CAST: Causal Modeling of Time-Varying Treatment Effects on Head and Neck Cancer

Everest Yang[1,2]*, Ria Vasishtha[1,3], Luqman K. Dad[4], Lisa A. Kachnic[4], Andrew Hope[5], Eric Wang[1], Xiao Wu[6], Yading Yuan[4,7], David J. Brenner[1], Igor Shuryak[1]

[1] Center for Radiological Research, Columbia University Irving Medical Center, USA
[2] Department of Computer Science, Brown University, USA
[3] Department of Applied Mathematics, Columbia University, USA
[4] Department of Radiation Oncology, Columbia University Irving Medical Center, USA
[5] Department of Radiation Oncology, Princess Margaret Cancer Centre, Canada
[6] Department of Biostatistics, Columbia University Irving Medical Center, USA
[7] Data Science Institute, Columbia University, USA

Abstract

Causal machine learning (CML) is quickly gaining recognition in medical research because it offers better strategies to estimate treatment effects in complex real-world data, helping guide treatment optimization. Causal survival forests (CSF) are a powerful CML method for estimating heterogeneous treatment effects on survival outcomes, which are essential for informed healthcare decision-making. However, they only estimate at a fixed horizon, rather than at multiple time points. We introduce Causal Analysis for Survival Trajectories (CAST), a novel extension of CSF that models treatment effects as continuous parametric and non-parametric effect trajectories over time. Applied to the RADCURE dataset [1] of 2,651 head and neck cancer patients, CAST reveals how the effects of chemotherapy (added to radiotherapy) evolve over time at the population and individual levels. By capturing the temporal dynamics of treatment response, CAST can help clinicians to determine when and for which patient subgroups treatment benefits are maximized.

1 Introduction

Methodological gap: Traditional statistical and machine learning models are based on correlations, making them unsuitable to answer causal questions critical to clinical research [2]. Such approaches cannot disentangle confounding factors or provide interpretable estimates of causal relationships, leaving a significant methodological gap [3].

Causal machine learning (CML) addresses this by modeling causal effects to estimate individualized or subgroup-specific treatment responses [4]. Causal survival forests (CSF) extend CML to survival outcomes and flexibly estimate heterogeneous treatment effects - critical for medical applications - but estimate effects only at a fixed time, missing how effects evolve over time [5, 6, 7].

Proposed approach: We present CAST (Causal Analysis for Survival Trajectories), which extends causal survival forests to model treatment effects continuously over time. We build on previous work by Shuryak et al. [8] to extend it to chemotherapy and continuous-time causal modeling. Combining parametric and nonparametric methods, CAST models the full trajectory of treatment response (Figure 1), which is critical to understanding the complex and time-varying effects of cancer therapy [9, 10, 11]. Figure 1 includes imaging for context, though only clinical and treatment data

*Corresponding Author: Everest Yang (everest_yang@brown.edu)

Accepted at the AI for Science Workshop (NeurIPS 2025).

were modeled here. This is key: biological responses unfold through complex temporal dynamics that include initial tumor control followed by potential diminishing returns due to repopulation, late toxicities, and other factors [9, 10, 11]. Because all patients received radiation therapy, the treatment being modeled is yes/no chemotherapy only.

Clinical motivation: We apply CAST to head and neck squamous cell carcinoma (HNSCC), a cancer with increased incidence [12] and changing demographics due to HPV. HPV-related tumors, more common in younger patients, differ in radiosensitivity and prognosis [13, 14, 15], highlighting the need for individualized treatment. Treatment typically combines chemotherapy and radiation, with responses that vary between patients and manifest gradually [16, 17, 18, 19, 20].

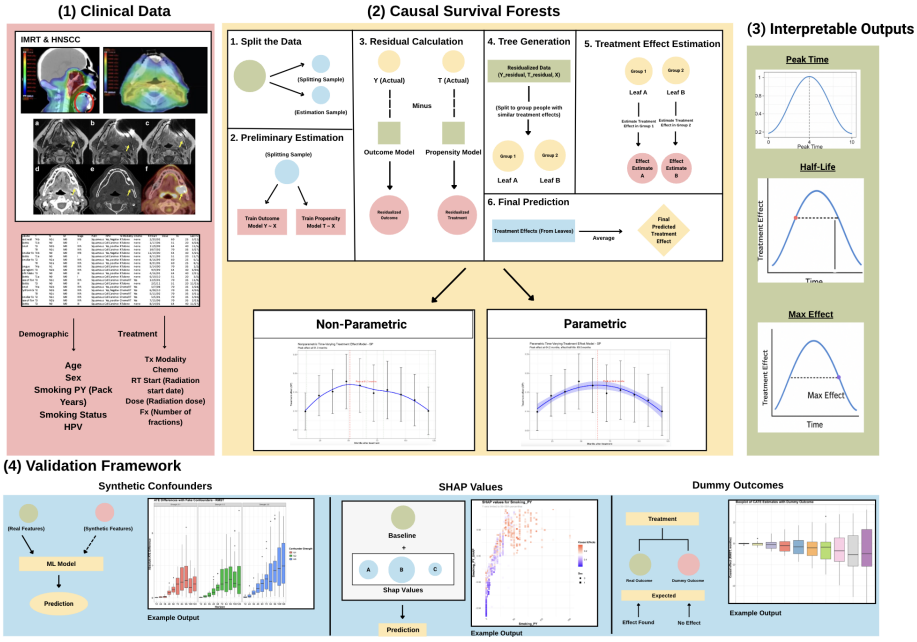


Figure 1: Overview of the CAST framework

Modeling philosophy: Our approach leverages causal survival forests to analyze high-dimensional data and identify heterogeneous treatment effects. Unlike standard survival methods, we explicitly estimate causal effects while controlling for confounders through propensity score modeling.

We used multiple validation methods, including overlap assessment and unmeasured confounding checks. We estimated propensity scores using elastic net logistic regression with cross-validation, trimming patients outside the range [0.1, 0.9] to ensure treatment group overlap. Refutation tests with dummy outcomes and negative controls verified robustness. SHAP values provided interpretable insights into how patient characteristics influence treatment outcomes for clinical application.

Significance: This study uses CSF and CAST temporal modeling to reveal how patient and disease factors shape the magnitude and timing of chemotherapy benefits in HNSCC. These insights highlight the potential in machine learning for personalized care.

Our contributions are as follows:

- **CAST is, to our knowledge, the first framework** to unify causal survival forests with *parametric* and *non-parametric* models for estimating **continuous-time treatment effects**, offering a new paradigm for temporal causal inference in survival analysis.
- CAST produces clinically interpretable metrics such as *peak effect time*, *maximum benefit*, and *effect half-life*, enabling richer understanding of treatment response dynamics.
- We introduce a rigorous validation framework incorporating *propensity score modeling*, *dummy outcome tests*, *synthetic tests*, and *SHAP-based heterogeneity analysis*.
- We apply CAST to a large real-world chemotherapy and radiotherapy dataset (RADCURE), uncovering actionable insights into when and for whom treatment benefits peak and decline.

2 Related Work

Clinical predictors of treatment response: Numerous studies have shown that treatment response in HNSCC patients is highly heterogeneous, influenced by clinical and demographic factors such as HPV status, gender, and disease stage. HPV-positive disease, more common in younger patients, is typically more treatment-sensitive with better survival [21]. This motivates methods, like CAST, that model heterogeneity beyond average effects.

Predictive survival models: Traditional models like the Cox proportional hazards model assume constant effects over time, limiting flexibility [22, 23]. More flexible models—including random survival forests (RSF), deep survival models (e.g., DeepSurv), and Bayesian additive regression trees (BART)—have improved prediction [24], notably in applications such as cervical cancer survival [25, 26], but remain predictive—not causal—unless adapted to address confounding.

Causal inference for survival analysis: New causal ML methods—including meta-learners (e.g., T-learner, S-learner) [27], G-formula-based two-learners [28], double robust estimators (e.g., AIPCW, AIPTW) [29], and causal survival forests [30]—estimate individualized effects from observational survival data [31, 32], but typically only at discrete time points, missing continuous treatment dynamics [33, 34].

Modeling time-varying treatment effects: Oncology treatments often show the early benefits, plateau and decline phases [35, 36]. Covariate effects also shift over time [37, 38], but most methods assume constant effects or treat follow-up intervals independently. CAST models effects continuously, blending parametric (e.g., quadratic fits) and non-parametric (e.g., smoothing splines) components to reveal the full temporal trajectory in treatment response.

3 Methodology

Problem Formulation: We address the challenge of estimating time-varying treatment effects in survival analysis, specifically focusing on how the impact of medical interventions evolves over time. Let $\mathcal{D} = \{(X_i, W_i, T_i, \delta_i)\}_{i=1}^n$ represent our dataset where:

- $X_i \in \mathbb{R}^p$ is a vector of covariates for subject i
- $W_i \in \{0, 1\}$ is the treatment indicator
- T_i is the observed survival time (either event time or censored time)
- δ_i is the event indicator (1 if event observed, 0 if censored)

The causal survival forest method is a powerful tool for estimating average and subgroup-specific treatment effects for survival outcomes, but it estimates the effects only at specific discrete times after treatment. This fails to capture the continuous temporal evolution of treatment responses, particularly in contexts like radiation therapy and chemotherapy where biological effects can substantially rise and fall over time.

3.1 Causal Machine Learning Framework

Our approach uses a CML framework to isolate treatment effects beyond traditional correlational methods. While conventional machine learning identifies correlations between variables, CML allows us to understand the causal impact of interventions [39]. This distinction is fundamental to our study: our goal is not just to predict outcomes but to dissect how treatments shape survival outcomes across patient subgroups.

Given the observational non-randomized nature of our clinical data, we rely on the following assumptions:

- **Unconfoundedness**: Treatment assignment is independent of potential outcomes conditional on observed covariates (also called ignorability or no unmeasured confounding)
- **Positivity (Overlap)**: Every subject has a non-zero probability of receiving each treatment
- **Consistency**: A subject's observed outcome under their received treatment equals their potential outcome for that treatment

- **Non-interference**: One subject's treatment does not affect another subject's outcome

To address selection bias in observational data, we performed propensity score modeling using elastic net logistic regression: $\hat{e}(X) = P(W = 1|X)$ with hyperparameters optimized through 10-fold cross-validation. Patients with extreme propensity scores (outside $[0.10, 0.90]$) are trimmed to ensure overlap between treatment groups. See Appendix C.1 for balance diagnostics.

3.2 CAST: Causal Analysis for Survival Trajectories

The theoretical foundation of CAST rests on modeling the effect trajectory as a function of time. Our target estimand is the conditional average treatment effect (CATE) at time t, given covariates X:

$$\tau(x, t) = \mathbb{E}[Y(1, t) - Y(0, t) \mid X = x] \tag{1}$$

where $Y(w, t)$ represents the potential outcome at time t under treatment w, and x denotes an individual's covariates. We consider two types of time-varying estimands: the difference in restricted mean survival time (RMST) and the difference in survival probability (SP) between treatment groups. RMST reflects the average time an individual is expected to survive up to a specified time horizon, calculated by integrating the survival probability from time 0 to that horizon.

Unlike prior methods that estimate effects at fixed time points, CAST models treatment effects as smooth functions of time. We use a smoothing spline to estimate the continuous effect trajectory and a quadratic fit to derive interpretable metrics.

3.2.1 Parametric Modeling Component

Our parametric modeling component employs a quadratic function: $\tau(t) = \beta_0 + \beta_1 t + \beta_2 t^2$ to capture the rise and fall of treatment effects. The parameters are estimated using weighted least squares:

$$\min_{\beta_0, \beta_1, \beta_2} \sum_t w(t)(\hat{\tau}(t) - (\beta_0 + \beta_1 t + \beta_2 t^2))^2 \tag{2}$$

where $w(t) = 1/\sigma^2(t)$ are weights based on the variance of the effect estimates at each timepoint. This approach yields clinically interpretable parameters, including the peak effect time ($t_{\text{peak}} = -\beta_1/2\beta_2$), the maximum effect magnitude ($\tau(t_{\text{peak}})$), and the treatment effect half-life, defined as the time it takes for the effect to diminish by 50% from its peak.

These parameters directly quantify key clinical aspects of the treatment response: when the maximum benefit occurs, how large that benefit is, and how quickly it diminishes—information critical for clinical decision-making that traditional methods cannot provide. See Appendix C.3 for fitted coefficients and summary statistics from the parametric model.

Algorithm 1 CAST-PARAMETRIC

Input: Horizons $\mathcal{H}$, ATEs $\{\hat{\tau}_h\}$, SEs $\{\hat{\sigma}_h\}$
Output: Temporal function $\hat{\tau}(t)$, peak time t^*, half-life λ
$\mathcal{W} \leftarrow \{w_h = 1/\hat{\sigma}_h^2\}$ ▷ Inverse-variance weights
$\hat{\tau}(t) \leftarrow \text{FITQUADRATICMODEL}(\mathcal{H}, \hat{\tau}, \mathcal{W})$
$\beta_1, \beta_2 \leftarrow$ coefficients from fit
if $\beta_2 \neq 0$ **then**
    $t^* \leftarrow -\beta_1/(2\beta_2)$ ▷ Time of peak effect
    $\lambda \leftarrow \text{SOLVE}(\hat{\tau}(t^* + \lambda) = \hat{\tau}(t^*)/2)$
else
    $t^*, \lambda \leftarrow$ NA ▷ Degenerate case
end if
return $\hat{\tau}(t), t^*, \lambda$

CAST-Parametric: This algorithm models treatment effects over time using a weighted quadratic fit to the estimated ATEs across discrete horizons. Inverse-variance weighting emphasizes more confident estimates. The peak effect time is derived analytically, while the half-life is computed by numerically solving for the point where the curve falls to half its maximum. This approach yields interpretable summaries of treatment dynamics, aligning with radiobiological phenomena such as delayed benefit and diminishing returns.

3.2.2 Non-parametric Modeling Component

Our non-parametric component employs cross-validated smoothing splines:

$$\tau(t) = g(t), \quad \text{where} \quad g = \arg\min_f \left\{ \sum_t w(t) \left(\hat{\tau}(t) - f(t)\right)^2 + \lambda \int f''(t)^2 \, dt \right\} \tag{3}$$

where λ is selected via cross-validation. This approach adapts to the data without imposing a predetermined functional form, revealing subtle inflection points in the effect trajectory that correspond to biological phase transitions in the treatment response.

We calculate the first and second derivatives of the fitted spline to identify key features of the treatment effect trajectory: local maxima and minima where $g'(t) = 0$, acceleration and deceleration phases based on sign changes in $g''(t)$, and inflection points where $g''(t) = 0$.

The non-parametric model complements the parametric fit by capturing complex, less predictable patterns—especially during later follow-up periods, when biological processes like accelerated repopulation and late toxicities may cause deviations from the smooth quadratic trend.

Algorithm 2 CAST-NONPARAMETRIC

Input: Horizons $\mathcal{H}$, ATEs $\{\hat{\tau}_h\}$, SEs $\{\hat{\sigma}_h\}$
Output: Spline $\hat{\tau}(t)$, peak t^*, inflections $\{t_i\}$
$\mathcal{W} \leftarrow \{w_h = 1/\hat{\sigma}_h^2\}$
$\hat{\tau}(t) \leftarrow \text{FITSPLINE}(\mathcal{H}, \hat{\tau}, \mathcal{W})$
$D_1(t), D_2(t) \leftarrow$ first and second derivatives of $\hat{\tau}(t)$
$t^* \leftarrow \text{ARGMAX}(\hat{\tau}(t))$ ▷ Peak effect
$\{t_i\} \leftarrow \text{ZEROCROSSINGS}(D_2(t))$ ▷ Inflection points
if t^* not in $[\min(\mathcal{H}), \max(\mathcal{H})]$ **then**
    $t^* \leftarrow$ NA
end if
return $\hat{\tau}(t), t^*, \{t_i\}$

CAST-Nonparametric: This algorithm fits a smoothing spline to the estimated treatment effects across time using inverse-variance weights. It computes the first and second derivatives of the spline to identify key dynamics: the peak effect time via the curve's global maximum and biological phase transitions via inflection points. This method captures delayed and non-monotonic effect trajectories often missed by parametric models, reflecting immune response, tissue adaptation, or timing heterogeneity.

CAST-Parametric and CAST-Nonparametric offer complementary modeling capabilities. The parametric method provides interpretable summary statistics such as peak effect timing and half-life, which are clinically intuitive and useful for hypothesis testing under smooth treatment dynamics. In contrast, the spline-based approach relaxes these assumptions and flexibly captures nonlinear, delayed, or multi-phase effects. Together, these models allow us to evaluate the robustness of temporal patterns and support a wide range of clinical interpretations.

Theoretical Guarantees: See Appendix A for theorem statements establishing consistency of CAST estimators and identifiability of time-varying treatment effects under standard causal assumptions.

4 Experiments

Dataset: We use the RADCURE observational dataset from The Cancer Imaging Archive (TCIA), a publicly accessible resource on multiple types of cancer. The dataset spans from 2005 to 2017 and contains clinical, demographic, and treatment metadata for 3,346 patients. We select 2,651 patients with pathologically confirmed HNSCC and a defined tumor site. While the dataset primarily focuses on oropharyngeal cancer, it also includes laryngeal, nasopharyngeal, and hypopharyngeal cases. The binary treatment variable used in CAST is chemotherapy (yes/no) with radiotherapy covariates.

Preprocessing: We filtered incomplete profiles and standardized continuous variables for comparability. We used radiotherapy data—dose/fraction, number of fractions, and total radiation treatment time duration in days—to calculate Biologically Effective Dose (BED) values, applying both dose-independent (DI) and dose-dependent (DD) models with established radiobiological parameters [8]. We then partitioned the dataset into training (75%) and testing (25%) sets, maintaining consistent

event rates across both subsets for unbiased evaluation of treatment effects. See Appendix B for more on data preprocessing and computing resources.

Propensity Score Modeling: To address selection bias, we used elastic net logistic regression to estimate the likelihood of a person receiving treatment, based on their characteristics. Hyperparameters were optimized through 10-fold cross-validation: elastic net mixing parameter $\alpha \in [0.01, 0.99]$ and regularization parameter λ chosen from a grid of 100 values. Propensity score distributions were assessed through both Pearson and Spearman correlation matrices ($\alpha = 0.05$, Bonferroni-corrected) and visualized using kernel density estimation. Patients with scores outside $[0.10, 0.90]$ were trimmed to ensure overlap, with sensitivity analyses conducted at thresholds $\{0.01, 0.03, 0.05, 0.07, 0.10\}$.

Implementation & Heterogeneity Analysis: We used causal survival forests with Nelson-Aalen estimation to handle right-censoring, estimating treatment effects over $12, 24, \ldots, 120$ months post-treatment. Our forest was constructed with 5,000 trees to ensure robust estimation of heterogeneous effects across the patient population. Sensitivity analyses using different numbers of trees showed similar results. For each time horizon, we independently trained a causal forest model using the training dataset, with covariates properly standardized and propensity scores incorporated through doubly-robust estimation. The forests were configured with tuning parameters selected through cross-validation, including minimum node size, split regularization, and sampling fraction. Prediction uncertainty was quantified through the infinitesimal jackknife method, providing variance estimates for each individual treatment effect. This approach allowed us to capture both average treatment effects and their heterogeneity across different patient subgroups at each follow-up time point, while properly accounting for the right-censoring inherent in survival data [40, 41].

Treatment effect heterogeneity was analyzed using approximate SHAP values calculated via Monte Carlo sampling with 1,000 iterations and a convergence threshold of $\epsilon = 0.01$. The SHAP values were normalized such that $\sum_i \text{SHAP}_i$ corresponds to the difference between the individual and mean model predictions. This approach revealed which patient characteristics most strongly influenced treatment response, with HPV status and smoking history emerging as particularly important predictors. We visualized the relationship between feature values and their SHAP contributions to identify subgroups with differential treatment benefits.

Validation Methods

We implemented several validation strategies as refutation tests for the causal effect estimates in our experiments. For each test, we computed summary statistics (mean, standard deviation, max deviation) to assess model robustness, using a consistent 5,000-tree specification and random seeds for reproducibility.

Dummy Outcome Tests: We shuffled treatment assignments and outcome times across 20 repetitions for each time horizon (12-120 months), generating a null distribution to assess false positive rates. Boxplots confirmed the null hypothesis centered around zero, showing that the causal effect estimates for each horizon were centered around zero as expected. The variance of these estimates increased with increasing horizon time due to the decreasing number of patients remaining at risk at longer times. The results suggested good reliability of the estimates for times ≤ 60 months.

Sensitivity to Additional Covariates: We introduced synthetic covariates with varying signal strengths of correlation with treatment assignment (0.1, 0.3, 0.5) that were unrelated to both treatment assignment and outcome, in order to assess the sensitivity of treatment effect estimates to irrelevant/spurious variables.

Negative Control Tests: Irrelevant binary treatments were randomly assigned to ensure the model did not detect spurious effects. Treatment effects for these were zero across all time horizons.

Robustness to Irrelevant Features: Five random noise variables were added, and changes in treatment effect estimates and feature importance were monitored to ensure no significant impact.

5 Results

We evaluate CAST on the RADCURE dataset, focusing on time-varying treatment effects, heterogeneity, and robustness.

As shown in Figure 2, chemotherapy benefit rises early, plateaus around 50–65 months, then declines—likely due to recurrence, toxicity, or competing risks. This indicates that chemotherapy is most impactful in the first few years post-treatment, with gradual tapering over time. On the test set, chemotherapy increased survival probability by $15.2 \pm 6.0\%$ at 36 months and $15.0 \pm 6.7\%$ at 60 months, with RMST gains of 3.6 ± 1.4 and 7.1 ± 2.6 months, respectively.

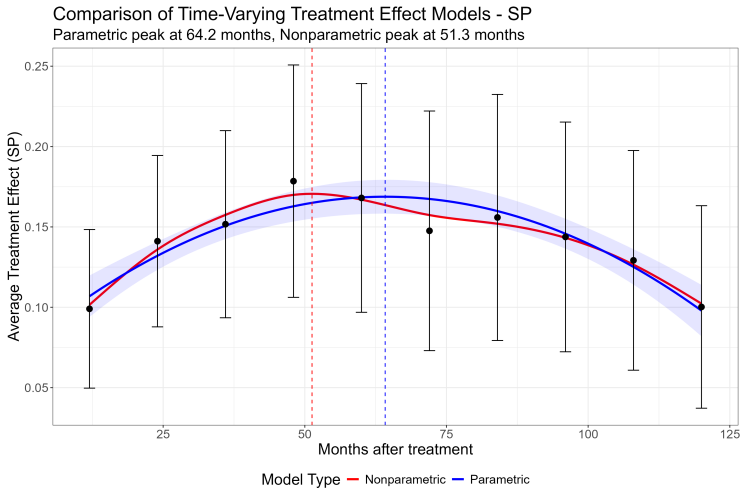
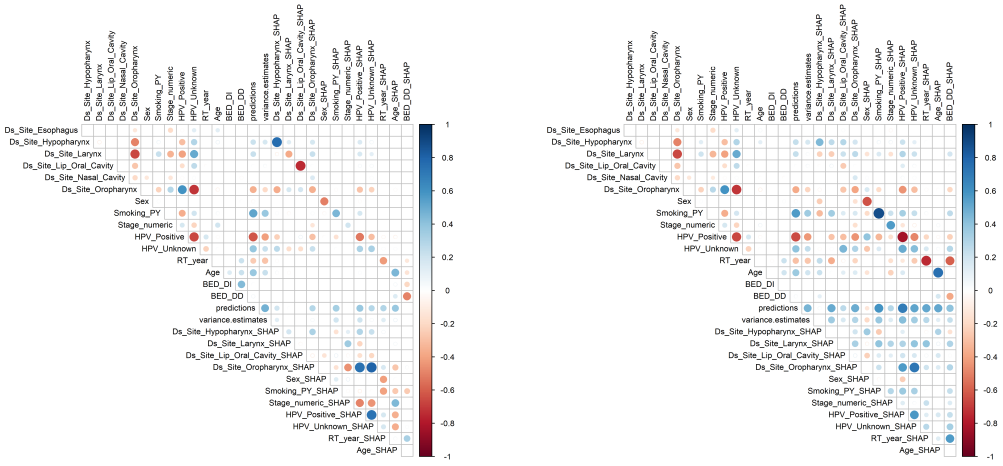


Figure 2: Comparison of time-varying treatment effect models using CAST. The red curve shows the parametric estimate with 95% CIs; the blue curve shows the non-parametric spline. Black dots denote average treatment effects $\pm$ standard errors on the survival probability scale.

Individualized effect distributions: Individual treatment effects varied across patients. While most experienced moderate benefit, CAST identified a long right tail of high responders and a subset with near-zero or negative effects.

Subgroup variation: Correlation matrices and SHAP analyses identified smoking pack-years as the strongest negative predictor of chemotherapy benefit, with HPV-negative patients showing greater benefit. Additional SHAP discussions are provided in Appendix C.2.



(a) Pearson Correlation

(b) Spearman Correlation

Figure 3: Correlation matrices between covariates, SHAP values, and treatment effects

Robustness & Effect Heterogeneity: CAST passed validation checks, including dummy outcome tests, synthetic confounders, and trimming sensitivity. Only strong confounder–treatment correlations

distorted estimates while weaker ones had little effect. The CSF largely ignored the noise variables. Individualized effects showed a long right tail of high responders and a subset with near-zero or negative benefit, showing potential for personalized treatment modeling. Additional visualizations are provided in Appendix C.4.

6 Discussion

The patterns uncovered by CAST have important clinical implications. The observed peak in survival benefit around 48 to 60 months post-treatment suggests that chemotherapy is most effective for short to mid-term local control but may not sustain long-term survival. This decline could reflect tumor repopulation, distant progression, or delayed toxicity [42]. However, since fewer patients remained at risk (did not experience a death or censoring event) at longer follow-up times, reliability of the causal effect estimates at long times is reduced compared with shorter times, as shown by our dummy tests.

These findings support the value of adaptive monitoring and adjunct strategies to extend therapeutic benefit. The heterogeneity revealed by CAST emphasizes the need for treatment personalization. Correlation and SHAP-based analysis together identified HPV positivity and smoking as the most influential factors. Favorable outcomes in HPV-positive patients align with known radiosensitivity and impaired DNA repair, while smoking was linked to reduced benefit—consistent with mechanisms like tumor hypoxia and immunosuppression. Age also showed a modest effect, with older patients generally benefiting more; an inflection point around 50–60 years may be clinically meaningful (Figure 3 and Figure 4 in Appendix C.2). In contrast, tumor site and TNM stage had limited influence on treatment effect heterogeneity, despite their prognostic relevance.

These findings align with efforts to tailor treatment by biologic subgroup. CAST offers a data-driven framework to support such stratifications and generate hypotheses for future trials. Rather than replacing existing tools, it complements them by modeling continuous-time dynamics and revealing patient-level variation. More broadly, this study shows how combining mechanistic modeling with causal machine learning can enhance the analysis of observational data. By embedding radiobiological insight into CAST using BED variants from different tumor repopulation models, we uncover treatment effects that align with known biology while also revealing discrepancies, such as stronger chemotherapy benefits than reported in prior meta-analyses. This offers a powerful way to complement clinical trials and generate new hypotheses.

Limitations and Broader Impacts

- **Data limitations:** The dataset exhibits substantial right-censoring: while 88.9% of patients remain in follow-up at one year, only 22.2% do so by year six. This may bias long-term survival estimates and obscure treatment effects that manifest later in time.
- **External validity:** The data come from a single institution (University Health Network, Toronto) and are predominantly male (80%), limiting generalizability to broader populations, especially women.
- **Causal assumptions:** Like all causal inference methods, CAST relies on the assumption of no unmeasured confounding. Important factors such as diet, lifestyle, or genetic risk—potentially related to both treatment and outcome—are not included.
- **Methodological scope:** From a machine learning perspective, CAST supports only binary treatment variables. Extending it to model continuous dosing, multi-arm comparisons, or longitudinal interventions remains an important direction for future work.

7 Conclusion

In this paper, we present CAST, a framework for modeling time-varying treatment effects in causal survival analysis using parametric and non-parametric methods. CAST extends causal survival forests to continuous-time modeling, estimating individualized treatment paths and highlighting effect peaks and declines. Applied to chemotherapy for HNSCC, CAST produces robust and interpretable insights, supporting personalized and adaptive care. Beyond cancer, CAST applies to settings with evolving treatment effects—such as infectious disease interventions—to pinpoint critical windows, tailor care, and adapt strategies as evidence grows.

References

[1] M. L. Welch, S. Kim, A. Hope, S. H. Huang, Z. Lu, J. Marsilla, M. Kazmierski, K. Rey-McIntyre, T. Patel, B. O'Sullivan, J. Waldron, J. Kwan, J. Su, L. Soltan Ghoraie, H. B. Chan, K. Yip, M. Giuliani, Neck Site Group Princess Margaret Head, S. Bratman, and T. Tadic. Computed tomography images from large head and neck cohort (radcure) (version 4). *The Cancer Imaging Archive*, 2023. doi: 10.7937/J47W-NM11.

[2] M. Hung, J. Bounsanga, and M. W. Voss. Interpretation of correlations in clinical research. *Postgraduate Medicine*, 129(8):902–906, November 2017. doi: 10.1080/00325481.2017.1383820.

[3] H. A. Miot. Correlation analysis in clinical and experimental studies. *Jornal Vascular Brasileiro*, 17(4):275–279, December 2018. doi: 10.1590/1677-5449.174118.

[4] K. Shiba and K. Inoue. Harnessing causal forests for epidemiologic research: key considerations. *American Journal of Epidemiology*, 193(6):813–818, June 2024. doi: 10.1093/aje/kwae003.

[5] A. Venkatasubramaniam, B. A. Mateen, B. M. Shields, A. T. Hattersley, A. G. Jones, S. J. Vollmer, and J. M. Dennis. Comparison of causal forest and regression-based approaches to evaluate treatment effect heterogeneity: an application for type 2 diabetes precision medicine. *BMC Medical Informatics and Decision Making*, 23(1):110, June 2023. doi: 10.1186/s12911-023-02207-2.

[6] G. Solana-Lavalle, M. D. Cusimano, T. Steeves, R. Rosas-Romero, and P. N. Tyrrell. Causal forest machine learning analysis of parkinson's disease in resting-state functional magnetic resonance imaging. *Tomography*, 10(6):894–911, June 2024. doi: 10.3390/tomography10060068.

[7] Yifan Cui, Michael R. Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society: Series B*, 85(2):380–403, 2023. doi: 10.1093/jrsssb/qkac020.

[8] I. Shuryak, E. Wang, and D. J. Brenner. Understanding the impact of radiotherapy fractionation on overall survival in a large head and neck squamous cell carcinoma dataset: A comprehensive approach combining mechanistic and machine learning models. *Frontiers in Oncology*, 14:1422211, August 2024. doi: 10.3389/fonc.2024.1422211.

[9] Z. Huang, N. A. Mayr, M. Gao, S. S. Lo, J. Z. Wang, G. Jia, and W. T. C. Yuh. The onset time of tumor repopulation for cervical cancer: first evidence from clinical data. *International Journal of Radiation Oncology*Biology*Physics*, 84(2):478–484, October 2012. doi: 10.1016/j.ijrobp.2011.12.037.

[10] C. Petersen and F. Würschmidt. Late toxicity of radiotherapy: a problem or a challenge for the radiation oncologist? *Breast Care (Basel)*, 6(5):369–374, October 2011. doi: 10.1159/000334220.

[11] I. Shuryak, E. J. Hall, and D. J. Brenner. Dose dependence of accelerated repopulation in head and neck cancer: Supporting evidence and clinical implications. *Radiotherapy and Oncology*, 127(1):20–26, April 2018. doi: 10.1016/j.radonc.2018.02.015.

[12] D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, and J. R. Grandis. Head and neck squamous cell carcinoma. *Nature Reviews Disease Primers*, 6(92):1–22, November 2020. doi: 10.1038/s41572-020-00224-3.

[13] M. E. Sabatini and S. Chiocca. Human papillomavirus as a driver of head and neck cancers. *British Journal of Cancer*, 122(3):306–314, February 2020. doi: 10.1038/s41416-019-0602-7.

[14] D. C. Beachler and G. D'Souza. Nuances in the changing epidemiology of head and neck cancer. *Oncology (Williston Park)*, 24(10):924–926, September 2010. PMID: 21138173.

[15] G. M. P. van Kempen, R. J. Baatenburg de Jong, and R. J. H. Borra. Hpv and head and neck cancers: Towards early diagnosis and prevention. *Oral Oncology*, 128:105214, September 2022. doi: 10.1016/j.oraloncology.2022.105214.

[16] Janet Tu. How long does it take chemotherapy to shrink tumors? *Cancerwise, MD Anderson Cancer Center*, 2024. https://www.mdanderson.org/cancerwise/how-long-does-it-take-chemotherapy-to-shrink-tumors.h00-159696756.html.

[17] UCSF Health. Coping with chemotherapy. *Patient Education, UCSF Health*, 2025. https://www.ucsfhealth.org/education/coping-with-chemotherapy.

[18] S. R. Rathod, S. Gupta, S. Ghosh-Laskar, V. Murthy, A. Budrukkar, J. Agarwal, and K. Kannan. Quality-of-life (qol) outcomes in patients with head and neck squamous cell carcinoma treated with intensity-modulated radiation therapy (imrt) compared to three-dimensional conformal radiotherapy (3d-crt): Evidence from a prospective randomized study. *Oral Oncology*, 49(6): 634–640, June 2013. doi: 10.1016/j.oraloncology.2013.02.013.

[19] A. Viganò, F. De Felice, N. A. Iacovelli, D. Alterio, R. Ingargiola, A. Casbarra, N. Facchinetti, O. Oneta, A. Bacigalupo, E. Tornari, S. Ursino, F. Paiar, O. Caspiani, A. Di Rito, D. Musio, P. Bossi, P. Steca, B. A. Jereczek-Fossa, L. Caso, N. Palena, A. Greco, and E. Orlandi. Quality of life changes over time and predictors in a large head and neck patients' cohort: secondary analysis from an italian multi-center longitudinal, prospective, observational study—a study of the italian association of radiotherapy and clinical oncology (airo) head and neck working group. *Supportive Care in Cancer*, 31(4):220, March 2023. doi: 10.1007/s00520-023-07661-2.

[20] R. Yang, A. C. Freeman-Cook, H. C. Kurnik, and D. C. Kirouac. Dissecting variability in responses to cancer chemotherapy through systems pharmacology. *Clinical Pharmacology & Therapeutics*, 88(1):34–38, July 2010. doi: 10.1038/clpt.2010.96.

[21] Y. Sun, Z. Wang, S. Qiu, and R. Wang. Therapeutic strategies of different hpv status in head and neck squamous cell carcinoma. *International Journal of Biological Sciences*, 17(4):1104–1118, March 2021. doi: 10.7150/ijbs.58077.

[22] N. Jiang, Y. Wu, and C. Li. Limitations of using cox proportional hazards model in cardiovascular research. *Cardiovascular Diabetology*, 23(219), June 2024. doi: 10.1186/s12933-024-02302-2.

[23] L. Xu, S. Jiang, T. Li, and Y. Xu. Limitations of the cox proportional hazards model and alternative approaches in metachronous recurrence research. *Gastric Cancer*, 27(6):1348–1349, November 2024. doi: 10.1007/s10120-024-01554-x.

[24] S. Saha. Survival analysis with bayesian additive regression trees and its application. https://huskiecommons.lib.niu.edu/allgraduate-thesesdissertations/5158/, 2017. Northern Illinois University Thesis.

[25] F. Zhai, S. Mu, Y. Song, M. Zhang, C. Zhang, and Z. Lv. A random survival forest model for predicting residual and recurrent high-grade cervical intraepithelial neoplasia in premenopausal women. *International Journal of Women's Health*, 16:1775–1787, October 2024. doi: 10.2147/IJWH.S485515.

[26] K. Matsuo, S. Purushotham, B. Jiang, R. S. Mandelbaum, T. Takiuchi, Y. Liu, and L. D. Roman. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *American Journal of Obstetrics and Gynecology*, 220(4):381.e1–381.e14, April 2019. doi: 10.1016/j.ajog.2018.12.030.

[27] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

[28] L. Wen, J. G. Young, J. M. Robins, and M. A. Hernán. Parametric g-formula implementations for causal survival analyses. *Biometrics*, 77(2):740–753, June 2021. doi: 10.1111/biom.13321.

[29] D. Lee, S. Yang, and X. Wang. Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *Journal of Causal Inference*, 10(1):415–440, December 2022. doi: 10.1515/jci-2022-0004.

[30] Erik Sverdrup and Stefan Wager. Treatment heterogeneity with right-censored outcomes using grf. *ASA Lifetime Data Science Newsletter*, 2024. arXiv:2312.02482.

[31] Y. Zhang, N. Kreif, V. S. Gc, and A. Manca. Machine learning methods to estimate individualized treatment effects for use in health technology assessment. *Medical Decision Making*, 44 (7):756–769, October 2024. doi: 10.1177/0272989X241263356.

[32] V. Chernozhukov, C. Hansen, N. Kallus, M. Spindler, and V. Syrgkanis. Applied causal inference powered by ml and ai. *arXiv preprint*, arXiv:2403.02467, March 2024. doi: 10.48550/arXiv.2403.02467.

[33] J. Sun and F. W. Crawford. The role of discretization scales in causal inference with continuous-time treatment. *arXiv preprint*, June 2023. doi: 10.48550/arXiv.2306.08840.

[34] A. Curth, C. Lee, and M. W. van der Laan. Survite: Learning heterogeneous treatment effects from time-to-event data. *arXiv preprint*, October 2021. doi: 10.48550/arXiv.2110.14001.

[35] W. J. Allard and L. W. M. M. Terstappen. Ccr 20th anniversary commentary: Paving the way for circulating tumor cells. *Clinical Cancer Research*, 21(13):2883–2885, July 2015. doi: 10.1158/1078-0432.CCR-14-2559.

[36] J. A. Langendijk, P. Doornaert, I. M. Verdonck de Leeuw, C. R. Leemans, N. K. Aaronson, and B. J. Slotman. Impact of late treatment-related toxicity on quality of life among patients with head and neck cancer treated with radiotherapy. *Journal of Clinical Oncology*, 26(22): 3770–3776, August 2008. doi: 10.1200/JCO.2007.14.6647.

[37] A. F. Brouwer, R. Meza, M. C. Eisenberg, C. H. Chapman, M. C. He, S. B. Chinn, A. M. Mondul, M. Banerjee, M. Ryser, and J. M. Taylor. Time-varying survival effects for squamous cell carcinomas at oropharyngeal and nonoropharyngeal head and neck sites in the united states, 1973–2015. *Cancer*, 126(23):5137–5146, December 2020. doi: 10.1002/cncr.33110.

[38] E. K. Roberts, L. Luo, A. M. Mondul, M. Banerjee, C. M. Veenstra, A. B. Mariotto, M. J. Schipper, K. He, J. M. G. Taylor, and A. F. Brouwer. Time-varying associations of patient and tumor characteristics with cancer survival: an analysis of seer data across 14 cancer sites, 2004–2017. *Cancer Causes & Control*, 35(10):1393–1405, May 2024. doi: 10.1007/s10552-024-01888-y.

[39] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68, January 2018. doi: 10.1093/ectj/uty017.

[40] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, July 2018. doi: 10.1080/01621459.2017.1319839.

[41] S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *Annals of Statistics*, 47(2): 1148–1178, April 2019. doi: 10.1214/18-AOS1709.

[42] I. Shuryak, E. J. Hall, and D. J. Brenner. Optimized hypofractionation can markedly improve tumor control and decrease late effects for head and neck cancer. *International Journal of Radiation Oncology, Biology, Physics*, 104(2):272–278, June 2019. doi: 10.1016/j.ijrobp.2019.02.025.

Ethics Statement

Existing at the intersection of machine learning (ML), healthcare, and causal inference, our work inevitably raises ethical considerations. By bringing ML methods to oncology research, we strive to advance personalized medicine and treatment strategies. However, our estimates are based on observational data and may be biased by unmeasured confounding. While the dataset includes a comprehensive description of variables including age, sex, smoking history, and HPV status, it omits race, ethnicity, and socioeconomic status data. These factors are key to understanding structural barriers to healthcare that could possibly affect outcomes. This risks amplifying existing biases in the data. ML models in oncology must be used cautiously and should not replace clinical judgment, but rather act as a supplement. Our findings require further clinical validation before integration into decision-making workflows.

A Theoretical Justification of CAST

We provide formal justification for the consistency and identifiability of the time-varying treatment effect estimator $\hat{\tau}(t)$ used in the CAST framework.

A.1 Problem Setting

Let $\mathcal{D} = \{(X_i, W_i, T_i, \delta_i)\}_{i=1}^n$ be a dataset of n i.i.d. samples where: - $X_i \in \mathbb{R}^p$ is a vector of observed covariates, - $W_i \in \{0, 1\}$ is a binary treatment indicator, - T_i is the observed event or censoring time, - $\delta_i \in \{0, 1\}$ is the event indicator (1 if the event occurred, 0 if censored).

Let $Y(w, t)$ denote the potential outcome (e.g., survival status at time t) under treatment $w \in \{0, 1\}$.

We define the time-varying Conditional Average Treatment Effect (CATE) as:

$$\tau(x, t) := \mathbb{E}[Y(1, t) - Y(0, t) \mid X = x].$$

CAST estimates $\tau(x, t)$ using a doubly-robust causal survival forest followed by a spline or quadratic fit across time.

A.2 Assumptions

We adopt standard causal inference and survival analysis assumptions:

- (A1) **Unconfoundedness:** $(Y(0, t), Y(1, t)) \perp W \mid X$ for all t.
- (A2) **Positivity:** $0 < P(W = 1 \mid X) < 1$ almost surely.
- (A3) **Consistency:** $Y = Y(W, t)$ if W is received.
- (A4) **Non-informative Censoring:** $C \perp (Y(0, t), Y(1, t)) \mid X, W$ for censoring time C.
- (A5) **Consistency of Forest Estimators:** The causal survival forests used yield consistent estimates of conditional survival functions $S_w(t \mid X)$.

A.3 Theorem: Pointwise Consistency of $\hat{\tau}(t)$

[Pointwise Consistency] Under assumptions (A1)–(A5), for each fixed t:

$$\hat{\tau}(t) := \mathbb{E}_X[\hat{S}_1(t \mid X) - \hat{S}_0(t \mid X)] \xrightarrow{p} \tau(t) := \mathbb{E}_X[S_1(t \mid X) - S_0(t \mid X)]$$

as $n \to \infty$, where $\hat{S}_w(t \mid X)$ is the estimated conditional survival function under treatment w from causal survival forests.

This follows from: 1. Consistency of $\hat{S}_w(t \mid X)$ (A5), 2. The continuous mapping theorem, since subtraction and expectation are continuous, 3. Trimming enforces overlap (A2), ensuring bounded inverse propensity weights.

A.4 Identifiability of $\tau(t)$ from Observational Data

[Identifiability] Under assumptions (A1)–(A4), the marginal time-varying treatment effect

$$\tau(t) := \mathbb{E}_X[\mathbb{E}[Y \mid W = 1, X, T \geq t] - \mathbb{E}[Y \mid W = 0, X, T \geq t]]$$

is identified from observational data using inverse probability weighting or doubly-robust estimation.

Under unconfoundedness and non-informative censoring, we can consistently estimate the conditional means $\mathbb{E}[Y(w, t) \mid X]$ from observed data. The difference in conditional expectations across treatment groups yields an identifiable estimator of $\tau(t)$.

A.5 Estimability of Peak Effect Time in CAST-Parametric

Let the parametric effect trajectory be:

$$\tau(t) = \beta_0 + \beta_1 t + \beta_2 t^2,$$

and suppose $\hat{\beta}_1, \hat{\beta}_2$ are estimated using weighted least squares.

[Consistency of Estimated Peak Time] If $\hat{\beta}_1 \xrightarrow{p} \beta_1$, $\hat{\beta}_2 \xrightarrow{p} \beta_2$ with $\beta_2 < 0$, then the estimated peak time

$$\hat{t}^* = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$$

is a consistent estimator of the true peak $t^* = -\frac{\beta_1}{2\beta_2}$.

This follows from Slutsky's theorem. Since both $\hat{\beta}_1$ and $\hat{\beta}_2$ converge in probability to non-zero limits, and the mapping $f(a, b) = -a/(2b)$ is continuous for $b \neq 0$, it follows that:

$$\hat{t}^* = -\frac{\hat{\beta}_1}{2\hat{\beta}_2} \xrightarrow{p} -\frac{\beta_1}{2\beta_2} = t^*.$$

B Expanded Dataset Subsection

Overview

Our analysis uses the RADCURE dataset from The Cancer Imaging Archive (TCIA), the largest to our knowledge publicly accessible head and neck cancer imaging dataset. The data spans from 2005 to 2017 and includes computed tomography (CT) images for 3,346 patients, from which we selected a subset of 2,651 patients after filtering for only HNSCC cases. These images are linked to clinical, demographic, and treatment metadata. Following standardized clinical imaging protocols, the RADCURE project includes CT images, pictured alongside manually-reviewed contours differentiating between the planning tumor volume (PTV) and the organs at risk (OARs). All patients in this dataset received radiotherapy, and some received chemotherapy.

The clinical data accounts for patient demographics, including age, gender, and HPV status. It also details tumor staging using the 7th edition TNM system to describe the cancer, in addition to treatment information. While the dataset primarily focuses on oropharyngeal cancer, it also covers laryngeal, nasopharyngeal, and hypopharyngeal cancers.

Data Preprocessing

In the preprocessing stage, we filtered out incomplete patient profiles to ensure the dataset included relevant variables and appropriately represented potential confounders. We standardized all continuous variables to have zero mean and unit variance to ensure comparability and optimize model performance. The dataset comprehensively describes treatment details—dose/fraction, number of fractions, and total days of radiotherapy—which we used to calculate Biologically Effective Dose (BED) values. We implemented both dose-independent (DI) and dose-dependent (DD) BED models to capture the biological effects of radiation therapy, using established radiobiological parameters ($\alpha = 0.2$ Gy^{-1}, $\alpha/\beta = 10$ Gy, accelerated repopulation rates and onset times). This allowed us to quantify the effective radiation dose accounting for different fractionation schedules. We employed a stratified data partitioning strategy, creating training (75%) and testing (25%) sets while maintaining consistent event rates across partitions. Both subsets contained similar proportions of survival events, allowing for unbiased evaluation of treatment effects.

Table 1 summarizes the estimated average treatment effects across time for both restricted mean survival time (RMST) and survival probability (SP) metrics. These values were computed using causal survival forests on held-out test data. We observe that the estimated effects generally increase with longer follow-up, particularly under the RMST metric, reflecting the accumulating benefit of treatment over time. Standard errors are included to reflect model uncertainty at each horizon.

Table 1: Summary statistics of the real dataset

Statistic	**Control Group**	**Treated Group**
Event Rate (%)	79.8	
Treatment Rate (%)	44.9	
Median Survival (months)	17.0	24.0
12-month Survival (%)	70.3	90.1
24-month Survival (%)	20.2	45.5
36-month Survival (%)	1.9	7.3
48-month Survival (%)	0.0	0.1
Age (mean)	60.42	59.23
TNM Stage (mean)	1.73	3.46
HPV Positivity Rate	0.68	0.51
Sex (Male = 1)	0.48	0.49

Computing Resources: All experiments were conducted with a 13th Gen Intel Core i7-1355U CPU, 16GB RAM, and integrated Intel Iris Xe Graphics. No discrete GPU or cloud resources were used, though such resources would significantly reduce runtime for large-scale extensions of this work.

C Additional Results

In this section, we present additional results that extend and validate the findings reported in the main paper. These include visualizations of treatment effect heterogeneity across time, a summary of average treatment effects, and robustness checks to support the reliability of our causal estimates.

C.1 Summary Table of Average Treatment Effects

Table 2 summarizes the estimated average treatment effects across time horizons using both RMST and survival probability metrics. These values were computed using causal survival forests on the held-out test set. The treatment effects tend to increase over time under both metrics, with RMST showing a steeper upward trend reflecting cumulative benefit. Standard errors are included for each estimate. The early rise in both SP and RMST suggests initial treatment efficacy, while the plateauing in later months reflects diminishing returns, possibly due to recurrence or late toxicity. The RMST gains—peaking at over 16 months—highlight how cumulative survival benefit continues to accrue even as survival probability differences taper off. These patterns support the biological intuition that treatment effects rise quickly post-intervention and then gradually attenuate.

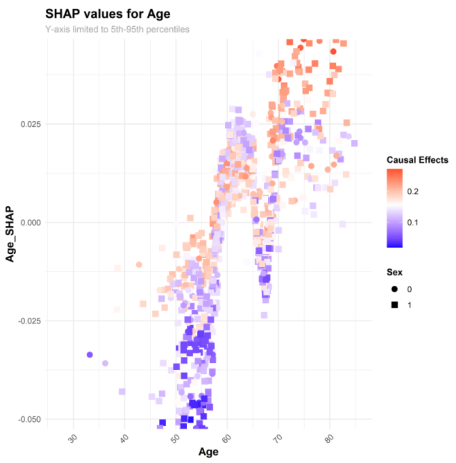
Table 2: Estimated average treatment effects (ATE) across time using RMST and survival probability (SP). SE represent standard errors

Months	ATE (SP)	SE (SP)	ATE (RMST)	SE (RMST)
12	0.099	0.049	0.44	0.26
24	0.141	0.053	1.88	0.80
36	0.152	0.058	3.58	1.46
48	0.178	0.072	5.80	2.31
60	0.168	0.071	7.39	2.73
72	0.148	0.075	8.38	3.52
84	0.156	0.077	11.08	4.76
96	0.143	0.071	13.89	5.90
108	0.129	0.068	14.76	6.16
120	0.100	0.063	16.11	6.92

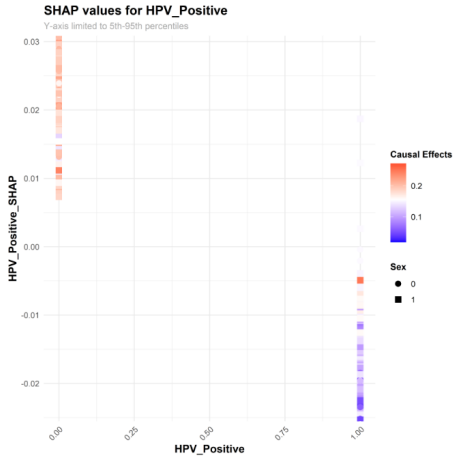
These summary statistics also inform the CAST modeling strategies described in Section 3.3. The steady increase followed by tapering motivates the use of both quadratic and spline-based approaches to flexibly capture the full temporal arc of treatment efficacy.

C.2 SHAP-Based Interpretability Analysis

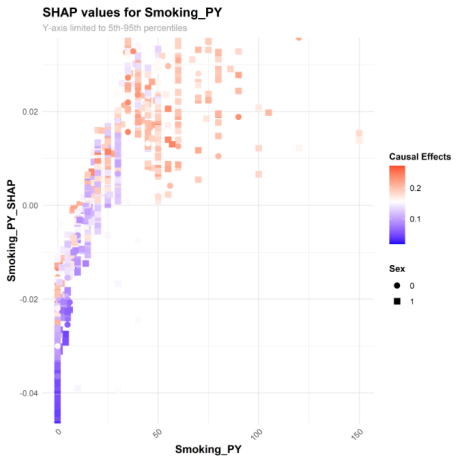
While SHAP provides valuable insights into feature influence, the estimates generated here using the fastshap R package are approximate and may be noisy, particularly in the context of survival analysis. We calculated approximate SHAP values because an exact SHAP explainer does not yet exist for the causal survival forest model. Figures 4(a–c) show SHAP plots for the three most influential variables—age, HPV status, and smoking pack-years—highlighting clear heterogeneity in treatment benefit across subgroups. Additional SHAP plots for other covariates—such as tumor site, treatment timing, dose metrics, and TNM stage—are also provided below. These variables had smaller contributions to the model, but are shown for completeness and transparency.



(a) Age

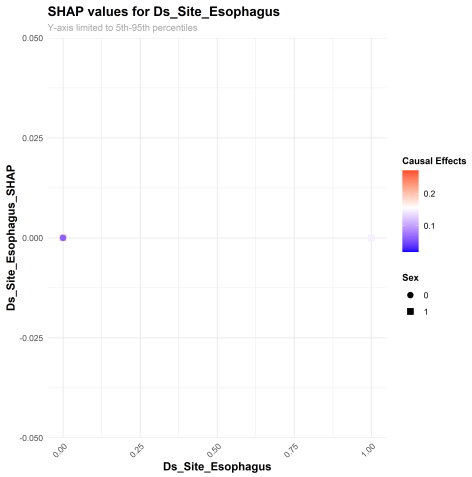


(b) HPV Status

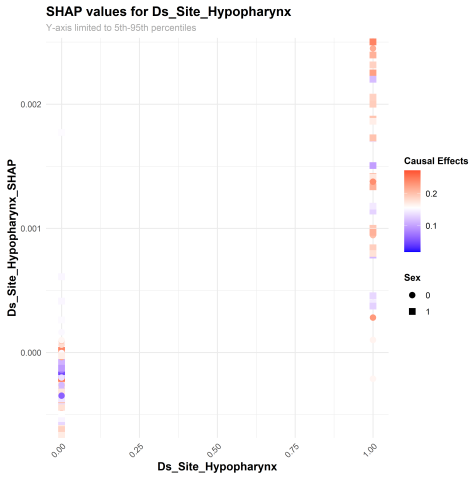


(c) Smoking Pack-Years

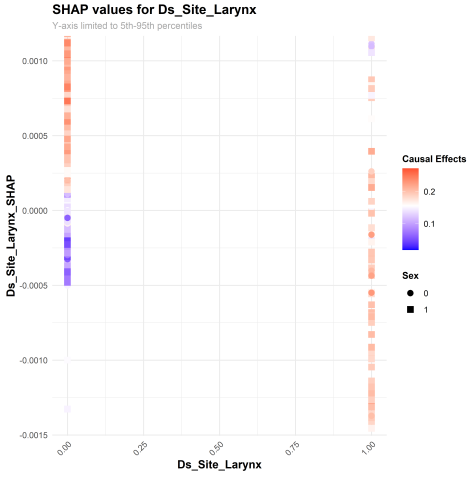
Figure 4: SHAP analysis of covariates driving treatment effect heterogeneity. (a) Older age is linked to greater chemotherapy benefit. (b) HPV-negative patients consistently show higher contributions. (c) Smoking history is positively associated with the chemotherapy benefit treatment.



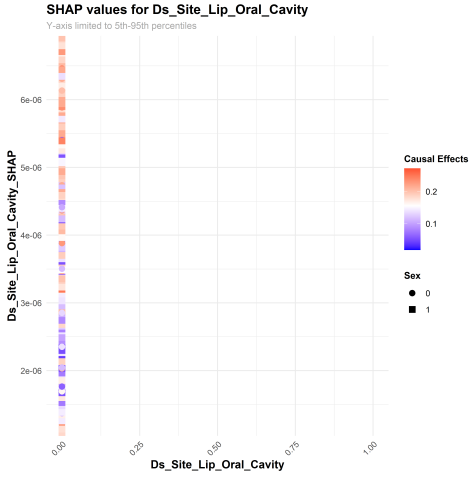
(a) Esophagus



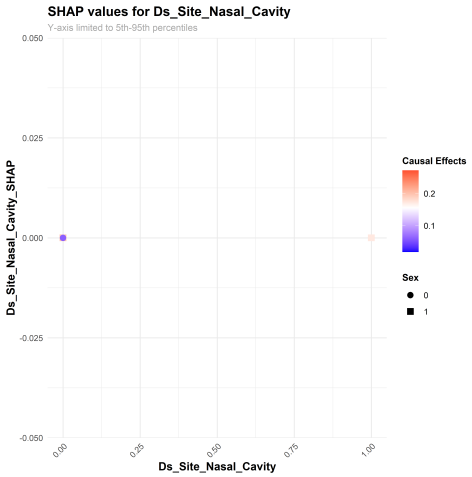
(b) Hypopharynx



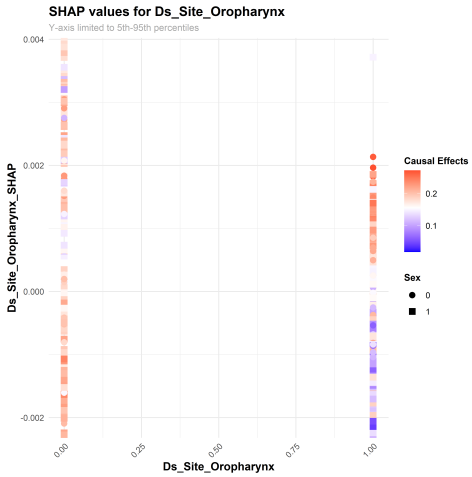
(c) Larynx



(d) Lip/Oral Cavity

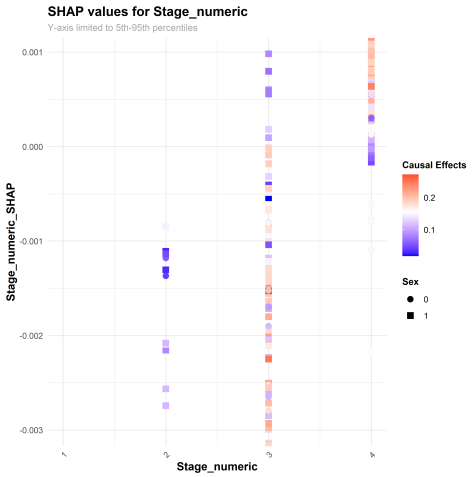


(e) Nasal Cavity

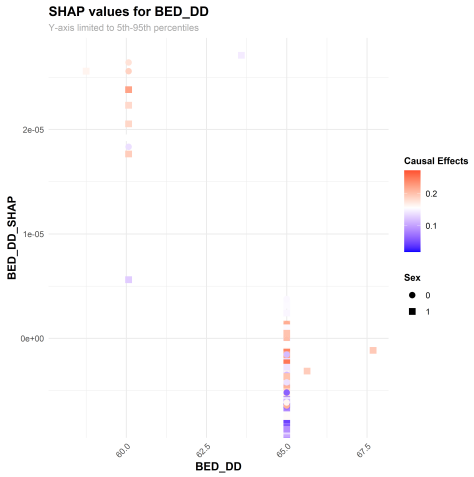


(f) Oropharynx

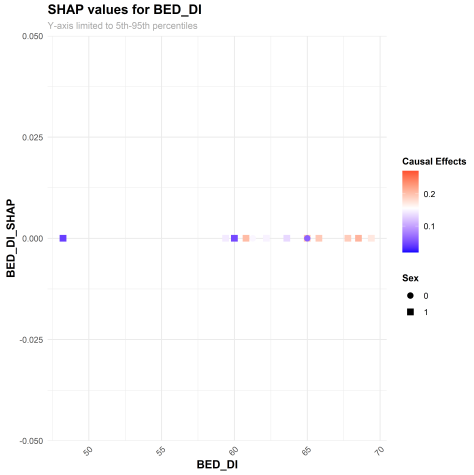
Figure 5: SHAP values for primary tumor site. These anatomical subgroups exhibited low or diffuse contributions to treatment effect heterogeneity, though subtle site-specific trends may still hold clinical value.



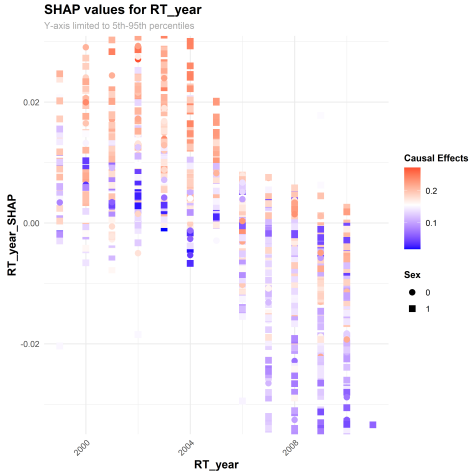
(a) TNM Stage



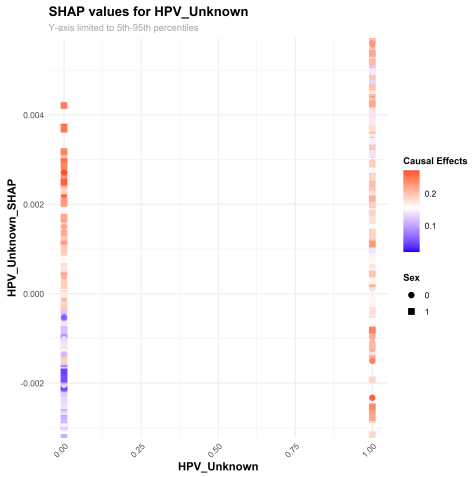
(b) BED (Dose-Dependent)



(c) BED (Dose-Independent)



(d) Year of RT



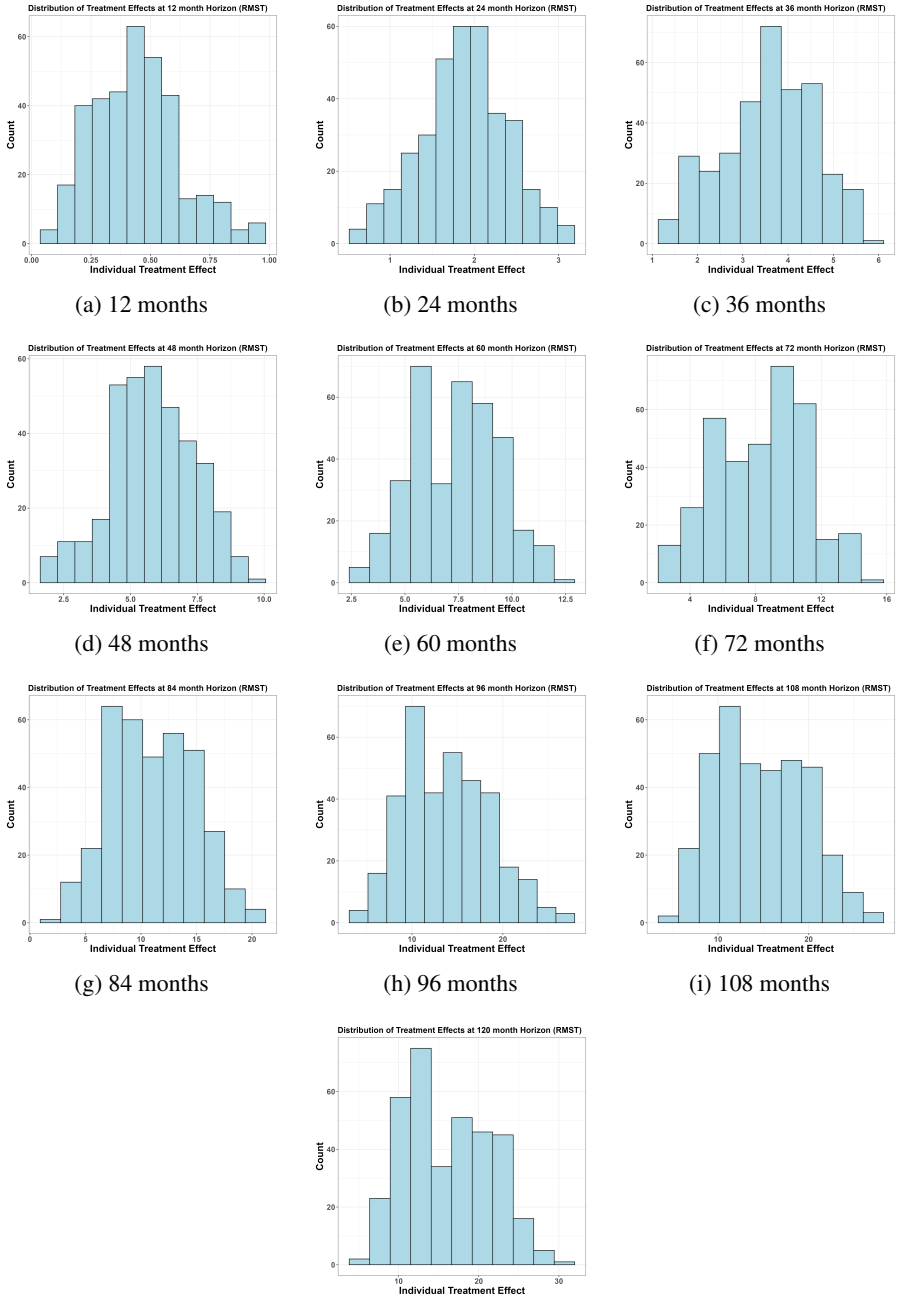
(e) HPV Unknown

Figure 6: SHAP values for additional covariates, including TNM stage, treatment year, and dose-related metrics. These features showed limited or context-specific contributions to treatment effect heterogeneity.

C.3 Distributions of Individualized Treatment Effects

We visualize the estimated treatment effect distributions for both RMST and survival probability (SP) at intervals ranging from 12 to 120 months. Figures 4 and 5 show individual-level causal effects derived from the causal survival forest at each time horizon.

RMST Treatment Effect Distributions



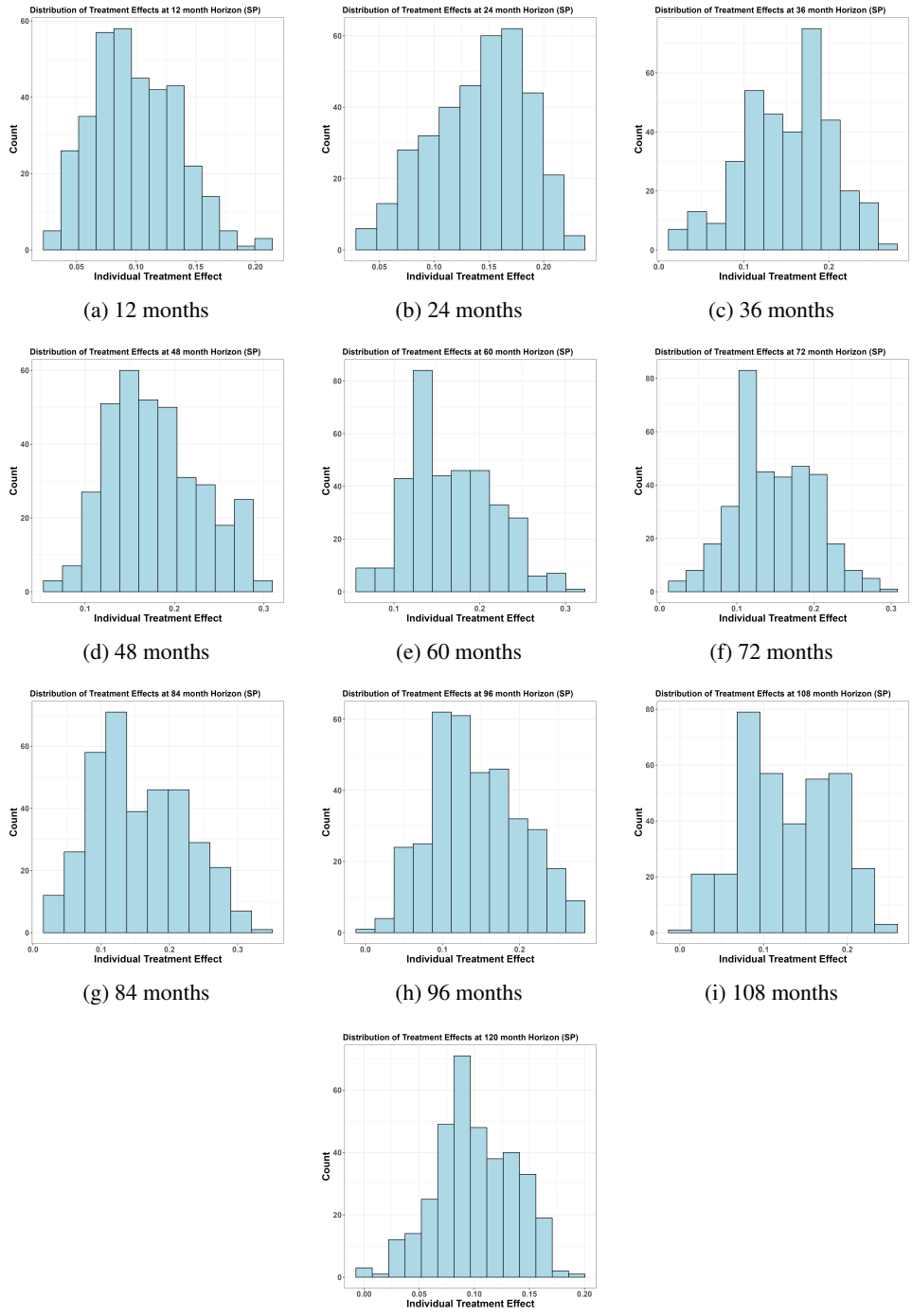
(a) 12 months (b) 24 months (c) 36 months

(d) 48 months (e) 60 months (f) 72 months

(g) 84 months (h) 96 months (i) 108 months

(j) 120 months

Figure 7: Distributions of estimated RMST-based treatment effects over time. Each panel shows the individual-level causal effect at a specific horizon as learned by the causal survival forest.



(a) 12 months (b) 24 months (c) 36 months

(d) 48 months (e) 60 months (f) 72 months

(g) 84 months (h) 96 months (i) 108 months

(j) 120 months

Figure 8: Distributions of estimated survival-probability-based treatment effects over time. Each panel shows the individual-level causal effect at a specific horizon as estimated by the causal survival forest.

C.4 Dummy Outcome Refutation Tests

To assess whether CAST detects spurious treatment effects in the absence of a true signal, we performed dummy outcome tests. For each time horizon, we randomly shuffled treatment assignments and outcome times across 20 repetitions to simulate a null setting. If the model was overfitting or improperly attributing causal structure, it would produce non-zero treatment effect estimates even under randomization. As shown in the boxplots below, the estimated treatment effects for both RMST and survival probability are centered around zero, especially at relatively short times ($\leq$ 60 months), when the number of patients still at risk was large. This confirms that CAST does not learn artifacts from the data and is robust to randomization of causal structure.

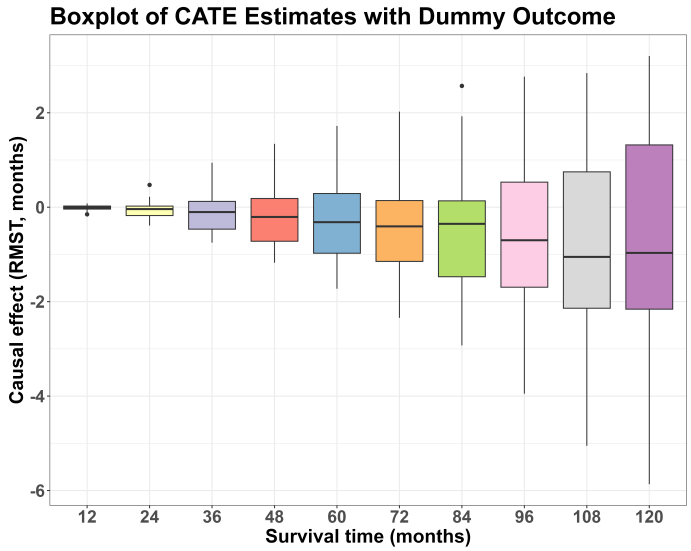


Figure 9: Dummy outcome test for RMST-based ATE estimates. Across 20 shuffles per horizon, treatment effects are centered near zero, consistent with the null.

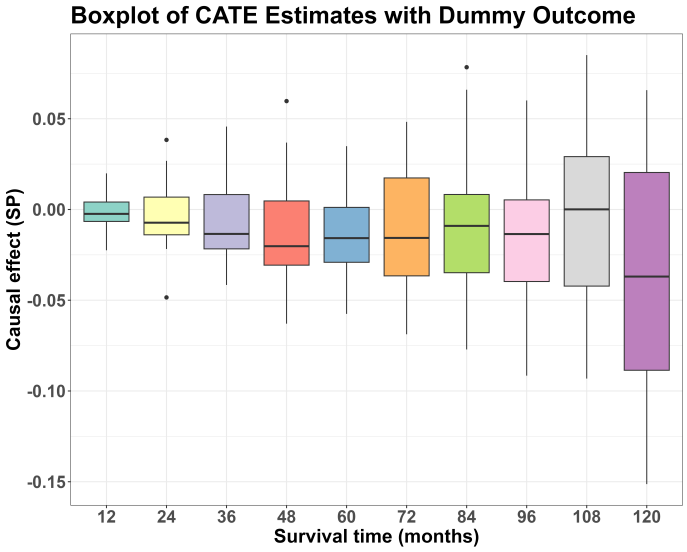


Figure 10: Dummy outcome test for survival probability-based ATE estimates. The model correctly reports no significant treatment effects under randomized labels.

To assess the robustness of CAST estimates to unobserved confounding, we performed a sensitivity analysis by injecting synthetic covariates with varying correlation to treatment assignment ($r = 0.1$, 0.3, 0.5). We then measured the resulting shifts in ATE estimates across time horizons for both RMST and survival probability outcomes.

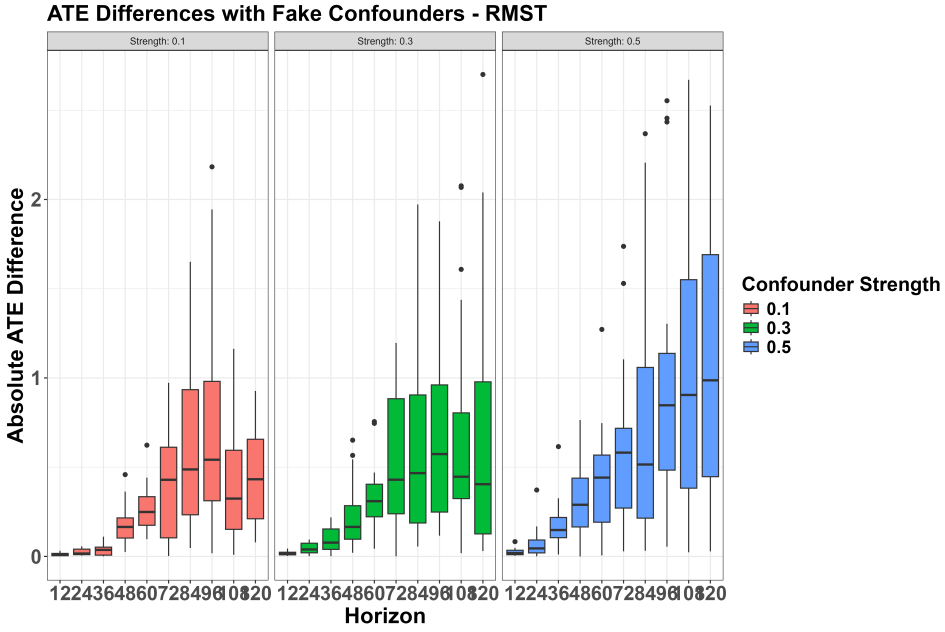


Figure 11: Absolute ATE differences in RMST under varying confounder strengths ($r = 0.1$, 0.3, 0.5). Estimates are stable under weak strengths but diverge at longer horizons and higher strengths.

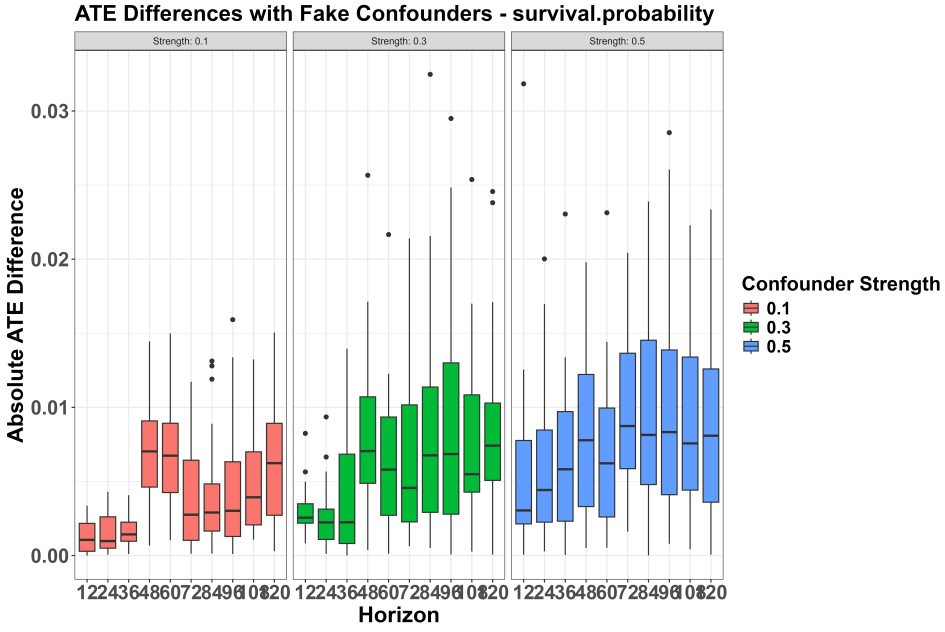


Figure 12: Absolute ATE differences in SP under varying confounder strengths ($r = 0.1$, 0.3, 0.5). CAST estimates remain stable under weak strengths, with modest shifts at stronger levels and longer horizons.