MATRICS: PROTOCOL-EQUIVALENT ATTRIBUTION OF CLINICAL TEXT PIPELINES FOR CAUSAL SUBGROUP DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

A text representation in causal subgroup discovery helps define the hypotheses that are ultimately tested: changing it can rotate proposal directions, move patients across score tails, and alter the effective search family. MATRICS isolates this consequential intervention by passing each representation pipeline through the same leakage-free operator for cross-fitted doubly robust scores, sparse signed-tail proposals, ranking, Benjamini–Hochberg screening, and complete-path stability selection. On 50 paired semi-synthetic MIMIC-IV replicates, the modifier-adapted ClinicalBERT pipeline recovers known note-borne effect modifiers at F1 0.691 versus 0.585 for TF–IDF, a gain of 0.106 (18.1% relative; paired 95% interval [0.061, 0.148]; 39/50 wins); CATE RMSE falls from 0.301 to 0.243 and matched-predicate stability rises from 0.73 to 0.86. The reference configuration and all six one-axis alternatives retain positive lexicalized intervals. Crossed information regimes explain the gain: masking literal modifier spans preserves a 0.056 advantage, every structured-only contrast lies within a four-point recovery-F1 compatibility band, and global-null runs produce at least one stable survivor in 10/50 versus 8/50 replicates. Together with a representation ladder that places domain-aligned biomedical encoders at the strongest observed operating points, these results show when contextual clinical text changes recoverable effect heterogeneity: it matters most when notes carry modifier information not duplicated in structured covariates.

1 INTRODUCTION

In text-conditioned subgroup discovery, the representation is part of the hypothesis generator. Replacing TF–IDF with a transformer can rotate the directions searched, change which patients occupy a candidate tail, and alter which predicates survive multiplicity correction. Even when admissions, treatments, outcomes, and random seeds are identical, two pipelines can therefore answer different subgroup questions. A defensible representation comparison must intervene on the text pipeline while holding the adaptive discovery operator fixed.

MATRICS (Matched Attribution of Text Representations for Inference of Causal Subgroups) makes that comparison an explicit estimand. Componentwise protocol equivalence covers the outer partition, nuisance class and doubly robust (DR) score, direction objective, candidate count and support, ranking, nominal tests, BH screening, and stability retention. Representation-specific geometry remains free to carry useful signal, but every geometry is converted into tested predicates by the same operator. The outer split keeps all adaptation and proposal learning inside proposal units, freezes a unique signed-tail family, and exposes evaluation outcomes only to estimation and screening.

The experiment is organized around where modifier information resides. We ask whether contextual representations improve recovery when clinical notes contain literal modifiers; whether that advantage survives after literal spans are masked; and whether it contracts when the same modifier is already available in structured covariates. A global-null regime measures family-level false survival, while timing-restricted Project STAR records probe whether independently learned text groups transfer to a second domain. These crossed regimes turn the location of modifier information into a mechanism test rather than treating one benchmark score as sufficient evidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The primary intervention is decisive. Across 50 paired MIMIC-IV replicates, modifier-adapted ClinicalBERT raises modifier-recovery F1 from 0.585 to 0.691, with a paired gain of 0.106 [0.061, 0.148] and 39/50 wins. CATE RMSE decreases from 0.301 to 0.243, empirical FDR from 0.09 to 0.07, and matched-predicate stability increases from 0.73 to 0.86. The lexicalized interval remains positive in the reference configuration and all six one-axis alternatives. Masking literal spans attenuates but preserves the gain at 0.056; structured-only contrasts remain inside a four-point compatibility band; and null family-survival rates are 0.20 and 0.16 for MATRICS and TF–IDF.

The technical contribution is the complete attribution object: a protocol-conditional pipeline contrast implemented by cross-fitted DR targets, modifier-directed adaptation, a fixed family of sparse signed-tail predicates, and reruns of the entire discovery path. The representation ladder, information-location regimes, null calibration, and structured-redundancy controls jointly separate domain alignment from search-protocol variation. They support a concrete modeling decision in the evaluated regime: contextual biomedical representations earn their added complexity when narrative records expose effect modifiers that structured covariates do not.

2 Related Work

Text representations for causal estimation. Representation balancing and text adjustment improve nuisance estimation, counterfactual prediction, and confounding control in high-dimensional settings (Johansson et al., 2016; Shalit et al., 2017; Yao et al., 2018; Veitch et al., 2020; Egami et al., 2022; Roberts et al., 2020; Pryzant et al., 2021). Recent methods use LLM-generated surrogates when structured confounders are unavailable at inference (Ma et al., 2025), extract interpretable clinical covariates from notes (Liu et al., 2026), or route pretrained semantics into assignment and heterogeneity pathways (Xu et al., 2026). Mozer et al. (Mozer et al., 2023) directly derive clinical subgroups from bag-of-words, principal components, and ClinicalBERT features, making it the closest text-conditioned subgroup workflow to our setting. These methods estimate effects or produce subgroups under their own representation-specific pipelines. MATRICS instead targets the paired recovery change under an invariant proposal, support, testing, and retention operator, making the complete discovery path the unit of representation attribution.

Heterogeneous effects and subgroup rules. Tree methods, meta-learners, uplift models, and policy learners estimate heterogeneous effects or treatment rules (Foster et al., 2011; Tian et al., 2014; Zhao et al., 2012; Athey & Imbens, 2016; Gutierrez & Gérardy, 2017; Kitagawa & Tetenov, 2018; Athey & Wager, 2021; Wager & Athey, 2018; Athey et al., 2019; Oprescu et al., 2019). DISCRET (Wu et al., 2024) synthesizes faithful per-sample rules for individual treatment-effect prediction, including text inputs, while causal distillation trees use a teacher model and stability diagnostics to recover interpretable subgroups (Huang et al., 2025). Their primary objects are faithful or stable subgroup rules. MATRICS uses executable signed-tail predicates as paired experimental units and asks whether changing the text pipeline changes recovery after candidate budget, support, ranking, screening, and retention are fixed.

Reliability after adaptive learning. Orthogonal scores, debiased estimation, influence-function validation, and multiplicity control address distinct sources of inferential error (Robins et al., 1994; Chernozhukov et al., 2018; Semenova & Chernozhukov, 2021; Kennedy, 2023; Alaa & van der Schaar, 2019; Benjamini & Hochberg, 1995). MATRICS composes these ingredients around a matched representation intervention and evaluates the learned family through positive recovery, null behavior, structured redundancy, and an exploratory transfer readout. BH is an empirical family screen rather than a selective-inference guarantee; the attribution evidence comes from paired recovery, complete-path reruns, and regime changes with distinct predicted outcomes.

3 Protocol-Equivalent Representation Attribution

3.1 Attribution Target and Protocol Equivalence

For units (X_i, T_i, Y_i, Z_i), let $X_i \in \mathbb{R}^p$ denote structured covariates, $T_i \in \{0, 1\}$ treatment, Y_i outcome, and Z_i text. Under unconfoundedness, overlap, and SUTVA, the conditional average

treatment effect is

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

A subgroup is an executable predicate $\phi(x, z)$ whose conditional treatment effect differs from the population average. In MIMIC-IV, semi-synthetic modifiers provide known recovery targets under observational identification. STAR is kept outside the attribution estimand and used only for descriptive transfer readouts of independently selected groups.

For regime e, representation pipeline r, search protocol Π, and shared replicate s, let $M_e(\Pi, r; s)$ denote recovery F1 for positive and structured regimes or an indicator of any stable survivor under no moderation. We define the protocol-conditional representation effect

$$\Delta_{\Pi,e}(r_a, r_b) = \mathbb{E}_s[M_e(\Pi, r_a; s) - M_e(\Pi, r_b; s)] .$$

The expectation is over the finite semi-synthetic replicate generator on a fixed MIMIC-IV cohort: s indexes treatment draws, outcome noise, and patient-disjoint folds, while the observed admissions and texts are not resampled. Pairing holds all three indexed objects fixed across pipelines. Thus $\Delta_{\Pi,e}$ is Monte Carlo uncertainty for this cohort and data-generating regime, not a population-sampling interval.

Let the subgroup-search protocol be

$$\Pi = (\mathcal{I}, \mathcal{F}, \mathcal{N}, \mathcal{D}, \mathcal{G}, K, \rho, \mathcal{R}, q, B, \kappa),$$

where $\mathcal{I}$ fixes unit splits, $\mathcal{F}$ cross-fitting, $\mathcal{N}$ the nuisance class and tuning budget, $\mathcal{D}$ DR-score construction, $\mathcal{G}$ the direction-dictionary objective, K candidate count, ρ tail support, $\mathcal{R}$ ranking and nominal testing, q the BH level, B bootstrap replicates, and κ the retention threshold. Representations r_a and r_b are protocol equivalent when $\Pi_a = \Pi_b$ componentwise. Encoder-specific directions $V(r)$ are outputs, not protocol changes: both representations face the same objective, sparsity budget, support, and number of tests.

The representation-attribution signature $\mathbf{A}_\Pi = (\Delta_{\Pi,e} : e \in \mathcal{E})$ collects lexicalized, semantic, null, and structured-redundancy evidence. Text-mediated value predicts positive recovery differences where text carries the modifier and contraction under structured redundancy; null family yield probes the selector itself. The separate STAR readout asks whether the learned group semantics recur in another domain without extending the causal attribution claim.

3.2 MATRICS Instantiation

Cross-fitted construction. Every replicate first fixes a patient-disjoint outer partition $(\mathcal{I}_{\text{prop}}, \mathcal{I}_{\text{eval}})$ with 70% proposal and 30% evaluation units. Five-fold cross-fitting inside $\mathcal{I}_{\text{prop}}$ fits propensity $\hat{e}(X_i, \mathbf{z}_i)$ and outcome regressions $\hat{\mu}_t(X_i, \mathbf{z}_i)$ with LightGBM on four folds and produces the following DR target only for the held-out fold:

$$\tilde{\tau}_i = \hat{\mu}_1 - \hat{\mu}_0 + \frac{T_i - \hat{e}_i}{\hat{e}_i(1 - \hat{e}_i)} (Y_i - \hat{\mu}_{T_i}) . \tag{1}$$

Stacking the five held-out folds gives one out-of-fold target per proposal unit. ClinicalBERT (Devlin et al., 2019; Alsentzer et al., 2019) maps each note to 768 dimensions; MATRICS adapts its final four blocks against these targets entirely within $\mathcal{I}_{\text{prop}}$. For each hyperparameter candidate, four proposal folds train the adapter and direction model and the fifth scores validation DR loss. Averaging the five held-out losses selects the hyperparameters and checkpoint rule; a single final adapter and direction family is then refit on all proposal units with the selected settings and frozen. This final refit is the only family passed to evaluation.

Nuisance functions are likewise refit on all proposal units after selection and applied once to $\mathcal{I}_{\text{eval}}$ to form evaluation DR scores. Evaluation outcomes enter contrasts and screening, but never cross-fitted proposal targets, hyperparameter choice, encoder adaptation, direction learning, signs, or thresholds. Propensities are clipped to $[0.025, 0.975]$. Every representation reuses the outer partition, inner folds, 500-tree nuisance class, 40-trial budget, and early-stopping rule.

Sparse proposal operator. On the proposal partition, MATRICS learns encoder parameters θ for the final four ClinicalBERT blocks, unit-norm directions $V = [v_1, \ldots, v_K]$, and coefficients (β, γ) by minimizing squared DR residuals plus an elastic-net penalty over projected text scores

Table 1: Leakage-free protocol flow. The outer split precedes nuisance fitting; no evaluation outcome, test statistic, or bootstrap decision feeds candidate construction.

Phase	Units	Learned object	Reused across representations
Outer split	all units	70% proposal / 30% evaluation	units and partition
Proposal score	$\mathcal{I}_{\text{prop}}$ only	five-fold $\hat{e}, \hat{\mu}_t$, cross-fit DR scores	folds, class, budget
Proposal	$\mathcal{I}_{\text{prop}}$ only	encoder, V, signs, thresholds	objective, K, support, ranking
Evaluation	$\mathcal{I}_{\text{eval}}$ only	DR contrasts, BH passes	tests, q, endpoints
Stability	separate partition resamples	recurrence of matched directions	resamples, match rule, cutoff

and structured covariates. Columns of V are initialized by supervised partial least squares after residualizing on X, with each accepted component deflated before the next is formed. Alternating gradient and proximal-gradient updates optimize $(\theta, V, \beta, \gamma)$; each V step is renormalized with sign chosen so its largest-magnitude coordinate is positive. Proposal-only cross-validation selects λ from 20 log-spaced values in $[10^{-4}, 10]$ with $\alpha = 0.5$ and selects the joint checkpoint by held-out DR loss. Optimization stops when relative objective improvement is below 10^{-5} for five updates or after 500 updates. Appendix B gives the objective and the status of its proposal coordinates.

Frozen family and empirical screen. Directions are ranked by absolute proposal-fold DR contrast. For $\beta_k \geq 0$, direction k contributes $\mathbf{1}\{\langle z, v_k \rangle \geq q_{0.80}^{\text{prop}}\}$; for $\beta_k < 0$, it contributes the corresponding lower-tail rule at $q_{0.20}^{\text{prop}}$. Thresholds are computed once on proposal units. If an optimized direction duplicates an earlier support mask, its slot is reinitialized from the next deflated PLS component and optimized again; this continues until the $K = 20$ frozen predicates have unique proposal supports. On disjoint evaluation units, the studentized contrast statistic tests $H_{0k} : \Delta(S_k) = 0$ using evaluation DR influence scores, and BH at $q = 0.10$ precedes stability filtering.

Each of 100 stratified reruns repeats nuisance fitting, proposal, testing, and BH on separately resampled proposal and evaluation partitions. Rerun predicates are matched to the frozen family by maximum-weight one-to-one assignment of signed proposal-unit scores; a match must preserve upper/lower-tail orientation and have correlation at least 0.8. Ties go to the lower original rank and then lower rerun rank. A predicate is retained when its assigned counterpart passes BH in at least 80% of reruns, so one rerun direction cannot stabilize multiple original predicates.

4 Experiments

4.1 Datasets, Regimes, and Endpoints

Each regime changes where modifier information resides while preserving the representation intervention. Lexicalized and semantic regimes quantify text-mediated recovery, the null records family yield without moderation, and structured-only regimes test redundancy. STAR is reported separately as a descriptive transfer study. Appendix A specifies the data-generating processes, and Appendix B gives the proposal objective and nominal test.

Semi-synthetic MIMIC-IV. We use 24,318 admissions, discharge summaries, and 14 structured covariates from MIMIC-IV (Johnson et al., 2023), with patient-disjoint outer splits. Treatment follows a logistic model of standardized structured covariates. Outcomes combine a nonlinear baseline response, Gaussian noise, and heterogeneous effects: mentions of non-compliance or poor adherence reduce benefit by 40%, while family-support documentation increases it by 25%. The semantic regime uses a fixed set of paraphrase clusters shared by every pipeline and masks all literal seed spans before representation learning. The null sets both modifier coefficients to zero. Two structured-only regimes move the target into baseline severity or a generated baseline-history score. All methods use the same 50 seeds, treatment draws, noise draws, outer partitions, and inner folds.

Table 2: Protocol-equivalent representation attribution on lexicalized MIMIC-IV over 50 paired replicates. Recovery F1 is the primary endpoint; paired intervals compare with TF–IDF under the same search protocol.

Pipeline / status / dimension	Recovery F1	Paired Δ [95% interval]	Wins / 50
TF–IDF / fixed / 20,000	0.585	0	—
GPT–2 / frozen / 768	0.573	-0.012 [-0.054, 0.030]	24
Generic BERT / frozen / 768	0.638	0.053 [0.011, 0.095]	34
ClinicalBERT / frozen / 768	0.641	0.056 [0.013, 0.098]	33
PubMedBERT / frozen / 768	0.676	0.091 [0.048, 0.134]	37
ClinicalBERT / outcome-tuned / 768	0.668	0.083 [0.038, 0.127]	36
MATRICS / modifier-adapted / 768	0.691	0.106 [0.061, 0.148]	39

Timing-restricted STAR. Project STAR contains 11,601 students assigned to small or regular classes (Krueger & Whitmore, 2001). We combine baseline covariates with teacher-evaluation text and reserve 30% for evaluation. The timing restriction retains 1,624 records linked to the pre-policy school-year form, 46.7% of the seed-42 evaluation fold. We designate STAR as a descriptive transfer study and reserve causal attribution for the known-target MIMIC regimes. MATRICS and TF–IDF independently execute proposal, ranking, BH, and stability on the same proposal units; groups are associated by selected-unit overlap and semantic descriptors without evaluation outcomes.

Comparators and compute. The primary comparator is a DR learner using 20,000 unigram/bigram TF–IDF features with identical Π. Generic BERT (Devlin et al., 2019), PubMedBERT (Gu et al., 2021), frozen ClinicalBERT (Alsentzer et al., 2019), outcome-tuned ClinicalBERT, and frozen GPT–2 (Radford et al., 2019) form a representation ladder under the same downstream operator. In structured-only regimes, causal forest is a paired text-free reference: it is fit on proposal units, its 20 highest-contrast support-qualified terminal-leaf predicates are frozen, and those predicates enter the same evaluation, BH, and complete-path stability stages. Causal-forest and R-learner estimates of fixed STAR predicates appear only as estimator checks in the appendix. The complete sweep used four NVIDIA A100 80GB GPUs for approximately six hours; CPU nuisance fitting ran on 64 AMD EPYC cores.

Metric hierarchy. Recovery F1 is the primary endpoint. Retained predicates are matched one-to-one to true modifiers by maximum unit-level F1; unmatched modifiers receive zero and unmatched discoveries determine empirical FDR. CATE RMSE, empirical FDR, stability, and survivor count diagnose how a recovery change arose. Primary intervals bootstrap the 50 paired replicate-level differences 10,000 times and cover the mean under the finite replicate generator; they do not resample patients. Stability is the within-replicate frequency over 100 complete-path reruns. STAR intervals instead resample evaluation units and are descriptive conditional spreads. Under the null, no positive target exists, so we report the probability and raw count of any stable survivor, survivor distribution, pre-stability BH passes, and maximum absolute DR contrast.

4.2 Modifier Recovery and the Representation Ladder

The 0.106 recovery-F1 increase, or 18.1% relative gain, is the lexicalized entry of $\mathbf{A}_\Pi$. Its 10,000-resample paired interval stays above zero, the median difference is 0.104, and MATRICS wins 39/50 replicates. Supporting endpoints move coherently: CATE RMSE decreases from 0.301 to 0.243, empirical FDR from 0.09 to 0.07, and matched-predicate stability rises from 0.73 to 0.86. Domain alignment is visible before modifier adaptation: frozen PubMedBERT gains 0.091 over TF–IDF and exceeds frozen ClinicalBERT by 0.035 descriptively. MATRICS records the largest observed recovery, 0.015 above PubMedBERT and 0.023 above outcome-tuned ClinicalBERT; the score-bearing paired intervals in this ladder are the displayed comparisons against TF–IDF. The central estimate is therefore the complete modifier-adapted pipeline contrast, while the ladder localizes its strongest operating region to biomedical contextual representations.

Table 3: Semantic and null entries of the attribution signature under the reference protocol. Positive regimes use recovery F1; null entries are descriptive family-yield diagnostics.

Regime and endpoint	TF–IDF	MATRICS	Contrast / status
Lexicalized recovery F1	0.585	0.691	0.106 [0.061, 0.148]
Semantic recovery F1	0.552	0.608	0.056 [0.014, 0.098]
Null: any stable survivor	0.16 (8/50)	0.20 (10/50)	observed difference $2/50$
Null: mean / 90th pct. survivors	0.4 / 1	0.5 / 2	out of 20 candidates
Null: mean BH passes / max $\lvert\hat{\Delta}\rvert$	1.20 / 0.084	1.42 / 0.089	descriptive readout

Table 4: One-axis-at-a-time protocol neighborhood over the same 50 paired seeds. Intervals are 95% bootstrap intervals of paired recovery-F1 differences; the center row is the reference configuration.

(K, tail, q)	Lexicalized Δ [95% interval]	Wins	Semantic Δ [95% interval]	Wins
$(10, 20\%, 0.10)$	0.093 [0.048,0.136]	37	0.047 [0.004,0.089]	33
$(20, 20\%, 0.10)$	0.106 [0.061,0.148]	39	0.056 [0.014,0.098]	34
$(40, 20\%, 0.10)$	0.099 [0.052,0.145]	38	0.061 [0.014,0.105]	35
$(20, 10\%, 0.10)$	0.084 [0.036,0.131]	35	0.041 [−0.006,0.087]	31
$(20, 30\%, 0.10)$	0.097 [0.055,0.139]	40	0.052 [0.009,0.094]	35
$(20, 20\%, 0.05)$	0.089 [0.047,0.132]	37	0.050 [0.007,0.092]	34
$(20, 20\%, 0.15)$	0.078 [0.031,0.124]	34	0.043 [−0.004,0.090]	31

4.3 Semantic Attenuation and Null Behavior

Semantic diffusion reduces the paired advantage from 0.106 to 0.056: contextual features retain useful modifier alignment after literal triggers are masked, while lexical anchoring supplies additional recoverable signal. The global-null family-level survivor probability is 0.16 for TF–IDF and 0.20 for MATRICS; Wilson 95% Monte Carlo intervals from the reported counts are [0.083, 0.285] and [0.112, 0.330]. MATRICS also has more pre-stability BH passes (1.42 versus 1.20) and a larger maximum absolute contrast (0.089 versus 0.084). Complete-path stability reduces this upstream ordering to an observed two-replicate difference in any-survivor count, but the 16–20% family-level risk remains the operative null diagnostic.

4.4 Protocol Robustness and Null Family Yield

Table 4 contains the reference setting and six one-axis alternatives spanning $K \in \{10, 20, 40\}$, tail support in $\{10\%, 20\%, 30\%\}$, and $q \in \{0.05, 0.10, 0.15\}$. Lexicalized paired gains range from 0.078 to 0.106, and every interval remains above zero. Five of seven semantic intervals also remain above zero, while the 10%-tail and $q = 0.15$ settings locate the semantic boundary. Appendix Table 9 separates selector stages: stable-survivor ordering reverses at two settings, but MATRICS has the higher mean BH-pass count and maximum absolute contrast in all seven. The result is a robust lexicalized recovery effect, a sharper semantic boundary, and an explicit family-level null cost for the adaptive pipeline.

4.5 Structured-Only Compatibility

We use a recovery-F1 compatibility margin of $\delta = 0.04$, the smallest difference that would change the representation choice in this study and 38% of the primary 0.106 gain. The assessment uses 90% percentile intervals over paired replicate differences; compatibility requires the entire interval to lie in $[-\delta, \delta]$.

Every MATRICS-minus-comparator interval lies inside the decision band, with sign reversals across modifiers and comparators. Methodwise recovery, empirical FDR, stability, and survivor intervals in Appendix Tables 10–11 agree with the paired analysis. In contrast to the 0.106 note-borne gain, moving the target into structured covariates removes a decision-relevant advantage for the text pipeline.

Table 5: Structured-only compatibility over 50 shared seeds. Every MATRICS-minus-comparator recovery-F1 interval lies inside the decision-relevant ± 0.04 margin.

Modifier	Comparison	Paired F1 difference (90% interval)	W / T / L
Baseline severity	MATRICS–TF–IDF	0.006 [−0.018, 0.029]	24/4/22
Baseline severity	MATRICS–causal forest	0.011 [−0.013, 0.035]	27/3/20
Baseline-history score	MATRICS–TF–IDF	−0.005 [−0.028, 0.017]	21/5/24
Baseline-history score	MATRICS–causal forest	−0.002 [−0.025, 0.021]	23/4/23

Table 6: Independent discovery on 1,624 timing-restricted STAR evaluation records. Both MATRICS rows are stable; for TF–IDF, Behavior is stable and Reading readiness is the remaining BH-passing rule. BH/stable gives family totals. CATE intervals are descriptive conditional readouts after screening.

Pipeline	Group readout	n (prev.)	BH / stable	Recur.	Eval. CATE
MATRICS	Behavior	188 (11.6%)	3/2	4/6	0.162 [0.071, 0.253]
TF–IDF	Behavior	171 (10.5%)	2/1	3/6	0.103 [0.021, 0.185]
MATRICS	Reading readiness	132 (8.1%)	3/2	3/6	0.058 [0.004, 0.112]
TF–IDF	Reading readiness	118 (7.3%)	2/1	2/6	0.026 [−0.020, 0.072]

4.6 Exploratory Transfer to Timing-Restricted STAR

The independently discovered behavior groups overlap at Jaccard 0.54 and have positive evaluation-set CATE readouts under both representations. Their membership-specific descriptive difference is 0.059 [0.014, 0.105]. Reading-readiness groups overlap less (Jaccard 0.41) and have a smaller readout difference of 0.032 [−0.006, 0.070]. Behavior is the more recurrent semantic pattern, appearing in 4/6 versus 3/6 additional splits. STAR therefore contributes a transfer observation—recurring text-defined groups with coherent ordering—while the paired MIMIC regimes carry the attribution claim.

5 Implications and Scope

The attribution signature turns representation selection into a testable decision. A pipeline earns credit when paired modifier recovery improves under an invariant operator, persists across neighboring specifications, and contracts when structured variables already expose the modifier. Domain-aligned encoders satisfy that pattern for note-borne modifiers: the ladder places PubMedBERT and modifier-adapted ClinicalBERT at the strongest observed operating points, while paraphrase masking separates contextual recovery from literal anchoring. The structured-redundancy result then identifies when the additional text path is unnecessary.

The evidential scope follows the design. The primary result is conditional on observational identification in one MIMIC-IV cohort, two curated modifier families, and the sparse signed-tail search class; the fixed paraphrase inventory defines the semantic regime, and the 16–20% null family-survival rate shows why retained predicates require downstream confirmation. STAR remains a selection-conditioned transfer diagnostic because field-level pretreatment timing is incomplete and screening shares outcomes with its readout. MIMIC-IV is handled under credentialed access, no note spans or identifiers are reproduced, and the same non-disclosure principle is applied to education records. Within these boundaries, paired regime changes support the representation-pipeline attribution without converting exploratory groups into clinical or educational decisions.

6 Conclusion

MATRICS makes the complete subgroup-search path the unit of representation attribution. Under a shared proposal, support, screening, and stability operator, the modifier-adapted clinical-text pipeline gains 0.106 recovery F1 over TF–IDF, retains positive lexicalized intervals across all six protocol

alternatives, and preserves a 0.056 gain after literal spans are masked. Moving the modifier into structured covariates contracts every comparison into the four-point decision band, pinpointing the source of value rather than merely ranking encoders. For causal subgroup searches in which narrative records contain otherwise unavailable modifiers, protocol-equivalent evaluation provides both a defensible representation choice and a sharper account of which heterogeneity the text makes recoverable.

AI Use Statement

The authors used generative language tools to assist manuscript preparation, including copyediting, presentation revisions, formatting, and citation and consistency checks. The authors reviewed the complete manuscript and take responsibility for its final content.

References

Ahmed Alaa and Mihaela van der Schaar. Validating causal inference models via influence functions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 191–201, 2019.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-1909.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *PNAS*, 113 (27):7353–7360, 2016.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47 (2):1148–1178, 2019.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652, 2022.

Jared C Foster, Jeremy M G Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, 2011.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021.

Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review of the literature. In *Proceedings of The 3rd International Conference on Predictive Applications and APIs*, volume 67 of *Proceedings of Machine Learning Research*, pp. 1–13, 2017.

Melody Huang, Tiffany M. Tang, and Ana M. Kenney. Distilling heterogeneous treatment effects: Stable subgroup estimation in causal inference, 2025.

Fredrik D Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 3020–3029, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV: a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Alan B Krueger and Diane M Whitmore. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project STAR. *Economic Journal*, 111(468):1–28, 2001.

Lei Liu, Jialin Chen, and Kathy Macropol. LLM-extracted covariates for clinical causal inference: Rethinking integration strategies, 2026.

Yuchen Ma, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. LLM-driven treatment effect estimation under inference time text confounding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0f3b62a67900418cbb8583839129ffc6-Abstract-Conference.html.

Reagan Mozer, Aaron R. Kaufman, Leo A. Celi, and Luke Miratrix. Leveraging text data for causal inference using electronic health records, 2023.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4932–4941, 2019.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. Causal effects of linguistic properties. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4095–4109, 2021.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903, 2020.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *JASA*, 89(427):846–866, 1994.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085, 2017.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 919–928, 2020.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Yinjun Wu, Mayank Keoliya, Kan Chen, Neelay Velingker, Ziyang Li, Emily J. Getzen, Qi Long, Mayur Naik, Ravi B. Parikh, and Eric Wong. DISCRET: Synthesizing faithful explanations for treatment effect estimation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53597–53618. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24n.html.

Jialu Xu, Mengkun Liang, Guannan Liu, Xiaojie Mao, and Junjie Wu. Uncertainty-guided LLM semantic augmentation for heterogeneous treatment effect estimation, 2026.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems 31*, 2018.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

A Semi-Synthetic Data-Generating Processes

All 50 replicates use the same observed MIMIC-IV cohort and pair their random seed across representation pipelines. Each seed draws a patient-disjoint 70/30 outer partition first; every nuisance fold and proposal update that can affect a candidate is nested inside the 70% partition. Standardized structured covariates X determine treatment through

$$T \sim \text{Bernoulli}\{\text{logit}^{-1}(0.15 + X^\top a)\},$$

where a has six nonzero coefficients in $[-0.45, 0.45]$. The baseline outcome is

$$m(X) = 0.4X_1 - 0.3X_2 + 0.25X_3X_4 + 0.2\sin(X_5), \qquad Y = m(X) + T\tau(X, Z) + \epsilon,$$

with $\epsilon \sim \mathcal{N}(0, 0.5^2)$. The lexicalized effect is $\tau = 0.30 - 0.12M_{\text{adherence}} + 0.075M_{\text{support}}$; prevalences are 18.7% and 14.2%, with 3.1% overlap. Trigger labels use curated mention patterns plus negation handling. The semantic-positive regime replaces direct phrases with a fixed paraphrase inventory, masks literal trigger spans, and uses the same target labels for every pipeline; the reported cosine range $[0.62, 0.82]$ characterizes this fixed inventory. The null sets both modifier coefficients to zero while preserving text, treatment, baseline outcomes, and nuisance structure.

The structured-severity regime replaces $M_{\text{adherence}}$ by $M_{\text{sev}} = \mathbf{1}\{S \geq Q_{0.813}(S)\}$, where S is standardized baseline severity, and uses $\tau = 0.30 - 0.12M_{\text{sev}}$. The structured-history regime uses $\tau = 0.30 + 0.075M_{\text{hist}}$ with $M_{\text{hist}} = \mathbf{1}\{H(X) \geq Q_{0.858}\{H(X)\}\}$; $H(X)$ is the fixed standardized score generated from pretreatment history covariates. The empirical quantiles reproduce the 18.7% and 14.2% text-modifier prevalences. Text remains available to the text pipelines but carries no target label in these regimes, so recovery measures whether redundant text machinery changes discovery once X exposes the target.

Matching and averaging. For each replicate, retained predicates and true modifiers form a bipartite graph weighted by unit-level F1. Maximum-weight one-to-one matching determines macro recovery: the mean F1 across true modifiers, assigning zero to any unmatched target. Micro F1 pools matched unit labels. Tables report macro F1; micro F1 gives the same method ordering. Empirical FDR is unmatched retained predicates divided by $\max(1, \text{retained predicates})$. Stability is the mean selection frequency of matched predicates. A replicate with no retained group receives recovery zero, FDR zero, and survivor count zero. Under the no-moderation regime, recovery and empirical FDR are omitted because no positive target exists.

B Proposal Optimization and Nominal Tests

Let $z_i(\theta)$ be the ClinicalBERT representation with trainable parameters θ restricted to the final four transformer blocks. For proposal units $\mathcal{I}_{\text{prop}}$, MATRICS solves

$$\min_{\theta,V,\beta_0,\beta,\gamma} \quad \sum_{i\in\mathcal{I}_{\text{prop}}} \left(\tilde{\tau}_i - \beta_0 - \{V^\top z_i(\theta)\}^\top \beta - X_i^\top \gamma\right)^2 + \lambda \left[\frac{1-\alpha}{2}\left(\|\beta\|_2^2 + \|\gamma\|_2^2\right) + \alpha\left(\|\beta\|_1 + \|\gamma\|_1 + \|V\|_1\right)\right],$$

subject to $\|v_k\|_2 = 1$. The ridge term excludes V because its Frobenius norm is constant under these constraints; sparsity of the proposal directions is controlled by $\|V\|_1$. Backpropagation through $z_i(\theta)$ alternates with proximal-gradient updates of (V, β, γ) and column renormalization of V; validation DR loss selects λ and the joint checkpoint using proposal units only. Frozen-encoder and TF–IDF comparators skip the θ update but use the identical direction and evaluation sequence. The factorization identifies the fitted predictor rather than a unique latent basis, so individual v_k are ordered algorithmic proposal coordinates rather than causal concepts. Deflated PLS initialization, sign normalization, duplicate-support removal, and deterministic rank ties define the frozen coordinate convention.

For candidate S_k, evaluation-unit influence scores are

$$\psi_{ik} = \frac{\mathbf{1}(i \in S_k)}{\hat{p}_k}\left(\tilde{\tau}_i - \hat{\tau}_{S_k}\right) - \left(\tilde{\tau}_i - \hat{\tau}_{\text{pop}}\right),$$

and $\hat{\Delta}_k = \hat{\tau}_{S_k} - \hat{\tau}_{\text{pop}}$. Nuisance functions fitted on $\mathcal{I}_{\text{prop}}$ generate every evaluation score. We use $t_k = \hat{\Delta}_k / \{\widehat{\text{Var}}(\psi_{ik})/n_{\text{eval}}\}^{1/2}$ with two-sided Gaussian nominal p-values. Evaluation units are independent of proposal units; dependence among tests is retained when BH is applied jointly to all 20 p-values. Stability follows BH. Each rerun resamples proposal and evaluation partitions separately and repeats proposal-side nuisance fitting, adaptation, direction learning, and evaluation-side testing. Maximum-weight one-to-one assignment uses signed proposal scores, requires the same tail orientation and correlation at least 0.8, and resolves ties by original then rerun rank before the 80% retention rule is applied.

C Configuration Matrix

D Representation-Ladder Diagnostics

This table places the primary recovery result beside its mechanism diagnostics. The common TF–IDF reference makes each displayed contrast comparable without treating adjacent point-estimate differences as separately tested effects.

Table 7: Reference and ablation configurations. "Outcome-tuned" adapts the encoder for nuisance outcome prediction; only MATRICS uses the modifier-directed proposal loss. All candidate-family and evaluation components remain matched.

Configuration	Encoder state	Proposal and nuisance inputs	Matched search
MATRICS	ClinicalBERT; proposal-only modifier adaptation	sparse DR directions; $X+$ embedding	full Π
Frozen ClinicalBERT	fixed checkpoint	sparse DR directions; $X+$ embedding	full Π
Outcome-tuned ClinicalBERT	proposal-fold outcome adaptation	sparse DR directions; $X+$ embedding	full Π
Frozen generic BERT	fixed checkpoint	sparse DR directions; $X+$ embedding	full Π
Frozen PubMedBERT	fixed checkpoint	sparse DR directions; $X+$ embedding	full Π
Frozen GPT–2	fixed checkpoint	sparse DR directions; $X+$ embedding	full Π
TF–IDF	fixed vocabulary	sparse DR directions; $X+$ TF–IDF	full Π
No text	none	sparse structured directions; X	matched downstream search

Table 8: Lexicalized and boundary diagnostics for the complete protocol-matched representation ladder over 50 shared seeds.

Pipeline	Lexical F1	RMSE	Emp. FDR	Stability	Semantic F1	Null any stable
TF–IDF	0.585	0.301	0.09	0.73	0.552	0.16
GPT–2 frozen	0.573	0.312	0.13	0.70	0.541	0.22
Generic BERT frozen	0.638	0.274	0.10	0.79	0.581	0.18
ClinicalBERT frozen	0.641	0.268	0.09	0.81	0.592	0.20
PubMedBERT frozen	0.676	0.251	0.08	0.84	0.603	0.18
ClinicalBERT outcome-tuned	0.668	0.254	0.08	0.83	0.600	0.18
MATRICS	0.691	0.243	0.07	0.86	0.608	0.20

E Protocol Robustness and Null Family Yield

The null rows use the same seven configurations as the positive-regime neighborhood and expose both pre-stability screening and final family survival. MATRICS has the higher mean BH-pass count and maximum absolute contrast in every configuration; the final any-survivor ordering reverses for $K = 40$ and $q = 0.15$, showing what the complete-path stability stage removes.

Table 9: Null family-yield details. Counts are replicates with any stable survivor out of 50. Survivor summaries are out of K; BH passes are measured before the 80% stability filter.

Setting, pipeline	Any stable (count)	Mean survivors	90th pct.	Mean BH passes	Max $\lvert\hat{\Delta}\rvert$
$(10, 20\%, .10)$ TF–IDF	.14 (7)	.32	1	.96	.078
$(10, 20\%, .10)$ MATRICS	.18 (9)	.42	1	1.18	.083
$(20, 20\%, .10)$ TF–IDF	0.16 (8/50)	0.40	1	1.20	.084
$(20, 20\%, .10)$ MATRICS	0.20 (10/50)	0.50	2	1.42	.089
$(40, 20\%, .10)$ TF–IDF	.20 (10)	.54	2	1.58	.092
$(40, 20\%, .10)$ MATRICS	.18 (9)	.48	2	1.67	.095
$(20, 10\%, .10)$ TF–IDF	.22 (11)	.58	2	1.74	.096
$(20, 10\%, .10)$ MATRICS	.24 (12)	.66	2	1.89	.101
$(20, 30\%, .10)$ TF–IDF	.12 (6)	.26	1	.82	.073
$(20, 30\%, .10)$ MATRICS	.16 (8)	.34	1	.95	.080
$(20, 20\%, .05)$ TF–IDF	.08 (4)	.18	1	.64	.069
$(20, 20\%, .05)$ MATRICS	.10 (5)	.22	1	.71	.074
$(20, 20\%, .15)$ TF–IDF	.24 (12)	.64	2	1.86	.101
$(20, 20\%, .15)$ MATRICS	.22 (11)	.60	2	1.92	.106

F Structured-Only Methodwise Results

Absolute methodwise results complement the paired compatibility assessment in the main text. The displayed MATRICS-minus-comparator differences remain inside the declared four-point margin.

Table 10: Baseline-severity modifier. Intervals summarize 50 shared replicates.

Method	Recovery F1	Emp. FDR	Stability	Stable survivors / 20
MATRICS	0.544 [0.512, 0.575]	0.10 [0.07, 0.14]	0.75 [0.69, 0.81]	2.4 [1.9, 2.9]
DR learner + TF–IDF	0.538 [0.505, 0.568]	0.11 [0.08, 0.15]	0.74 [0.68, 0.80]	2.3 [1.8, 2.8]
Causal forest	0.533 [0.501, 0.564]	0.12 [0.08, 0.16]	0.72 [0.66, 0.78]	2.2 [1.7, 2.7]

Table 11: Baseline-history-score modifier. Intervals summarize 50 shared replicates.

Method	Recovery F1	Emp. FDR	Stability	Stable survivors / 20
MATRICS	0.517 [0.486, 0.547]	0.13 [0.09, 0.18]	0.71 [0.64, 0.78]	1.8 [1.3, 2.4]
DR learner + TF–IDF	0.522 [0.491, 0.553]	0.12 [0.08, 0.17]	0.72 [0.65, 0.79]	1.9 [1.4, 2.5]
Causal forest	0.519 [0.488, 0.551]	0.12 [0.08, 0.17]	0.73 [0.66, 0.80]	1.9 [1.4, 2.6]

Table 12: Paired supporting-diagnostic differences (MATRICS minus comparator), with 90% intervals. All differences remain practically small and change sign across regimes.

Modifier	Comparator	Emp. FDR difference	Stability difference	Survivor difference
Baseline severity	TF–IDF	−0.010 [−0.040,0.020]	0.010 [−0.040,0.060]	0.10 [−0.30,0.50]
Baseline severity	causal forest	−0.020 [−0.050,0.010]	0.030 [−0.020,0.080]	0.20 [−0.20,0.60]
Baseline-history score	TF–IDF	0.010 [−0.020,0.040]	−0.010 [−0.060,0.040]	−0.10 [−0.50,0.30]
Baseline-history score	causal forest	0.010 [−0.020,0.040]	−0.020 [−0.070,0.030]	−0.10 [−0.50,0.30]

G Ablations and Sensitivity

The reference MATRICS system uses modifier-directed adaptation on proposal units. Frozen ClinicalBERT removes all encoder updates; outcome-tuned ClinicalBERT replaces the modifier objective with nuisance outcome prediction; frozen GPT–2 changes the pretrained representation; and no text reruns the downstream protocol on structured variables. The ladder separates three effects under common candidate construction and evaluation: contextual pretraining raises recovery over TF–IDF, biomedical domain alignment supplies the strongest frozen checkpoints, and modifier adaptation produces the highest observed F1 and lowest CATE RMSE.

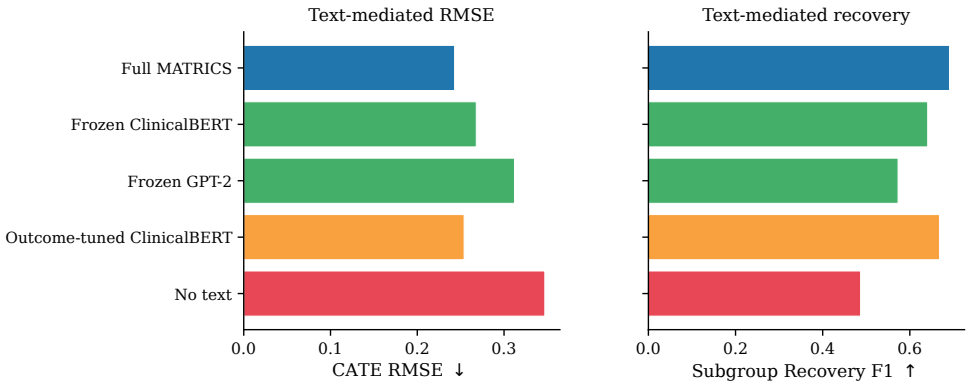


Figure 1: Representation and adaptation ablations under the lexicalized operating point. CATE RMSE and recovery F1 identify the effect of each encoder intervention; Table 5 supplies paired intervals for the structured-only control.

H Full STAR Readouts

The two STAR displays separate estimator sensitivity for fixed MATRICS predicates (Table 13) from representation-specific discovery outputs (Table 6). They are reported as conditional transfer diagnostics after screening.

Across six additional 70/30 splits, the full-fold behavior and reading groups recur 4/6 and 2/6 times with median CATE readouts 0.121 [0.058, 0.181] and 0.029 [−0.008, 0.067]. The full-fold TF–IDF readouts are 0.109 [0.032, 0.187] and 0.027 [−0.019, 0.074]. These values reproduce the stronger behavior and weaker reading-readiness ordering in the timing-restricted matrix.

Exact predicates. Each reported note group is an upper-tail rule $\mathbf{1}\{\langle z, v_k\rangle \geq q_{0.80}^{\text{prop}}\}$ with its score and threshold frozen on proposal units. Behavior-group selected units are characterized by stems including *disrupt*, *attention*, and *behavior*; reading-readiness units emphasize *reading readiness* and *literacy preparation*. These descriptors label the learned embedding rules and do not replace them.

Table 13: Seed-42 evaluation-set effects on frozen, MATRICS-discovered STAR groups. CF and R columns evaluate the same predicates and are estimator checks rather than independent discovery outputs.

#	Frozen subgroup predicate	MATRICS DR	CF evaluation	R evaluation	Bootstrap frequency
(i)	Behavior note pattern	0.178 [0.091, 0.265]	0.142 [0.061, 0.218]	0.135 [0.052, 0.219]	0.91
(ii)	Reading-readiness note pattern	0.064 [0.013, 0.115]	0.038 [−0.021, 0.097]	0.041 [−0.019, 0.101]	0.87
(iii)	Low baseline achievement	0.149 [0.070, 0.228]	0.151 [0.073, 0.229]	0.148 [0.071, 0.225]	0.72 / 0.69 / 0.71

Decision-rule diagnostic. For comparability, we evaluate the arithmetic rule that assigns the small-class treatment to evaluation units satisfying a frozen subgroup and regular-class treatment otherwise. Stacked DR readouts with 1,000 evaluation-unit bootstraps are 0.31 [0.24, 0.38] for MATRICS, 0.22 [0.15, 0.29] for causal forest, and 0.24 [0.17, 0.31] for the R-learner; the matched full-fold TF–IDF value is 0.26 [0.19, 0.33]. These values rank selected-group rules within the exploratory STAR transfer study and are not used for the paper's causal attribution.

I Nuisance and Overlap Diagnostics

On MIMIC-IV, the estimated propensity 5th–95th percentile range is [0.112, 0.887], with 96.8% of estimates in $(0.05, 0.95)$. On STAR, the BH pass count remains 3/20 under both nuisance specifications, while evaluation outcome RMSE changes from 0.297/0.274 to 0.304/0.281 under a shallower-tree rerun. These diagnostics support the overlap and nuisance stability of the operating point.