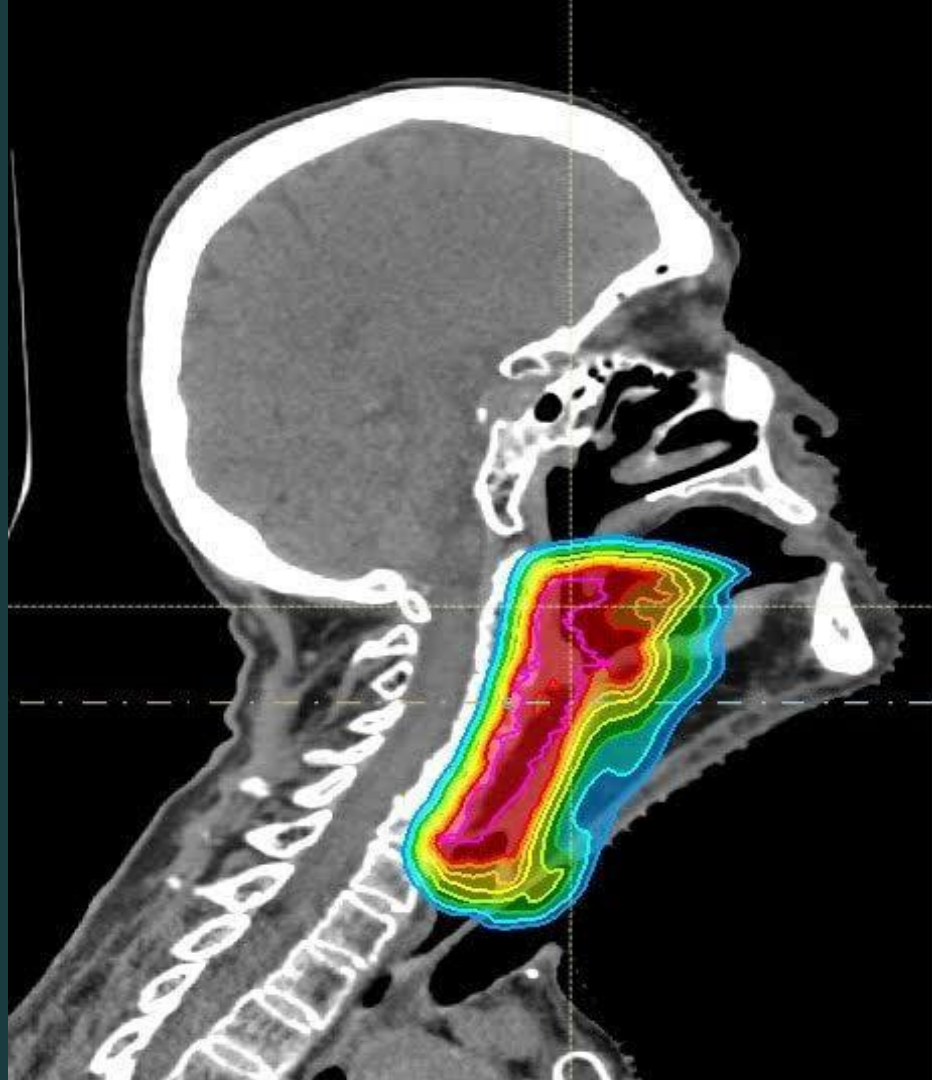
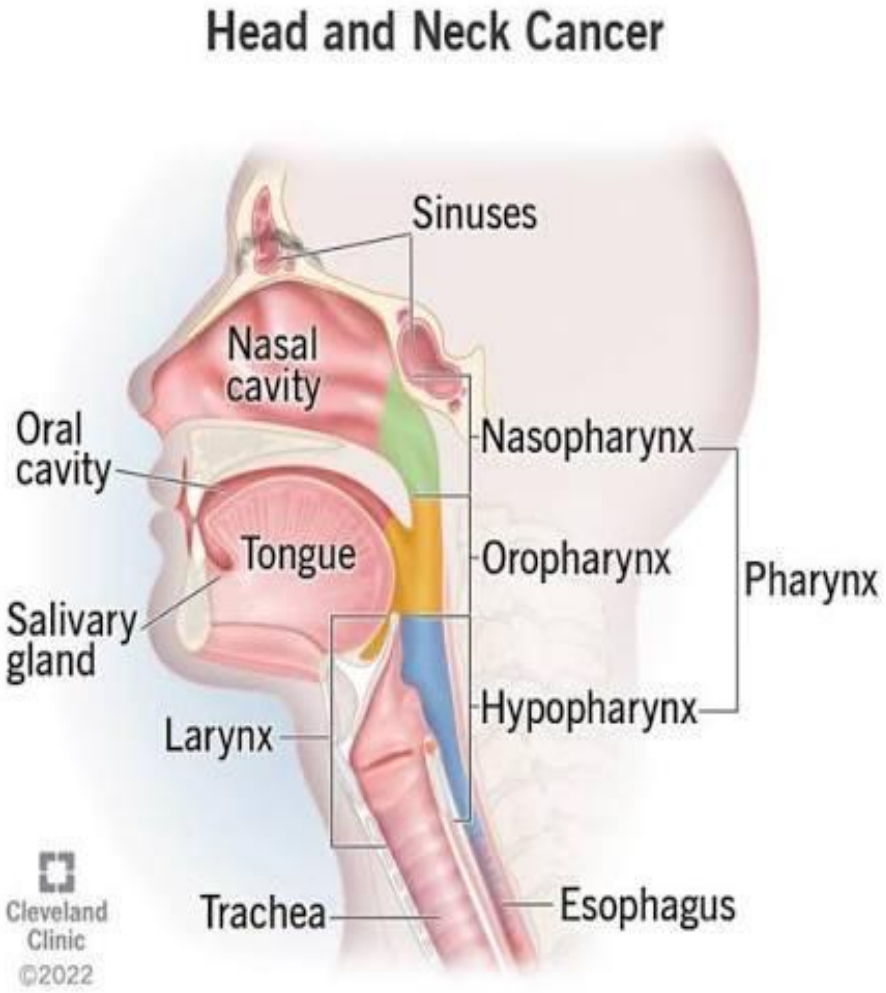
14/11/2025

TITLE:CAUSAL MACHINE LEARNING IN ONCOLOGY - CHEMOTHERAPY/RADIOTHERAPY HEAD AND NECK CANCER TREATMENT RESPONSE

Does adding chemotherapy to radiotherapy improve survival for real-world patients for whom does it help or harm at what time ?





Head and Neck Cancers(HNC)

Cancers that affect your mouth, throat or other parts of your head and neck

N

Head and Neck Cancers (HNCs) account for > 900,000 cases and 390,000 deaths annually worldwide(Iwatsubo et al, 2019).

N

- Radiotherapy (RT) is a foundational treatment.
- many patients receive chemoradiotherapy (ChemoRT), adding systemic chemotherapy to improve tumor control.

Problem Statement

THE CHALLENGE

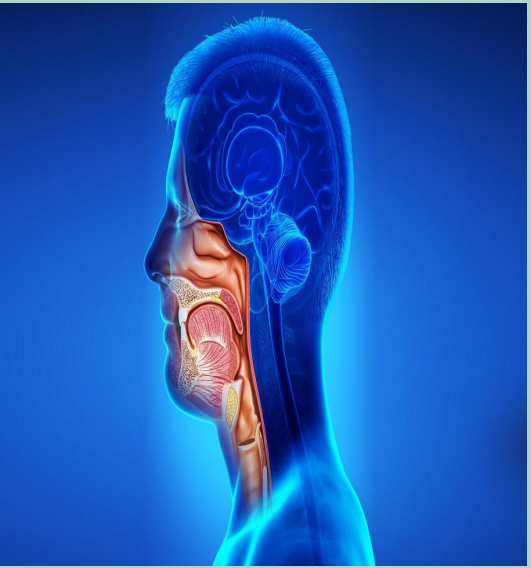
- Management of head and neck cancers—especially HPV-associated OPSCC—varies widely across centers,
- Randomized trials are scarce or outdated → clinicians rely on real-world data.
- Observational data are confounded → standard ML fails to estimate causal effects.

Content *overview*

Causal Machine Learning

Causal Machine Learning enables us to reason about the effects of changes to this process (interventions) and what would have happened in hindsight (counterfactuals)

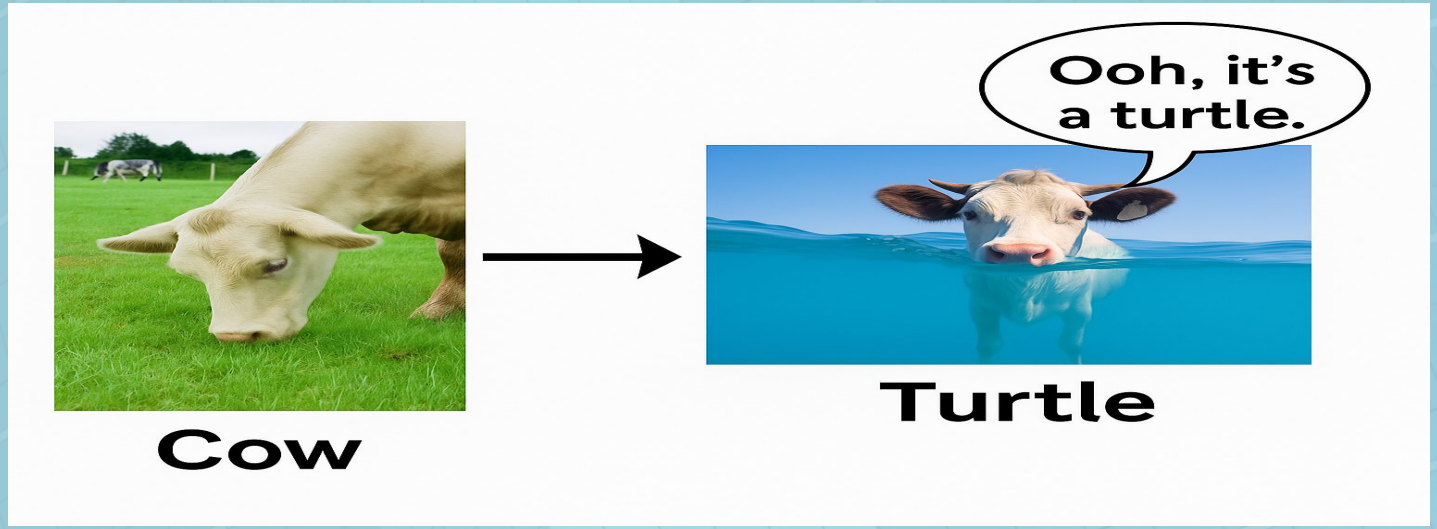
Aspect	Regular ML	Causal ML
Goal	**Predict outcomes**	**Estimate treatment effects**
Focus	**Associations (`P(Y)**	Estimate effect Y(1) −Y(0)Y(1)
Key Question	**“What *is* likely to happen?” No counterfactuals**	**“What *would* happen if we change treatment?”**
Methods	**Classification, regression, deep nets**	**Propensity scores, , causal forests, DML**
	Learns spurious correlations	**Removes confounding- via Propensity Scores & IPW**



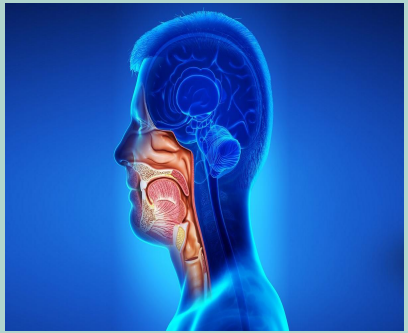
THE PROBLEM

A chest X-ray deep learning study (Zech et al., PNAS 2018) showed that the model was not learning lung pathology it was learning hospital-specific artifacts, such as metallic tokens/markers and scanner signatures.

A model might latch onto patterns that don't genuinely reflect the underlying reality.



THE PROBLEM

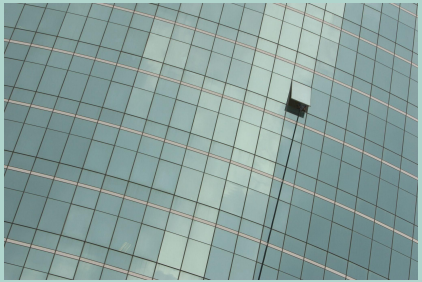


- A chest X-ray deep learning study (Zech et al., PNAS 2018) showed that the model was not learning lung pathology it was learning hospital-specific artifacts, such as metallic tokens/markers and scanner signatures.

THE PROBLEM

- X-rays from Hospital A had more severe pneumonia cases and a distinct metallic marker in the corner of the image.
- The model learned:

metal marker = severe pneumonia, instead of learning actual disease patterns in the lungs.



The Challenge Remains

In health care this would appear highly accurate in one setting but fail catastrophically elsewhere, leading to:

- delayed treatment,
- inappropriate admissions,
- false reassurance or unnecessary dose escalation

THE CHALLENGE

Recommendations exist:

- RT alone for T1N1
- RT or CRT for T2N1
- CRT for T1–T2 N2a–N2b
- But these recommendations are based on low-level evidence

Recommendations exist:

- A multi-center cohort, CRT did not improve OS for T1–T2 N1–N2a patients.
- Only younger patients (≤70) with T2 N2b showed improved locoregional control—not overall survival.
- Other studies show no difference in survival between older and younger patient on ChemoRT
- Older patients derived no clear benefit, raising concern about toxicity burden

Goal:

- Overall treatment effect (ATE)
- Individualized effects (CATE)
- Time-varying treatment effects
- Survival differences between RT and ChemoRT

- Causal machine learning (ML) offers flexible, data-driven methods for predicting treatment outcomes including efficacy and toxicity, thereby supporting the assessment and safety of drugs.
- A key benefit of causal ML is that it allows for estimating individualized treatment effects.
- Hence that clinical decision-making can be personalized to individual patient profiles (Feuerriegel et al, 2023),

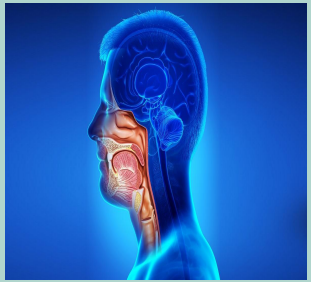
Our goal is not prediction; it is estimating treatment effects

- We are not just asking “Who will respond?”

- We are asking:

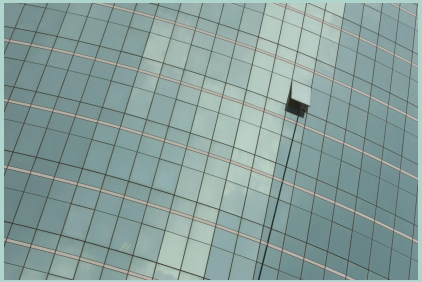
“What difference does chemo + radiotherapy make compared to radiotherapy alone for this patient group after adjusting the confounders?”

-Only causal ML can answer this counterfactual question



- **Confounder :** A variable that affects both the dependent variable and the independent variable, creating a spurious relationship

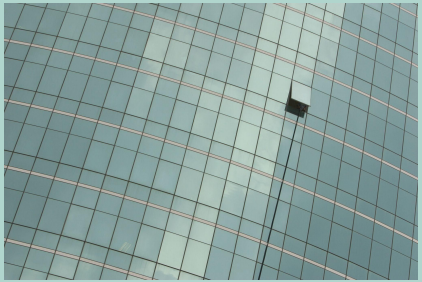
Perfect Example



Younger patients more likely to get ChemoRT

Standard ML will wrongly learn “younger age = better survival”

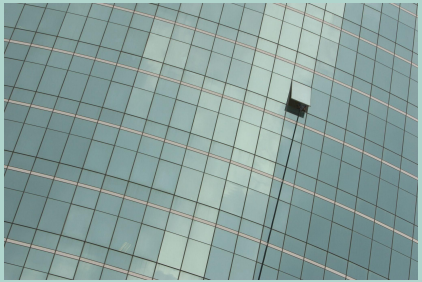
Not because of treatment → but confounding.



KEY ASSUMPTION OF CAUSAL

- **Conditional Exchangeability**: Given baseline confounders X treatment assignment A is independent of potential outcomes:

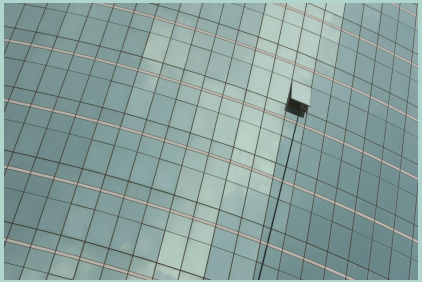
$$Y(a) \perp A \mid X$$



KEY ASSUMPTIONS OF CAUSAL ML

Consistency (SUTVA):
Each patient's outcome corresponds to the treatment they received, with no interference across patients.

KEY ASSUMPTION OF CAUSAL ML

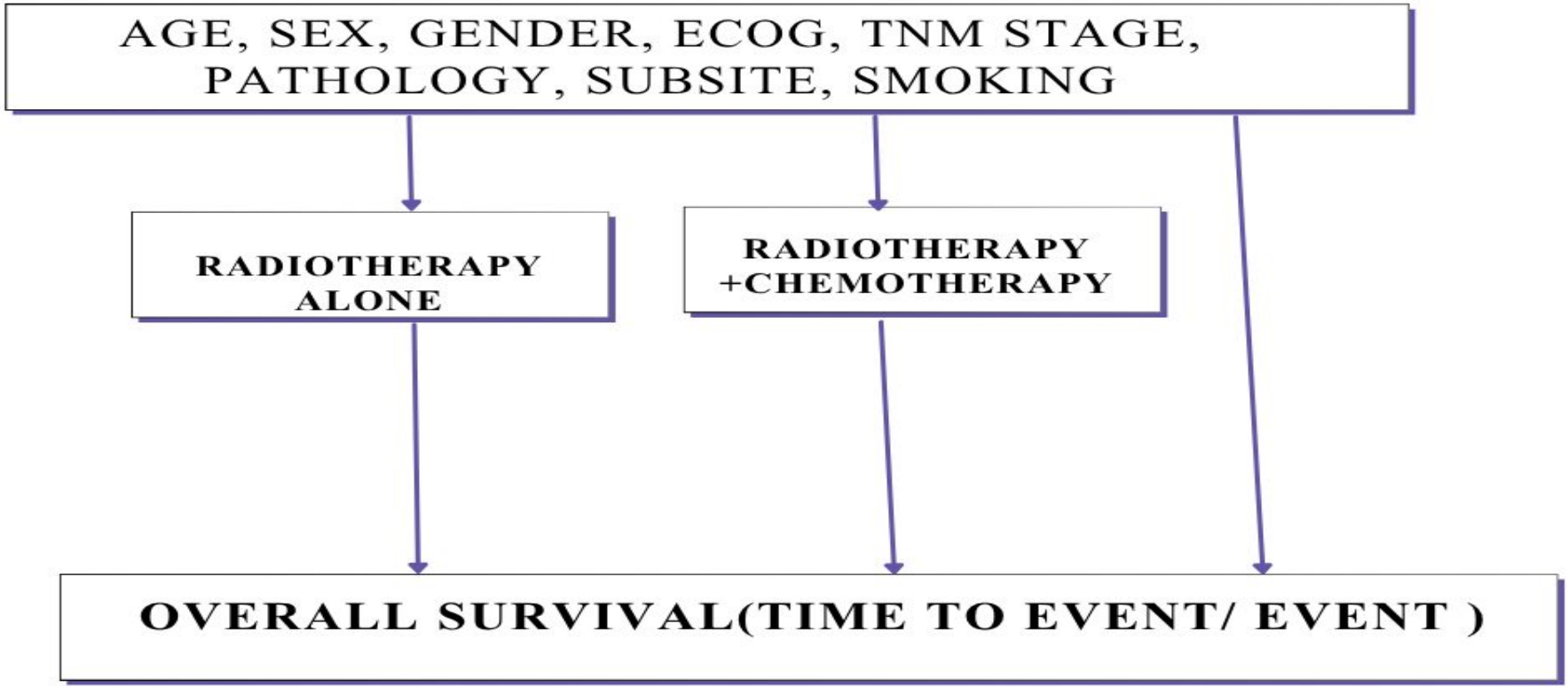


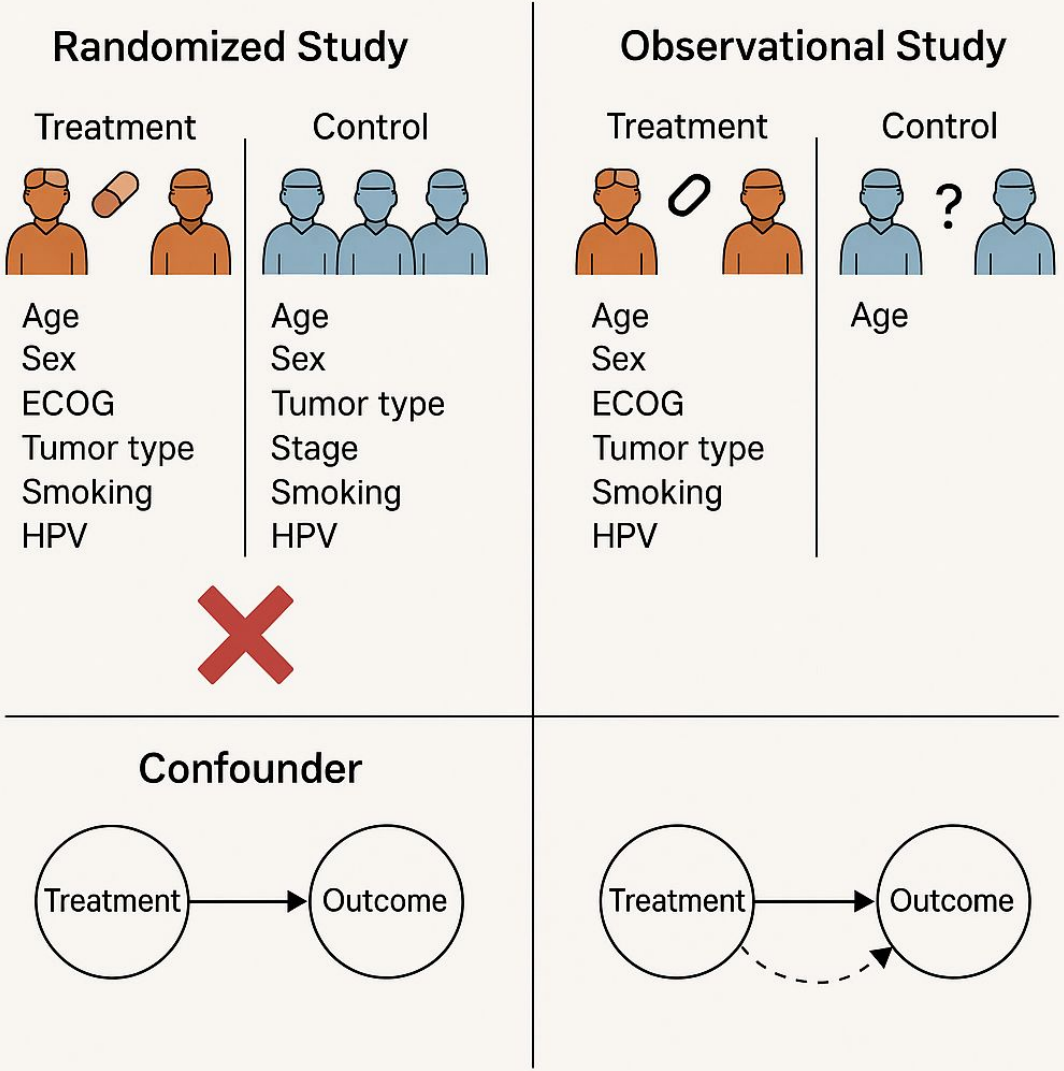
Positivity:
Every combination of confounders must have a non-zero probability of receiving either treatment.

Ignorability

- Ignorability assumes that after adjusting for all measured confounders,
  treatment assignment is ‘as good as random.’
- This assumption allows us to estimate causal effects from observational data

CONFOUNDERS, TREATMENT, OUTCOME,

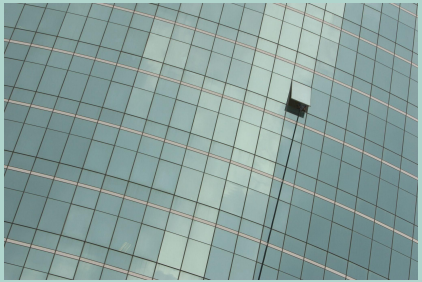




What are we missing?

The counterfactual!

AVERAGE TIME EFFECT

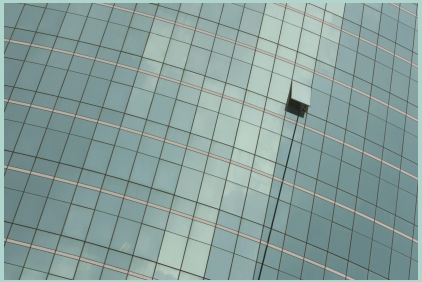


{ATE} = E[Y^{A=1}] - E[Y^{A=0}]

where:

Y(1) represents survival time had all patients received ChemoRT, and

Y(0) represents survival time had all patients received RT alone.



Conditional Average Treatment Effect (CATE)

(CATE) for a patient with covariates $X=xX = xX=x$ is:
$CATE(x)=E[Y(1)−Y(0)∣X=x]$

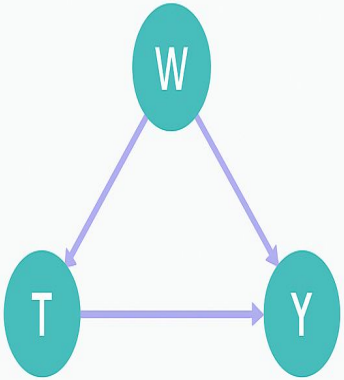
Where:

- $Y(1)Y(1)Y(1)$ = potential outcome **if treated**

- $Y(0)Y(0)Y(0)$ = potential outcome **if untreated**

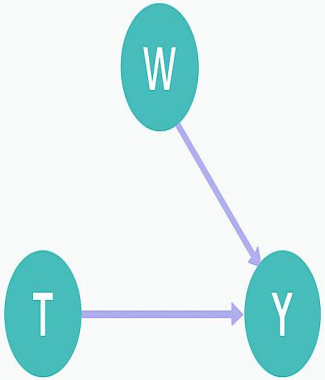
Inverse probability weighting

Regular Population



$$P(T|W) \neq P(T)$$

Pseudopopulation

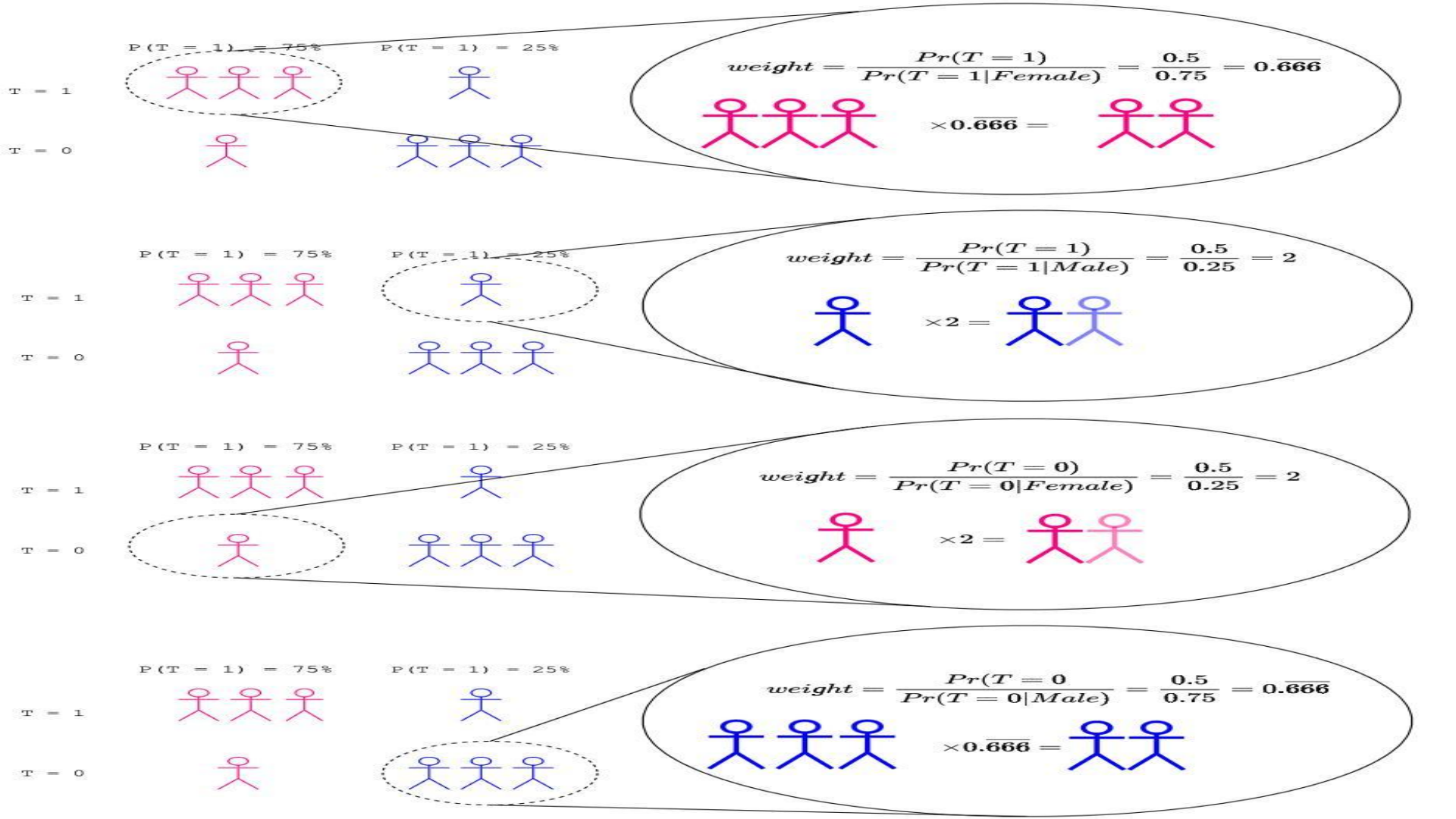


$$P(T|W) = P(T)$$

or

$$P(T|W) = 1$$

$$w_i = 1/(e(X_i)) + 1/(1 - e(X_i))$$


P(T = 1) = 75%
P(T = 1) = 25%
T = 1
T = 0
$weight = \frac{Pr(T=1)}{Pr(T=1|Female)} = \frac{0.5}{0.75} = 0.\overline{666}$
$\times 0.\overline{666} =$
P(T = 1) = 75%
P(T = 1) = 25%
T = 1
T = 0
$weight = \frac{Pr(T=1)}{Pr(T=1|Male)} = \frac{0.5}{0.25} = 2$
$\times 2 =$
P(T = 1) = 75%
P(T = 1) = 25%
T = 1
T = 0
$weight = \frac{Pr(T=0)}{Pr(T=0|Female)} = \frac{0.5}{0.25} = 2$
$\times 2 =$
P(T = 1) = 75%
P(T = 1) = 25%
T = 1
T = 0
$weight = \frac{Pr(T=0}{Pr(T=0|Male)} = \frac{0.5}{0.75} = 0.\overline{666}$
$\times 0.\overline{666} =$

Modules

-Streamlit app

Streamlit

plotly

-File handling

joblib

-For parsing and cleaning

python-dateutil

-Web deployment utilities

requests

Modules

- numpy
- pandas
- scikit-learn
- statsmodels
- scipy

-Causal inference and ML

- econml
- causalml

-Survival analysis

- lifelines

Step1: Data collection

RADCURE clinical dataset (N > 3,400), a multi-institutional real-world registry of head and neck cancer patients treated with radiotherapy alone, radiotherapy+chemotherapy and radiotherapy+EGFRI:https://www.cancerimagingarchive.net/collection/radcure/

- Demographics (age, sex)
- Clinical staging (T, N, M, Overall Stage)
- HPV status
- Smoking history
- Pathology subtype
- Treatment fields (RT, ChemoRT, RT+EGFRI, PostOp radiotherapy)
- Dose & fractionation (Gy, fractions)

Step 2: Define key variables

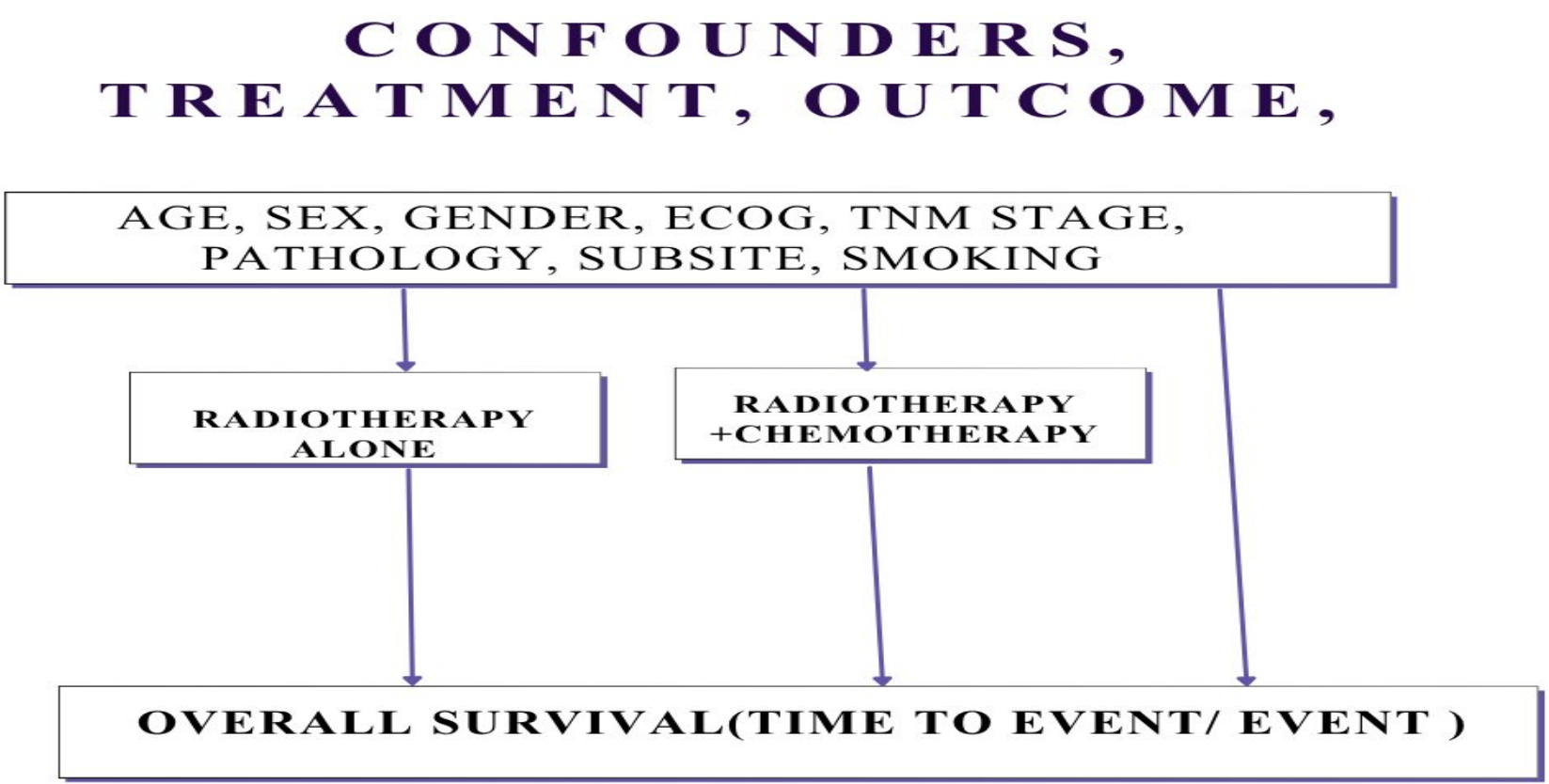
Treatment definition (binary exposure)

- `treatment = 1` → Concurrent ChemoRT
- `treatment = 0` → RT alone

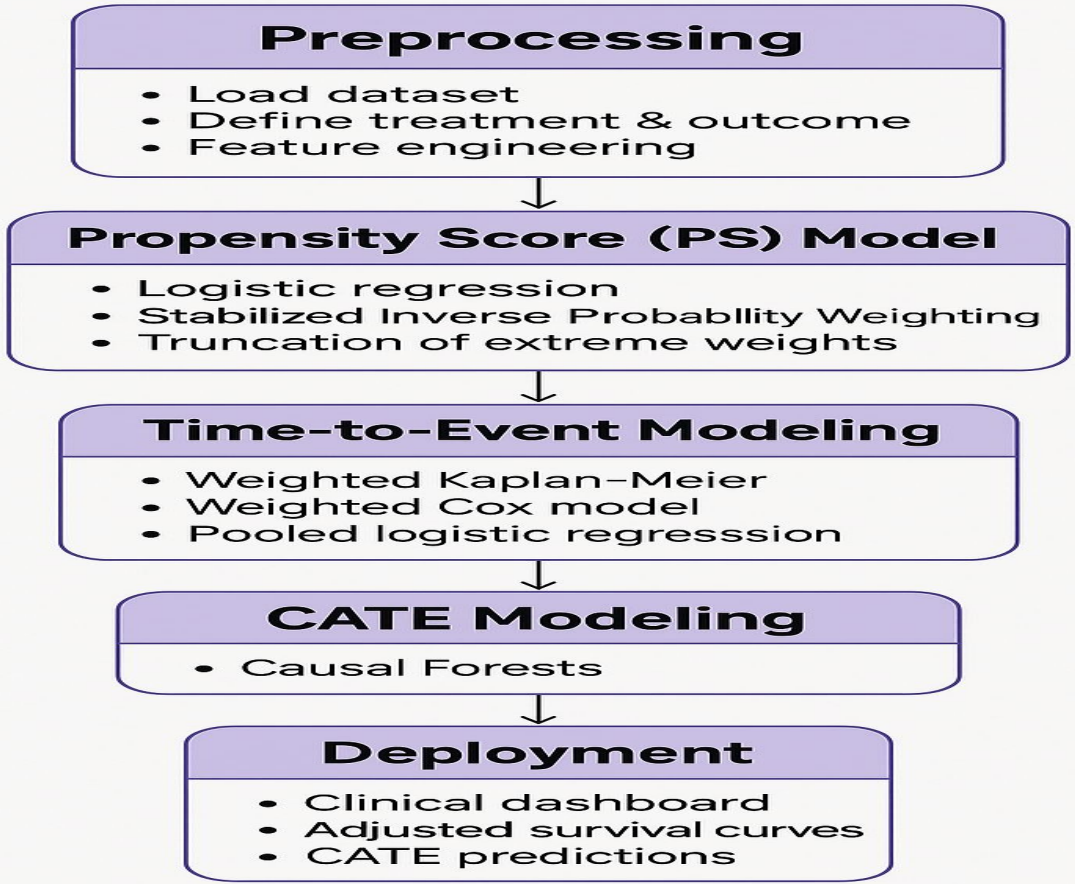
Outcome variables

- `time_os_days`
- `event_os` (1 = death, 0 = censored)

Step 3: Causal DAG(Our Assumptions)



Methodology Workflow



4. Data Cleaning & Harmonization

Data loading: original_radcure_data = pd.read_excel("/content/RADCURE_Clinical_v04_20241219.xlsx")

4. Data Cleaning & Harmonization

- Checked for missingness deduced that the missingness was not at random :
- Why We Did NOT Drop Missing Rows
  - Would introduce severe selection bias
  - Drops clinically important subgroups (HPV-, old patients, poor documentation groups)
  - Reduces overlap → PS extremes & unstable weights

4. Data Cleaning & Harmonization

Collapsed Rare Categories

High-cardinality clinical variables → unstable models.

We collapsed rare categories to avoid:

- Sparse dummy variables
- Noisy coefficients
- Unstable propensity scores

Variables collapsed:

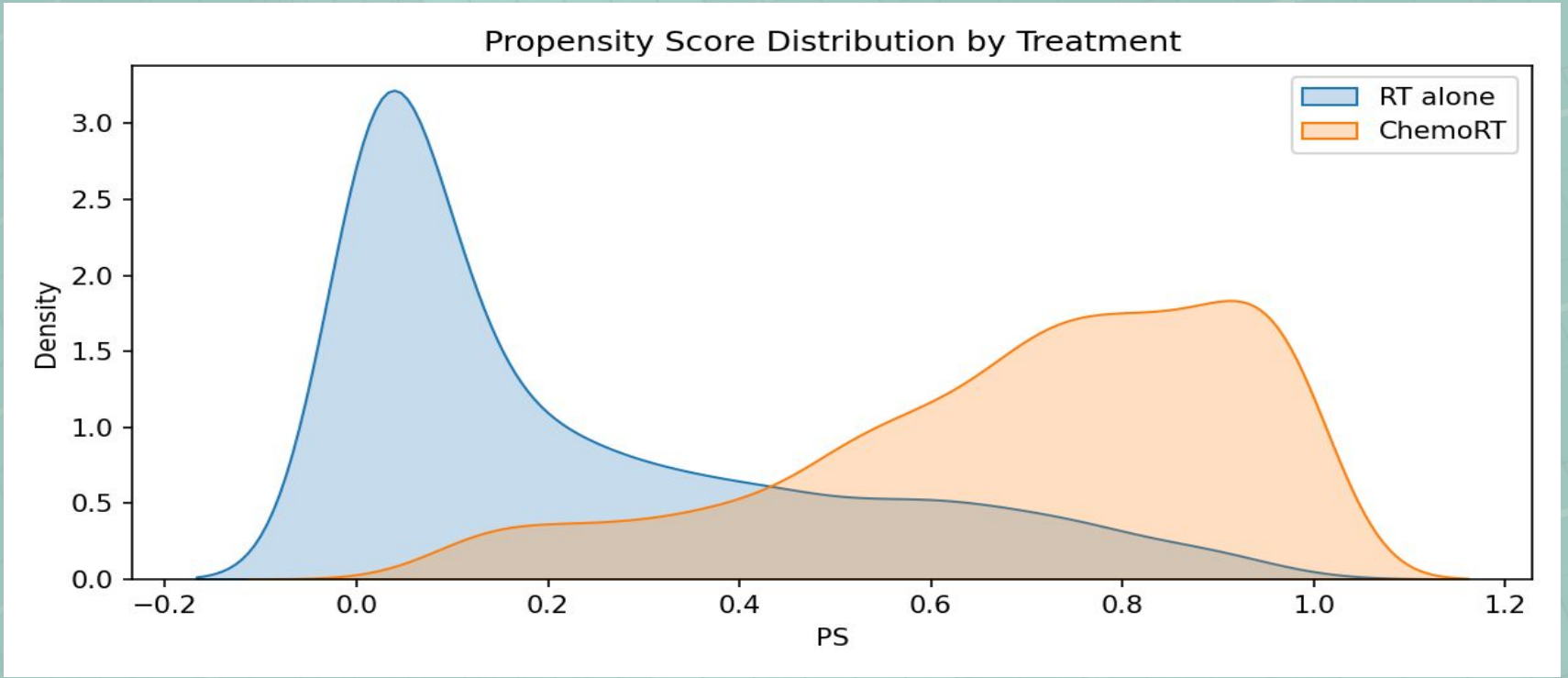
- `subsite_clean`
- `pathology_group`
- `primary_site_group`
- `stage`

4. Data Cleaning & Harmonization

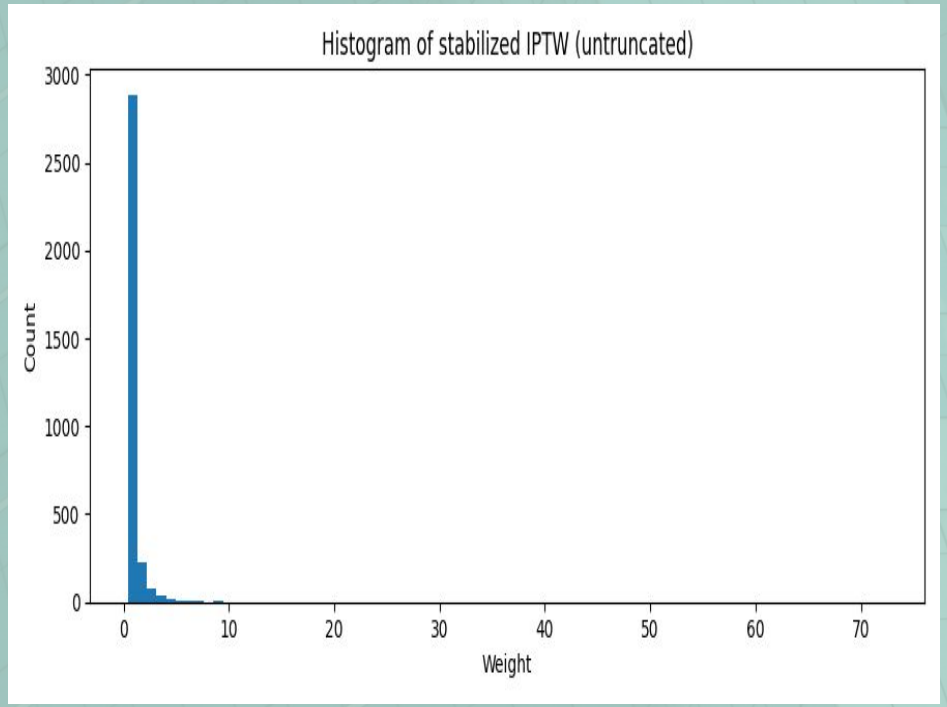
Created or cleaned variables

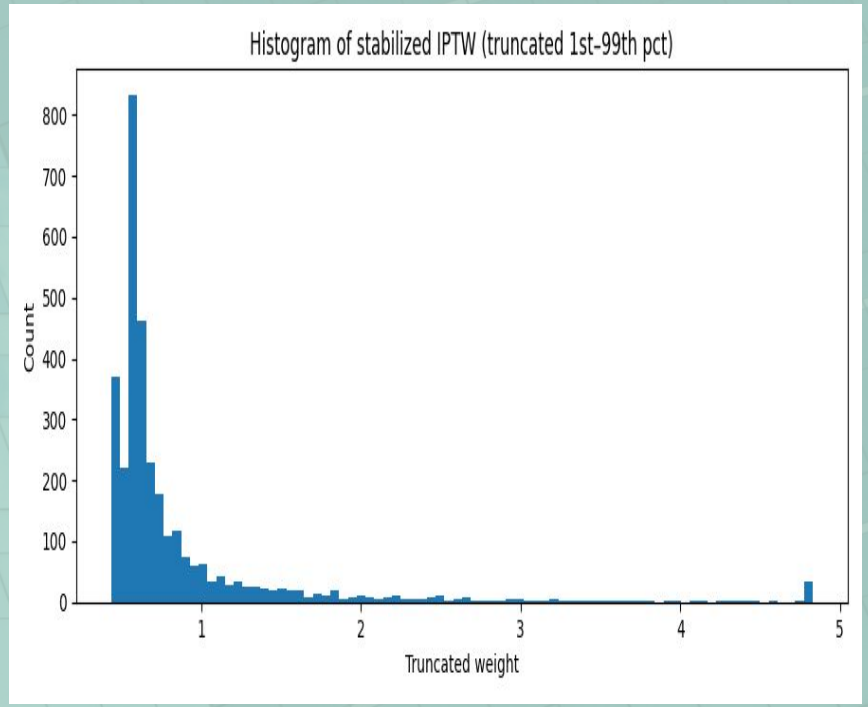
- **Smoking variables**:
  a. `Smoking_status_clean`, `smoking_missing`
  b. `smoking_py_clean`
- **HPV status:**
  a. `hpv_clean` (+ missing indicator)
- **T/N/M staging recoded-** t,n,m ,stage, t,n,m_missing
- **Primary site groups:** OPX, Larynx, Hypopharynx, Oral cavity
- **ECOG performance status**
  - Created `ecog_ps_missing` indicator
- **Treatment indicator** (`treatment`) normalized to 0/1
- **Outcome variables** : event_os, time_to_event

Propensity Distribution

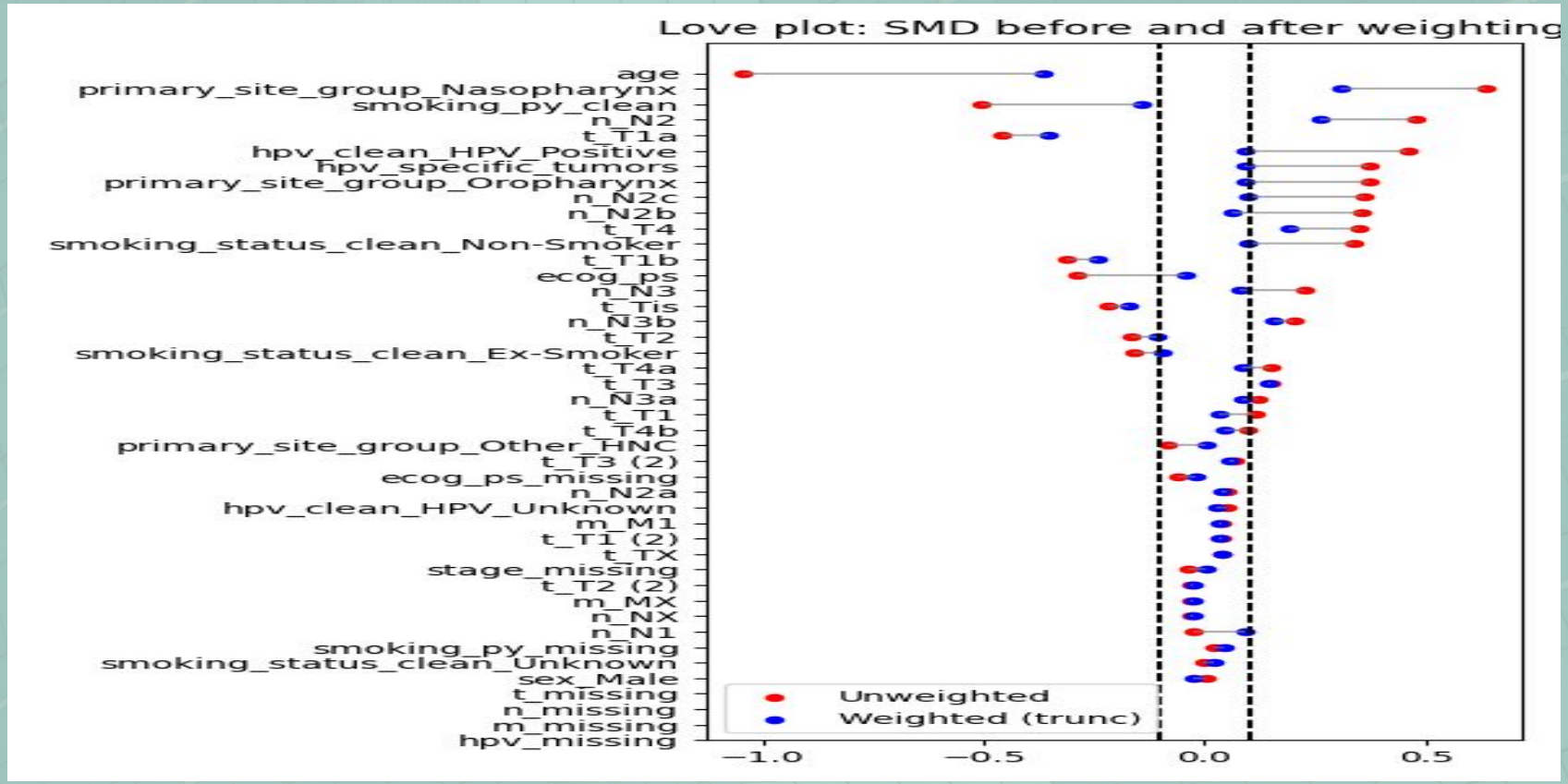


(Inverse Probability Treatment Weighting IPTW

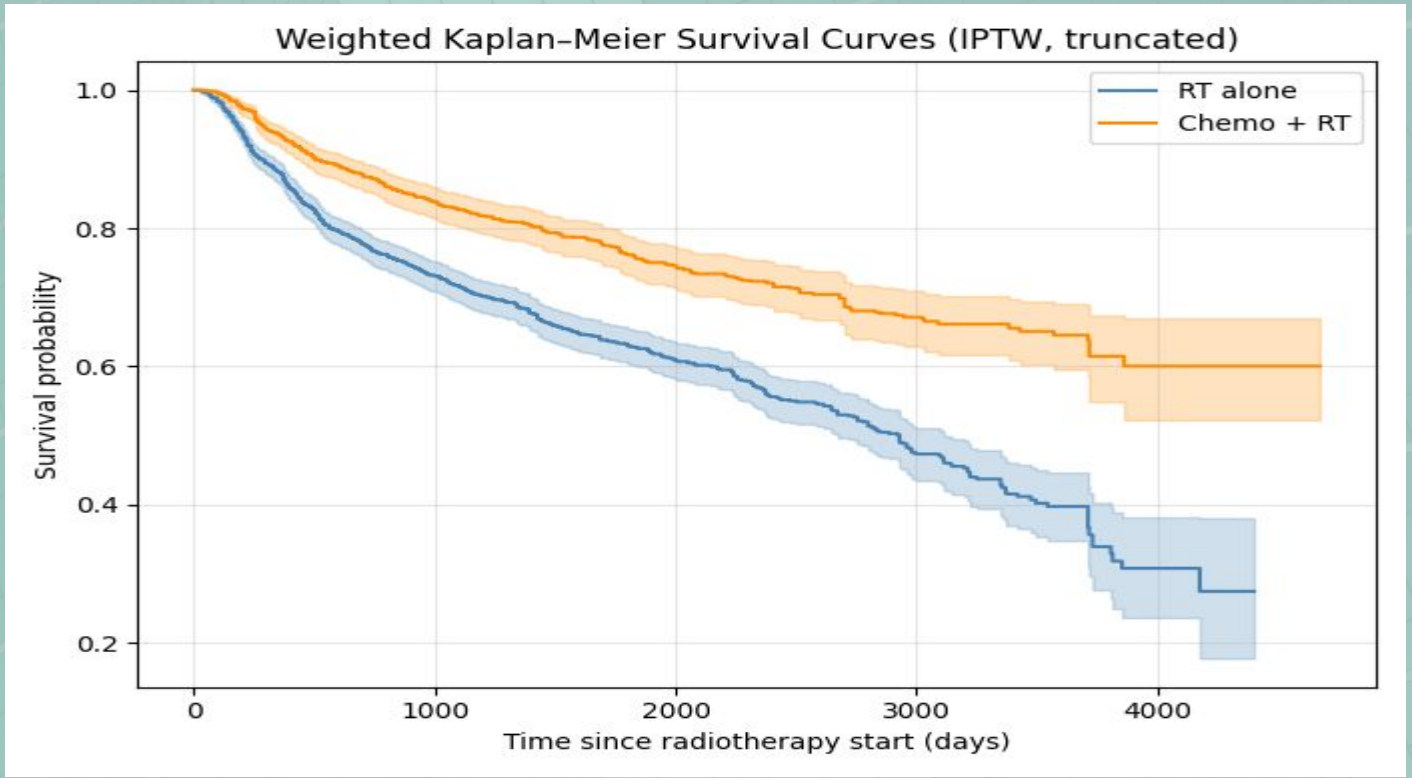




Love plot SMD before and after weighting



Weighted Kaplan -Meier survival curves:



Cox proportional Hazard

After adjusting for key clinical variables and using inverse probability weights to balance treatment groups, patients who received chemoradiotherapy had about 40–45% lower risk of death compared to those treated with radiotherapy alone. Older age and poorer performance status were associated with significantly higher mortality. Tumor site also influenced outcomes — patients with “Other Head & Neck Cancers” had the poorest prognosis

Cox proportional Hazard - Proportional Hazard assumption

Schoenfeld residuals revealed significant violations of the proportional hazards (PH) assumption for several key covariates including: treatment, age, ECOG performance status, and primary tumor site ($p < 0.05$). For treatment ($p = 0.025$), indicating that the treatment effect varied over time. Hence, we interpret the HR as an average effect

Variable 'treatment' failed the non-proportional test: p-value is 0.0318.

Discrete-Time Survival Modeling- Pooled Logistic Regression

Pooled Logistic Regression (PLR) estimates **time-varying hazards** when survival time is measured in discrete intervals

- Converts each patient into multiple “person–period” rows

- Models the risk of the event *within each interval*

$$\text{logit}[P(Y_{it}=1 \mid Y_{i,t-1}=0, X_i)] = \beta_0 + \beta_t t + \beta_{trt} T_i + \beta_{int}(T_i \times t) + \beta^\top X_i$$

where: Y_{it} = event indicator in interval t

T_i = treatment assignment

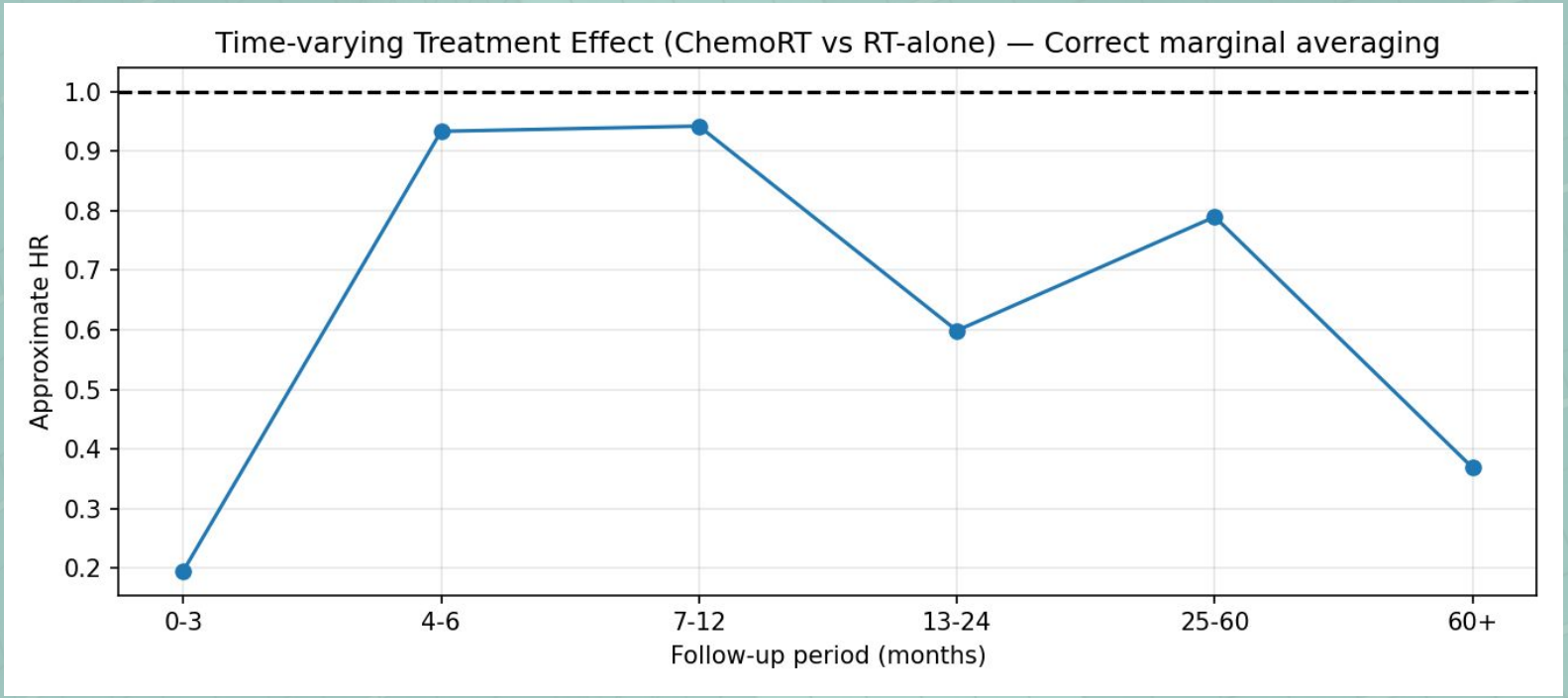
Xi = baseline covariates

t or period_bin = time effects (flexible specification)

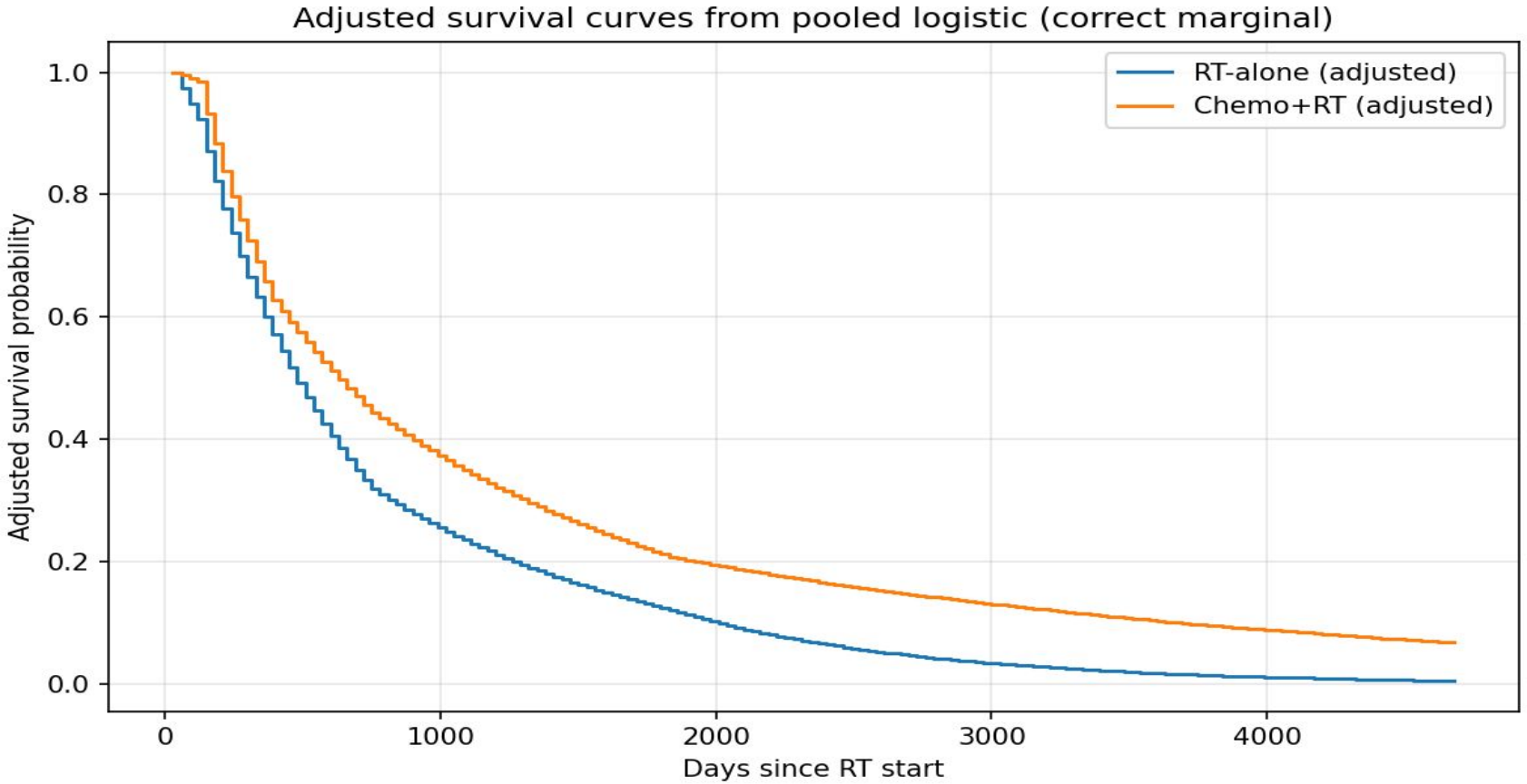
Weighted Kaplan -Meier survival probabilities:

Time point	RT Alone	Chemo+RT	Difference
1-year	0.88	0.93	+0.05
3-year	0.72	0.83	+0.11
5-year	0.63	0.76	+0.13

Time-Varying Treatment Effect (Pooled Logistic Regression)



Weighted Kaplan -Meier survival curves:



Causal Forests (Heterogeneous Treatment Effect Estimation)

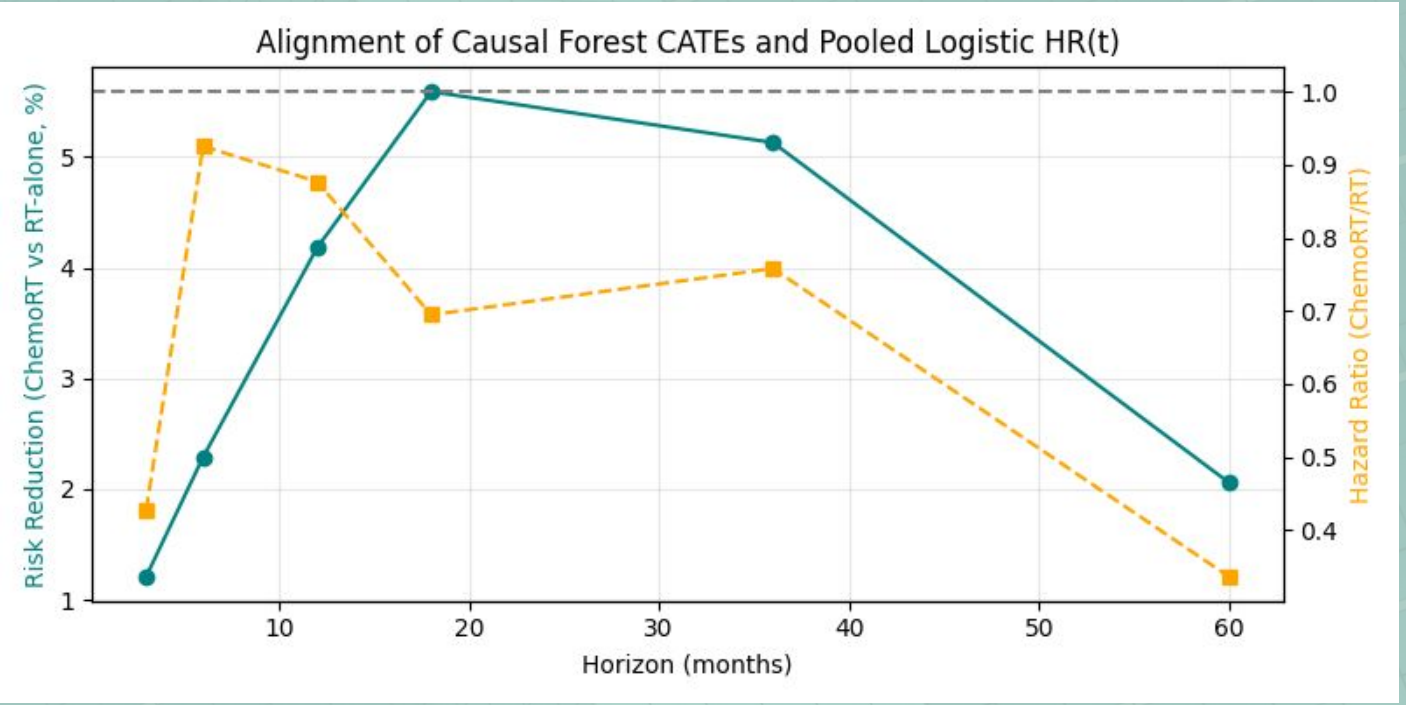
- Causal Forests (CF), an extension of generalized random forests, estimate the **Conditional Average Treatment Effect (CATE)** as a function of patient covariates. The method employs *honest splitting*: one part of the data determines the tree structure, while another part estimates treatment effects, reducing bias and enhancing asymptotic consistency.
- It captures complex nonlinearities and high-order interactions among clinical variables.
- Produces individual-level treatment effect estimates and subgroup-level summaries (e.g., HPV+, T/N stage, smoking burden)

Causal Forests (Heterogeneous Treatment Effect Estimation)

For each patient with covariates X=x the CATE is:

$$\tau(x) = E[Y \mid T = 1, X = x] - E[Y \mid T = 0, X = x]$$

Causal Forest CATE vs Pooled Logistic Regression:



Time-Varying Treatment Effect (Pooled Logistic Regression)

Model Performance

AUC (Train/Test): 0.760 / 0.735

Brier Score: 0.197

Max Coefficient Magnitude: 3.19 (well-controlled; no exploding effects

The pooled logistic model reveals a time-varying benefit, strongest early and very late, milder in mid-follow-up.

Adjusted Survival (Weighted Kaplan–Meier)

Median Survival

- RT alone: 2,929 days
- Chemo + RT: Not reached (survival curve does not fall below 0.5)

Log-rank p-value < 0.00001 → Strong evidence of overall benefit in unadjusted survival.

Heterogeneous Treatment Effects (Causal Forest CATE)

Horizon 3 months:
Mean CATE: -0.0121
SD(CATE): 0.0340

Horizon 6 months:
Mean CATE: -0.0229
SD(CATE): 0.0624

Horizon 12 months:
Mean CATE: -0.0418
SD(CATE): 0.1303

Horizon 18 months:
Mean CATE: -0.0559
SD(CATE): 0.1568

Horizon 36 months:
Mean CATE: -0.0513
SD(CATE): 0.1558

Horizon 60 months:
Mean CATE: -0.0206
SD(CATE): 0.1508

Heterogeneous Treatment Effects (Causal Forest CATE)

Not all HNSCC patients benefit equally from concurrent chemotherapy.

Benefit depends strongly on tumor biology and patient fitness.

Heterogeneous Treatment Effects (Causal Forest CATE)

Groups with higher benefit from supporting by literature):

HPV-positive OPSCC

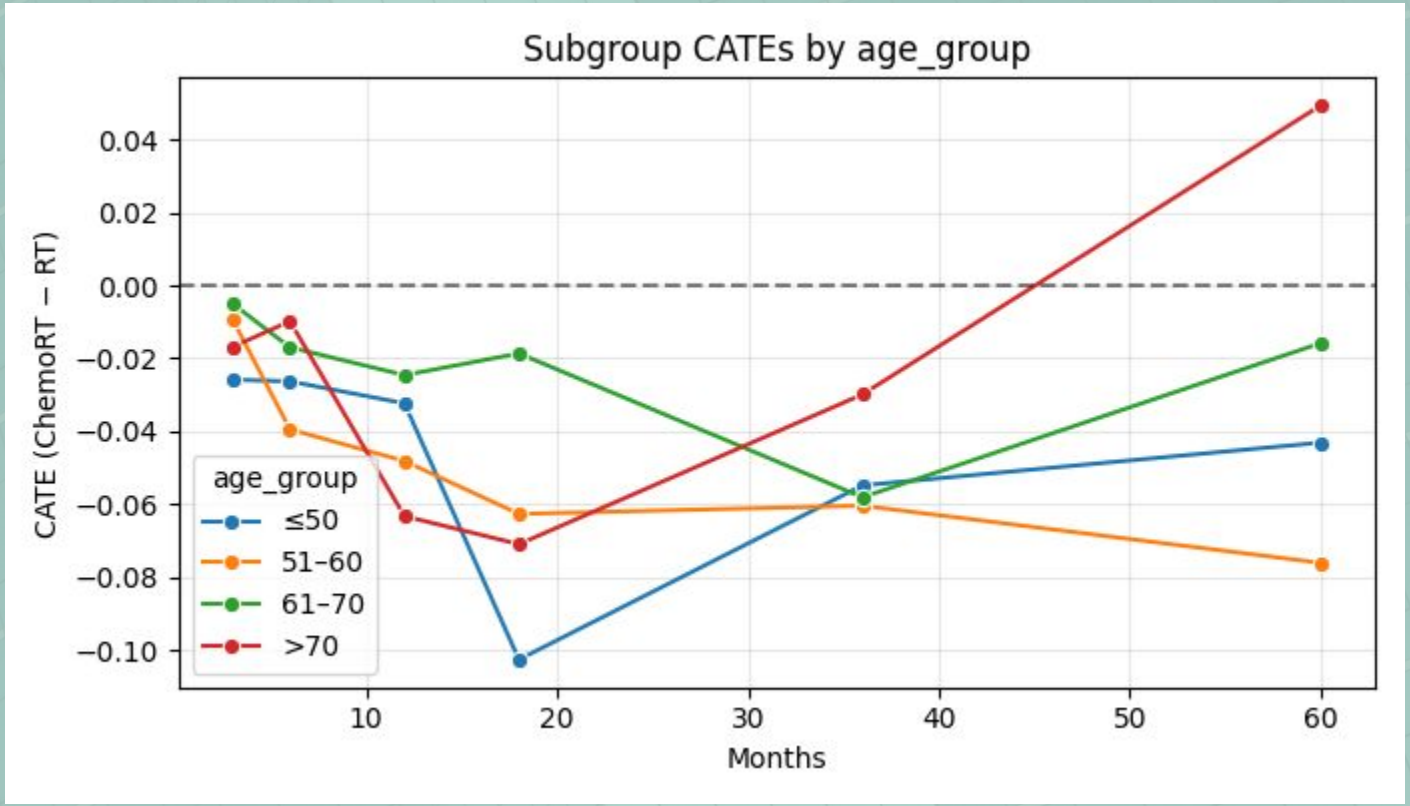
T2–T3 tumors

N2b/N2c nodal disease

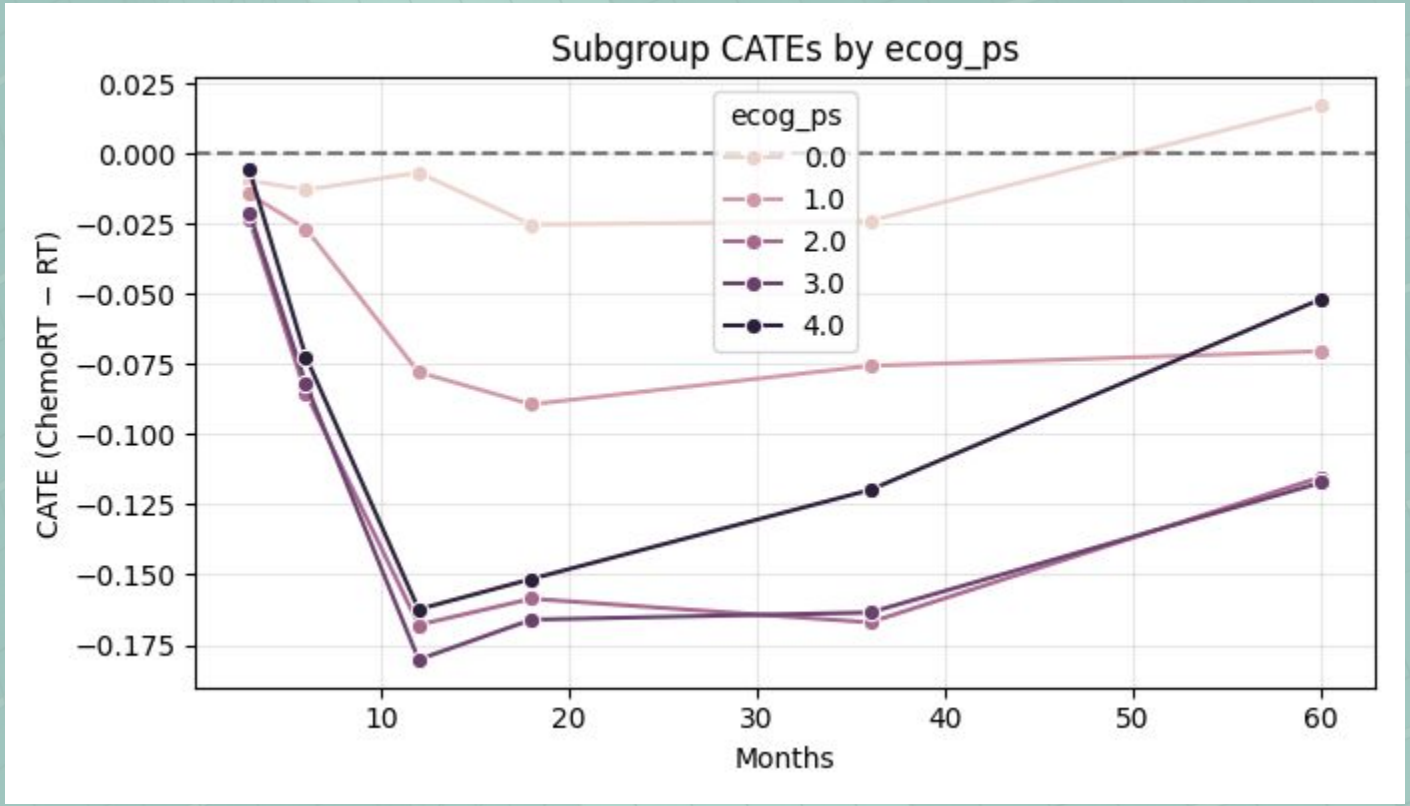
Patients ≤70 years

Low comorbidity burden (ECOG 0–1)

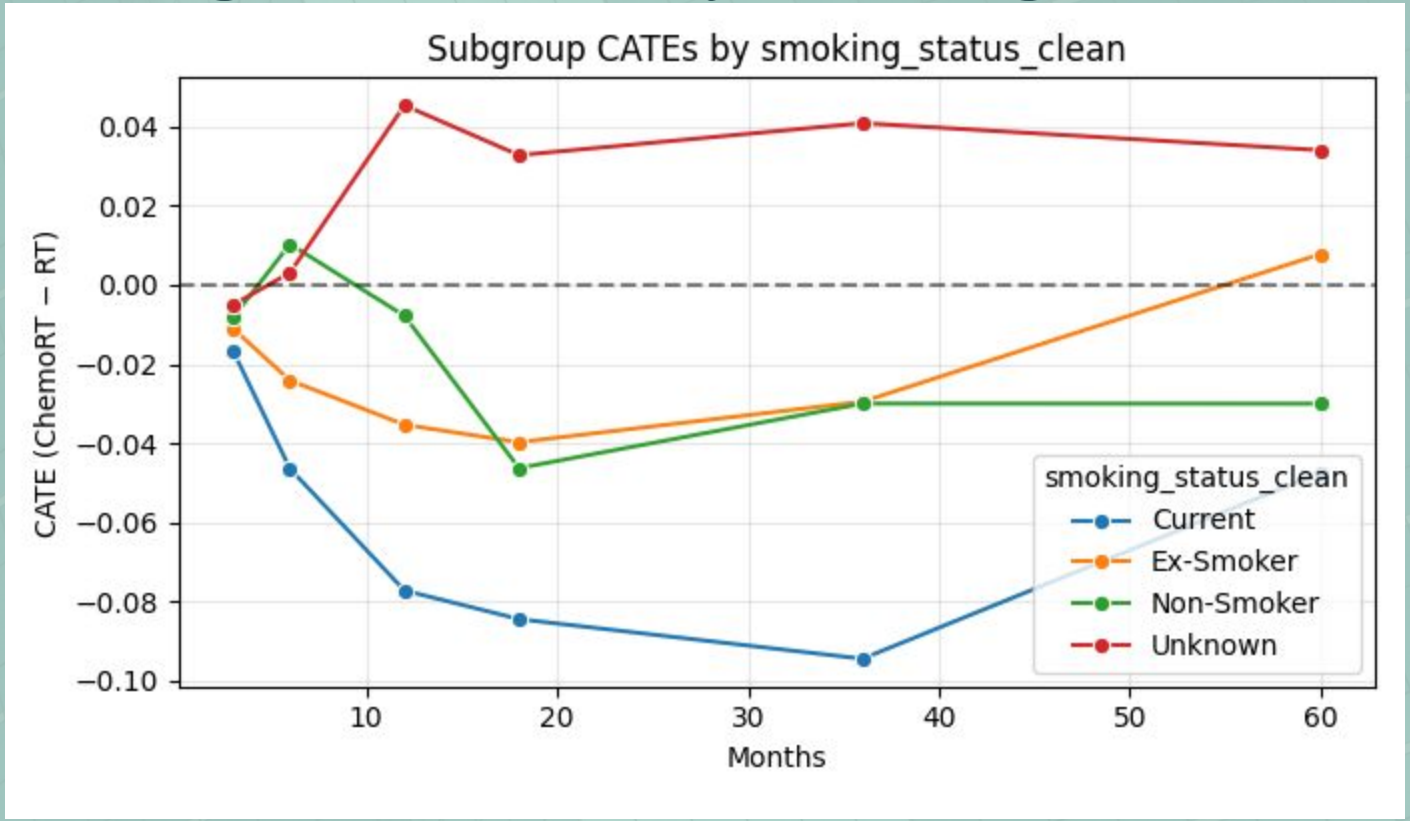
Subgroup CATE By Age:



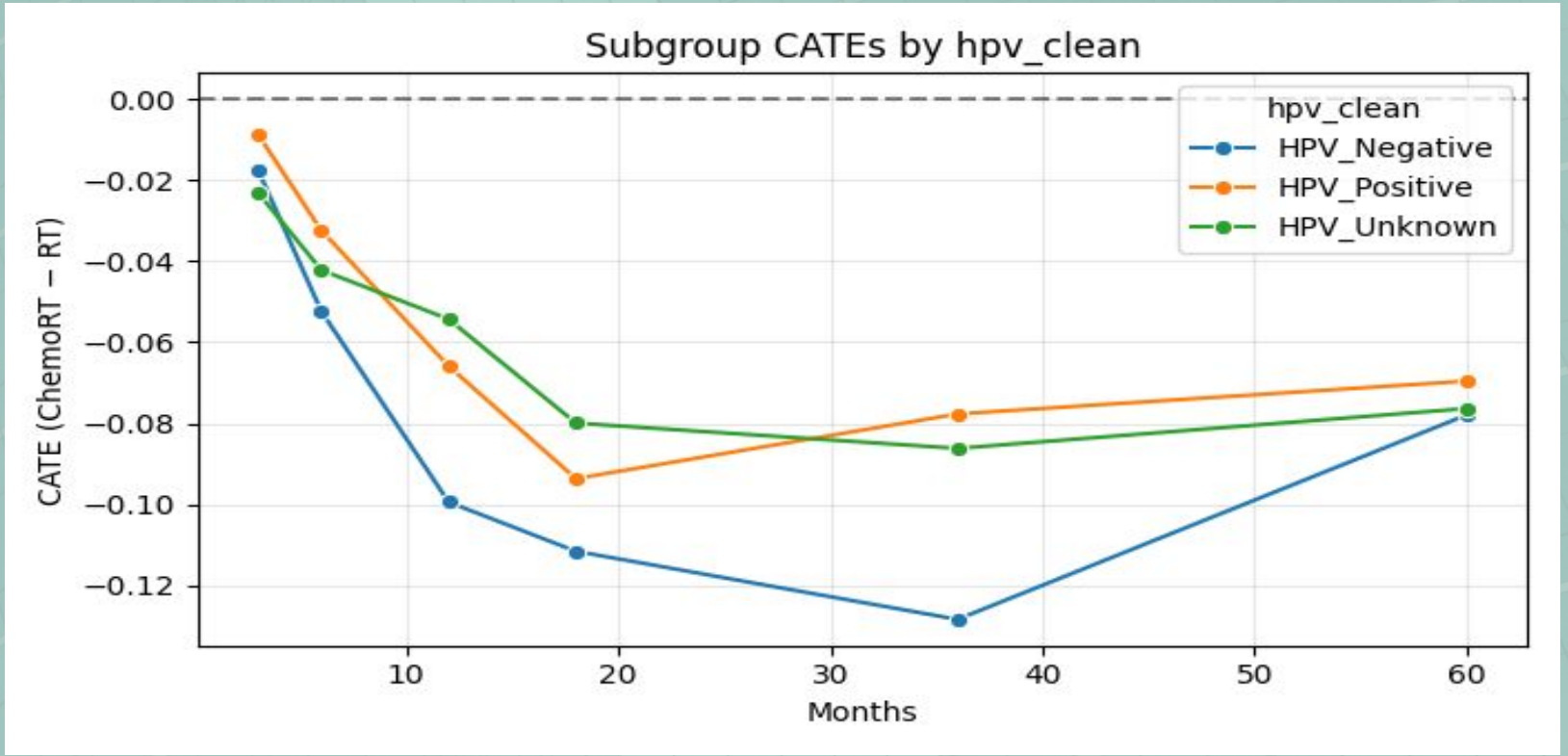
Subgroup CATE By ECOG:



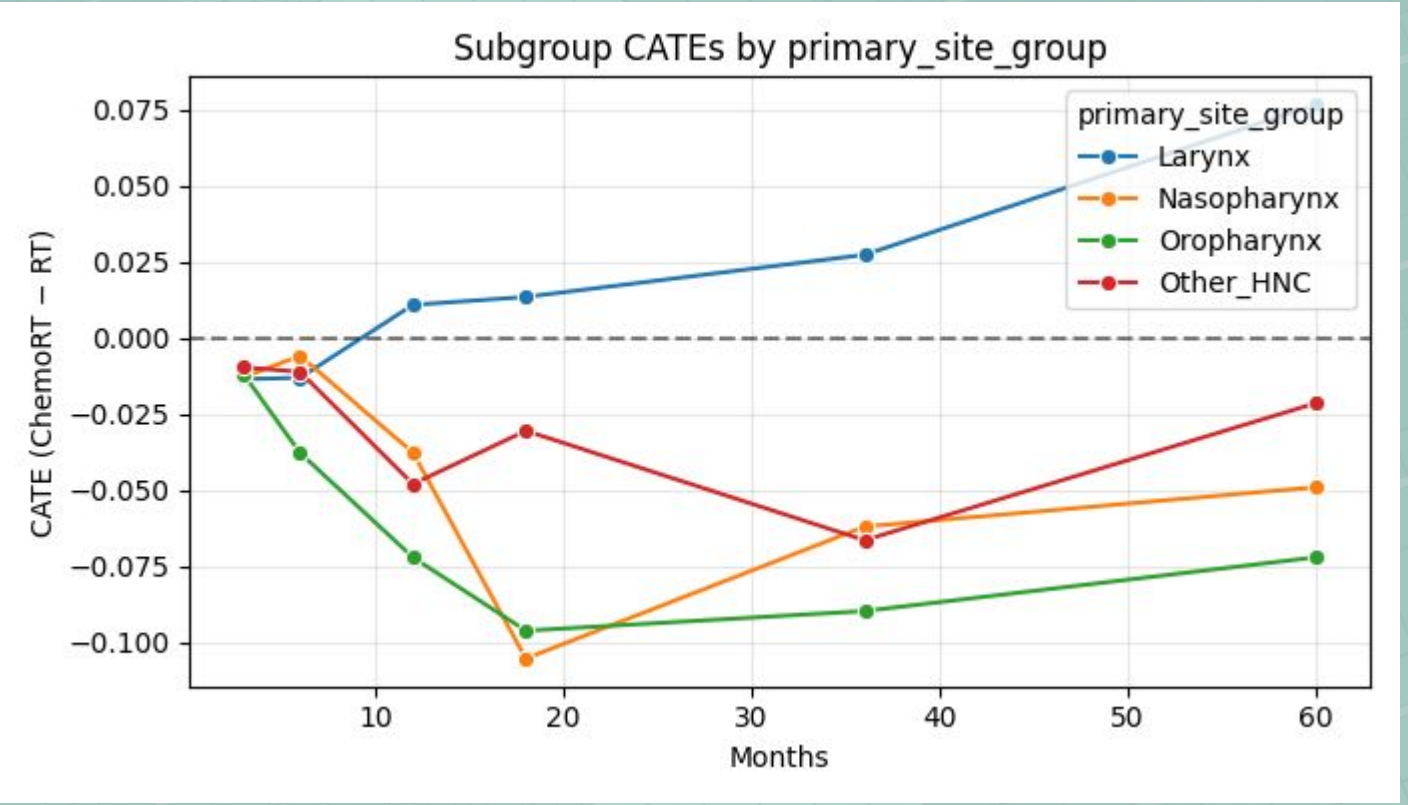
Subgroup CATE By smoking status:



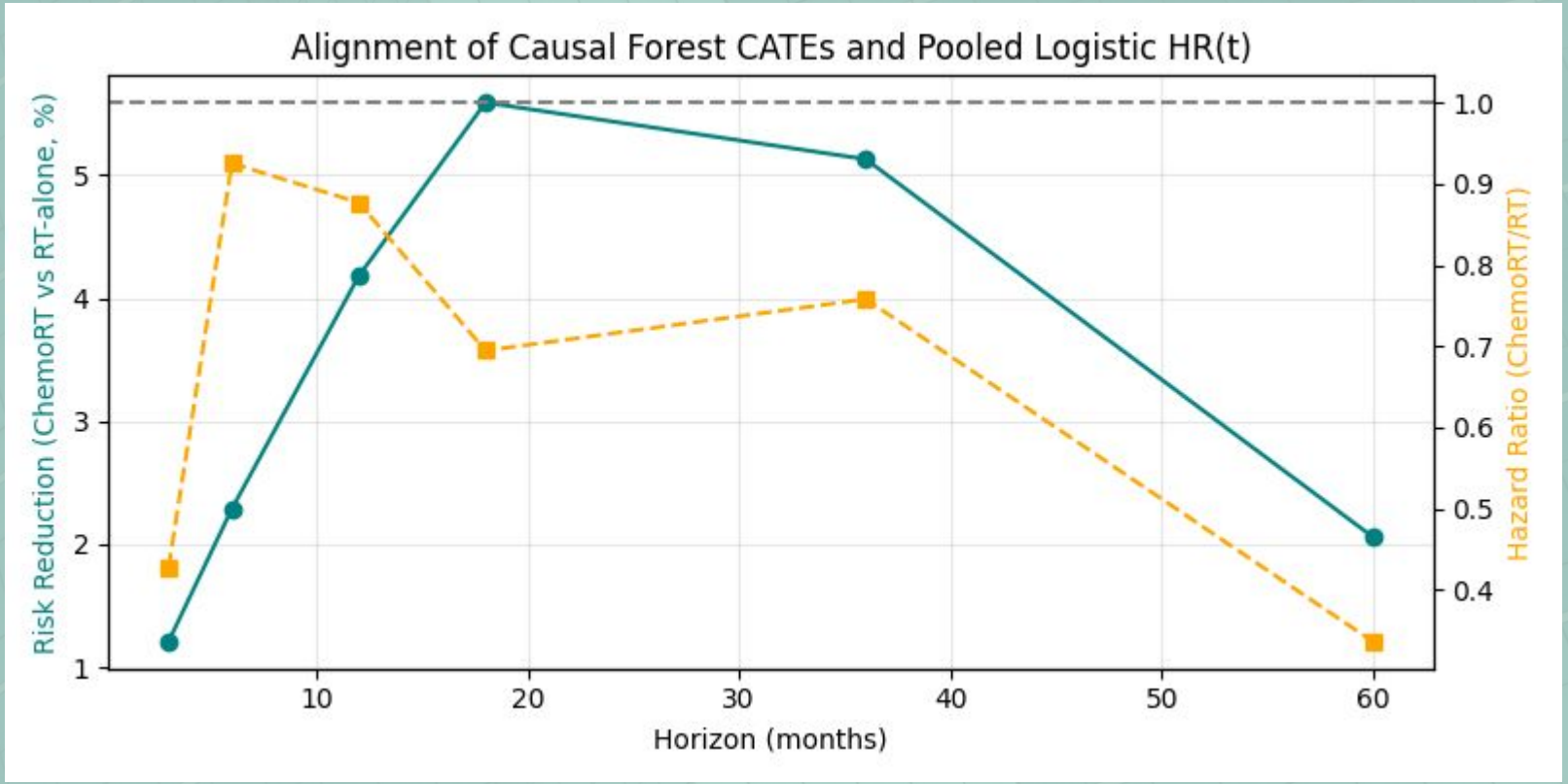
Subgroup CATE By HPV status:



Subgroup CATE By primary tumor site:



Compare curves: Causal Forest CATE and Pooled Logistic HR(t):



“Correlation does imply causation”

Future Directions

Multimodal Patient-Level Digital Twin for HNSCC

Using causal ML + survival models + radiomics:

simulate outcomes under RT alone vs CRT

estimate recurrence risks

Deployment

Multimodal Patient-Level Digital Twin for HNSCC

Using causal ML + survival models + radiomics:

simulate outcomes under RT alone vs CRT

estimate recurrence risks

References

Feuerriegel, S., Frauen, D., Melnychuk, V. *et al.* Causal machine learning for predicting treatment outcomes. *Nat Med* 30, 958–968 (2024). https://doi.org/10.1038/s41591-024-02902

Buchberger, A. M. S., Strzelczyk, E. A., Wollenberg, B., Combs, S. E., Pickhard, A., & Pigorsch, S. U. (2021). *Report on late toxicity in head-and-neck tumor patients with long-term survival after radiochemotherapy. Cancers, 13*(17), 4292. https://doi.org/10.3390/cancers13174292

Adrian, G., Rudolfson Falklind, N., McDowell, L., & Gebre-Medhin, M. (2025). *Chemoradiation therapy versus radiation therapy alone in T1–2 oropharyngeal cancer with low-volume neck disease: A population-based cohort study using the Swedish Head and Neck Cancer Register. International Journal of Radiation Oncology, Biology, Physics.* Advance online publication. https://doi.org/10.1016/j.ijrobp.2025.08.040

Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018 Nov 6;15(11):e1002683. doi: 10.1371/journal.pmed.1002683. PMID: 30399157; PMCID: PMC6219764.

Thank you

Danke!