Specimen Weight (gram), Mean (SD)		607.80 (408.2)	602.15 (394.9)	604.56 (400.6)
Age at Diagnosis (yr), Mean (SD)		3.50 (2.6)	3.56 (2.5)	3.53 (2.6)

Table 2. Expected Temporal Progression Between Relapse and Follow-Up Events Across Relapse and Vital Status Groups

Relapse and Vital Status	Study 3 (N=1671)	Study 4 (N=2244)	Total (N=3915)
No relapse, alive at last follow-up	0 / 1364 (0.00)	0 / 1882 (0.00)	0 / 3246 (0.00)
Relapsed, alive at last follow-up	87 / 87 (1.00)	134 / 138 (0.97)	221 / 225 (0.98)
Deceased	151 / 220 (0.69)	163 / 224 (0.73)	314 / 444 (0.71)

Table 3. Categorized Distribution of Tumor Diameter and Age at Diagnosis in Study 3, Study 4, and the Combined Cohorts

Variables	Category	Study 3 (N=1671) n (%)	Study 4 (N=2244) n (%)	Total (N=3915) n (%)
Tumor Diameter	**< 10 cm**	549 (32.9)	773 (34.4)	1322 (33.8)
	10-15 cm	906 (54.2)	1227 (54.7)	2133 (54.5)
	> 15 cm	216 (12.9)	244 (10.9)	460 (11.7)
Age at Diagnosis	**< 2 years**	534 (32.0)	705 (31.4)	1239 (31.7)
	2-4 years	572 (34.2)	752 (33.5)	1324 (33.8)
	> 4 years	565 (33.8)	787 (35.1)	1352 (34.5)

Table 4. Log-Rank Test Results for Time-to-Event Outcomes by Key Clinical Variables ($\alpha = 0.05$)

Variables	Endpoint	Test Statistic (Chi-Square)	Degree of Freedom (DF)	P-value
NWTS Study Group	**Time to Relapse**	1.30	1	0.255
	Time to Death	3.63	1	0.057
	Time from Relapse to Death	6.27	1	0.012
Stage of Disease	**Time to Relapse**	118.45	3	< 0.001
	Time to Death	182.64	3	< 0.001
	Time from Relapse to Death	91.70	3	< 0.001
Central Path Histology	**Time to Relapse**	328.80	1	< 0.001
	Time to Death	492.72	1	< 0.001
	Time from Relapse to Death	107.12	1	< 0.001
Age Group at Diagnosis	**Time to Relapse**	37.93	2	< 0.001
	Time to Death	23.81	2	< 0.001
	Time from Relapse to Death	4.25	2	0.120
Tumor Diameter Category	**Time to Relapse**	12.83	2	0.002
	Time to Death	5.14	2	0.076
	Time from Relapse to Death	2.07	2	0.355

Table 5. Akaike Information Criterion (AIC) Values for Cox Models with Varying Specifications of Study Effect Across Three Time-to-Event Outcomes

Outcome	Study as Covariate	Stratified by Study	Stratified + Study Interaction
Time to Relapse	10557	9646	9635
Time to Death	6767	6160	6166
Time from Relapse to Death	5233	4619	4625

Table 6. p-values from Grambsch–Therneau Test for Proportional Hazards Assumption ($\alpha = 0.05$) Across Time-to-Event Outcomes

Covariates	Time to Relapse	Time to Death	Time from Relapse to Death
Age Group at Diagnosis	0.4341	0.015	0.0664
Tumor Diameter Category	0.1063	0.078	0.3725
Central Path Histology	<0.0001	<0.0001	0.0113
Stage of Disease	0.0002	<0.0001	0.0002
Global	<0.0001	<0.0001	<0.0001

Table 7. BIC Values for AFT Models Across Three Clinical Endpoints by Study

Distribution	Time to Relapse		Time to Death		Time from Relapse to Death	
	Study 3	Study 4	Study 3	Study 4	Study 3	Study 4
Weibull	2717	3141	1957	1949	901	906
Exponential	3106	3404	2071	1929	1029	931
Log-normal	2670	3079	**1858**	**1839**	**870**	**880**
Log-logistic	2698	3118	1870	1855	875	884
Normal	3646	4017	2608	2490	1755	1607
Gamma	**2519**	**2999**	1855	1902	867	883

Table 8. AFT Model Estimates for Best-Fitting Distributions Based on BIC Across Three Clinical Endpoints (Study 3 and Study 4): Time Ratios, Standard Errors, 95% Confidence Intervals, and p-values (α = 0.05)

	Log (Time Ratios)	Time Ratios	SEs	95% CIs		P-vals	Log (Time Ratios)	Time Ratios	SEs	95% CIs		P-vals	Log (Time Ratios)	Time Ratios	SEs	95% CIs		P-vals
3	**Time to Relapse (Lognormal)**						**Time to Death (Gamma)**						**Time from Relapse to Death (Lognormal)**					
Intercept	0.5830	1.79	0.47	-0.33	1.49	0.21	8.0299	3071	0.48	7.08	8.98	<0.001	1.9296	6.89	0.43	1.08	2.78	<0.001
Age: 2-4 yrs	0.0328	1.03	0.21	-0.39	0.45	0.88	0.8877	2.43	0.37	0.16	1.61	0.02	0.4670	1.60	0.46	-0.43	1.37	0.31
Age: >4 yrs	0.0681	1.07	0.24	-0.41	0.55	0.78	0.4246	1.53	0.36	-0.29	1.14	0.24	0.0732	1.08	0.44	-0.79	0.93	0.87
Diameter: 10-15 cm	-0.1473	0.86	0.21	-0.56	0.27	0.48	0.2518	1.29	0.33	-0.39	0.89	0.44	0.8434	2.32	0.38	0.10	1.59	0.03
Diameter: >15 cm	-0.2916	0.75	0.31	-0.89	0.31	0.34	-0.3295	0.72	0.45	-1.21	0.55	0.46	1.1142	3.05	0.50	0.13	2.30	0.03
Histology: Unfavorable	-1.1744	0.31	0.31	-1.78	-0.57	<0.001	-3.7994	0.02	0.37	-4.52	-3.08	<0.001	-1.8979	0.15	0.37	-2.62	-1.17	<0.001
Stage: 2	0.1605	1.17	0.24	-0.31	0.63	0.51	-1.0830	0.34	0.41	-1.90	-0.27	0.01	-1.3313	0.26	0.49	-2.29	-0.37	0.01
Stage: 3	-0.2555	0.77	0.24	-0.73	0.22	0.29	-2.2118	0.11	0.39	-2.98	-1.44	<0.001	-1.3022	0.27	0.44	-2.16	-0.44	0.003
Stage: 4	-1.4248	0.24	0.32	-2.05	0.80	<0.001	-3.5158	0.03	0.46	-4.41	-2.62	<0.001	-1.3585	0.26	0.57	-2.48	-0.24	0.02
Scale (Gamma, Lognormal) Estimate	3.1651		0.29	2.64	3.80		3.4683		0.20	3.10	3.88		2.2712		0.14	2.01	2.57	
Shape (Gamma) Estimate	-6.0718		0.76	-6.07	0.76													
4	**Time to Relapse (Lognormal)**						**Time to Death (Gamma)**						**Time from Relapse to Death (Lognormal)**					
Intercept	0.7335	2.08	0.38	-0.01	1.48	0.05	6.9180	1010	0.39	6.16	7.68	<0.001	2.4718	11.84	0.34	1.80	3.15	<0.001
Age: 2-4 yrs	0.3603	1.43	0.19	-0.01	0.73	0.06	0.5289	1.70	0.27	0.00	1.05	0.049	0.3881	1.47	0.33	-0.26	1.04	0.24
Age: >4 yrs	-0.0467	0.95	0.21	-0.46	0.36	0.82	-0.1191	0.89	0.26	-0.62	0.38	0.64	0.3847	1.46	0.31	-0.23	1.00	0.22
Diameter: 10-15 cm	0.1548	1.17	0.15	-0.15	0.46	0.31	-0.0209	0.98	0.22	-0.46	0.41	0.92	-0.0151	0.99	0.26	-0.52	0.49	0.95
Diameter: >15 cm	0.3357	1.40	0.25	-0.15	0.82	0.18	0.3768	1.46	0.36	-0.33	1.08	0.29	0.0900	1.09	0.39	-0.68	0.86	0.82
Histology: Unfavorable	-0.9597	0.38	0.24	-1.43	-0.49	<0.001	-2.9306	0.05	0.26	-3.45	-2.41	<0.001	-1.6875	0.18	0.25	-2.18	-1.19	<0.001
Stage: 2	-0.5691	0.57	0.17	-0.91	-0.23	<0.001	-1.0536	0.35	0.29	-1.62	-0.49	<0.001	-0.5882	0.56	0.31	-1.21	0.03	0.06
Stage: 3	-0.2150	0.81	0.19	-0.59	0.16	0.26	-1.2352	0.29	0.30	-1.82	-0.65	<0.001	-1.4640	0.23	0.35	-2.14	-0.79	<0.001
Stage: 4	-0.8619	0.42	0.24	-1.33	-0.39	<0.001	-2.6450	0.07	0.33	-3.29	-2.00	<0.001	-2.0353	0.13	0.38	-2.78	-1.29	<0.001
Scale (Gamma, Lognormal) Estimate	3.0229		0.21	2.64	3.46		2.6215		0.15	2.35	2.93		1.7438		0.11	1.55	1.96	
Shape (Gamma) Estimate	-5.6457		0.58	-6.78	-4.51													

Table 9. Supremum and Kolmogorov–Smirnov Test Results for Time-Varying Effects in the Additive Aalen Model across Clinical Endpoints (α = 0.05)

	Supremum-test Statistics	P-vals H(t): B(t)=0	KS-test Statistics	P-vals H(t): constant effect	Supremum-test Statistics	P-vals H(t): B(t)=0	KS-test Statistics	P-vals H(t): constant effect	Supremum-test Statistics	P-vals H(t): B(t)=0	KS-test Statistics	P-vals H(t): constant effect
3	**Time to Relapse**				**Time to Death**				**Time from Relapse to Death**			
Intercept	5.29	<0.001	0.068	<0.001	4.17	<0.001	0.040	0.021	4.87	<0.001	0.800	<0.001
Age: 2-4 yrs	4.70	<0.001	0.077	0.001	4.48	<0.001	0.062	0.001	2.21	0.273	0.335	0.179
Age: >4 yrs	3.68	0.007	0.051	0.104	4.15	<0.001	0.063	0.002	2.73	0.127	0.303	0.287
Diameter: 10-15 cm	2.57	0.118	0.038	0.110	2.04	0.395	0.018	0.640	2.39	0.201	0.315	0.143
Diameter: >15 cm	2.23	0.275	0.080	0.069	2.94	0.050	0.034	0.533	2.06	0.400	0.409	0.166
Histology: Unfavorable	7.30	<0.001	0.372	<0.001	7.66	<0.001	0.321	<0.001	4.29	<0.001	0.825	<0.001
Stage: 2	1.98	0.507	0.039	0.180	2.85	0.075	0.041	0.039	2.71	0.091	0.490	0.042
Stage: 3	4.73	<0.001	0.123	<0.001	5.65	<0.001	0.093	<0.001	2.87	0.082	0.312	0.142
Stage: 4	5.37	<0.001	0.214	<0.001	6.20	<0.001	0.205	<0.001	4.85	<0.001	0.954	<0.001
4	**Time to Relapse**				**Time to Death**				**Time from Relapse to Death**			
Intercept	3.62	0.002	0.036	0.024	2.58	0.119	0.015	0.500	4.42	<0.001	0.406	0.001
Age: 2-4 yrs	2.80	0.073	0.026	0.315	4.39	<0.001	0.043	<0.001	3.83	0.004	0.427	0.007
Age: >4 yrs	2.92	0.041	0.052	0.023	2.87	0.051	0.035	0.055	3.78	0.004	0.350	0.023
Diameter: 10-15 cm	1.81	0.574	0.022	0.384	1.69	0.677	0.014	0.665	1.46	0.823	0.102	0.790
Diameter: >15 cm	2.63	0.133	0.044	0.274	4.21	0.001	0.042	0.113	2.13	0.379	0.273	0.217
Histology: Unfavorable	7.39	<0.001	0.307	<0.001	8.22	<0.001	0.269	<0.001	5.37	<0.001	0.974	<0.001
Stage: 2	4.38	0.001	0.080	<0.001	4.21	0.001	0.034	0.008	1.90	0.522	0.156	0.337
Stage: 3	3.12	0.031	0.046	0.056	4.37	0.001	0.156	<0.001	3.57	0.010	0.477	0.003
Stage: 4	5.09	<0.001	0.159	<0.001	6.51	<0.001	0.269	<0.001	5.49	<0.001	1.020	<0.001

Table 10. Supremum and Kolmogorov–Smirnov Test Results for Time-Varying Effects in Additive Aalen Models by Transition and Study (α = 0.05)

	Supremum-test Statistics	P-vals H(t): B(t)=0	KS-test Statistics	P-vals H(t): constant effect	Supremum-test Statistics	P-vals H(t): B(t)=0	KS-test Statistics	P-vals H(t): constant effect	Supremum-test Statistics	P-vals H(t): B(t)=0	KS-test Statistics	P-vals H(t): constant effect
3	**Transition 1 → 2 (Relapse)**				**Transition 1 → 3 (Direct Death)**				**Transition 2 → 3 (Death after Relapse)**			
Intercept	4.66	0.001	0.061	<0.001	2.64	0.081	0.017	0.239	3.91	0.002	1.450	0.006
Age: 2-4 yrs	3.62	0.006	0.052	0.029	3.02	0.025	0.028	0.035	2.95	0.029	0.981	0.032
Age: >4 yrs	2.71	0.085	0.025	0.621	3.07	0.022	0.034	0.004	2.94	0.015	1.070	0.009
Diameter: 10-15 cm	2.15	0.363	0.038	0.088	1.76	0.494	0.011	0.519	3.52	0.001	1.430	0.009
Diameter: >15 cm	2.51	0.130	0.080	0.035	2.75	0.075	0.015	0.649	3.32	0.008	1.770	0.006
Histology: Unfavorable	5.86	<0.001	0.269	<0.001	4.30	<0.001	0.097	<0.001	3.95	<0.001	1.390	<0.001
Stage: 2	1.57	0.759	0.029	0.355	2.16	0.287	0.013	0.310	4.44	<0.001	1.580	<0.001
Stage: 3	4.26	0.001	0.093	0.002	2.92	0.047	0.030	0.002	4.28	<0.001	1.600	<0.001
Stage: 4	2.15	0.281	0.067	0.070	5.52	<0.001	0.146	<0.001	3.93	<0.001	0.954	<0.001
4	**Transition 1 → 2 (Relapse)**				**Transition 1 → 3 (Direct Death)**				**Transition 2 → 3 (Death after Relapse)**			
Intercept	2.42	0.194	0.016	0.401	3.43	0.004	0.015	0.125	2.21	0.238	0.154	0.795
Age: 2-4 yrs	1.42	0.761	0.014	0.723	4.33	0.001	0.029	0.001	1.72	0.308	0.362	0.267
Age: >4 yrs	4.66	<0.001	0.065	<0.001	3.15	0.021	0.028	0.007	1.78	0.275	0.379	0.233
Diameter: 10-15 cm	2.06	0.412	0.029	0.069	1.72	0.571	0.010	0.555	1.09	0.643	0.218	0.629
Diameter: >15 cm	1.78	0.568	0.034	0.340	2.94	0.054	0.016	0.265	1.59	0.483	0.305	0.417
Histology: Unfavorable	6.13	<0.001	0.184	<0.001	4.30	<0.001	0.079	<0.001	6.43	<0.001	1.220	<0.001
Stage: 2	4.14	0.001	0.052	0.004	2.93	0.036	0.013	0.052	2.20	0.130	0.317	0.139
Stage: 3	2.70	0.089	0.035	0.083	2.85	0.050	0.021	0.003	3.67	0.002	0.929	0.005
Stage: 4	2.77	0.075	0.057	0.047	5.38	<0.001	0.093	0.025	3.80	0.001	1.006	0.005

Table 11. Risk Set Sizes and Cumulative Event Counts at Key Time Points for Transitions 1→2 (Relapse), 1→3 (Direct Death), and 2→3 (Death After Relapse), Stratified by Study

Time Points, Study 3	At Risk: 1 → 2	Events: 1 → 2	At Risk: 1 → 3	Events: 1 → 3	At Risk: 2 → 3	Events: 2 → 3
1	1484	133	1577	43	93	40
2	1395	199	1510	49	115	84
3	1354	219	1461	52	107	111
5	1301	230	1395	54	94	134
8	1229	234	1315	55	86	143
10	1158	234	1241	58	83	144
15	817	237	873	65	56	148
20	111	238	120	69	9	151
25	0	238	0	69	0	151
Time Points, Study 4	**At Risk: 1 → 2**	**Events: 1 → 2**	**At Risk: 1 → 3**	**Events: 1 → 3**	**At Risk: 2 → 3**	**Events: 2 → 3**
1	2024	164	2158	32	134	28
2	1891	250	2049	41	158	86
3	1800	288	1958	44	158	121
5	1623	296	1755	53	132	142
8	1155	301	1238	55	83	159
10	746	301	800	57	54	162
15	6	301	7	61	1	163
20	0	301	0	61	0	163
25	0	301	0	61	0	163

Table 12. Selected Covariate Profiles Used in Multi-State Additive Hazards Modeling, Ranked by Potential Clinical Severity

Potential Rank by Severity	Age Group	Tumor Diameter	Stage	Histology	Profile Type Description
1	<2 yrs	<10 cm	1	Favorable	All favorable (reference)
2	2-4 yrs	10-15 cm	2	Favorable	Mildly elevated risk in all dimensions
3	>4 yrs	>15 cm	2	Favorable	Oldest age, largest tumor, but favorable histology
4	<2 yrs	>15 cm	3	Favorable	Young child with large tumor and moderate stage
5	2-4 yrs	>15 cm	4	Favorable	Advanced tumor size and stage, but favorable histology
6	>4 yrs	10-15 cm	1	Unfavorable	High age and poor histology, but favorable stage and size
7	>4 yrs	<10 cm	2	Unfavorable	Older age, unfavorable histology, small tumor
8	2-4 yrs	<10 cm	3	Unfavorable	Intermediate overall risk; small tumor, advanced stage
9	<2 yrs	10-15 cm	4	Unfavorable	Very young child with high stage and unfavorable histology
10	>4 yrs	>15 cm	4	Unfavorable	Worst combination: oldest, largest tumor, highest stage, poor histology

Profiles represent clinically relevant combinations of age group, tumor diameter, disease stage, and histology. Rankings reflect increasing risk burden across transitions in the multi-state framework.

Table 13. Estimated Transition Probabilities in a Three-State Model for Selected Covariate Profiles, by Study and Time

		Study 3					Study 4				
Profile Selected	**Time (yrs)**	**1→1**	**1→ 2**	**1→ 3**	**2 → 2**	**2 → 3**	**1→1**	**1→ 2**	**1→ 3**	**2 → 2**	**2 → 3**
Age: <2, Diameter: <10 cm, Stage: 1, Histology: Favorable	**1**	0.950	0.037	0.013	0.321	0.679	0.974	0.014	0.012	0.899	0.101
	2	0.933	0.054	0.013	0.250	0.750	0.967	0.016	0.017	0.823	0.176
	3	0.916	0.070	0.014	0.209	0.791	0.949	0.029	0.022	0.777	0.223
	5	0.906	0.080	0.014	0.209	0.791	0.945	0.030	0.025	0.776	0.224
	8	0.902	0.083	0.015	0.143	0.856	0.937	0.038	0.024	0.641	0.359
	10	0.897	0.083	0.020	0.146	0.854	0.937	0.038	0.024	0.572	0.427
	15	0.894	0.083	0.023	0.143	0.857	0.928	0.038	0.034	0.566	0.434
	20	0.871	0.087	0.042	0.100	0.900	0.928	0.038	0.034	0.566	0.434
Age: 2-4 yrs, Diameter: 10-15 cm, Stage: 2, Histology: Favorable	**1**	0.943	0.057	0.000	1.000	0.000	0.912	0.088	0.000	1.000	0.000
	2	0.854	0.146	0.000	0.693	0.307	0.819	0.181	0.000	0.946	0.054
	3	0.837	0.163	0.000	0.614	0.386	0.795	0.205	0.000	0.900	0.100
	5	0.823	0.177	0.000	0.394	0.606	0.787	0.213	0.000	0.803	0.197
	8	0.819	0.181	0.000	0.341	0.659	0.783	0.217	0.000	0.763	0.237
	10	0.819	0.181	0.000	0.341	0.659	0.783	0.217	0.000	0.740	0.260
	15	0.817	0.173	0.009	0.289	0.711	0.783	0.217	0.000	0.605	0.395
	20	0.821	0.179	0.000	0.225	0.775	0.783	0.217	0.000	0.605	0.395
Age: >4 yrs, Diameter: >15 cm, Stage: 2, Histology: Favorable	**1**	0.960	0.040	0.000	0.824	0.176	0.932	0.068	0.000	1.000	0.000
	2	0.932	0.068	0.000	0.589	0.411	0.888	0.112	0.000	0.957	0.043
	3	0.935	0.065	0.000	0.465	0.535	0.860	0.140	0.000	0.893	0.106
	5	0.938	0.062	0.000	0.355	0.645	0.859	0.141	0.000	0.807	0.193
	8	0.941	0.059	0.000	0.365	0.635	0.858	0.142	0.000	0.670	0.330
	10	0.941	0.059	0.000	0.361	0.639	0.857	0.141	0.002	0.638	0.362
	15	0.930	0.064	0.006	0.357	0.643	0.857	0.141	0.002	0.644	0.356
	20	0.923	0.068	0.009	0.300	0.700	0.857	0.141	0.002	0.644	0.356
Age: <2 yrs, Diameter: >15 cm, Stage: 3, Histology: Favorable	**1**	0.838	0.125	0.037	0.243	0.757	0.934	0.046	0.020	0.570	0.430
	2	0.796	0.165	0.039	0.184	0.816	0.918	0.061	0.021	0.341	0.659
	3	0.775	0.184	0.041	0.165	0.835	0.909	0.069	0.022	0.274	0.726
	5	0.771	0.192	0.037	0.108	0.892	0.888	0.071	0.041	0.232	0.768
	8	0.759	0.202	0.039	0.106	0.894	0.890	0.071	0.039	0.232	0.768
	10	0.757	0.201	0.042	0.102	0.898	0.890	0.072	0.038	0.203	0.797
	15	0.758	0.205	0.037	0.107	0.893	0.890	0.072	0.038	0.216	0.784
	20	0.752	0.208	0.040	0.083	0.917	0.890	0.072	0.038	0.216	0.784
Age: 2-4 yrs, Diameter: >15 cm, Stage: 4, Histology: Favorable	**1**	0.801	0.081	0.118	0.936	0.064	0.810	0.110	0.080	0.746	0.254
	2	0.743	0.124	0.133	0.580	0.420	0.746	0.162	0.092	0.421	0.579
	3	0.735	0.132	0.133	0.367	0.633	0.729	0.179	0.092	0.279	0.721
	5	0.738	0.129	0.133	0.304	0.696	0.721	0.181	0.098	0.223	0.777
	8	0.738	0.130	0.132	0.271	0.729	0.721	0.181	0.098	0.185	0.815
	10	0.734	0.129	0.137	0.263	0.737	0.721	0.181	0.098	0.183	0.817
	15	0.716	0.131	0.153	0.255	0.745	0.667	0.167	0.166	0.175	0.825
	20	0.710	0.128	0.162	0.265	0.735	0.667	0.167	0.166	0.175	0.825
Age: >4 yrs, Diameter: 10-15 cm, Stage: 1, Histology: Unfavorable	**1**	0.741	0.191	0.068	1.000	0.000	0.820	0.147	0.033	0.791	0.209
	2	0.626	0.295	0.078	1.000	0.000	0.689	0.255	0.056	0.434	0.566
	3	0.604	0.306	0.090	0.878	0.122	0.673	0.271	0.056	0.282	0.718
	5	0.570	0.320	0.110	0.824	0.176	0.674	0.276	0.050	0.183	0.817
	8	0.571	0.319	0.110	0.743	0.257	0.664	0.272	0.064	0.155	0.845
	10	0.565	0.315	0.120	0.777	0.223	0.667	0.274	0.059	0.169	0.831
	15	0.566	0.315	0.119	0.574	0.426	0.667	0.274	0.059	0.155	0.845
	20	0.567	0.310	0.123	0.575	0.425	0.667	0.274	0.059	0.155	0.845
Age: >4 yrs, Diameter: <10 cm, Stage: 2, Histology: Unfavorable	**1**	0.727	0.198	0.075	0.070	0.930	0.747	0.195	0.058	0.500	0.500
	2	0.656	0.258	0.086	0.031	0.969	0.643	0.275	0.082	0.265	0.735
	3	0.625	0.274	0.101	0.021	0.979	0.608	0.311	0.081	0.163	0.837
	5	0.598	0.290	0.112	0.011	0.989	0.605	0.313	0.082	0.108	0.892
	8	0.600	0.286	0.114	0.009	0.991	0.592	0.318	0.090	0.077	0.923
	10	0.594	0.283	0.123	0.009	0.991	0.588	0.316	0.096	0.074	0.926
	15	0.588	0.282	0.130	0.008	0.992	0.596	0.321	0.083	0.075	0.925
	20	0.581	0.288	0.131	0.006	0.994	0.596	0.321	0.083	0.075	0.925

Age: 2-4 yrs, Diameter: <10 cm, Stage: 3, Histology: Unfavorable	**1**	0.658	0.258	0.084	0.064	0.936	0.720	0.219	0.061	0.317	0.683
	2	0.581	0.323	0.096	0.021	0.979	0.621	0.292	0.087	0.143	0.857
	3	0.536	0.353	0.111	0.015	0.985	0.596	0.316	0.088	0.066	0.934
	5	0.503	0.372	0.125	0.008	0.992	0.578	0.321	0.101	0.039	0.961
	8	0.497	0.375	0.128	0.006	0.994	0.569	0.326	0.105	0.033	0.967
	10	0.493	0.371	0.136	0.006	0.994	0.569	0.326	0.105	0.034	0.966
	15	0.495	0.372	0.133	0.005	0.995	0.566	0.324	0.110	0.036	0.964
	20	0.497	0.377	0.126	0.004	0.996	0.566	0.324	0.110	0.036	0.964
Age: <2 yrs, Diameter: 10-15 cm, Stage: 4, Histology: Unfavorable	**1**	0.574	0.211	0.215	0.118	0.882	0.692	0.162	0.146	0.286	0.714
	2	0.460	0.303	0.237	0.099	0.901	0.558	0.257	0.185	0.086	0.914
	3	0.438	0.315	0.247	0.064	0.936	0.541	0.275	0.184	0.047	0.953
	5	0.417	0.323	0.260	0.033	0.967	0.535	0.274	0.191	0.023	0.977
	8	0.417	0.325	0.258	0.029	0.971	0.535	0.274	0.191	0.023	0.977
	10	0.407	0.317	0.276	0.030	0.970	0.539	0.276	0.185	0.021	0.979
	15	0.405	0.314	0.281	0.023	0.977	0.517	0.265	0.218	0.018	0.982
	20	0.399	0.315	0.286	0.026	0.974	0.517	0.265	0.218	0.018	0.982
Age: >4 yrs, Diameter: >15 cm, Stage: 4, Histology: Unfavorable	**1**	0.602	0.201	0.197	0.285	0.715	0.677	0.188	0.135	0.387	0.613
	2	0.518	0.263	0.219	0.181	0.819	0.570	0.265	0.165	0.131	0.869
	3	0.502	0.268	0.231	0.097	0.903	0.550	0.287	0.163	0.062	0.938
	5	0.488	0.272	0.240	0.073	0.927	0.547	0.284	0.169	0.036	0.964
	8	0.491	0.268	0.240	0.068	0.932	0.540	0.282	0.178	0.026	0.974
	10	0.485	0.265	0.250	0.069	0.931	0.541	0.282	0.176	0.026	0.974
	15	0.476	0.267	0.257	0.059	0.941	0.514	0.268	0.218	0.027	0.973
	20	0.470	0.263	0.267	0.066	0.933	0.514	0.268	0.218	0.027	0.973

Note: Probabilities correspond to transitions between states 1 (initial), 2 (relapse), and 3 (death), based on fitted additive hazards models from Studies 3 and 4. Covariate profiles include Age, Tumor Diameter, Stage, and Histology.

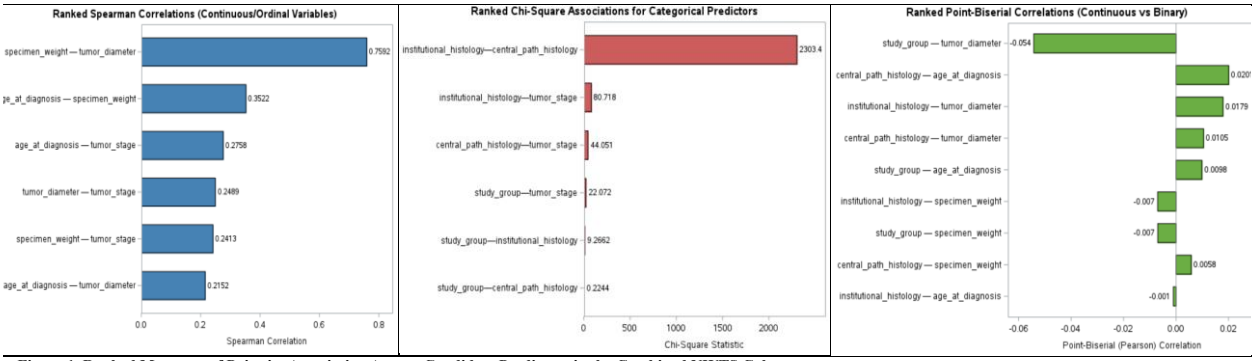
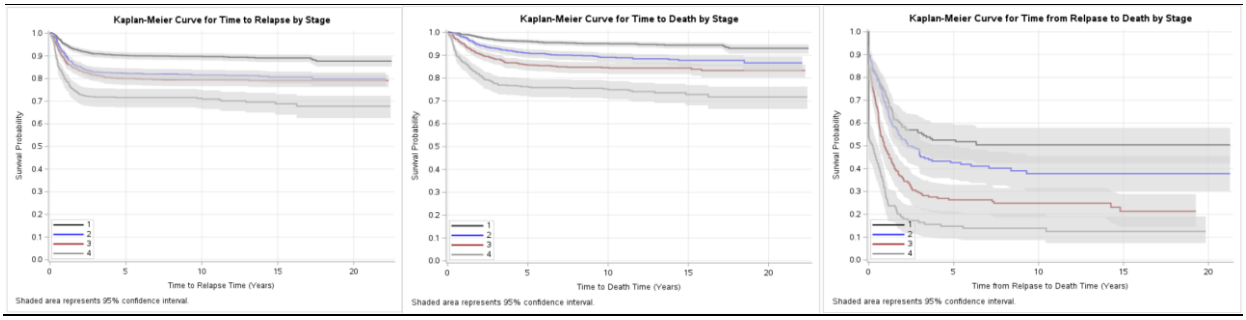


Figure 1. Ranked Measures of Pairwise Association Among Candidate Predictors in the Combined NWTS Cohort.
Spearman correlations (Panel A) reflect monotonic associations among continuous and ordinal variables (e.g., age at diagnosis, tumor diameter, and tumor stage). Chi-square statistics (Panel B) quantify dependencies between categorical predictors (e.g., study group, institutional and central pathology histology classifications). Point-biserial (Pearson) correlations (Panel C) capture linear associations between continuous and binary variables. These analyses support variable screening and redundancy assessment for subsequent multivariable modeling.



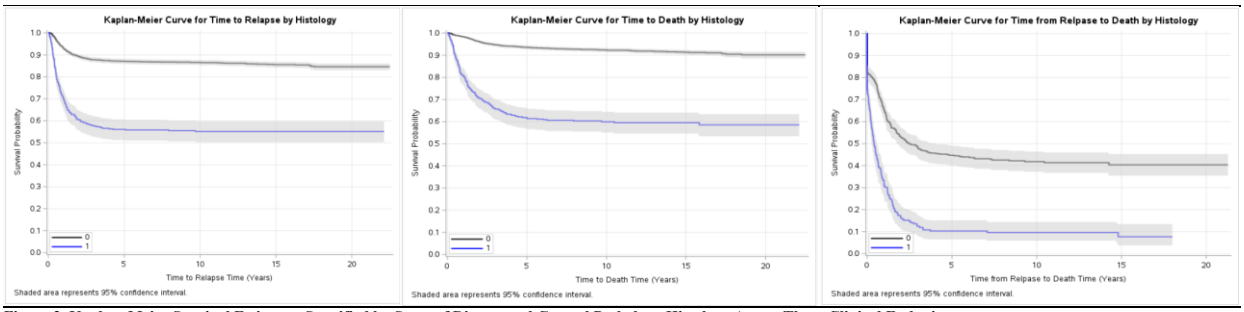


Figure 2. Kaplan–Meier Survival Estimates Stratified by Stage of Disease and Central Pathology Histology Across Three Clinical Endpoints

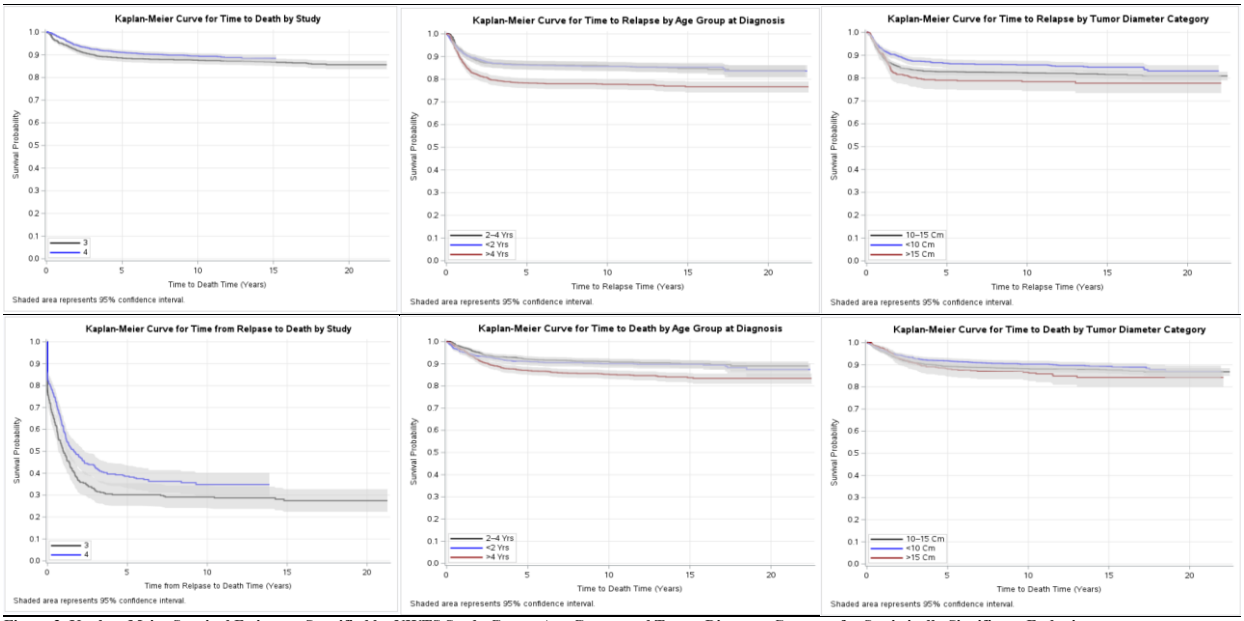
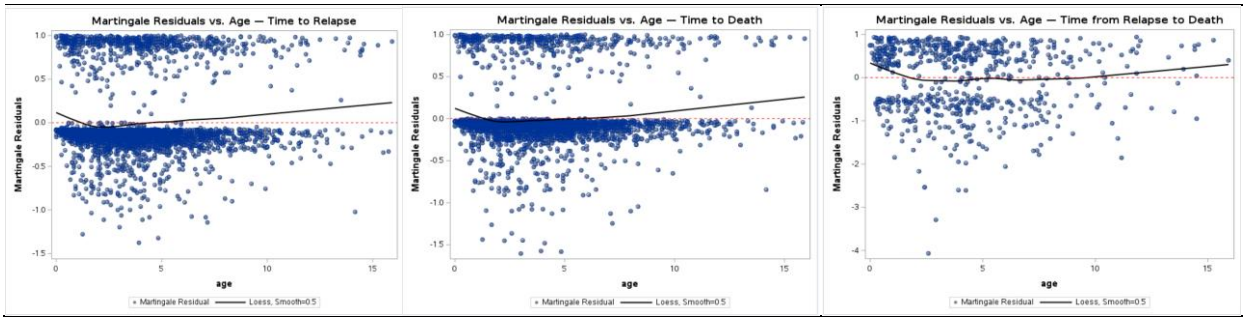


Figure 3. Kaplan–Meier Survival Estimates Stratified by NWTS Study Group, Age Group, and Tumor Diameter Category for Statistically Significant Endpoints



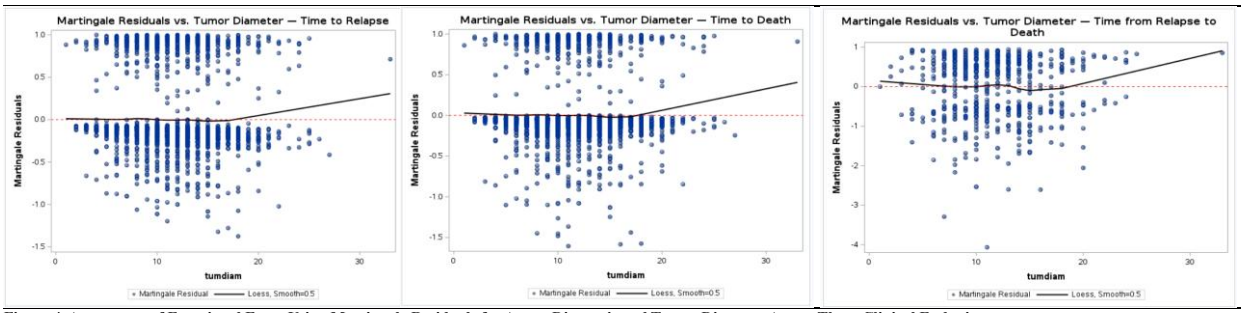


Figure 4. Assessment of Functional Form Using Martingale Residuals for Age at Diagnosis and Tumor Diameter Across Three Clinical Endpoints

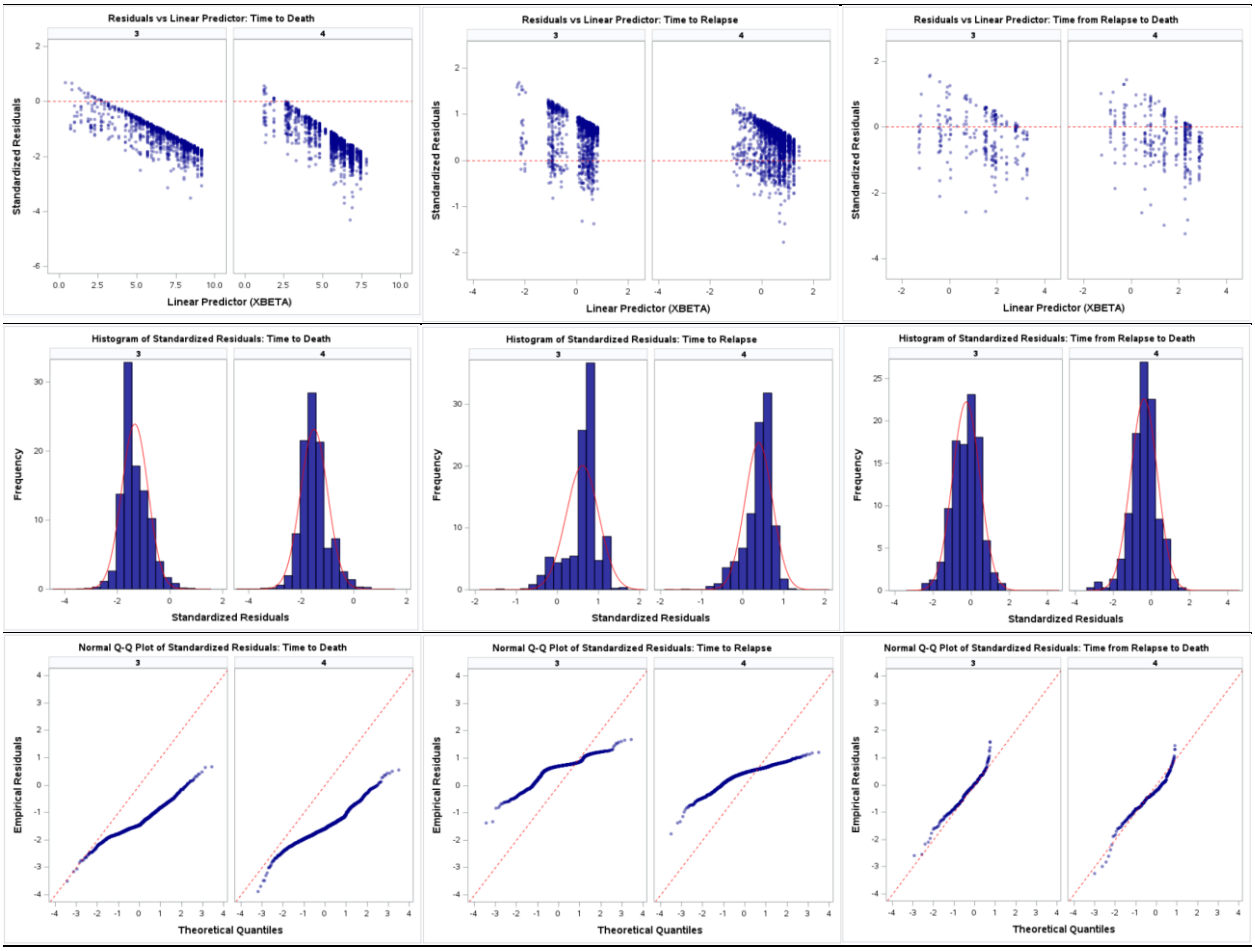


Figure 5. Diagnostics for Best-Fitting AFT Models based on BIC Across Three Clinical Endpoints: Standardized Residual Plots, Residual Histograms, and Normal Q-Q Plots

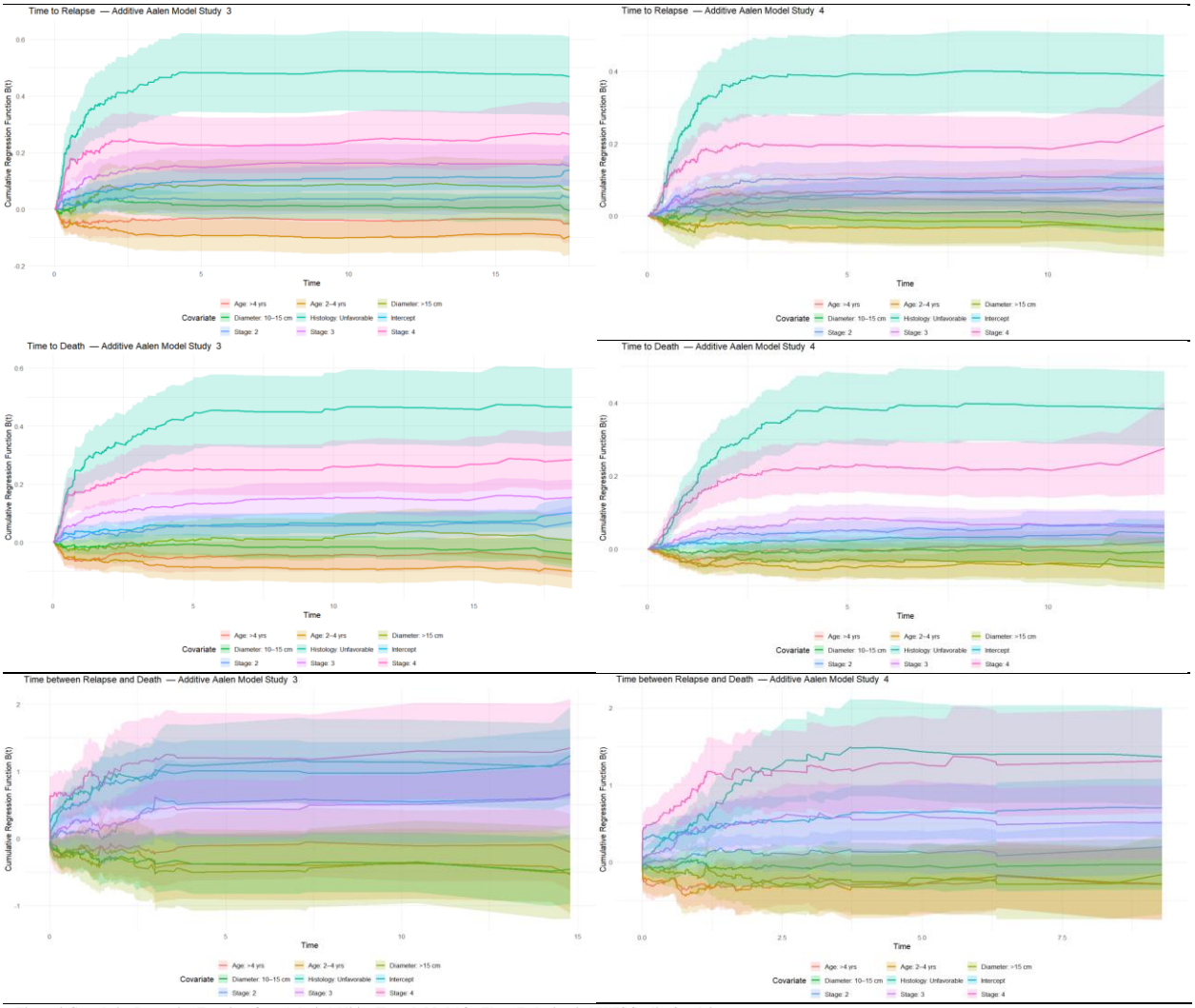


Figure 6. Cumulative Regression Functions from Aalen's Additive Hazards Model for Three Endpoints in NWTS Studies 3 and 4

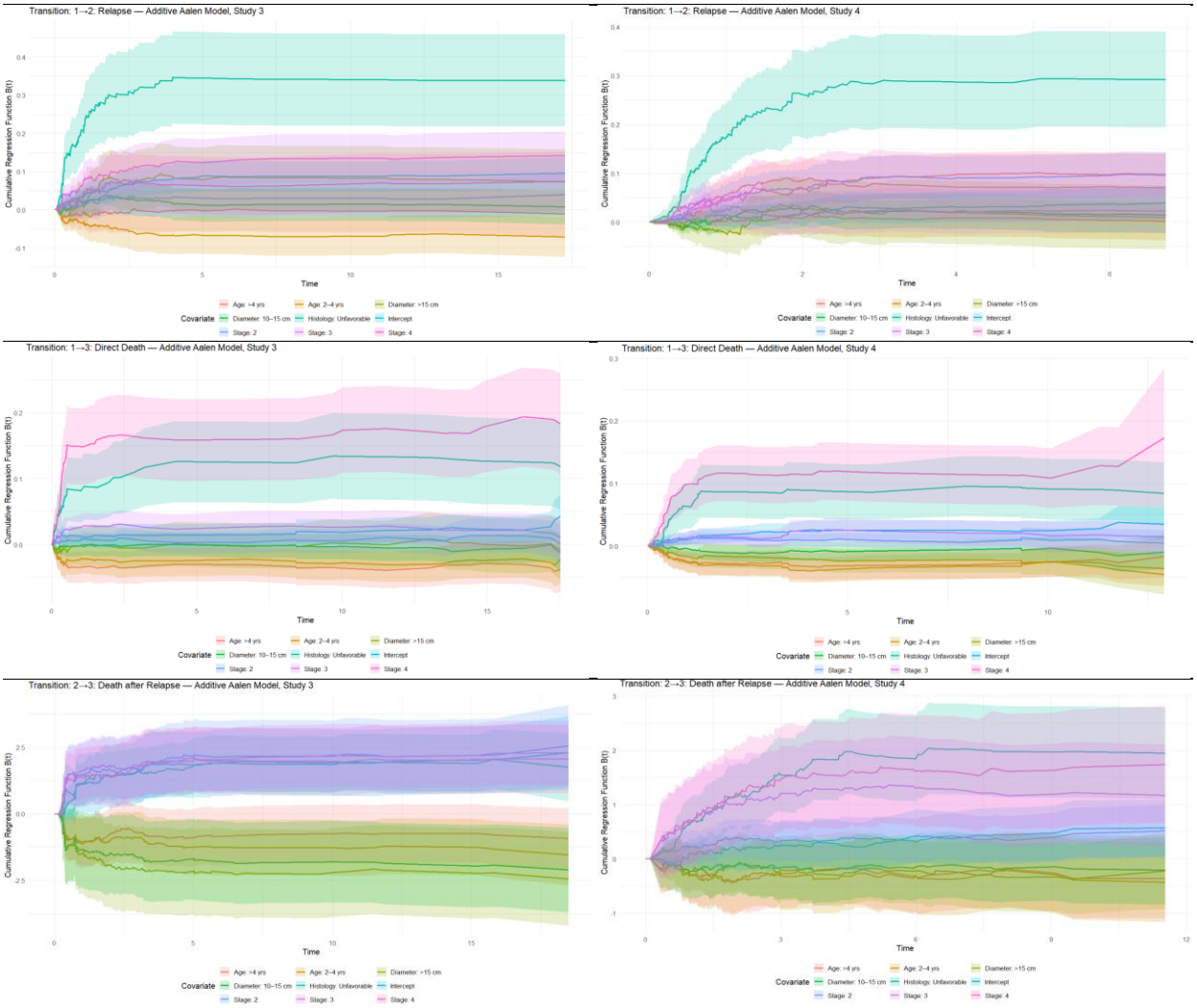


Figure 7. Cumulative Regression Functions from Aalen's Additive Hazards Model for Multi-State Transitions in NWTS Studies 3 and 4

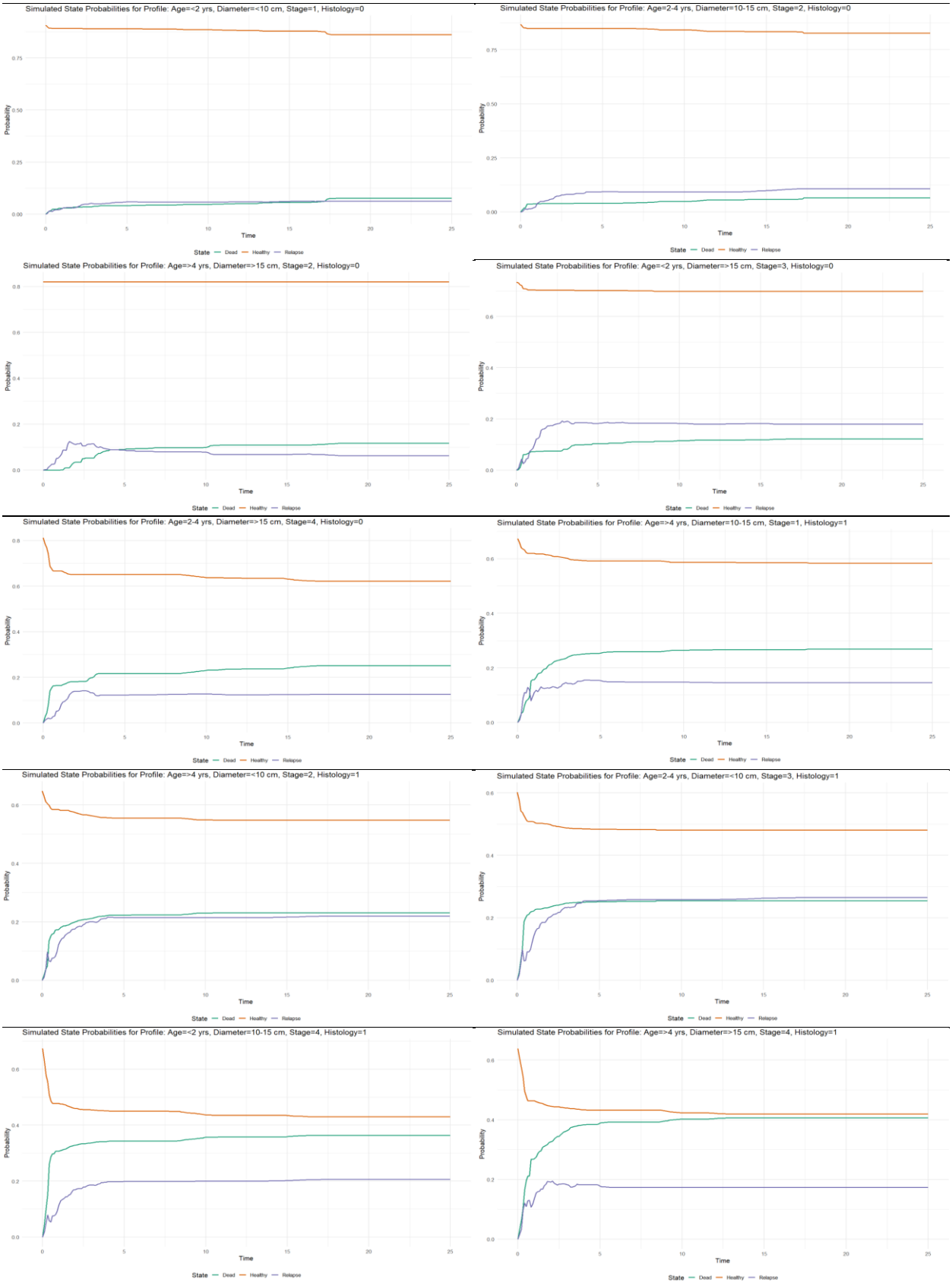


Figure 8. Estimated state occupancy probabilities over time for ten representative covariate profiles in Study 3. Profiles vary by age, tumor diameter, disease stage, and histology to capture a spectrum of clinical severity. Each panel shows the predicted probability of being in the Healthy, Relapse, or Dead state over 25 years, based on 5,000 simulations from a multi-state additive hazards model.

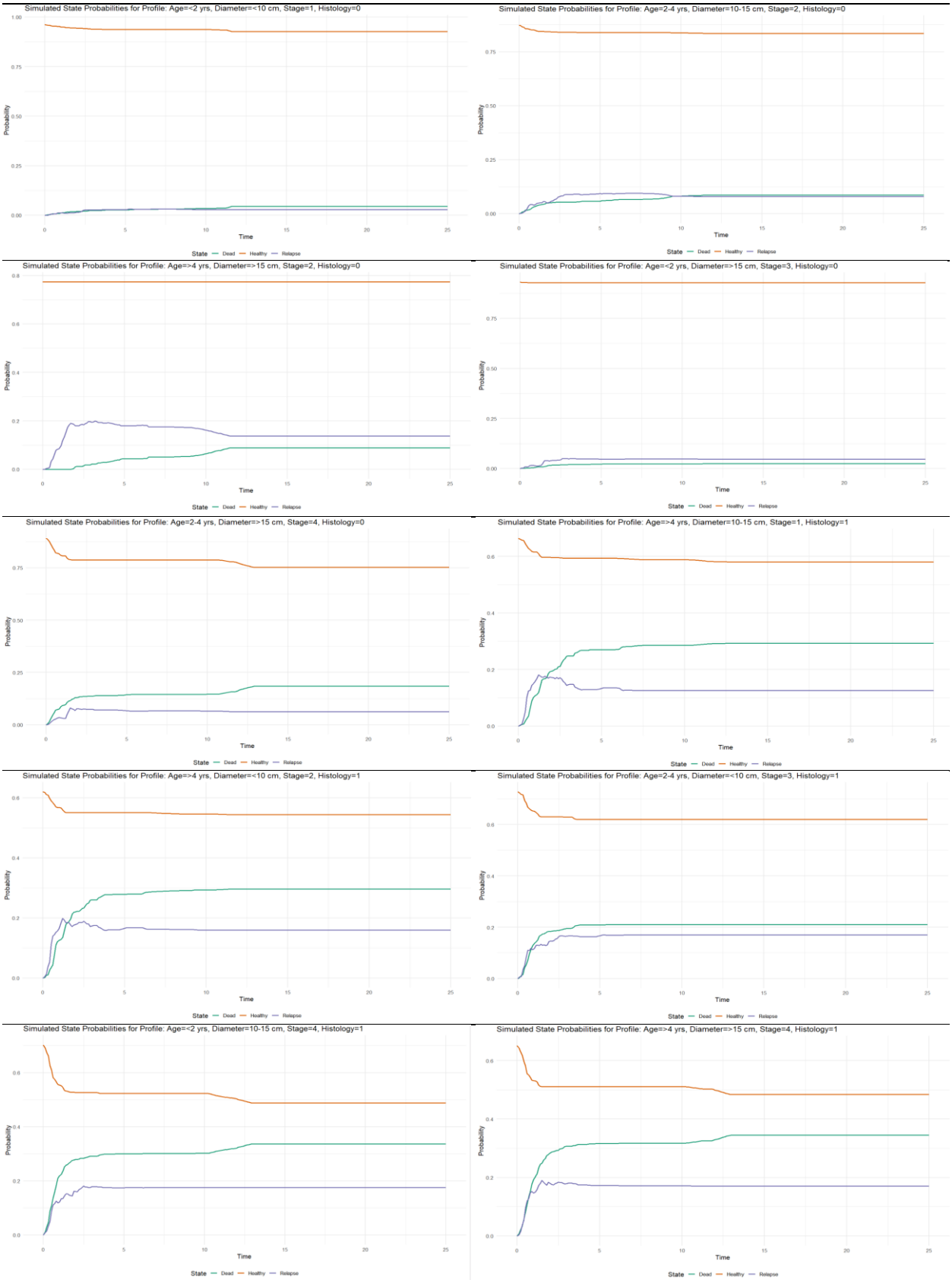


Figure 9. Estimated state occupancy probabilities over time for ten representative covariate profiles in Study 4. Profiles vary by age, tumor diameter, disease stage, and histology to capture a spectrum of clinical severity. Each panel shows the predicted probability of being in the Healthy, Relapse, or Dead state over 25 years, based on 5,000 simulations from a multi-state additive hazards model.

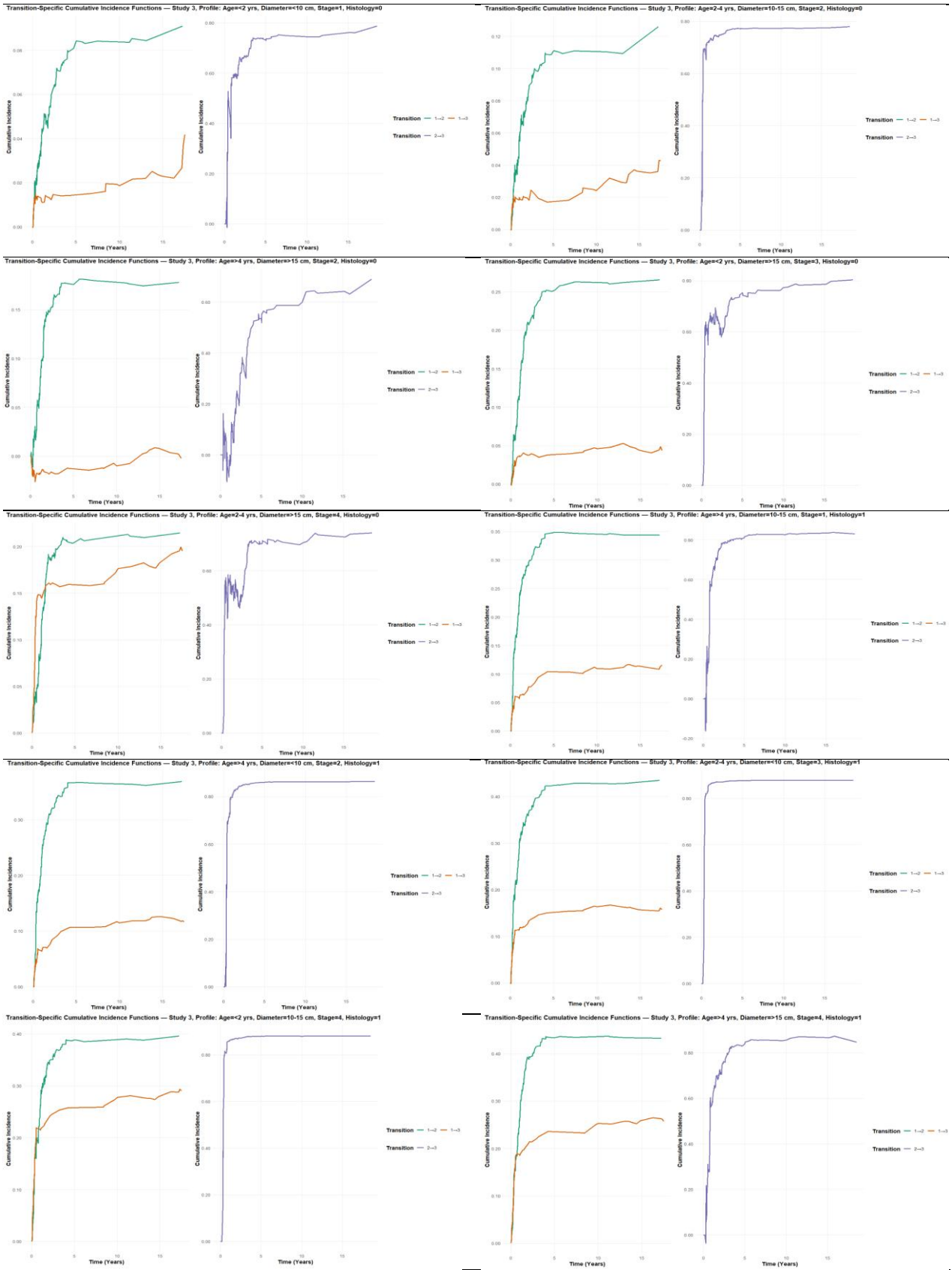


Figure 10. Estimated cumulative incidence functions for selected transitions in Study 3, based on a multi-state additive hazards model. Each panel presents the transition-specific cumulative probability over specified years for a representative covariate profile, capturing clinical variability in age, tumor diameter, disease stage, and histology. Curves reflect model-based estimates without simulation.

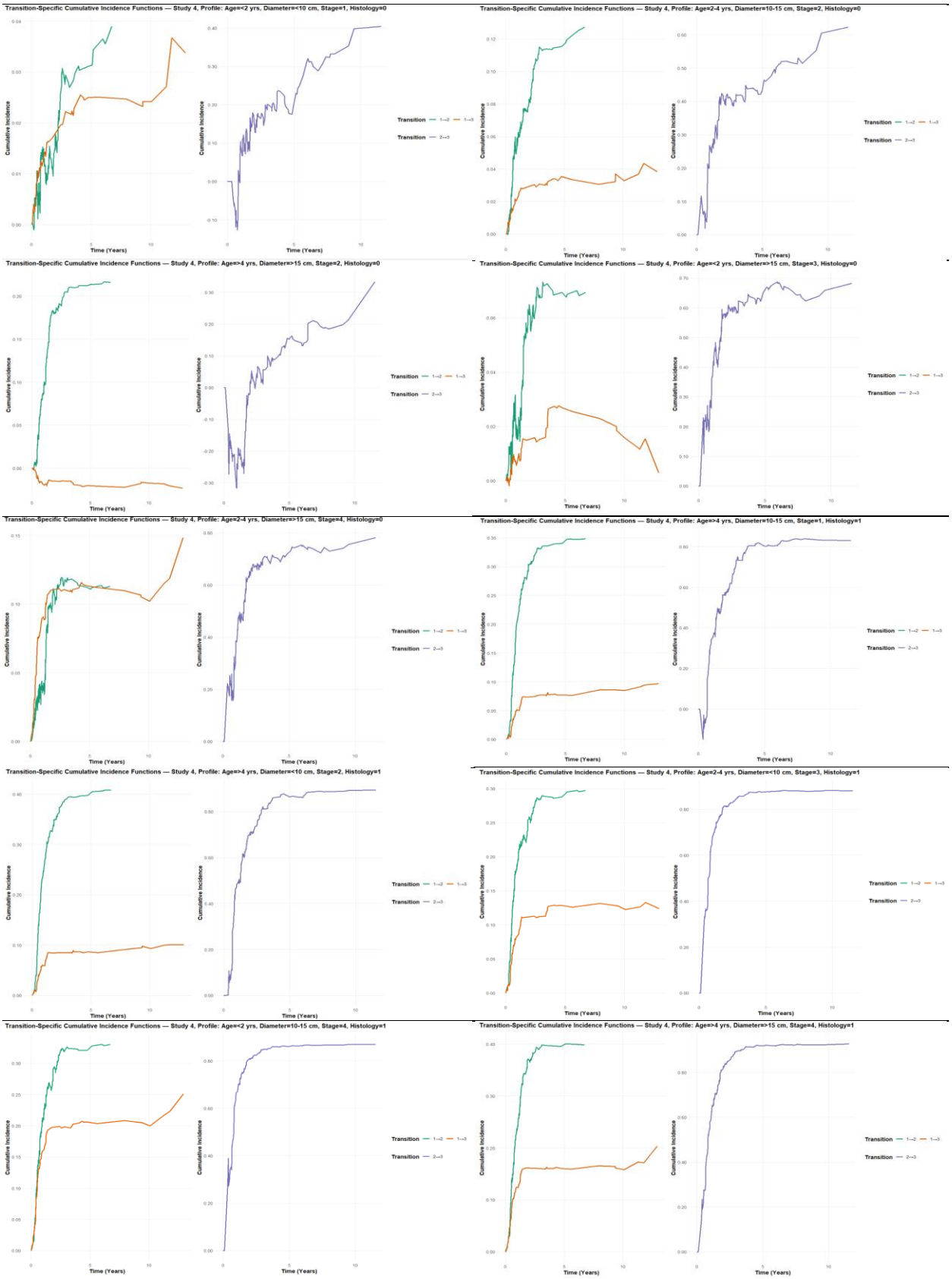


Figure 11. Estimated cumulative incidence functions for selected transitions in Study 4, based on a multi-state additive hazards model. Each panel presents the transition-specific cumulative probability over specified years for a representative covariate profile, capturing clinical variability in age, tumor diameter, disease stage, and histology. Curves reflect model-based estimates without simulation.