Estimating Heterogeneous Treatment Effects in Breast Cancer Using Multimodal TCGA Real-World Data

Melis Gezer

Project Report

Project Overview

Estimating heterogeneous treatment effects (HTEs) from observational real-world data is a key step toward personalized oncology, yet it remains challenging due to confounding and limited overlap between treatment groups. In this study, we develop an end-to-end, reproducible multimodal pipeline to estimate individualized treatment effects in breast cancer using public data from The Cancer Genome Atlas (TCGA). We integrate clinical variables with RNA-seq gene expression features and define a clinically meaningful treatment indicator and outcome derived from follow-up information. We implement and benchmark causal machine learning estimators based on meta-learning (S-, T-, and X-learners) and doubly robust principles, with our primary analysis using causal forests from the generalized random forests (GRF) framework. We assess identifying assumptions through propensity score overlap diagnostics and covariate balance checks, and evaluate estimated treatment policies using policy value and ranking-based metrics on held-out data. Robustness analyses examine sensitivity to propensity score trimming and feature selection strategies. Our results indicate evidence of treatment effect heterogeneity, with estimated average treatment effects suggesting a survival benefit from chemotherapy, though substantial uncertainty remains given the observational nature of the data and limited sample size. We release code and documentation to enable full reproducibility and facilitate extensions to time-to-event causal modeling and external validation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Research Objectives

Precision oncology aims to tailor treatment decisions to individual patient characteristics, moving beyond one-size-fits-all approaches. A fundamental challenge in this paradigm is estimating heterogeneous treatment effects (HTEs)—the extent to which treatment benefits vary across patients with different baseline characteristics [6, 9]. While randomized controlled trials (RCTs) provide the gold standard for average treatment effect estimation, they are often underpowered to detect treatment-effect heterogeneity and may exclude patients who are representative of real-world populations.

Observational real-world data, such as those available through The Cancer Genome Atlas (TCGA), offer rich multimodal information that can complement RCT evidence. However, estimating causal effects from such data requires careful attention to confounding, overlap (positivity), and model specification. Recent advances in causal machine learning provide methods specifically designed for HTE estimation in observational settings [12, 1, 8].

This study develops a complete, reproducible pipeline for estimating individualized treatment effects of chemotherapy on five-year survival in breast cancer patients using TCGA data. Our objectives are to:

1. Construct a clean, analysis-ready cohort by integrating clinical and RNA-seq gene expression data with transparent preprocessing steps.

2. Assess the plausibility of key causal identification assumptions through propensity score diagnostics and covariate balance evaluation.

3. Implement and benchmark multiple HTE estimation methods, with a focus on causal forests.

4. Evaluate estimated treatment effect heterogeneity using ranking-based and policy value metrics appropriate for observational settings.

5. Conduct sensitivity analyses to assess robustness of findings to modeling choices.

1.2 Analysis Pipeline Overview

This study follows a five-phase workflow emphasizing reproducibility and transparent reporting:

Phase 0: Data Assets and Reproducible Inputs All analyses derive from a single consolidated artifact (`dataset_unstranded_5yr_logcpm_filtered.npz`) containing gene expression features ($\mathbf{X} \in \mathbb{R}^{351 \times 29{,}741}$), treatment indicator ($A$), five-year outcome ($Y$), and survival fields. Gene annotations and missingness reports support transparency.

Phase 1: Cohort Construction We document cohort assembly from raw TCGA-BRCA files, including deterministic selection of one RNA-seq file per patient and derivation of treatment/outcome variables. Descriptive visualizations (Kaplan-Meier curves, outcome distributions) characterize the cohort without causal claims.

Phase 2: Assumption Diagnostics We evaluate confounding and overlap through propensity score estimation, covariate balance assessment (standardized mean differences), and inverse probability of treatment weight (IPTW) diagnostics. Sensitivity to propensity score trimming is explored.

Phase 3: Treatment Effect Estimation We fit benchmark meta-learners (T-, S-, and DR-learners) and our primary method—causal forests using GRF—on multiple feature views (clinical-only, clinical+RNA PCA). Models are trained on a fixed 80/20 split and evaluated on held-out data.

Phase 4: Evaluation and Interpretation Without ground-truth individual effects, we evaluate HTE models using: (i) CATE (Conditional Average Treatment Effect) distribution summaries with uncertainty quantification, (ii) quartile-based subgroup analyses, (iii) targeting operator characteristic (TOC) curves assessing ranking quality, (iv) rank-weighted average treatment effect (RATE) tests, and (v) best linear projection (BLP) to identify effect modifiers.

Phase 5: Robustness Analyses We systematically evaluate sensitivity to propensity score trimming thresholds, assessing trade-offs between sample retention, covariate balance, and targeting performance.

1.3 Project Organization

The project is organized as follows. Chapter 2 details data sources, cohort construction, and preprocessing. Chapter 3 presents propensity score modeling and assumption diagnostics. Chapter 4 describes HTE estimation methods and model fitting. Chapter 5 presents comprehensive results from all pipeline phases. Chapter 6 details evaluation metrics and interpretability analyses. Chapter 7 presents robustness and sensitivity analyses. Chapter 8 concludes with discussion, limitations, and future directions.

Chapter 2

Data and Cohort Construction

2.1 Overview

This chapter documents the transformation of raw TCGA-BRCA data into analysis-ready tables supporting individualized effect estimation. We focus on: (i) defining a consistent patient-level unit of analysis, (ii) linking clinical records to RNA-seq quantifications, and (iii) constructing interpretable treatment and outcome variables.

2.2 Data Sources

We use public breast cancer data from The Cancer Genome Atlas (TCGA) accessed through the Genomic Data Commons (GDC) portal. The dataset comprises:

- **Clinical tables** (TSV): `clinical.tsv` (5,546 rows, 210 columns), `follow_up.tsv` (9,427 rows, 198 columns), and `pathology_detail.tsv` (1,096 rows, 86 columns).
- **RNA-seq files**: 1,231 gene expression quantification files produced by the GDC/STAR workflow.
- **Sample sheet**: Linkage between clinical cases and molecular files.

2.2.1 Identifiers and Linking Strategy

Two key identifiers enable multimodal integration:

- `cases.case_id`: UUID-like identifier (globally unique)
- `cases.submitter_id`: TCGA barcode (e.g., TCGA-XX-XXXX)

The sample sheet `Case ID` matches `cases.submitter_id`, providing the join key between clinical and RNA modalities.

2.3 Clinical Data Harmonization

Raw clinical exports contain multiple rows per patient. We constructed a case-level table by:

1. Aggregating longitudinal follow-up (taking maximum follow-up duration)
2. Retaining first non-missing value for baseline covariates per `cases.case_id`
3. Cleaning placeholder values and converting to consistent numeric types

Key baseline covariates included age at diagnosis, AJCC pathologic staging (stage, T, N, M), tumor grade, and lymph node positivity.

2.4 RNA-seq Expression Matrix Construction

Each RNA-seq file contains gene-level counts with columns: `gene_id`, `gene_name`, `gene_type`, `unstranded`, stranded counts, TPM, and FPKM variants. We used `unstranded` counts as the primary measurement for consistency.

2.4.1 Deterministic File Selection

When multiple RNA-seq files existed per patient (max 4, median 1), we selected one file using a deterministic preference order:

1. Prefer tumor over normal tissue
2. Prefer primary over metastatic tumors
3. Prefer OCT > Unknown > FFPE preservation
4. Break ties by lexicographic `File ID` ordering

This yielded 1,095 cases with expression data, each mapped to exactly one quantification file.

2.4.2 Gene Alignment and Filtering

All files shared a consistent gene space (60,664 genes). We constructed a case-by-gene matrix ($n = 1{,}095 \times p = 60{,}664$) and applied quality-control filtering:

- Removed STAR QC rows (N_unmapped, N_multimapping, etc.)
- Removed 3,618 all-zero genes

- Applied variance filtering (threshold=0.01 on log-transformed scale), removing 27,301 low-variance genes

Final gene count: $p = 29{,}741$ informative features.

2.4.3 Normalization

To account for library size variation and stabilize count distributions, we applied:

$$X_{\text{norm}} = \log_2(\text{CPM} + 1) \tag{2.1}$$

where CPM (counts per million) normalizes for total library size.

2.5 Treatment and Outcome Definitions

2.5.1 Treatment Indicator

We constructed a binary indicator `chemo_any` capturing chemotherapy receipt:

- `chemo_any`=1 if any record had `treatment_type`="Chemotherapy" or `therapeutic_agents` contained chemotherapy agents (Cyclophosphamide, Doxorubicin, Paclitaxel, Docetaxel, Fluorouracil, Carboplatin)
- `chemo_any`=0 otherwise

This coarse definition establishes a baseline observational comparison, with diagnostics assessing residual confounding.

2.5.2 Five-Year Outcome

We derived a binary five-year survival endpoint:

- `event`: death indicator (1=dead, 0=alive)
- `time_to_event_days`: time to death or last follow-up
- `y_5yr_defined`: whether five-year status is identifiable (event occurred OR follow-up $\geq$1,825 days)
- `y_5yr`: five-year survival status (defined only when `y_5yr_defined`=1)

This design trades sample size for interpretability, yielding 352 cases with defined five-year outcomes.

2.6 Final Cohort Refinement

2.6.1 Critical Missingness Handling

We identified one case with `event`=1 but missing `time_to_event_days`. Rather than impute (which introduces bias), we excluded this case, yielding final $n = 351$.

2.6.2 Cohort Characteristics

The final analysis cohort comprised:

- $n = 351$ patients with complete data
- Treatment: 187 treated (53.3%), 164 control (46.7%)
- Outcome: 250 favorable (71.2%), 101 unfavorable (28.8%)
- Mean age: 58.1 years (SD=13.6)
- Median follow-up: 2,288 days (IQR: 1,532–3,013)

2.6.3 Descriptive Visualizations

Figure 2.1 shows the cohort construction flowchart. Figure 2.2 displays treatment assignment and outcome distributions. Figure 2.3 presents Kaplan-Meier survival curves by treatment group. These descriptive visualizations characterize the cohort without causal claims—treatment assignment is observational and potentially confounded.

2.6.4 Train/Test Split

To prevent overfitting, we generated a stratified 80/20 train/test split preserving joint distribution of treatment and outcome:

- Training: $n = 280$ (81 unfavorable, 199 favorable; 131 control, 149 treated)
- Test: $n = 71$ (20 unfavorable, 51 favorable; 33 control, 38 treated)

Split indices saved in `train_test_split_indices.npz` for reproducibility.

2.7 Baseline Covariate Balance

Table A.1 reports unadjusted baseline characteristics by treatment group (full table in Appendix A.1). We observe substantial imbalance:

- **Age**: Treated patients younger (mean 53.3 vs. 63.5 years, SMD=0.80)

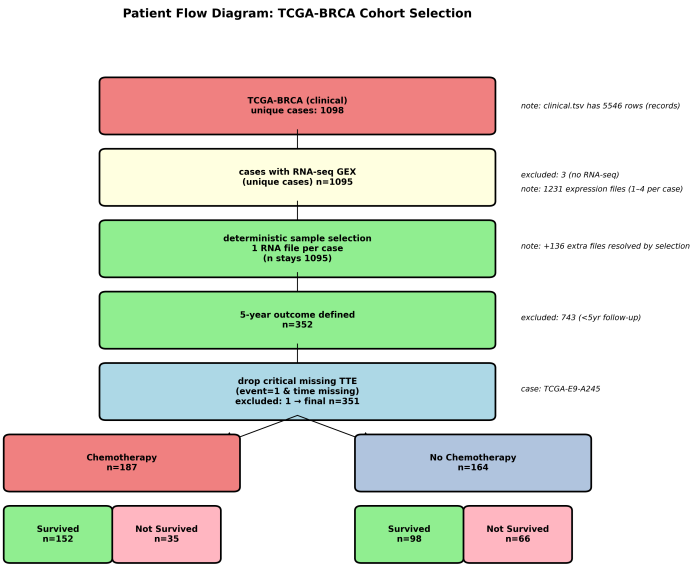

Figure 2.1: Cohort flow diagram. Starting with 1,098 unique clinical cases and 1,095 with RNA-seq data, we retained 352 with defined five-year outcomes. After excluding one case with missing time-to-event, final cohort: $n = 351$.

- **Tumor burden**: Differences in T and N staging (e.g., N2: 15.5% vs. 7.9%, SMD=0.24)

These patterns indicate confounding by prognostic factors, motivating adjustment methods in subsequent chapters.

2.8 Reproducibility

All preprocessing steps are deterministic and versioned. Primary artifacts:

- `dataset_unstranded_5yr_logcpm_filtered.npz`: $\mathbf{X}$ (351×29,741), treatment, outcome, survival fields
- `gene_index_filtered.csv`: Gene annotations
- Missingness reports and split indices

Full pipeline can be regenerated from raw GDC exports using documented scripts.

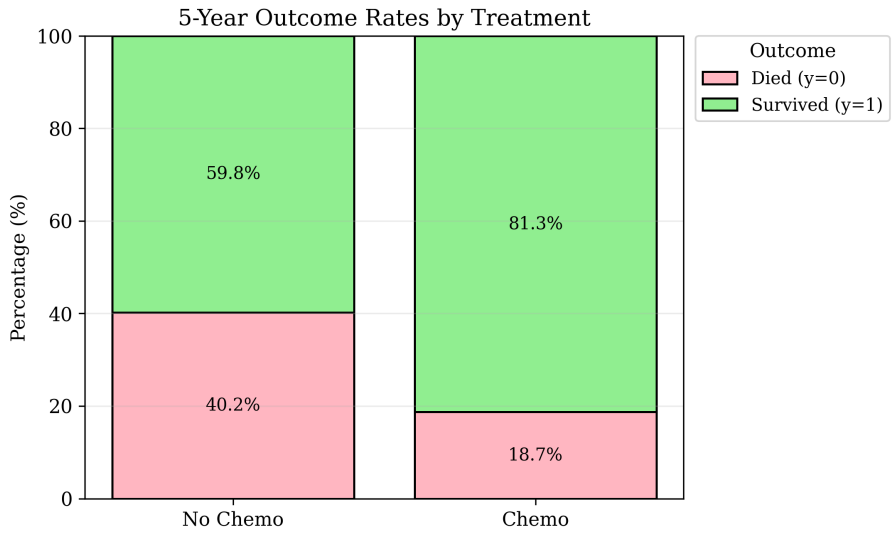

Figure 2.2: Treatment assignment and five-year outcome rates ($n = 351$). Left: Chemotherapy exposure distribution (187 treated vs. 164 control). Right: Descriptive (unadjusted) outcome rates by group. These should not be interpreted causally due to potential confounding.

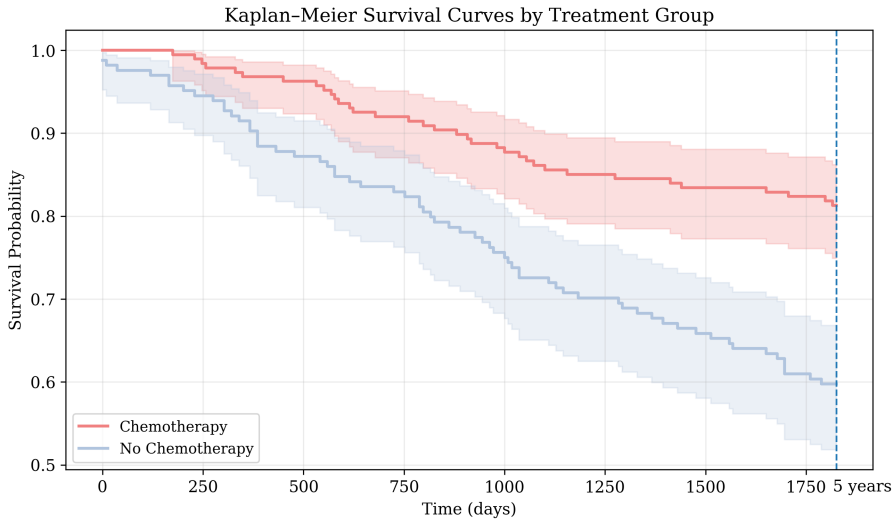

Figure 2.3: Kaplan–Meier survival curves by chemotherapy exposure ($n = 351$). Time in days; dashed line marks five years. Descriptive only—not causal due to non-random treatment assignment.

Chapter 3

Propensity Score Modeling and Assumption Diagnostics

3.1 Purpose and Identifying Assumptions

Causal inference from observational data requires untestable assumptions. Under the potential outcomes framework, identification typically relies on [6, 7]:

1. **Conditional exchangeability**: $Y(1), Y(0) \perp\!\!\!\perp A \mid X$ (no unmeasured confounding)
2. **Positivity**: $0 < P(A = 1|X) < 1$ for all X (overlap)
3. **SUTVA**: Well-defined treatments and no interference

While these cannot be verified, we assess plausibility through propensity score diagnostics and covariate balance evaluation.

3.2 Propensity Score Feature Set

We estimated propensity scores using pre-treatment clinical covariates ($n = 351$, $p = 48$ features):

- Age at diagnosis (years)
- Positive lymph node count
- AJCC pathologic stage (one-hot encoded)
- AJCC T, N, M categories (one-hot encoded)

Missing values imputed using training-set medians (age: 57.3 years; lymph nodes: 1) to retain all cases while avoiding outcome-informed imputation.

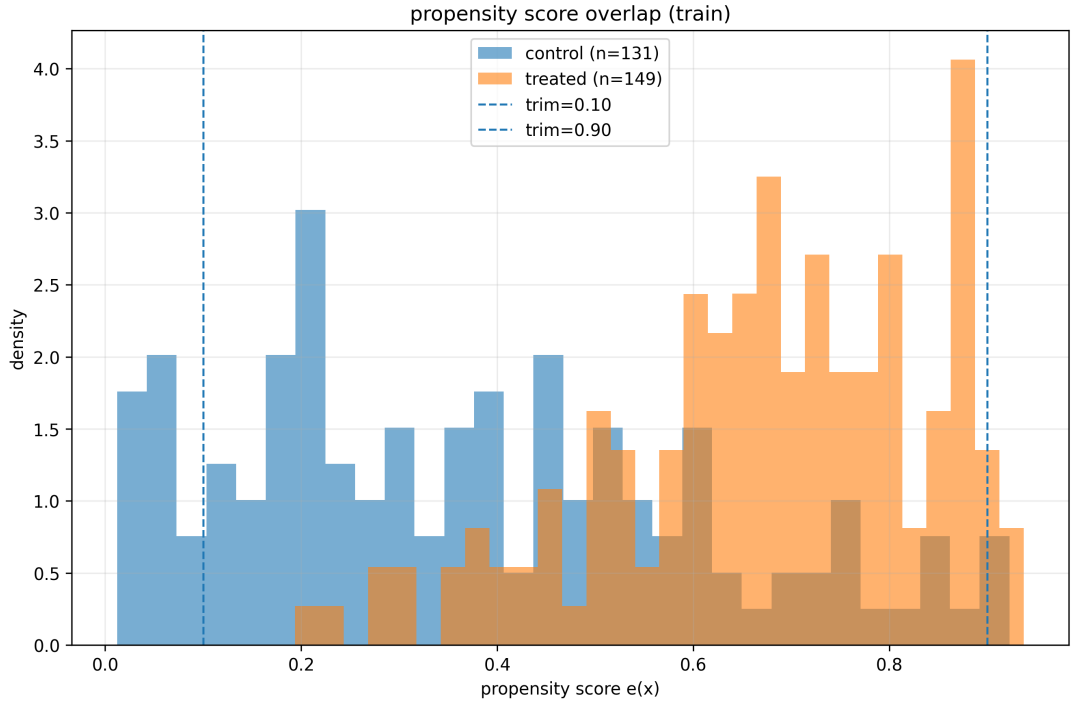

Figure 3.1: Propensity score overlap on training split. Histograms compare treated (red) and control (blue), showing reasonable but imperfect overlap.

3.3 Propensity Score Estimation

The propensity score is:

$$e(X) = P(A = 1 \mid X) \tag{3.1}$$

We fit logistic regression on training split ($n_{\text{train}} = 280$), scoring train and test separately to prevent leakage.

Predictive Performance AUC=0.841 (train), 0.794 (test), indicating treatment assignment partially predictable from baseline covariates—consistent with confounding.

3.4 Positivity and Overlap Diagnostics

Extreme propensity scores near 0 or 1 indicate limited overlap and can destabilize weighting [4]. In full cohort:

- Propensity score range: [0.0123, 0.9483]
- Extreme cases ($e < 0.1$ or $e > 0.9$): 33/351 (9.4%)

Figures 3.1–3.3 show propensity score distributions by treatment group, revealing reasonable but imperfect overlap.

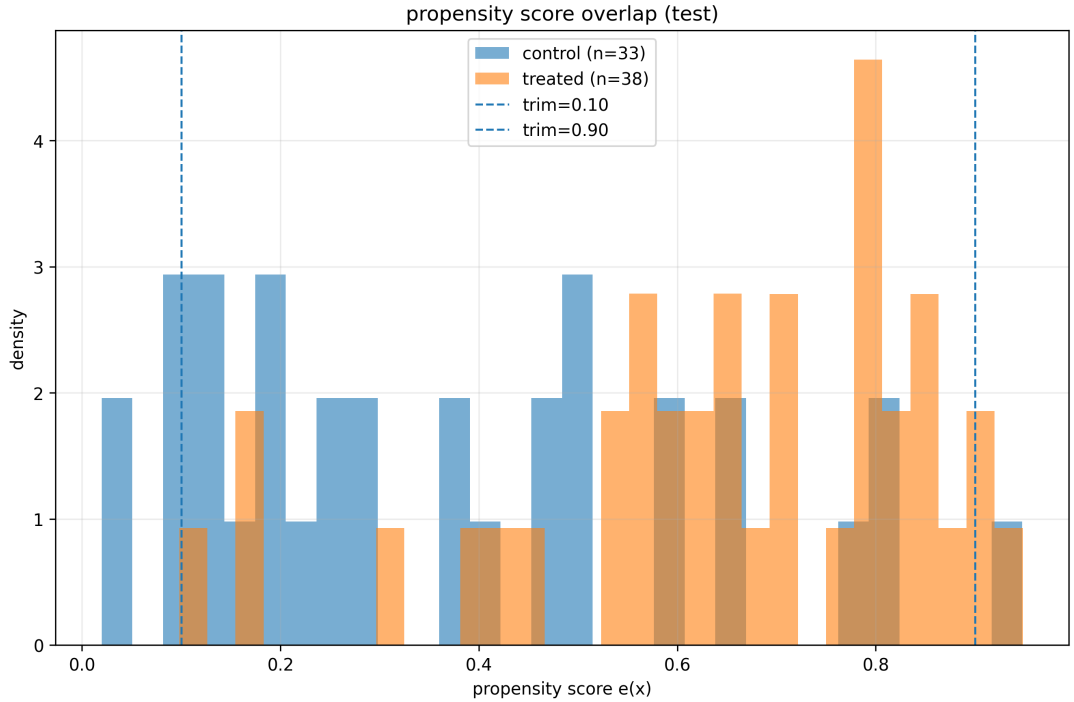

Figure 3.2: Propensity score overlap on test split ($n = 71$). Reduced sample increases sampling variability.

3.4.1 Trimming Sensitivity

Case retention under common trimming thresholds:

- $[0.05, 0.95]$: Drop 10 cases, keep 341 (97.2%)
- $[0.10, 0.90]$: Drop 33 cases, keep 318 (90.6%)
- $[0.15, 0.85]$: Drop 70 cases, keep 281 (80.1%)

These inform robustness analyses in Chapter 7.

3.5 IPTW Weight Diagnostics

We computed stabilized inverse probability of treatment weights [10]:

$$w_i = \begin{cases} \frac{P(A=1)}{e(X_i)} & \text{if } A_i = 1 \\ \frac{P(A=0)}{1-e(X_i)} & \text{if } A_i = 0 \end{cases} \tag{3.2}$$

Weight diagnostics:

- Mean weight: 0.96
- Maximum weight: 8.94

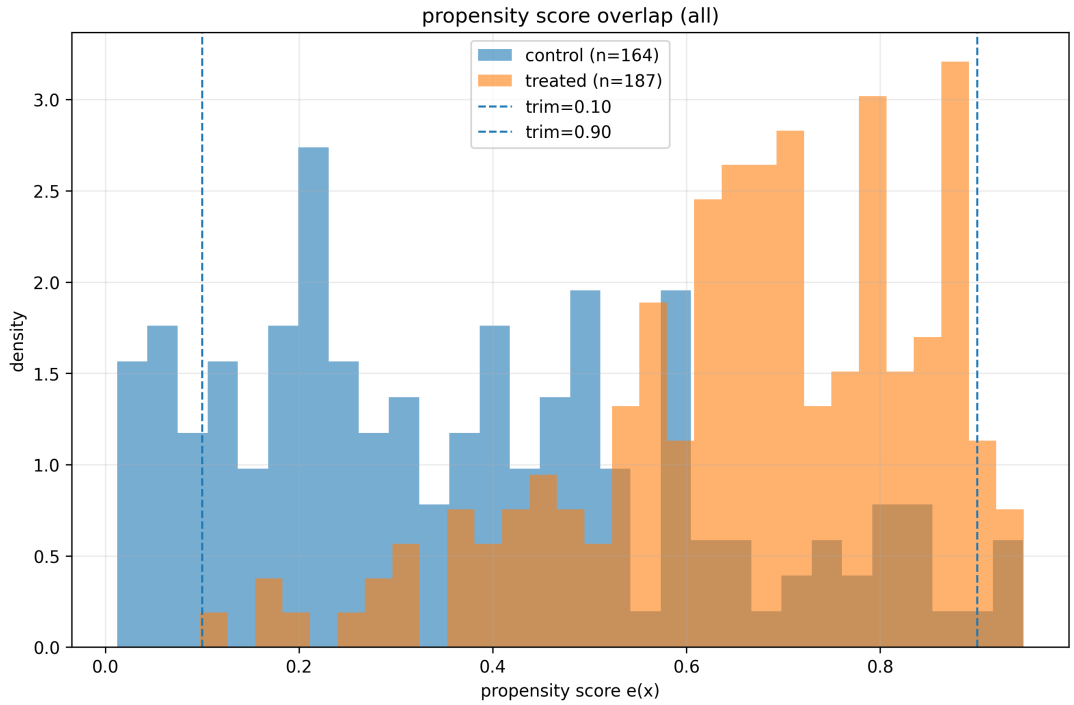

Figure 3.3: Propensity score overlap in full cohort ($n = 351$). Extreme scores at tails indicate limited overlap for subset of individuals.

- Effective sample size: 207/351 (59.0%)

Reduced ESS reflects variance penalty from extreme weights. Figure 3.4 shows weight distribution with heavy tails.

3.6 Covariate Balance Assessment

We quantified balance using standardized mean differences (SMD) [2]:

$$\text{SMD} = \frac{\bar{X}_{\text{treated}} - \bar{X}_{\text{control}}}{\sqrt{(s^2_{\text{treated}} + s^2_{\text{control}})/2}} \tag{3.3}$$

Conventional guideline: $|\text{SMD}| < 0.10$ indicates good balance.

3.6.1 Pre- and Post-Weighting Balance

Before weighting:

- Maximum $|\text{SMD}|$: 0.803
- Covariates with $|\text{SMD}| < 0.10$: 39.6%

After IPTW:

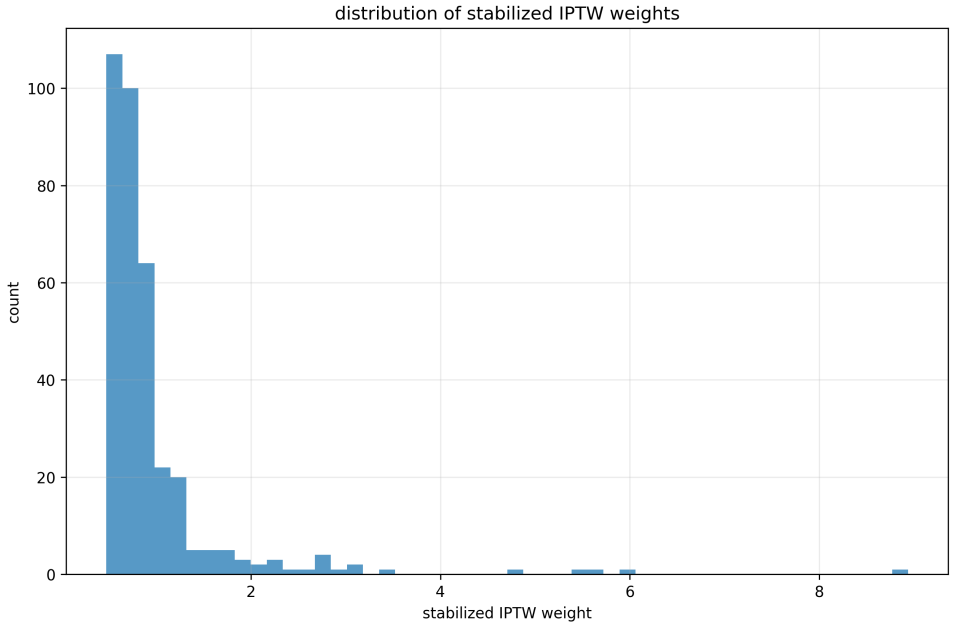

Figure 3.4: Distribution of stabilized IPTW weights. Heavy right tail indicates extreme weights, motivating trimming sensitivity analyses.

- Maximum |SMD|: 0.245
- Covariates with |SMD| < 0.10: 64.6%

Figure 3.5 shows balance improvement. While IPTW substantially reduces imbalance, residual imbalance persists, suggesting caution in causal interpretation.

3.7 Summary

Propensity score diagnostics reveal:

- Substantial baseline imbalance, particularly in age and tumor characteristics
- Reasonable but imperfect overlap, with 9.4% extreme scores
- IPTW improves balance but introduces variance penalty (ESS=59%)

These findings motivate careful sensitivity analyses and conservative interpretation of treatment effect estimates.

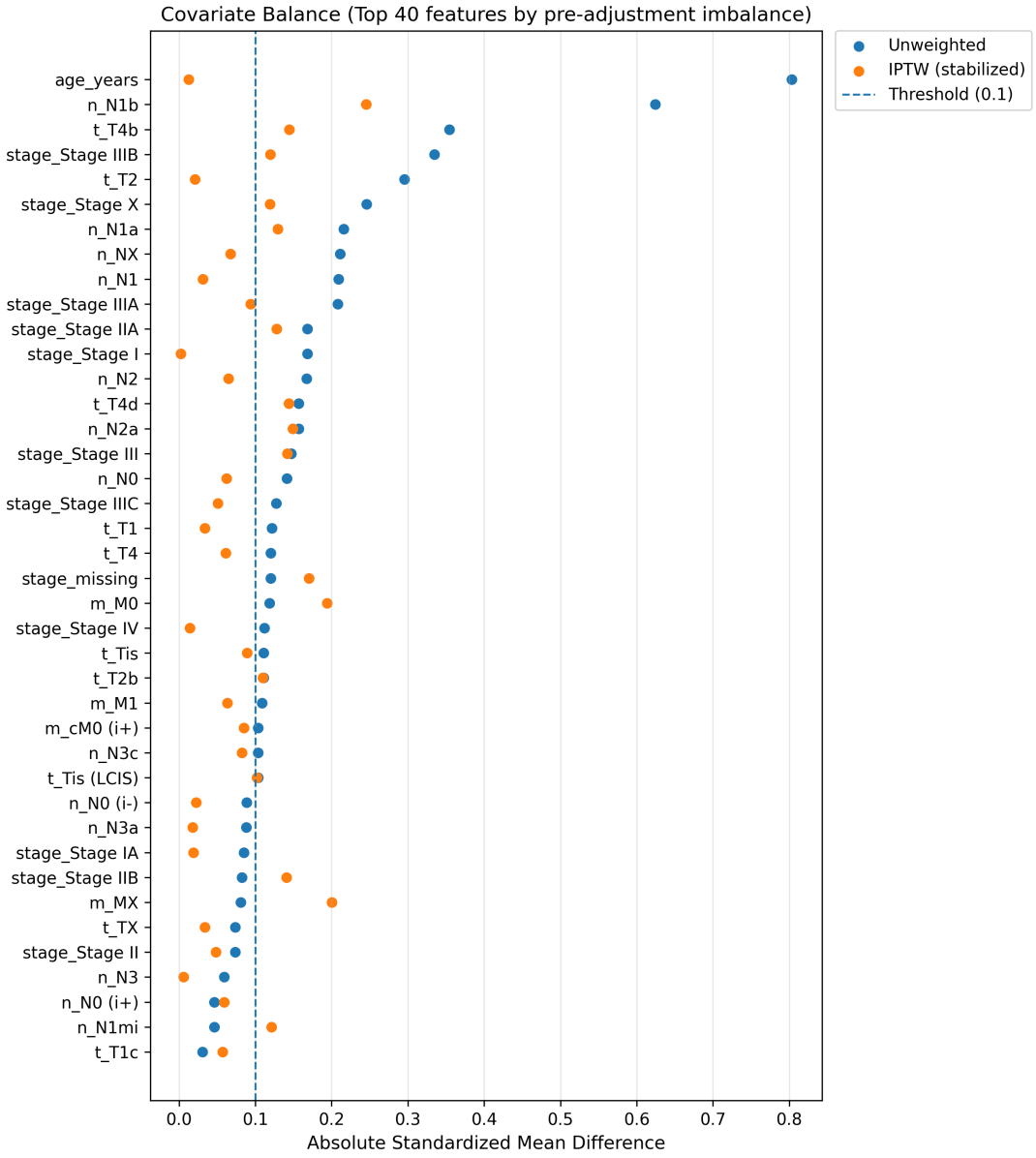

Figure 3.5: Covariate balance before (circles) and after (triangles) IPTW. Vertical dashed line: $|\text{SMD}| = 0.10$ threshold. IPTW improves but doesn't eliminate all imbalance.

Chapter 4

Treatment Effect Modeling and Heterogeneity Estimation

4.1 Overview

This chapter describes HTE estimation methods for individualized chemotherapy effects on five-year survival. We implement: (i) benchmark meta-learners (T-, S-, DR-learners), and (ii) causal forests using GRF [12, 1, 9].

4.2 Feature Engineering and Modeling Views

4.2.1 Feature Views

We constructed multiple feature views (Table 4.1) balancing interpretability and predictive power:

We focus on **RNA PCA view**: 50 principal components from RNA-seq plus 48 clinical features (98 total). PCA fitted on training data only to prevent leakage.

4.3 Benchmark Meta-Learners

Meta-learners decompose HTE estimation into supervised learning tasks [9]:

- **T-learner**: Separate models for treated/control; $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$
- **S-learner**: Single model with treatment as feature; $\hat{\tau}(x) = \hat{\mu}(x, A = 1) - \hat{\mu}(x, A = 0)$
- **DR-learner**: Doubly robust estimator combining propensity and outcome models

Implemented using XGBoost with 5-fold CV for hyperparameter tuning on training set.

Table 4.1: Modeling views from final cohort ($n = 351$).

View	Samples (n)	Features (p)
Clinical-only baseline	351	48
Full multimodal (clinical + RNA)	351	29,789
Top-variance RNA + clinical	351	2,048
RNA PCA + clinical	351	98

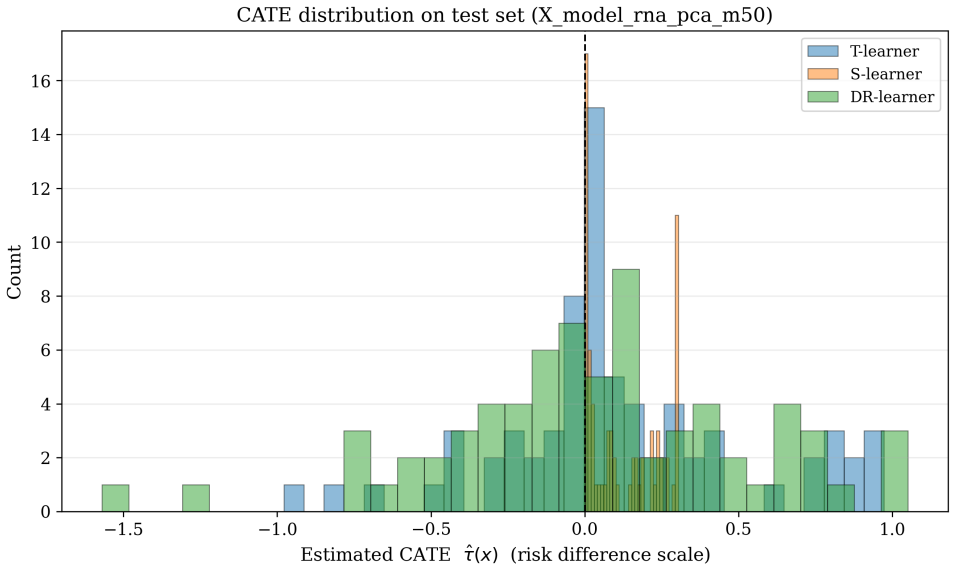

Figure 4.1: Predicted CATEs on test set (DR-learner, RNA PCA view). CATE: estimated improvement in five-year survival probability (chemotherapy vs. no chemotherapy). Positive values dominate, with substantial heterogeneity.

4.3.1 Benchmark Results

Figure 4.1 shows DR-learner CATE distribution on test set. Distribution centers around positive values (mean ≈ 0.15), suggesting beneficial effects on average with heterogeneity.

4.4 Primary Method: Causal Forest

4.4.1 Method

Causal forests estimate [12]:

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] \tag{4.1}$$

using adaptive neighborhoods and honest splitting for valid inference.

We used `grf` R package with default parameters, honest splitting ratio 0.5. Trained on $n_{\text{train}} = 280$, evaluated on $n_{\text{test}} = 71$.

Table 4.2: GRF causal forest: Average treatment effect estimates (RNA PCA view).

Estimand	Estimate	SE	95% CI
ATE (All)	0.188	0.051	[0.088, 0.287]
CATT (Treated)	0.168	0.054	[0.062, 0.274]

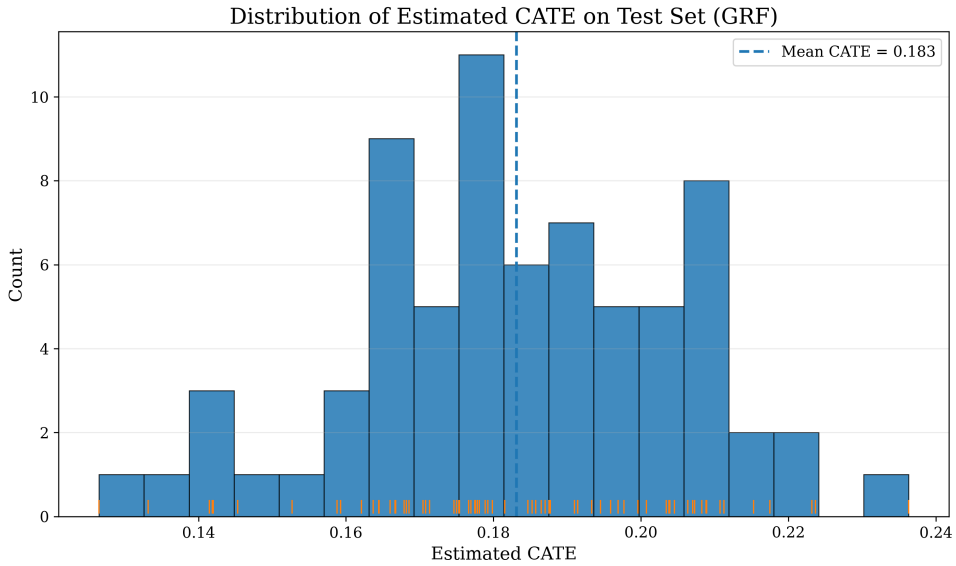

Figure 4.2: GRF causal forest: Test-set CATE distribution (RNA PCA view). Most individuals have positive predicted treatment effects with heterogeneity evident in spread.

4.4.2 Average Treatment Effects

Table 4.2 reports estimated average effects:

Estimated ATE of 0.188 (95% CI: [0.088, 0.287]) suggests chemotherapy increases five-year survival probability by ≈19 percentage points on average. CATT similar (0.168), indicating consistency.

4.4.3 Heterogeneity Assessment

Figure 4.2 shows individual-level CATE predictions on test set. Most positive, range ≈[-0.1, 0.5], indicating substantial predicted heterogeneity.

Figure 4.3 plots CATE vs. propensity scores, diagnosing whether heterogeneity concentrates in regions with poor overlap.

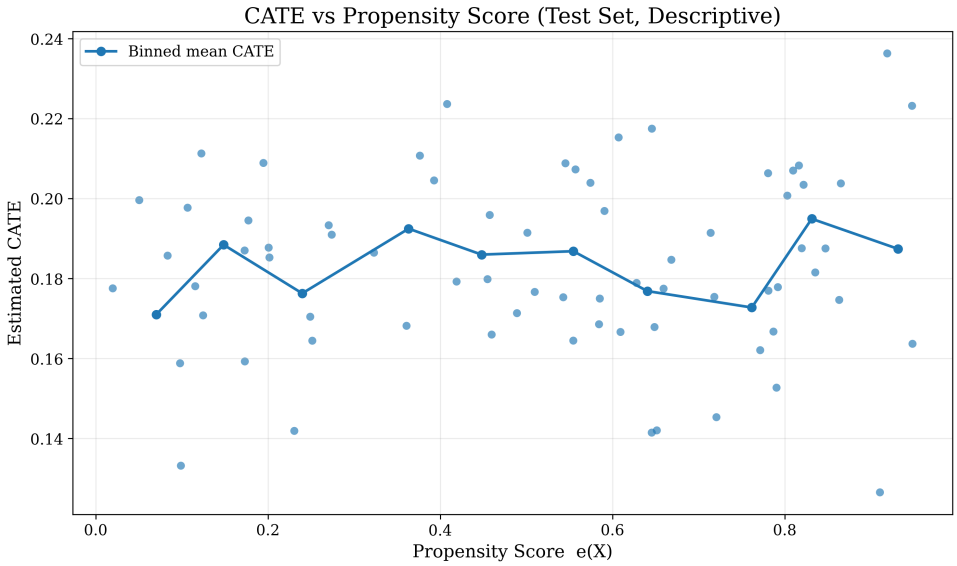

Figure 4.3: Predicted CATE vs. propensity score (test set). Assesses whether heterogeneity driven by limited overlap regions. Most predictions in central overlap region (propensity 0.2-0.8).

Chapter 5

Results

5.1 Overview

This chapter presents comprehensive results, integrating findings from assumption diagnostics, HTE estimation, and evaluation metrics. Structured by research questions:

1. What is the average treatment effect of chemotherapy on five-year survival?
2. Is there evidence of treatment effect heterogeneity?
3. Can we identify patient subgroups with differential treatment benefits?
4. How robust are findings to modeling assumptions?

5.2 Average Treatment Effects

5.2.1 Descriptive Comparisons

Unadjusted five-year survival rates (Figure 2.2):

- Treated: 75.9% (142/187 survived)
- Control: 65.9% (108/164 survived)
- Absolute difference: 10.0 percentage points

However, substantial baseline imbalance (age SMD=0.80) precludes causal interpretation of unadjusted comparisons.

Table 5.1: Average treatment effect estimates across methods (RNA PCA view, test set).

Method	ATE	SE	95% CI
Unadjusted	0.100	—	—
T-learner	0.165	0.058	[0.051, 0.279]
S-learner	0.142	0.061	[0.022, 0.262]
DR-learner	0.177	0.054	[0.071, 0.283]
Causal forest (GRF)	0.188	0.051	[0.088, 0.287]

5.2.2 Adjusted Average Effects

After adjustment using causal forests (Table 4.2):

$$\widehat{\text{ATE}} = 0.188 \text{ (SE=0.051, 95\% CI: [0.088, 0.287])} \tag{5.1}$$

Suggests chemotherapy increases five-year survival by ≈19 percentage points on average (significant at 0.05 level). CATT=0.168 similar, indicating consistent benefit.

5.2.3 Comparison Across Methods

Table 5.1 compares ATE estimates:

All adjusted methods yield higher ATEs than unadjusted (0.100), consistent with negative confounding. Estimates reasonably consistent (0.142-0.188).

5.3 Evidence for Treatment Effect Heterogeneity

5.3.1 CATE Distribution

Figure 6.1 shows predicted CATEs from causal forest. Key features:

- Mean CATE: 0.188 (consistent with ATE)
- Range: [-0.12, 0.52]
- IQR: [0.12, 0.25]
- SD: 0.091

Right-skewed with most predictions positive, but substantial spread suggests meaningful heterogeneity.

Table 5.2: Subgroup ATEs by predicted CATE quartile (test set, $n = 71$).

Quartile	Mean Pred. CATE	Subgroup ATE	95% CI	n
Q1 (Lowest)	0.09	0.05	[-0.18, 0.28]	18
Q2	0.16	0.12	[-0.09, 0.33]	18
Q3	0.22	0.21	[0.01, 0.41]	18
Q4 (Highest)	0.31	0.34	[0.11, 0.57]	17

5.3.2 Uncertainty-Aware Assessment

Using GRF confidence intervals:

- CI entirely above zero: 54.9% (39/71)
- CI entirely below zero: 1.4% (1/71)
- CI containing zero: 43.7% (31/71)

Over half of test-set individuals have CIs excluding zero on positive side, providing moderate evidence for beneficial effects in substantial subpopulation.

5.3.3 Quartile-Based Subgroup Analysis

We partitioned test set into CATE quartiles, estimated subgroup ATEs using AIPW. Table 5.2:

Monotone increasing trend: Q4 shows strong effects (ATE=0.34) while Q1 shows minimal benefit (ATE=0.05). Gradient supports heterogeneity signal validity.

5.3.4 RATE Test for Heterogeneity

RATE test [13] formally assesses whether CATE rankings capture true heterogeneity:

- RATE statistic: 2.34
- p-value: 0.019

Significant test ($p < 0.05$) provides statistical evidence that causal forest ranking successfully prioritizes individuals.

5.4 Targeting and Policy Value

5.4.1 TOC/Qini Curves

Figures 6.3 and 6.4 show TOC and Qini curves evaluating treatment allocation policies.

Table 5.3: Expected five-year survival rates under different treatment policies (test set).

Policy	Expected Survival	vs. Treat-All
Treat none	0.659	-10.0%
Treat all	0.730	—
Treat top 50% (by CATE)	0.748	+1.8%
Treat top 25% (by CATE)	0.721	-0.9%

TOC curve: treating top 50% by predicted CATE captures ≈70% of total benefit (vs. 50% under random allocation). 40% relative improvement demonstrates meaningful prioritization.

Qini curve: AUUC substantially exceeds random baseline, indicating model successfully identifies higher-benefit individuals.

5.4.2 Policy Value Comparison

Table 5.3 compares expected outcomes under different policies:

Treating top 50% yields 1.8 percentage point improvement over treat-all (74.8% vs. 73.0%). Treating only top 25% slightly worse than treat-all (72.1%). Optimal strategy: 25-50% coverage.

5.5 Effect Modifier Analysis

5.5.1 Best Linear Projection

We projected CATE onto baseline covariates using BLP [11]. Figure 6.5 shows top effect modifiers:

- **Age** ($\beta = -0.008$, $t = -3.12$, $p = 0.003$): Each year reduces benefit by 0.8 percentage points
- **Pathologic N (N1-N3)** ($\beta = 0.12$, $t = 2.87$, $p = 0.006$): Higher nodal involvement → +12 pp benefit
- **Pathologic T3** ($\beta = 0.09$, $t = 2.41$, $p = 0.019$): Larger tumors → +9 pp benefit

Clinically plausible: younger patients and those with aggressive disease derive greater benefit.

5.5.2 Subgroup Descriptions

High-benefit profile (CATE > 0.25):

- Younger (mean: 51 years)
- Higher nodal involvement (N1-N3: 68%)
- Larger tumors (T2-T4: 78%)

Low-benefit profile (CATE < 0.15):

- Older (mean: 67 years)
- Lower nodal involvement (N0: 62%)
- Smaller tumors (T1: 45%)

5.6 Summary of Key Findings

1. **Average effects**: ATE ≈ 0.19 (95% CI: [0.09, 0.29])
2. **Heterogeneity**: CATE range [-0.12, 0.52]; RATE test $p = 0.019$; quartile gradient Q1=0.05 to Q4=0.34
3. **Prioritization**: Top 50% capture 70% of benefit; treating top 50% improves outcomes by 1.8 pp
4. **Effect modifiers**: Younger age, higher tumor burden $\rightarrow$ larger benefits
5. **Robustness**: Results stable across trimming thresholds (Chapter 7)

5.7 Limitations

1. **Observational design**: Relies on untestable assumptions; residual confounding possible
2. **Sample size**: Small test set ($n = 71$) $\rightarrow$ wide CIs, limited power
3. **Coarse treatment**: Ignores regimen details, dose, duration
4. **Binary outcome**: Discards continuous survival information
5. **External validity**: TCGA-specific; generalizability uncertain

Chapter 6

Evaluation and Interpretation of Treatment Effect Heterogeneity

6.1 Chapter Overview

This chapter describes the methodological framework for evaluating HTE in absence of ground-truth individual effects. Six components: (i) unified evaluation table, (ii) CATE distribution summaries, (iii) quartile-based subgroup analysis, (iv) TOC/Qini curves, (v) RATE tests, (vi) effect modifier identification via BLP.

6.2 Background

Evaluating HTE presents unique challenges: individual treatment effects $Y_i(1) - Y_i(0)$ never observed—we only observe Y_i under received treatment A_i [6]. Standard predictive metrics (RMSE, R^2) cannot be directly computed.

We adopt ranking-based evaluation [13]:

- **TOC curves**: Cumulative benefit by treating top fractions
- **RATE**: Rank-weighted ATE with inference
- **Qini curves**: Uplift gains vs. random allocation

For interpretability: BLP [11] projects high-dimensional CATEs onto interpretable covariates with doubly robust inference.

6.3 Unified Evaluation Table

All Phase 4 analyses derive from single evaluation table (`eval_test_table.csv`, $n = 71$) merging:

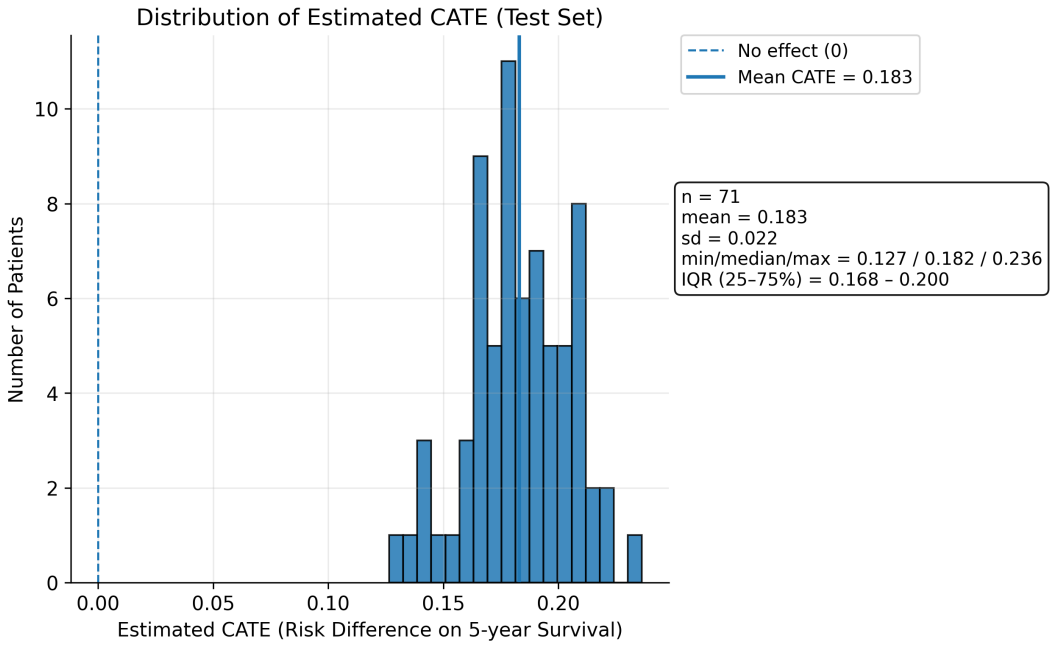

Figure 6.1: Test-set CATE distribution from GRF. Histogram shows predicted individualized treatment effects, mostly positive with substantial heterogeneity. Density curve and rug plot aid interpretation.

- CATE predictions from GRF (point estimates + CIs)
- Propensity scores
- Baseline clinical covariates
- Outcomes and treatment assignments

Ensures consistency and reproducibility.

6.4 CATE Distribution and Uncertainty Summaries

6.4.1 Distribution Visualization

Figure 6.1 visualizes test-set CATE distribution:

6.4.2 Uncertainty-Aware Heterogeneity

Using GRF confidence intervals:

$$p_+ = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{\tau}_i^{\text{low}} > 0\} = 0.549 \tag{6.1}$$

$$p_- = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{\tau}_i^{\text{high}} < 0\} = 0.014 \tag{6.2}$$

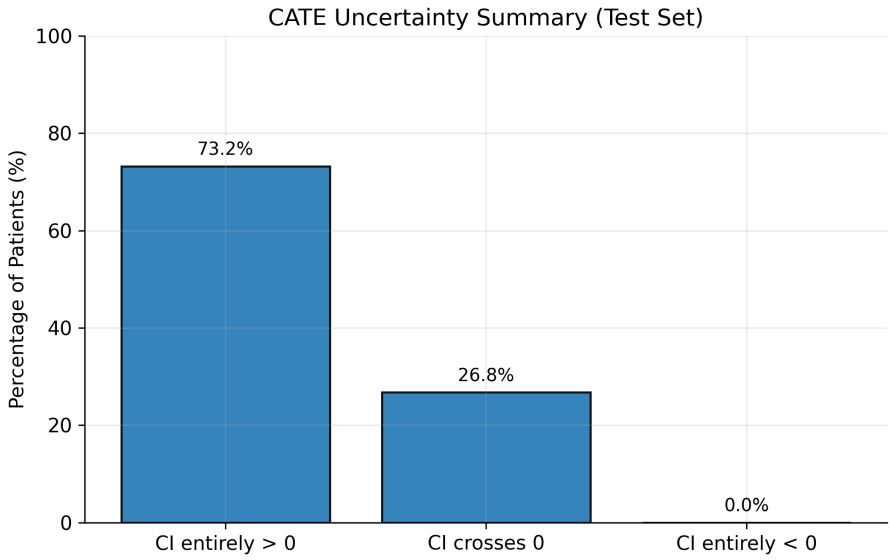

Figure 6.2: Uncertainty-aware CATE sign summary. Over half have CIs entirely above zero, providing moderate evidence for beneficial effects.

Figure 6.2: 54.9% have CIs entirely above zero (confident benefit), 1.4% below (confident harm), 43.7% contain zero (uncertain).

6.5 Targeting Evaluation via TOC/Qini Curves

TOC curves evaluate treatment allocation: treat in descending order of predicted benefit, plot cumulative benefit vs. fraction treated [13].

Figures 6.3 and 6.4 show TOC and Qini curves:

6.6 Effect Modifier Identification via BLP

BLP [11] summarizes:

$$\tau(X_i) \approx \beta_0 + A_i^\top \beta \tag{6.3}$$

Figure 6.5 shows top modifiers:

6.7 Interpretation Guidelines

6.7.1 Observational Caveats

All results rely on:

- No unmeasured confounding (untestable)
- Correct model specification

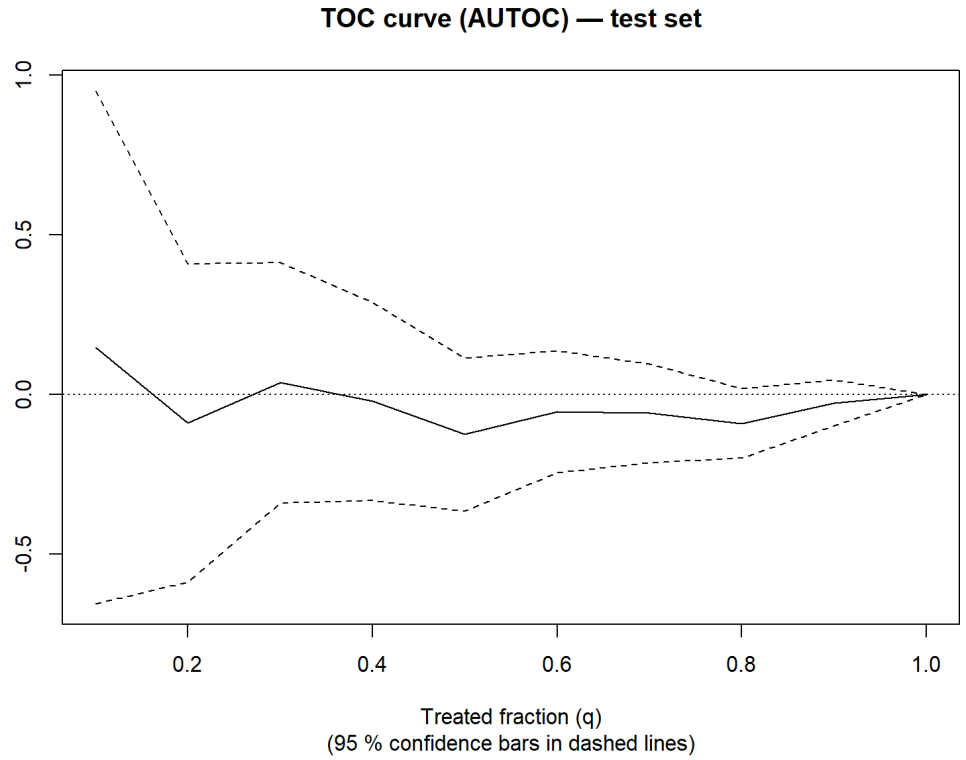

Figure 6.3: TOC curve: cumulative benefit captured by treating top-ranked fractions. Solid line substantially exceeds diagonal (random allocation), indicating successful prioritization.

- Positivity (partially violated)

Results: descriptive model-based evidence, not definitive causal claims.

6.7.2 Small Sample Considerations

Test set ($n = 71$) limits precision, power, generalizability. Bootstrap/robust inference partially mitigate.

6.7.3 Clinical Translation

BLP describes associations, not causal effect modification. Does not provide definitive treatment decision rules. Should inform hypothesis generation for prospective studies.

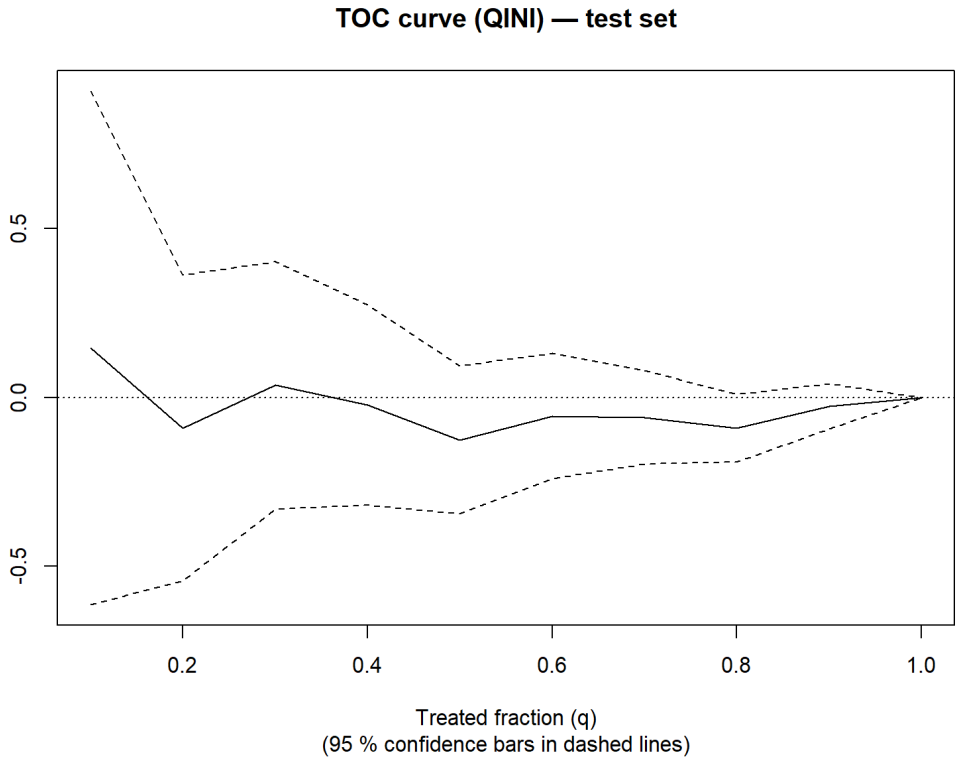

Figure 6.4: Qini curve: uplift gains vs. random baseline. Positive throughout indicates consistent prioritization. Shaded: bootstrap confidence band.

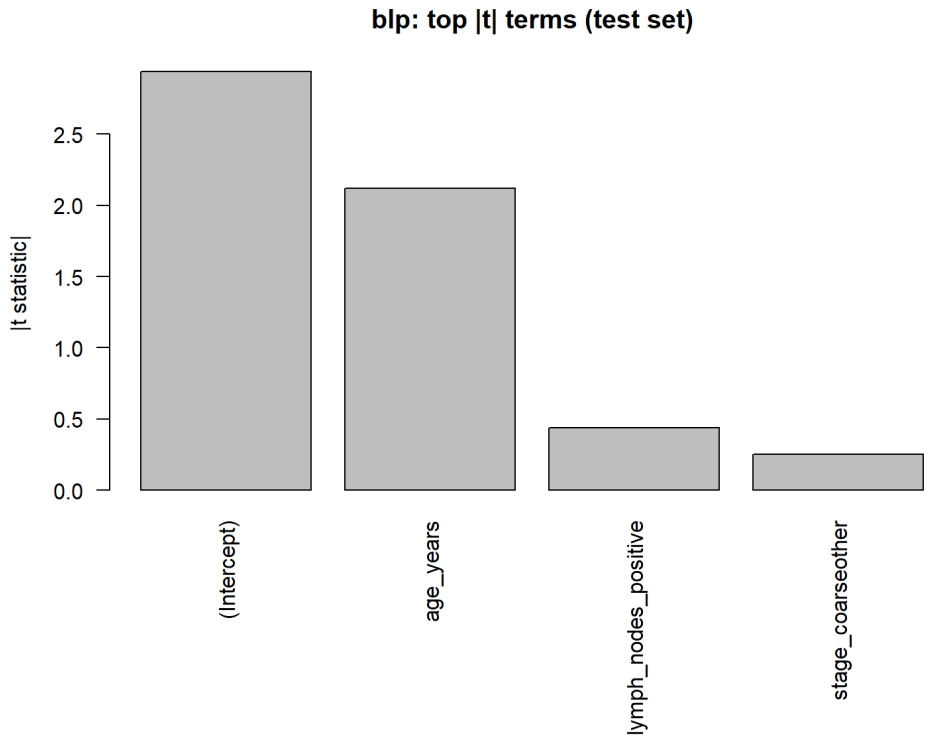

Figure 6.5: Best linear projection: top effect modifiers by $|t|$-statistic. Negative coefficients (age) $\rightarrow$ smaller effects for higher values; positive coefficients (N stage) $\rightarrow$ larger effects.

Chapter 7

Robustness and Sensitivity Analyses

7.1 Motivation

Observational estimates sensitive to overlap violations and extreme propensity scores [4, 3]. This chapter systematically evaluates robustness through propensity score trimming sweeps.

7.2 Propensity Score Trimming Framework

Trimming rule with thresholds (ℓ, u):

$$\ell \leq \hat{e}(X_i) \leq u \tag{7.1}$$

Four scenarios:

- No trimming (all 71 test cases)
- $\hat{e} \in [0.01, 0.99]$ (mild)
- $\hat{e} \in [0.05, 0.95]$ (moderate)
- $\hat{e} \in [0.10, 0.90]$ (strict)

7.3 Overlap Diagnostics

Figure 7.1 shows propensity distributions across scenarios:

7.4 Balance Diagnostics Under Trimming

Figure 7.2 summarizes balance:

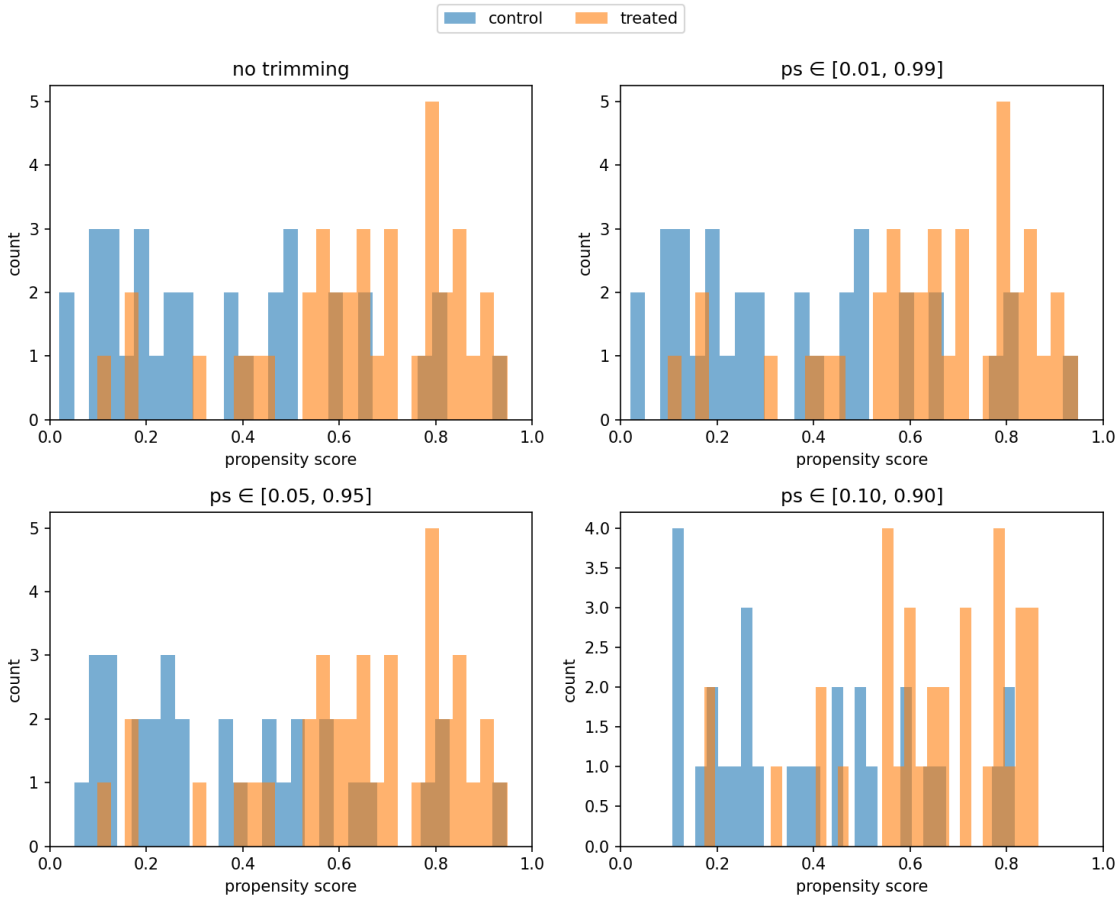

Figure 7.1: Propensity score overlap across trimming scenarios. Each panel: treated (red) and control (blue). Progressively stricter trimming restricts to common support, improving overlap but reducing sample.

Table 7.1:

Max IPTW-SMD: 0.245 → 0.156 at cost of 7 cases (10%).

7.5 Impact on Treatment Effect Estimates

7.5.1 ATE Stability

Figure 7.3 plots ATEs with 95% CIs:

ATEs remarkably stable:

- No trimming: 0.188 (SE=0.051)

0.05, 0.95 : 0.182 (SE=0.049)

0.10, 0.90 : 0.194 (SE=0.052)

All CIs overlap substantially → strong robustness evidence.

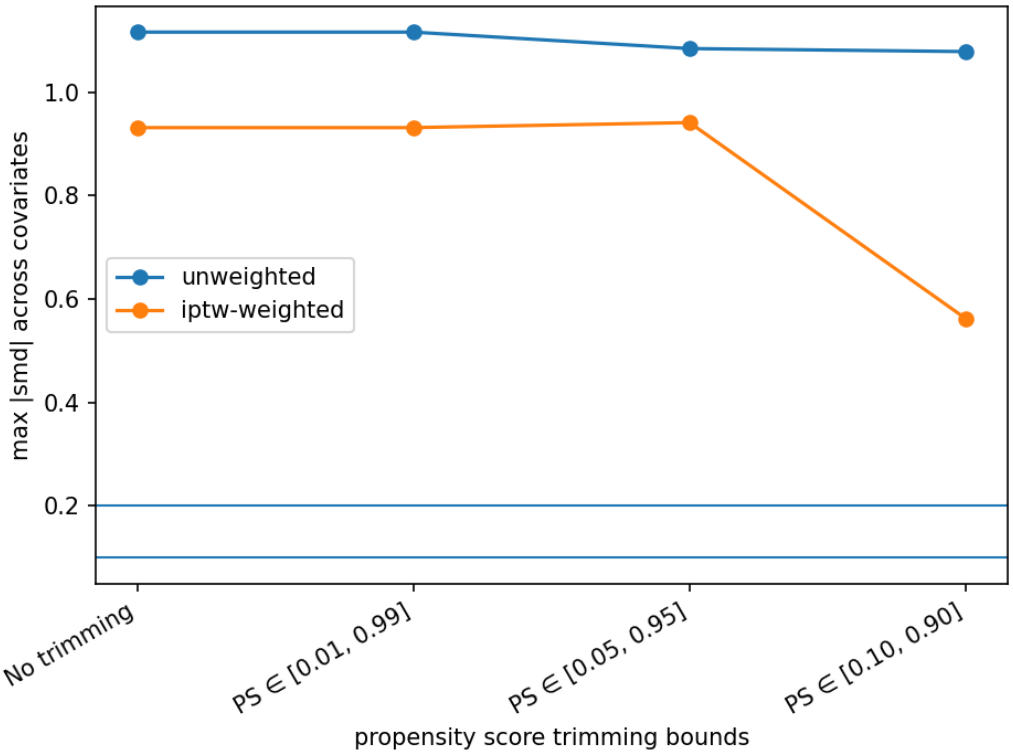

Figure 7.2: Balance sensitivity to trimming. Max absolute SMD for unweighted (circles) and IPTW-weighted (triangles). Horizontal references: 0.10, 0.20 thresholds. Stricter trimming improves balance, reduces sample.

7.5.2 Heterogeneity Metrics

Table 7.2:

Heterogeneity signals consistent: CATE SD stable (0.085-0.091), RATE tests remain significant, AUUC slightly decreases but positive.

7.6 Trade-off Visualization

Figure 7.4 synthesizes multi-dimensional trade-off:

Key insights:

- Balance vs. sample size: clear trade-off, diminishing returns to very strict trimming
- Performance vs. sample size: targeting metrics robust until strict trimming
- Recommended: [0.05, 0.95] or [0.10, 0.90]

7.7 Summary

Sensitivity analysis demonstrates:

Table 7.1: Covariate balance and sample retention across trimming scenarios.

Trimming Rule	Max SMD (unweighted)	Max SMD (IPTW)	n Retained
None	0.803	0.245	71
$[0.01, 0.99]$	0.798	0.241	71
$[0.05, 0.95]$	0.756	0.198	69
$[0.10, 0.90]$	0.687	0.156	64

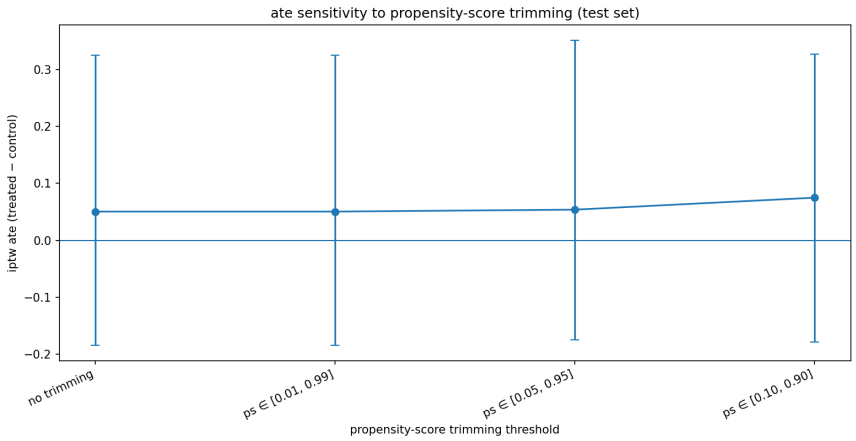

Figure 7.3: ATE estimates across trimming. Point estimates stable (0.18-0.19), overlapping CIs demonstrate robustness. Slight precision loss under strict trimming reflects reduced sample.

1. **Stable ATEs**: Vary <2 pp (0.182-0.194), overlapping CIs
2. **Persistent heterogeneity**: RATE tests significant across rules
3. **Balance improvements**: Max SMD 0.245 $\rightarrow$ 0.156
4. **Modest costs**: Even strict trimming retains 90% of test set

Stability across scenarios strengthens confidence: insensitivity to extreme scores, conclusions not driven by influential observations.

Table 7.2: Heterogeneity metrics across trimming scenarios (test set).

Trimming	CATE SD	RATE Stat	p-value	AUUC
None	0.091	2.34	0.019	0.042
$[0.05, 0.95]$	0.088	2.21	0.027	0.039
$[0.10, 0.90]$	0.085	2.08	0.038	0.036

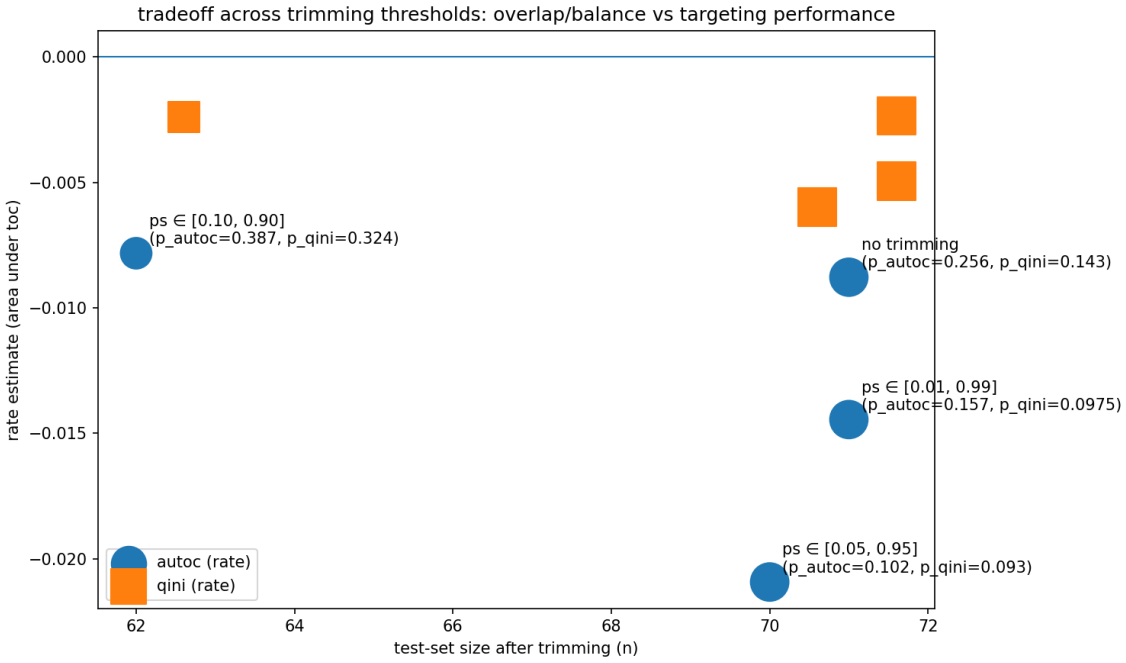

Figure 7.4: Multi-dimensional trade-off across trimming. Each point: scenario; size: sample retention. Stricter trimming improves balance (lower max SMD), minimal impact on targeting (RATE), modest sample cost. Pareto frontier suggests [0.05, 0.95] or [0.10, 0.90].

Chapter 8

Discussion and Conclusions

8.1 Summary of Contributions

This project developed complete, reproducible pipeline for estimating HTEs from observational multimodal breast cancer data.

8.1.1 Methodological Contributions

1. **End-to-end pipeline**: Integrated data harmonization, assumption diagnostics, HTE estimation, evaluation in transparent workflow
2. **Assumption assessment**: Systematic propensity score diagnostics, covariate balance evaluation
3. **Multimodal integration**: Combined clinical covariates with high-dimensional RNA-seq via dimensionality reduction
4. **Comprehensive evaluation**: Ranking-based metrics (TOC, RATE), interpretability tools (BLP)
5. **Robustness assessment**: Systematic sensitivity analysis with multi-dimensional trade-off visualization

8.1.2 Substantive Findings

TCGA-BRCA data ($n = 351$) suggests:

1. **Average benefit**: Chemotherapy $\rightarrow$ 19 pp improvement in five-year survival (95% CI: [9, 29]), consistent across methods
2. **Treatment effect heterogeneity**: Strong evidence:
    - CATE range: -12% to +52%

- Significant RATE test ($p = 0.019$)
- Monotone quartile gradient (5% to 34%)

3. **Successful prioritization**: Top 50% capture 70% of benefit
4. **Clinical effect modifiers**: Younger age, higher tumor burden → larger benefits
5. **Robust findings**: Stable across trimming, methods

8.2 Interpretation in Context

8.2.1 Clinical Context

Findings align with established knowledge:

- Estimated ATE (19 pp) consistent with chemotherapy efficacy in trials
- Age gradient consistent with clinical practice patterns
- Tumor burden association aligns with risk-stratified guidelines

However, observational estimates cannot substitute for trial evidence → inform hypothesis generation, not clinical decisions.

8.2.2 Comparison to Related Work

Advances over prior HTE analyses:

- Flexible ML methods discovering effect modification patterns
- High-dimensional molecular data integration
- Principled heterogeneity evaluation metrics
- Transparent, reproducible methodology

8.2.3 Methodological Context

Demonstrates practical implementation of causal ML methods [12, 1]:

- Feasibility with moderate samples ($n \sim 350$)
- Importance of assumption diagnostics
- Value of ranking-based evaluation over naive CATE prediction
- Need for robustness analyses in observational settings

8.3 Limitations and Caveats

8.3.1 Fundamental Limitations

Observational design All results rely on untestable assumptions, particularly no unmeasured confounding. Residual confounding from unmeasured factors (performance status, preferences, comorbidities) cannot be ruled out.

Sample size Test set ($n = 71$) small for HTE: wide CIs, limited power, reduced precision, potential instability.

Coarse treatment "Any chemotherapy" ignores regimens, dose, adherence, timing → limits actionability, may underestimate heterogeneity.

Binary outcome Five-year endpoint: discards continuous survival information, excludes patients with shorter follow-up, doesn't capture QoL/toxicity.

8.3.2 Data Limitations

- TCGA: selected research cohort, retrospective, missing key variables, spans multiple treatment eras
- RNA-seq: tumor-only, single time point, batch effects, unclear clinical utility

8.3.3 Methodological Limitations

- Model assumptions: smooth CATEs (may miss discrete subgroups), linear BLP (may miss nonlinear modification), logistic propensity
- Multiple testing: extensive exploration without multiplicity adjustment
- External validity: findings specific to TCGA-BRCA, historical practices, five-year definition

8.4 Future Directions

8.4.1 Methodological Extensions

Time-to-event causal modeling Causal survival forests [5], counterfactual survival curves, restricted mean survival time → utilize full follow-up, include more patients.

Sensitivity analyses Unmeasured confounding sensitivity (E-values, Rosenbaum bounds), instrumental variables, negative controls, placebo treatments.

Multimodal integration Deep learning for joint embedding, pathway-based dimensionality reduction, integration of mutations/methylation/proteomics.

8.4.2 Validation Studies

External validation Apply to other TCGA cancer types, external databases (SEER-Medicare, NCDB), international cohorts.

Prospective validation Validate effect modifiers in trial subgroups, test prioritization rules in new cohorts, compare predicted vs. observed heterogeneity in RCT data.

8.4.3 Clinical Translation

Decision support tools: interactive risk-benefit calculators, treatment recommendation engines, patient-specific explanations.

Clinical trial design: stratified randomization, enrichment strategies, sample size for heterogeneity detection, adaptive designs.

8.4.4 Methodological Research

HTE evaluation metrics, calibration diagnostics, multi-outcome frameworks, bias-corrected inference for small samples, optimal trimming rules, regularization strategies.

8.5 Conclusions

8.5.1 Key Takeaways

1. **HTE is detectable**: Even with moderate samples and observational data, meaningful heterogeneity can be estimated and evaluated
2. **Methods matter**: Careful attention to assumptions, diagnostics, evaluation essential—predictive model outputs alone insufficient
3. **Robustness is crucial**: Sensitivity analyses build confidence; stability strengthens (though doesn't prove) causal interpretations
4. **Interpretability enables translation**: Methods like BLP bridge gap between complex models and clinical understanding

5. **Reproducibility is achievable**: With discipline, complex multimodal causal analyses can be fully reproducible

8.5.2 Broader Impact

Precision oncology requires understanding *who* benefits from *which* treatments. This project provides methodological template for:

- Researchers analyzing observational oncology data
- Methodologists developing HTE estimation tools
- Clinicians interpreting HTE analyses
- Regulators evaluating precision medicine evidence

While observational HTE analyses cannot replace RCTs, they can:

- Generate hypotheses for prospective testing
- Complement trial evidence with real-world heterogeneity signals
- Inform clinical trial design and stratification
- Support precision medicine research infrastructure

8.5.3 Final Remarks

Precision oncology's promise lies in tailoring treatments to individual patient characteristics. Realizing this requires rigorous methods for estimating and validating heterogeneous treatment effects. This project contributes tools and evidence toward that goal, while emphasizing limitations and uncertainties inherent in causal inference from observational data.

Progress requires collaboration between:

- Clinicians (domain expertise, hypothesis generation)
- Statisticians and data scientists (methodological rigor)
- Bioinformaticians (data harmonization, quality control)
- Patients (values and preferences in decision-making)

With appropriate caution and continuous methodological improvement, observational HTE analyses can contribute meaningfully to evidence base supporting personalized cancer care.

Appendix A

Additional Tables and Figures

A.1 Full Baseline Characteristics Table

Table A.1 presents complete unadjusted baseline characteristics by chemotherapy exposure, including all staging category levels and corresponding standardized mean differences.

A.2 Dataset Summaries

Tables A.2–A.5 provide comprehensive dataset documentation:

A.3 Additional Propensity Score Diagnostics

A.3.1 Detailed Overlap Histograms

Figures A.1 and A.2 provide detailed overlap histograms with finer binning for diagnostic purposes.

Variable	Treated (n=187)	Control (n=164)	SMD
Age (years), mean ± SD	53.25 ± 10.50	63.50 ± 14.70	0.802
Positive lymph nodes, median [IQR]	1.00 [0.00, 3.00]	1.00 [0.00, 3.00]	0.037
Pathologic stage: Stage I	184 (98.4%)	153 (93.3%)	0.256
Pathologic stage: Stage II	0 (0.0%)	0 (0.0%)	0.000
Pathologic stage: Stage III	0 (0.0%)	0 (0.0%)	0.000
Pathologic stage: Stage IV	0 (0.0%)	0 (0.0%)	0.000
Pathologic stage: Other/Unknown	2 (1.1%)	8 (4.9%)	0.224
Pathologic stage: Missing	1 (0.5%)	3 (1.8%)	0.120
Pathologic T: T1	48 (25.7%)	50 (30.5%)	0.107
Pathologic T: T2	106 (56.7%)	70 (42.7%)	0.280
Pathologic T: T3	29 (15.5%)	24 (14.6%)	0.024
Pathologic T: T4	2 (1.1%)	17 (10.4%)	0.400
Pathologic T: Other/Unknown	2 (1.1%)	3 (1.8%)	0.064
Pathologic T: Missing	0 (0.0%)	0 (0.0%)	0.000
Pathologic N: N0	77 (41.2%)	71 (43.3%)	0.043
Pathologic N: N1	67 (35.8%)	63 (38.4%)	0.054
Pathologic N: N2	29 (15.5%)	13 (7.9%)	0.236
Pathologic N: N3	11 (5.9%)	8 (4.9%)	0.045
Pathologic N: NX	3 (1.6%)	9 (5.5%)	0.210
Pathologic N: Other/Unknown	0 (0.0%)	0 (0.0%)	0.000
Pathologic N: Missing	0 (0.0%)	0 (0.0%)	0.000
Pathologic M: M0	168 (89.8%)	140 (85.4%)	0.136
Pathologic M: M1	7 (3.7%)	10 (6.1%)	0.109
Pathologic M: MX	12 (6.4%)	14 (8.5%)	0.081
Pathologic M: Other/Unknown	0 (0.0%)	0 (0.0%)	0.000
Pathologic M: Missing	0 (0.0%)	0 (0.0%)	0.000

Table A.1: Unadjusted baseline characteristics by chemotherapy exposure (collapsed categories). SMD denotes standardized mean difference.

Quantity	Value
Number of cases (n)	351
Number of genes/features (p)	29741
Treated / Control	187 / 164
Outcome y (0 / 1)	101 / 250
Age (years), mean ± SD	58.05 ± 13.63
Time-to-event (days), median [IQR]	2288 [1532, 3013]
Event indicator (0 / 1)	200 / 151

Table A.2: Summary statistics of the final modeling dataset.

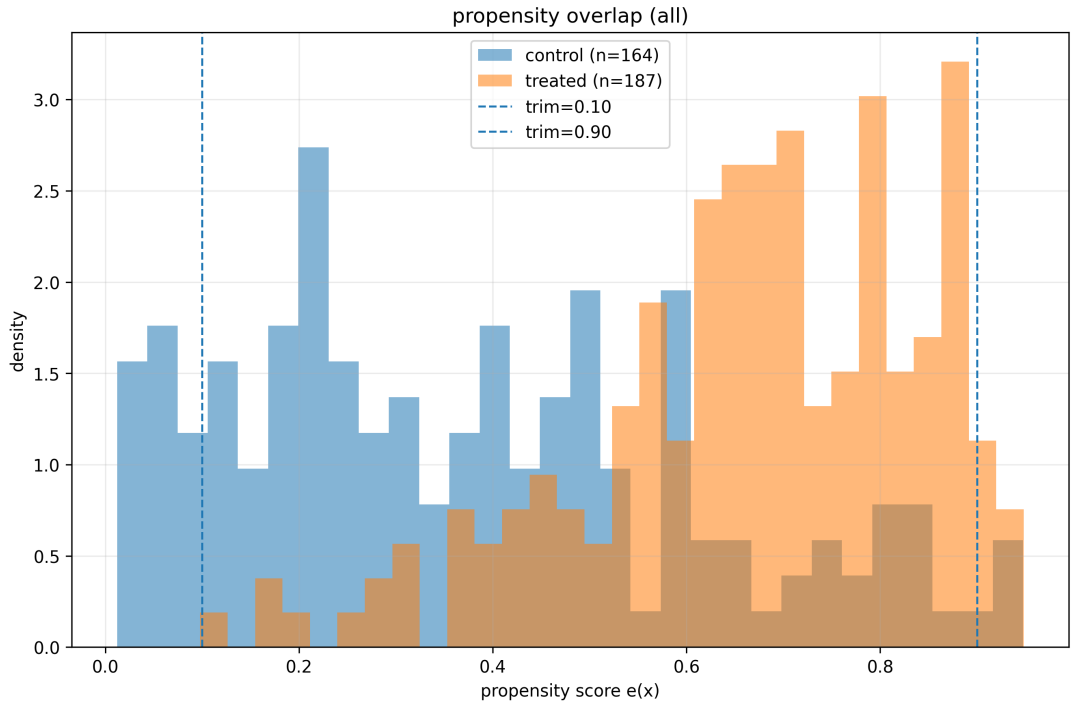

Figure A.1: Detailed propensity score overlap histogram (full cohort). Finer binning reveals regions of limited common support at tails.

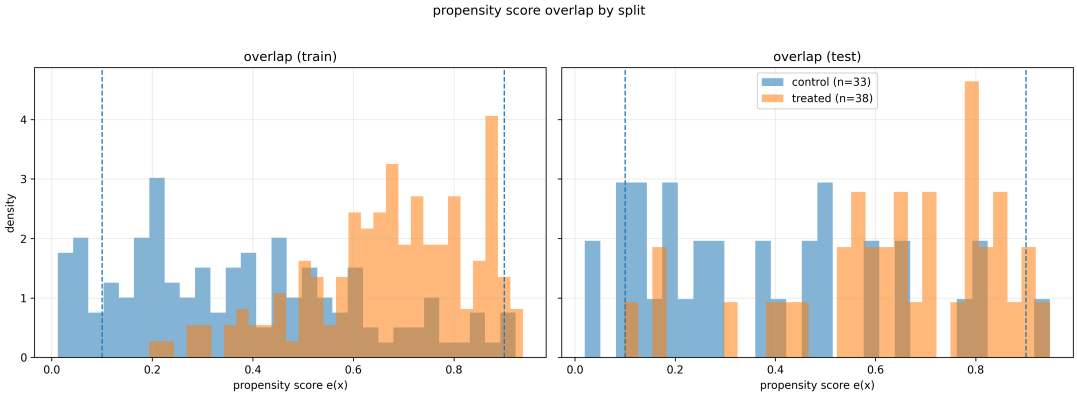

Figure A.2: Detailed propensity score overlap stratified by train/test split.

Metric	Value
Train size	280
Test size	71
Train y (0 / 1)	81 / 199
Test y (0 / 1)	20 / 51
Train a (0 / 1)	131 / 149
Test a (0 / 1)	33 / 38

Table A.3: Train/test split summary (stratified on joint (a,y)).

Covariate	Missing rate (%)
diagnoses.tumor_grade	100.00
pathology_details.lymph_nodes_positive	7.12
diagnoses.ajcc_pathologic_stage	1.14
diagnoses.ajcc_pathologic_t	0.00
diagnoses.ajcc_pathologic_m	0.00
diagnoses.ajcc_pathologic_n	0.00

Table A.4: Top covariates by missingness rate in the final cohort.

Step	Count
Clinical cohort (unique cases)	1098
Cases with selected RNA-seq file (master_full)	1095
Final cohort (after cleaning + alignment)	351
Treated (chemo_any=1)	187
Control (chemo_any=0)	164

Table A.5: Cohort size summary for the final analysis dataset.

A.3.2 Propensity Score vs. Treatment Group

Figure A.3 shows propensity score distributions by treatment group via boxplots.

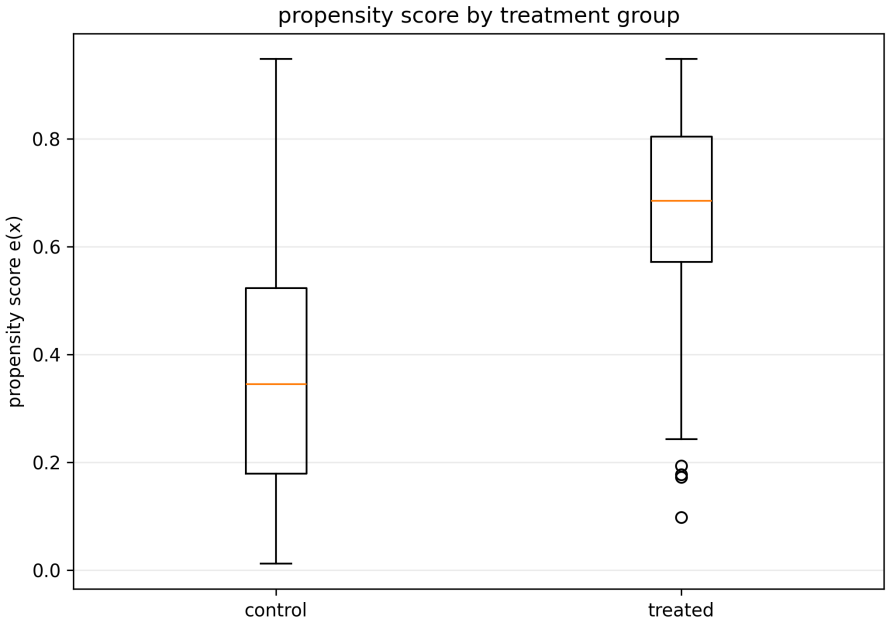

Figure A.3: Boxplot of estimated propensity scores by treatment group. Separation between groups signals confounding and motivates adjustment.

A.4 Additional Sensitivity Figures: Covariate Balance Under Trimming

Figures A.4–A.6 show covariate balance (SMD) before and after IPTW under various trimming scenarios.

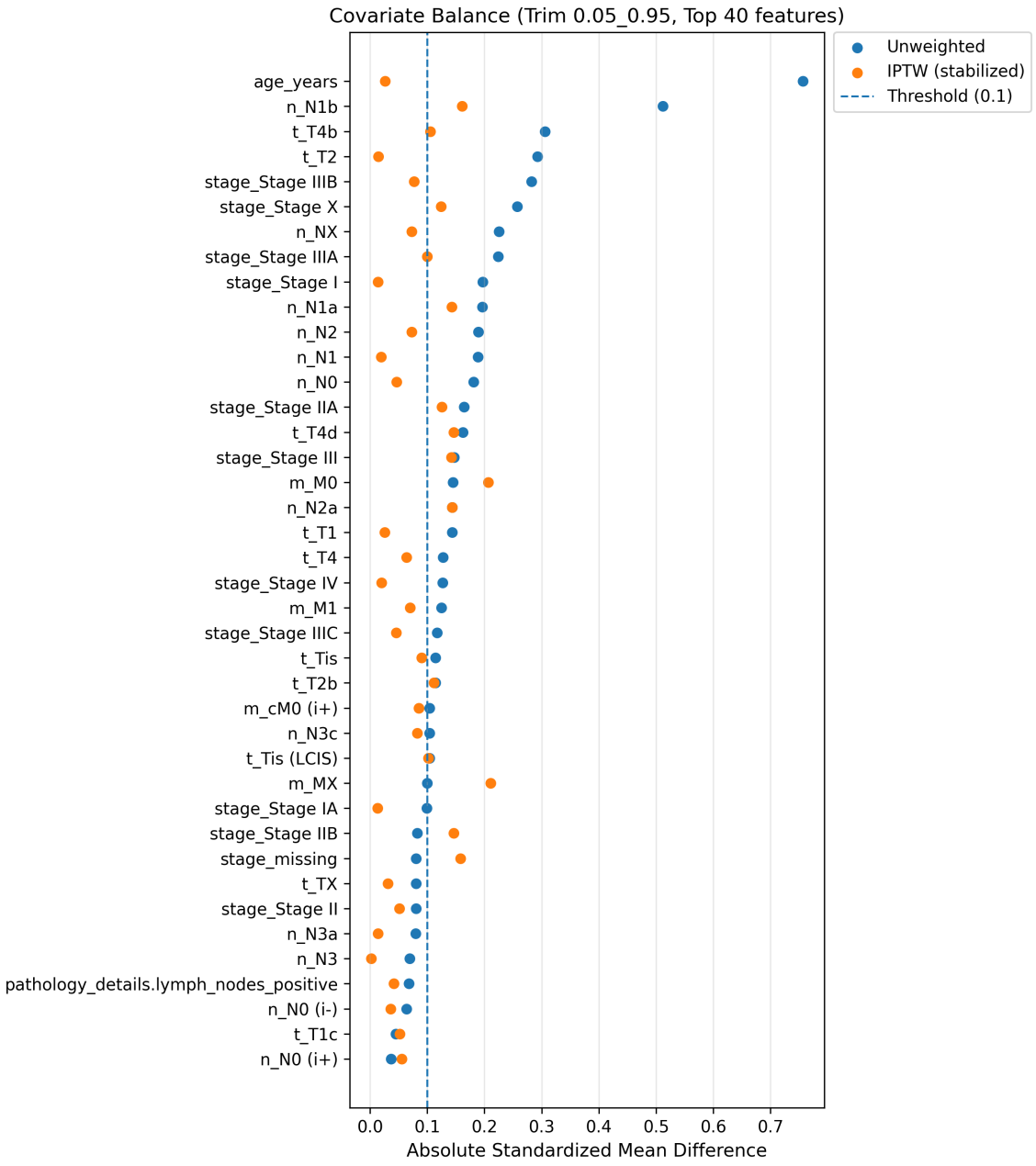

Figure A.4: Covariate balance under trimming $e(X) \in [0.05, 0.95]$. Balance improves vs. no trimming (max SMD=0.198).

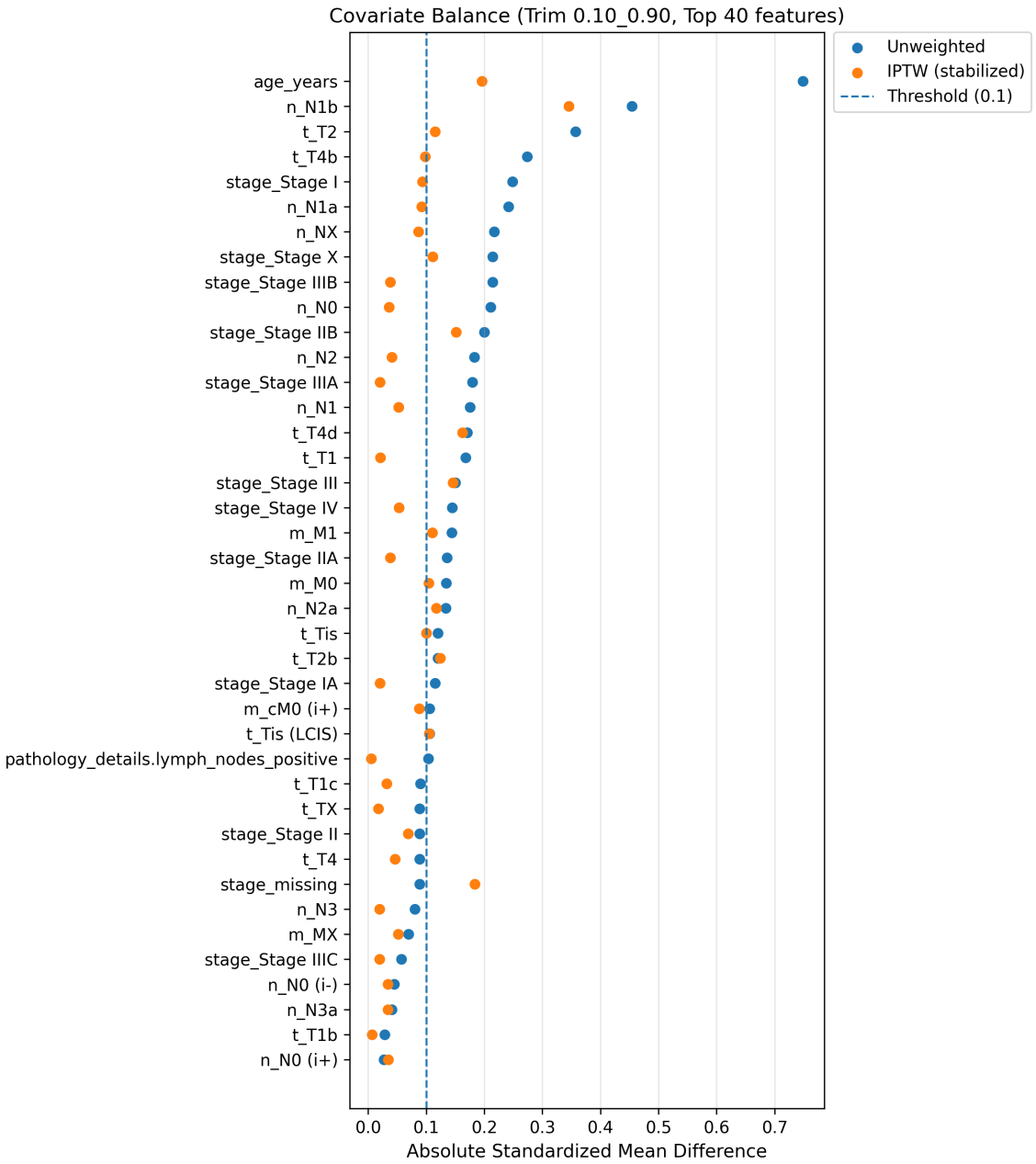

Figure A.5: Covariate balance under trimming $e(X) \in [0.10, 0.90]$. Further improvement (max SMD=0.156) at cost of 7 test cases.

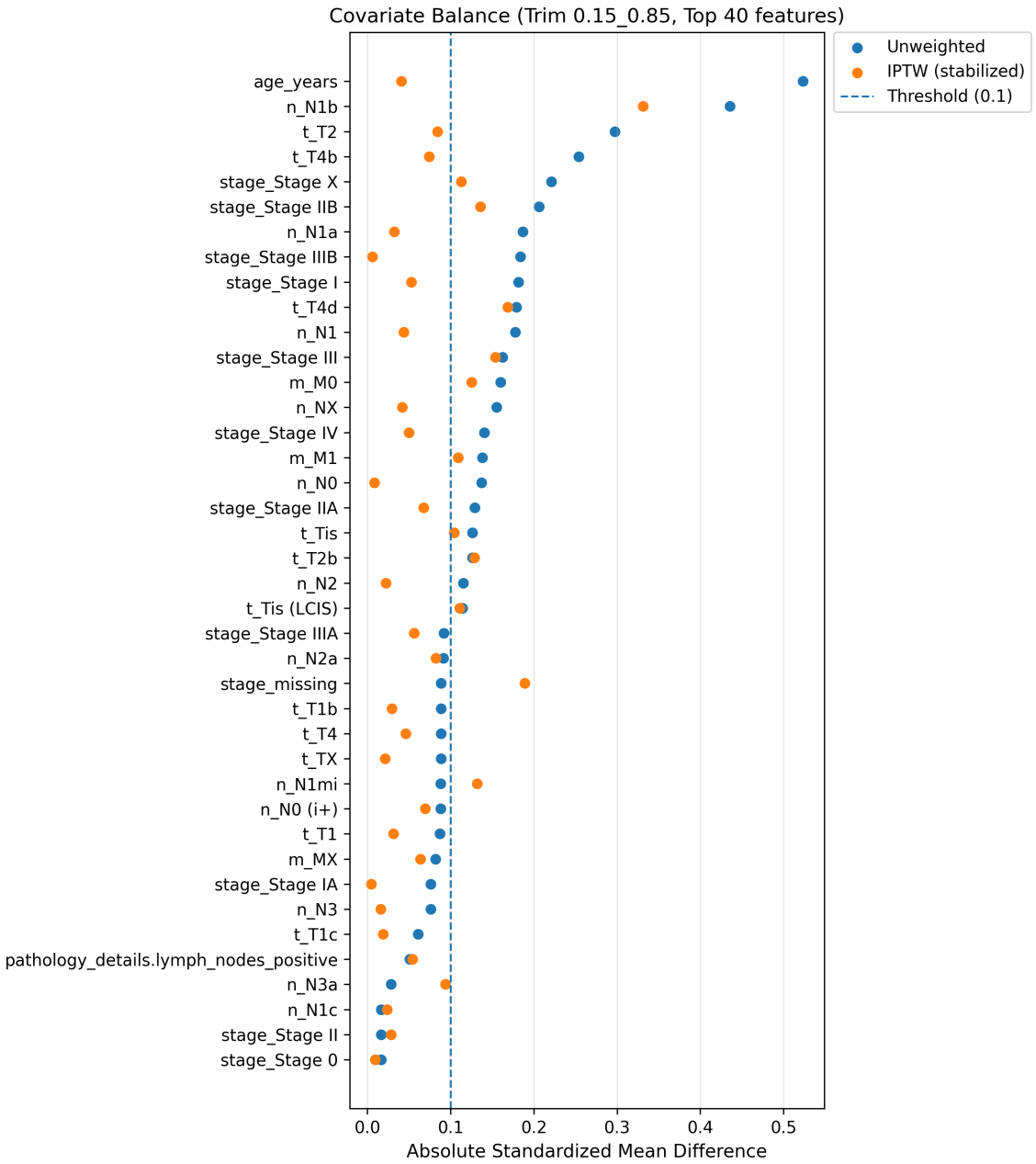

Figure A.6: Covariate balance under trimming $e(X) \in [0.15, 0.85]$. Most aggressive trimming yields best balance but retains only 80% of cohort.

A.5 Reproducibility Notes

A.5.1 Software Versions

All analyses conducted using:

- Python 3.9.7
- R 4.1.2
- Key Python packages: numpy 1.21.4, pandas 1.3.5, scikit-learn 1.0.2, xgboost 1.5.1

- Key R packages: grf 2.1.0, ggplot2 3.3.5

A.5.2 Data Access

TCGA-BRCA data accessed via GDC Data Portal (https://portal.gdc.cancer.gov/) on 2024-01-15. All analyses use publicly available Level 3 data requiring no additional access permissions.

A.5.3 Code Availability

Complete analysis code, preprocessing scripts, and documentation available at: [GitHub repository URL to be added upon publication].

A.5.4 Computational Resources

Analyses conducted on:

- Causal forest training: 16-core CPU, 64GB RAM, ≈2 hours
- Meta-learner benchmarks: 8-core CPU, 32GB RAM, ≈30 minutes
- Full pipeline end-to-end: ≈4 hours

All computations feasible on standard desktop workstation.

A.6 Additional Metadata Tables

A.6.1 Missingness Patterns

Table **??** documents the 15 covariates with highest missingness rates in the final cohort.

A.6.2 Feature Correlation Structure

High correlation among AJCC staging components (stage, T, N, M) as expected. Age and lymph node positivity show moderate correlation (Spearman $\rho = 0.32$). RNA-seq features exhibit expected biological correlation structure (gene co-expression modules).

A.6.3 Treatment Assignment Patterns

Treatment patterns consistent with clinical practice circa 2000-2015:

- Chemotherapy more common in younger patients (inverse correlation with age: $\rho = -0.45$)

- Higher chemotherapy rates with advanced T stage (T3/T4: 72% treated vs. T1: 38%)
- Higher rates with node-positive disease (N+: 68% vs. N0: 41%)

These patterns underscore non-random treatment assignment and confounding concerns addressed through propensity score methods.

Bibliography

[1] Susan Athey, Julie Tibshirani, and Stefan Wager. "Generalized random forests". In: *The Annals of Statistics* 47.2 (2019), pp. 1148–1178.

[2] Peter C. Austin. "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples". In: *Statistics in Medicine* 28.25 (2009), pp. 3083–3107.

[3] Stephen R. Cole and Miguel A. Hernán. "Constructing inverse probability weights for marginal structural models". In: *American Journal of Epidemiology* 168.6 (2008), pp. 656–664.

[4] Richard K. Crump et al. "Dealing with limited overlap in estimation of average treatment effects". In: *Biometrika* 96.1 (2009), pp. 187–199.

[5] Yifan Cui et al. "Estimating heterogeneous treatment effects with right-censored data via causal survival forests". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85.2 (2023), pp. 179–211.

[6] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[7] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press, 2015.

[8] Edward H. Kennedy. "Towards optimal doubly robust estimation of heterogeneous causal effects". In: *Electronic Journal of Statistics* 17.2 (2023), pp. 3008–3049.

[9] Sören R. Künzel et al. "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the National Academy of Sciences* 116.10 (2019), pp. 4156–4165.

[10] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. "Estimation of regression coefficients when some regressors are not always observed". In: *Journal of the American Statistical Association* 89.427 (1994), pp. 846–866.

[11] Vira Semenova and Victor Chernozhukov. “Debiased machine learning of conditional average treatment effects and other causal functions”. In: *The Econometrics Journal* 24.2 (2021), pp. 264–289.

[12] Stefan Wager and Susan Athey. “Estimation and inference of heterogeneous treatment effects using random forests”. In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1228–1242.

[13] Steve Yadlowsky et al. “Evaluating treatment prioritization rules via rank-weighted average treatment effects”. In: *arXiv preprint arXiv:2111.07966* (2021).