FROM CAUSAL DIAGNOSTICS TO REPORTING DECISIONS: TRIALVETO FOR OMOP TARGET-TRIAL EMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal diagnostics expose mistimed treatment, inadequate overlap, negative-control falsification, and censoring instability in OMOP target-trial emulations, but convergence alone does not resolve whether an effect estimate supports an unqualified claim. We formulate reportability as a three-action decision—report, soft flag, or abstain—and introduce TrialVeto, a calibration-locked precedence rule with active-reason and estimand-drift traces. Evaluation scores every analysis in the declared workload by reporting coverage, risk among released estimates, risk among withheld estimates, and their difference. On 610 held-out drug–outcome pair–database analyses, TrialVeto emits 418 (68.5%) with median/P90 absolute log-HR disagreement of 0.12/0.31, compared with 0.19/0.58 for full-report Standard TTE and 0.15/0.43 after OHDSI calibration. The 192 abstentions have 0.34/0.71 error, yielding median/P90 gaps of 0.22 [0.16,0.29]/0.40 [0.29,0.53] and containing 75 of the 88 severe full-report disagreements. With thresholds re-locked, the median gap remains 0.24 on 153 unseen pairs and 0.25 under leave-one-database-out evaluation; at the identical 418-output budget, TrialVeto improves on the evaluated squared-error gradient-boosted selector by 0.025, 0.072, and 0.033 in median error, P90 error, and AURCC. Blinded clinical review recovers the intended reportability gradient (adjusted OR 10.4), and utility favors TrialVeto over OHDSI calibration whenever withholding costs less than 0.17 of a severe released error. In active-comparator, time-to-event OMOP emulations, TrialVeto turns post-estimation diagnostics into a transferable reporting decision with explicit error and availability costs.

1 INTRODUCTION

An observational effect estimate can converge while failing the conditions needed for an unqualified causal claim. Target-trial emulation aligns eligibility, treatment strategies, assignment, time zero, follow-up, and outcomes (Hernán et al., 2016; Hernán & Robins, 2016; Hernán et al., 2022; Cashin et al., 2025). In a distributed OMOP analysis, however, mistimed treatment, poor overlap, negative-control failures, censoring instability, or a population change after trimming can remain after the protocol is specified (Suissa, 2008; Yadav & Lewis, 2021; Danaei et al., 2012; Crump et al., 2009; Petersen et al., 2012; Lipsitch et al., 2010; Cole & Hernán, 2008). These diagnostics are routinely inspectable, but existing workflows do not turn their joint state into a reproducible decision about whether to report, flag, or withhold the effect estimate.

Selective reporting creates a second statistical problem: accuracy among released estimates says nothing about the analyses removed from that score. A reporting rule should therefore be evaluated as a partition of one declared workload. We report coverage, emitted risk, withheld risk, and the withheld-minus-emitted gap together, while preserving each policy's native ordering for risk–coverage comparison. This object asks whether abstention locates concentrated error, whether that separation transfers to new clinical questions and databases, and whether the reduction in erroneous claims justifies its availability cost.

TrialVeto supplies the corresponding decision mechanism. Whether follow-up begins at treatment initiation, treated and comparator cohorts overlap, negative controls falsify the contrast, and censoring

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.

adjustment is stable enters a deterministic precedence map with report, soft-flag, and abstain actions. Hard failures dominate warnings; every active reason remains in the trace; and trimming that changes the treated population triggers an estimand-drift disclosure. Thresholds and the operating point are selected on calibration units, then frozen for each test regime. The same cohorts and survival-AIPW estimates feed TrialVeto, its ablations, and the learned selection comparator, isolating which analyses each rule exposes.

Contributions and evidence.

- **A measurable reportability object.** The complete-denominator profile is instantiated on 218 protocol-matched RCT drug–outcome pairs crossed with MarketScan, Optum, IQVIA, and CPRD. Unit-, pair-, and database-separated evaluations keep the released and withheld regions in a common ledger.
- **An inspectable causal decision rule.** TrialVeto composes four failure families rather than treating a generic confidence score as permission to report. Single-family removals, three-action gradients, reason co-occurrence, and independently curated negative controls identify how the composed boundary operates.
- **Transferable error separation at a quantified cost.** The primary median/P90 gaps are 0.22/0.40 with pair-bootstrap lower bounds above zero, and the abstentions contain 75 of Standard TTE's 88 severe disagreements. Pair- and database-separated tests preserve the ordering; equal-budget comparison isolates selection quality from abstention rate; and utility states the corresponding availability price.

Blinded clinical judgments supply action-semantic concordance, while the severe-error utility curve states the availability price at which selective reporting improves on full-report practice. Together, these results turn causal diagnostics from a post hoc display into a testable reporting decision.

2 Related work

Target-trial design and reporting. Target-trial guidance requires alignment of eligibility, treatment assignment, time zero, follow-up, outcome, and causal contrast; TARGET and RECORD-PE further require these choices and their mapping to routinely collected data to be visible (Hernán et al., 2016; Hernán & Robins, 2016; Hernán et al., 2022; Cashin et al., 2025; Langan et al., 2018). OHDSI infrastructure standardizes cohort construction, balance assessment, negative controls, and cross-database execution (Hripcsak et al., 2015; Voss et al., 2015; Reps et al., 2018; Suchard et al., 2019). TrialVeto operates after these design choices: it turns observed diagnostic states into a frozen output decision and then evaluates every resulting action.

Calibration and RCT-anchored evaluation. Empirical calibration models systematic error to adjust an estimate and its interval while retaining the estimate (Schuemie et al., 2014; 2020). RCT-anchored benchmarks compare emulated and randomized effects at the aggregate level (Schuemie et al., 2020; Wang et al., 2023). Our primary object is instead the partition induced by selective reporting: the same anchor loss is evaluated among outputs the policy releases and among analyses it withholds, under calibration, pair, and database separation.

Selective prediction and risk control. Reject-option and selective-classification methods optimize predictive risk under abstention, learned confidence, ensembles, or set-valued outputs (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Jiang et al., 2018; Geifman & El-Yaniv, 2019; Sadinle et al., 2019; Guo et al., 2017; Lakshminarayanan et al., 2017; Corbière et al., 2019). Conformal risk control supplies finite-sample control for monotone losses under its exchangeability conditions (Angelopoulos et al., 2024). Our object is a causal-effect output governed by failure-specific actions, reason traces, estimand-change disclosure, and explicit risk on both sides of the reporting decision. The learned comparison is a squared-error gradient-boosted selector fit on calibration units with cohorts, diagnostics, and the effect estimator held fixed; matched coverage tests whether the prespecified causal rule selects a lower-risk subset than this error-regression baseline.

3 BENCHMARK AND EVALUATION PROTOCOL

3.1 RCT-TO-OMOP ALIGNMENT AND STUDY UNITS

Two clinical epidemiologists independently mapped and verified 218 active-comparator, time-to-event RCT anchors (Cohen's $\kappa = 0.91$; 14 disagreements resolved by consensus). For each candidate, the mapping form records the trial population and observational eligibility rule, active treatment and comparator strategies, assignment and time zero, outcome definition, follow-up horizon, censoring rule, target population, published HR and interval, and treatment-direction normalization. Inclusion requires an incident-user active-comparator emulation, an ascertainable endpoint, compatible follow-up, and a published HR that can be oriented to the same treatment contrast; ambiguous endpoint or protocol mappings remain excluded. The resulting anchor error is therefore a protocol-matched agreement endpoint rather than an assertion that an RCT point estimate is ground truth for every database population.

Crossing every retained pair with MarketScan, Optum, IQVIA, and CPRD produces 872 analysis units across six therapeutic areas. The primary split contains 262 calibration and 610 untouched pair–database test units. A harder grouped split assigns all four database units for 65 pairs to calibration and all units for 153 unseen pairs to a 612-unit test, stratified by therapeutic area. Four leave-one-database-out folds each calibrate without one database and test on its 218 units. Each design has its own calibration/test assignment and is scored separately.

3.2 ENDPOINTS, LOCKING, AND COMPARISONS

The operating point is selected only on calibration units from the fixed grid in Appendix A.1. Among settings with 65–75% reporting coverage, selection minimizes P90 anchor error, then median error, then maximizes coverage; a final tie favors the lexicographically simpler threshold tuple. Test anchors are unavailable to threshold selection. Standard TTE and OHDSI-calibrated are full-report baselines, overlap-only and NC-only are separately locked one-family comparators, and the learned selector supplies native abstention semantics. These one-family comparators retain their native actions; the single-guardrail removals in Appendix Table 20 instead delete one branch from the full TrialVeto rule.

Primary endpoints are median and P90 absolute log-HR anchor error. Sign disagreement, severe disagreement $|\Delta \log \mathrm{HR}| > 0.50$ (an HR-ratio discrepancy above $e^{0.50} = 1.65$), reporting coverage, and native AURCC are secondary endpoints; sign and severe counts and their rate gaps are descriptive. Point-anchor coverage is supporting agreement evidence (Wang et al., 2023). Coverage and estimand-drift percentages are standardized to the benchmark's six-area composition with fixed inverse split-probability weights; action counts retain raw denominators. Reported 95% intervals for continuous-risk gaps use 1,000 percentile-bootstrap replicates sampled by drug–outcome pair; learned-selector intervals also include between-seed variation. No estimator failure occurred among the scored benchmark units, so each full-report denominator equals its complete test set.

The experiments ask whether withheld risk exceeds emitted risk, whether that separation survives pair and database transfer, whether squared-error ranking or a single diagnostic family explains the result, and when the reduction in severe emitted errors exceeds the cost of withholding. Blinded action adjudication measures concordance with the intended reportability semantics, and independently curated negative controls test sensitivity to the original control panel without using anchor error.

Ethics and governance. The four retrospective databases were analyzed under the institutional approvals, waivers, or non-human-subject-research determinations and data-use authorizations applicable to each contributing data custodian. The study involved no participant contact or intervention. The clinical epidemiologists performed mapping and adjudication in their professional roles; adjudication was blinded to RCT-anchor error and TrialVeto action.

4 TRIALVETO AND COMPLETE-DENOMINATOR SCORING

4.1 COMPLETE-DENOMINATOR RISK PROFILE

For held-out analysis units $i = 1, \ldots, N$ and policy p, let $\widehat{\beta}_{ip}$ be the effect estimate exposed by that policy, $e_{ip} = |\widehat{\beta}_{ip} - \beta_i^{\mathrm{RCT}}|$ its absolute log-HR anchor error, a_{ip} its action, and $z_{ip} = \mathbb{1}[a_{ip} \neq \texttt{abstain}]$. TrialVeto, its ablations, Standard TTE, and the learned selector share the same $\widehat{\beta}_i$; OHDSI calibration uses $\widehat{\beta}_{i,\mathrm{OHDSI}} = \widehat{\beta}_i - \widehat{\mu}$. For $m \in \{\mathrm{med}, \mathrm{p90}, \mathrm{sign}, \mathrm{sev}\}$, metric-specific unit losses are $\ell_{ip}^{\mathrm{med}} = \ell_{ip}^{\mathrm{p90}} = e_{ip}$, $\ell_{ip}^{\mathrm{sign}} = \mathbb{1}[\mathrm{sign}(\widehat{\beta}_{ip}) \neq \mathrm{sign}(\beta_i^{\mathrm{RCT}})]$, and $\ell_{ip}^{\mathrm{sev}} = \mathbb{1}[e_{ip} > 0.50]$. Let A_m be the corresponding sample median, 90th percentile, or arithmetic mean. Reporting coverage is $c_p = N^{-1} \sum_i z_{ip}$. For metric m, define

$$\begin{aligned} R_{E,p}(m) &= A_m(\{\ell_{ip}^m : z_{ip} = 1\}), \quad R_{W,p}(m) = A_m(\{\ell_{ip}^m : z_{ip} = 0\}), \\ \Delta_{\mathrm{auth},p}(m) &= R_{W,p}(m) - R_{E,p}(m). \end{aligned} \tag{1}$$

The risk-routing profile is $\mathcal{Q}_p(m) = (c_p, R_{E,p}(m), R_{W,p}(m), \Delta_{\mathrm{auth},p}(m))$; policy subscripts are suppressed when the comparison is unambiguous. Emitted risk alone cannot identify selective reliability: for a fixed emitted set and fixed $R_{E,p}$, arbitrarily high-loss units can be placed outside the reported score without changing it. Recording c_p, $R_{W,p}$, and $\Delta_{\mathrm{auth},p}$ exposes that degree of freedom. For the primary median and P90 losses, evidence for authorization requires improvement over the declared comparator, a pair-bootstrap lower bound for $\Delta_{\mathrm{auth},p}$ above zero, and the accompanying coverage cost. Sign and severe-disagreement gaps retain complete counts but are interpreted descriptively. Native risk–coverage compares each policy through its own score; L_λ in Equation 2 evaluates the reporting decision under explicit withholding costs.

4.2 TRIALVETO DECISION RULE

TrialVeto is $C = (d, \pi, \rho)$, where $d(u)$ is the diagnostic state of unit u, $\pi(d(u))$ is its mandatory action, and $\rho(u)$ contains every active reason code and estimand-drift disclosure. Identical states receive identical actions, hard failures dominate warnings, and all active restrictions remain visible. A time-zero or overlap failure forces abstention. A censoring range above 0.30, at least two Bonferroni-significant negative controls, or one such control combined with a censoring range in $[0.20, 0.30]$ also forces abstention. A single remaining warning produces a soft flag with mandatory manual review; otherwise the estimate is reported. Soft flags remain in the emitted denominator. Appendix Table 4 gives the complete precedence map.

Each emulation uses an incident-user time zero at treatment or active-comparator initiation. The three-day quantity is a diagnostic tolerance between the recorded eligibility date and initiation, not exposed pre-initiation person-time: follow-up and both treatment risk sets begin at initiation, and any outcome timestamped before that point is a hard failure. TrialVeto abstains if more than 10% of treated subjects exceed the tolerance or any pre-follow-up outcome occurs. Five-fold cross-fitted XGBoost propensity scores (Chen & Guestrin, 2016) define trimming to $[0.05, 0.95]$; retained sample below 70% or any post-trim SMD above 0.1 forces abstention, and removing more than 5% of treated subjects requires a drift note. Twelve pair-specific negative-control outcomes are fixed across databases and tested with Bonferroni $\alpha = 0.05$. The censoring branch re-estimates the effect over $\gamma \in \{0, 0.1, 0.2, 0.5, 1.0\}$ and uses the resulting log-HR range; Appendix A defines the tilt.

4.3 SHARED ESTIMATOR AND LEARNED COMPARATOR

Every selection policy and full-report baseline begins from the same five-fold cross-fitted survival-AIPW estimate in the propensity-overlap population. XGBoost estimates treatment propensity, penalized Cox models estimate treatment-specific event survival, and Cox models estimate treatment-specific censoring survival. The exact update in Appendix A applies calibration-locked 1st/99th-percentile clipping to arm-specific stabilized censoring weights before standardizing over the retained eligible cohort. A marginal structural Cox projection then produces the log-HR matched to the RCT contrast, with within-unit sandwich variance propagated from the cross-fitted one-step contributions. Standard TTE emits this estimate directly. OHDSI calibration fits a database- and area-stratified Gaussian systematic-error model on calibration negative controls, subtracts estimated mean bias, and

Table 1: Primary risk-routing profile on 610 held-out units. Error cells report median/P90 absolute log-HR error; bracketed values are pair-bootstrap 95% intervals. Positive gaps show how much risk the authorization boundary isolates, and disagreements retain their complete slice denominators.

Slice	n	Error	Sign	Severe
Emitted	418	0.12/0.31 [0.09,0.15] / [0.24,0.39]	24/418	13/418
Withheld	192	0.34/0.71 [0.28,0.41] / [0.59,0.85]	87/192	75/192
Authorization gap	–	0.22/0.40	39.6 pp	36.0 pp

adds estimated systematic variance to the interval variance. Non-convergence, non-finite weights, or non-positive information invokes estimator-failure abstention before diagnostic precedence.

The learned comparator minimizes squared error to absolute anchor error on calibration units using the TrialVeto diagnostics, effective sample size, interval width, cohort and event counts, retained fraction, and maximum post-trim SMD. Five gradient-boosted models use pair-grouped cross-fitting and distinct seeds; their median predicted error is the native rejection score. Its threshold is locked under the same coverage constraint as TrialVeto, and a matched comparison emits exactly 418 of 610 units. Native AURCC orders TrialVeto by its prespecified action-and-diagnostic severity tuple and the selector by predicted error, then integrates median risk over the same 20 coverage knots (Appendix A.1).

5 Results

5.1 The reportable set excludes concentrated error

TrialVeto's primary risk-routing profile emits 418 of 610 test analyses and withholds 192. Table 1 shows separation in central error, tail error, and raw decision disagreements. The median authorization gap is 0.22 ([0.16, 0.29]); the P90 gap is 0.40 ([0.29, 0.53]). Both intervals remain well above zero under pair-level resampling. Descriptively, sign and severe-disagreement rates are 39.6 pp and 36.0 pp higher among abstentions. The complete profile therefore shows that the 31.5% no-claim cost buys risk isolation across continuous and decision-facing losses.

The withheld errors arise from complementary causal failure states. Overlap, negative-control, censoring, and time-zero triggers occur in 58.3%, 41.1%, 28.1%, and 18.8% of withheld units, respectively, with 34.9% activating multiple branches. Removing time-zero validation admits more analyses and raises median/P90 error from 0.12/0.31 to 0.16/0.44; removing overlap diagnostics yields 0.14/0.39; removing either remaining family also increases error (Appendix Table 20). The full rule thus gains from composition rather than one dominant veto.

5.2 Causal diagnostics improve on the matched squared-error selector

At exactly 418 emitted units, the squared-error gradient-boosted selector reaches median/P90 error 0.145/0.382, native AURCC 0.151, sign disagreement 34/418 (8.1%), and severe disagreement 21/418 (5.0%). TrialVeto retains lower central and tail error (0.12/0.31), lower native AURCC (0.118), and fewer disagreements at the identical budget. Paired pair-bootstrap differences (selector minus TrialVeto) are 0.025 [0.006,0.046] for median error, 0.072 [0.021,0.129] for P90 error, and 0.033 [0.014,0.055] for AURCC. Because TrialVeto and this selector act on identical cohorts and shared survival-AIPW estimates, these differences measure which analyses each policy chooses to expose.

Table 2: Competitive primary-split authorization profiles. TrialVeto and the learned selector each use their own score at the same 418-output budget, and AURCC integrates each policy's native ordering; full-report baselines retain all 610 units. Emitted error, authorization gap, and AURCC are lower, higher, and lower when better, respectively.

Policy	Emitted	Cov.	Emitted med./P90	Gap med./P90	AURCC
TrialVeto	418	68.5%	0.12/0.31	0.22/0.40	0.118
Learned risk selector	418	68.5%	0.145/0.382	0.165/0.278	0.151
OHDSI-calibrated	610	100.0%	0.15/0.43	–	–
Standard TTE	610	100.0%	0.19/0.58	–	–

Table 3: Transferred risk-routing profiles. Error is median/P90 and gaps compare withheld minus emitted risk; brackets are pair-bootstrap 95% intervals. Positive gaps across all regimes show that the authorization boundary survives new pairs and new databases.

Regime	Emitted / withheld	Emitted error	Withheld error	Gap
Unit-held-out	418 / 192	0.12 / 0.31	0.34 / 0.71	0.22 [0.16,0.29] / 0.40 [0.29,0.53]
Pair-held-out	409 / 203	0.13 / 0.35	0.37 / 0.78	0.24 [0.17,0.32] / 0.43 [0.31,0.57]
Leave-one-database-out	568 / 304	0.15 / 0.40	0.40 / 0.84	0.25 [0.18,0.33] / 0.44 [0.32,0.59]

5.3 The partition persists for new pairs and databases

On 153 unseen drug–outcome pairs, TrialVeto emits 409 of 612 units and preserves a median authorization gap of 0.24. Leave-one-database-out evaluation pools only after all four 218-unit folds are scored separately; emitted error increases under site transfer, while withheld error remains higher in every fold. Pooled emitted/withheld medians are 0.15/0.40, with gap 0.25. MarketScan, Optum, IQVIA, and CPRD emit 154/218, 148/218, 140/218, and 126/218; the varying coverage reflects site-specific diagnostic states rather than a fixed quota.

5.4 Clinical review and decision cost support the actions

The primary authorization audit pools soft flags with reports, and the three-action analysis adds a graded-risk test within that emitted set. Of the 418 emitted outputs, 334 are clean reports and 84 are warning-bearing soft flags. Median/P90 error rises from 0.10/0.26 for reports to 0.20/0.47 for soft flags and 0.34/0.71 for abstentions. Raw sign/severe disagreements similarly rise from 14/334 and 6/334, to 10/84 and 7/84, and then to 87/192 and 75/192. Appendix Table 9 reports pair-clustered intervals and review obligations.

A prespecified stratified, blinded two-reviewer adjudication of 180 held-out units labels 70/90 abstentions, 11/30 soft flags, and 9/60 reports as non-reportable ($\kappa = 0.79$). The adjusted abstain-versus-report odds ratio is 10.4. Because reviewers apply the same four reportability concepts encoded by TrialVeto, this result establishes blinded concordance with the intended action semantics rather than an additional causal-effect accuracy test. Separately, 60 pairs yielded 168 pair–database units with complete independently curated negative-control panels; the remaining units were outside this paired action comparison. Raw agreement over the 168 eligible units is 88.7% with $\kappa = 0.78$, and the recomputed emitted/withheld median errors are 0.13/0.35. The independent panel therefore tests the stability of negative-control-driven actions, while RCT-anchor analyses supply the external error comparison.

Finally, we evaluate the availability tradeoff with

$$L_\lambda = \frac{N_{\text{severe, emitted}} + \lambda N_{\text{withheld}}}{N}, \tag{2}$$

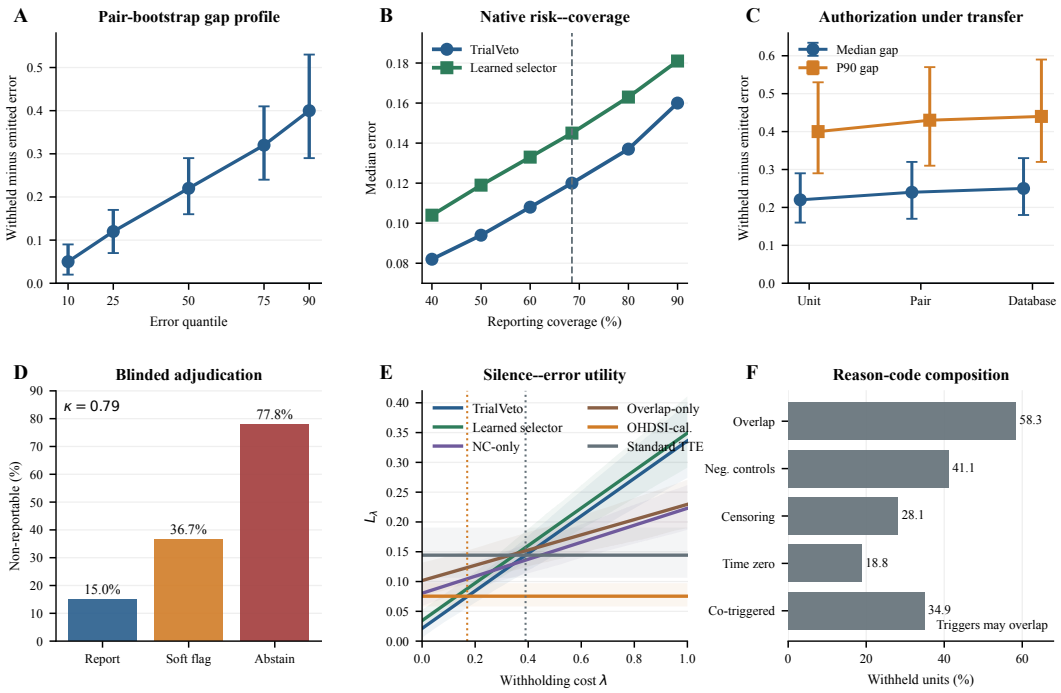


Figure 1: Report authorization consistently routes risk away from emitted causal claims. (A) The withheld-minus-emitted gap is positive from the 10th through 90th error percentiles. (B) TrialVeto has lower native risk than the learned selector over the plotted 40–90% selective-coverage range. (C) Median and P90 gaps persist under unit, pair, and database holdouts. (D) Blinded adjudication reproduces the three-action gradient. (E) Utility curves identify withholding-cost regimes where routing improves decisions. (F) Reason traces reveal complementary, co-occurring failure states. Intervals and bands use pair-level bootstrap resampling.

where $\lambda \in [0, 1]$ is the cost of withholding relative to one severe emitted error. TrialVeto has lower severe-error loss than OHDSI-calibrated for $\lambda < 0.17$ and than Standard TTE for $\lambda < 0.39$; sign-disagreement analysis gives the same ordering over a broad interval (Appendix Table 16). The operating point is therefore supported across a transparent range of decision preferences rather than one chosen exchange rate.

6 Discussion

The evidence separates the roles of statistical accuracy, mechanism, and action semantics. Positive median and tail gaps locate concentrated anchor disagreement beyond the reporting boundary. Equal-budget paired comparisons isolate selection quality from the amount of abstention, and guardrail removals plus co-trigger frequencies attribute the gain to complementary causal diagnostics. Pair and database holdouts establish transfer beyond the calibration units. Blinded adjudication then measures whether the reason-coded actions match their intended clinical meaning, while the utility thresholds translate the RCT-anchor error reduction into an explicit availability tradeoff.

RCT anchors are clinically grounded external comparators rather than ground truth for every database population. Residual differences in population, follow-up, proportional-hazards fit, and outcome measurement can contribute to anchor error, while unmeasured confounding can survive all four diagnostic families. The learned comparison establishes an advantage over the evaluated squared-error gradient-boosted selector; objectives tailored directly to quantile or coverage risk define a different comparison. Negative controls and censoring perturbations add computation, and soft flags require expert review. The represented regime comprises active-comparator, time-to-event questions in four databases and six therapeutic areas; protocol matching, sampled-anchor sensitivity, action-semantic

adjudication, and transfer across pairs and databases provide complementary support within that regime.

The main implication is specific to selective causal reporting: a lower error among released estimates is persuasive only when the displaced analyses, coverage loss, and action semantics are observable. TrialVeto supplies that evidence with a deterministic policy whose reasons can be inspected case by case, and the fixed causal rule outperforms the evaluated squared-error gradient-boosted comparator under limited calibration data. The same profile can compare future policies without requiring them to share TrialVeto's diagnostic rule.

7 Conclusion

For OMOP target-trial emulation, reportability can be defined as a reproducible decision over causal diagnostics rather than an automatic consequence of estimator convergence. TrialVeto shows that a deterministic, reason-traced rule can route severe anchor disagreement away from released causal claims while keeping every abstention in the evaluation ledger. Positive gaps under pair and database transfer, an advantage over the matched squared-error selector, and explicit utility break-even points quantify that tradeoff for the evaluated workload. Jointly accounting for released risk, withheld risk, and availability makes the consequence of each reporting decision visible.

AI Use Statement

Generative AI tools assisted language editing, formatting, and bibliographic checks during manuscript preparation.

References

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg. Unpaginated proceedings.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

Aidan G. Cashin, Harrison J. Hansford, Miguel A. Hernán, Sonja A. Swanson, Hopin Lee, Matthew D. Jones, Issa J. Dahabreh, Barbra A. Dickerman, Matthias Egger, Xabier García-Albéniz, Robert M. Golub, Nazrul Islam, Sara Lodi, Margarita Moreno-Betancur, Sallie-Anne Pearson, Sebastian Schneeweiss, Melissa K. Sharp, Jonathan A. C. Sterne, Elizabeth A. Stuart, and James H. McAuley. Transparent reporting of observational studies emulating a target trial: The TARGET statement. *BMJ*, 390:e087179, 2025. doi: 10.1136/bmj-2025-087179.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16:41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Stephen R. Cole and Miguel A. Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168:656–664, 2008. doi: 10.1093/aje/kwn164.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2902–2913, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/757f843a169cc678064d9530d12a1881-Abstract.html.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96:187–199, 2009. doi: 10.1093/biomet/asn055.

Goodarz Danaei, Mohammad Tavakkoli, and Miguel A. Hernán. Bias in observational studies of prevalent users: Lessons for comparative effectiveness research from a meta-analysis of statins. *American Journal of Epidemiology*, 175(4):250–262, 2012. doi: 10.1093/aje/kwr301.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Miguel A. Hernán and James M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016. doi: 10.1093/aje/kwv254.

Miguel A. Hernán, Babette Brumback, and James M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11:561–570, 2000. doi: 10.1097/00001648-200009000-00012.

Miguel A. Hernán, Brian C. Sauer, Sonia Hernández-Díaz, Robert Platt, and Ian Shrier. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, 79:70–75, 2016. doi: 10.1016/j.jclinepi.2016.04.014.

Miguel A. Hernán, Wei Wang, and David E. Leaf. Target trial emulation. *JAMA*, 328(24):2446–2447, 2022. doi: 10.1001/jama.2022.21383.

George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R. Rijnbeek, Johan van der Lei, Nicole Pratt, G. Niklas Norén, Yu-Chuan Li, Paul E. Stang, David Madigan, and Patrick B. Ryan. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216:574–578, 2015. doi: 10.3233/978-1-61499-564-7-574.

Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5546–5557, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/7180cffd6a8e829dacfc2a31b3f72ece-Abstract.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Sinéad M. Langan, Sigrún A. J. Schmidt, Kevin Wing, Vera Ehrenstein, Stuart G. Nicholls, Kristian B. Filion, Olaf Klungel, Irene Petersen, Henrik T. Sørensen, William G. Dixon, Astrid Guttmann, Katie Harron, Lars G. Hemkens, David Moher, Sebastian Schneeweiss, Liam Smeeth, Miriam Sturkenboom, Erik von Elm, Shirley V. Wang, and Eric I. Benchimol. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ*, 363:k3532, 2018. doi: 10.1136/bmj.k3532.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21:383–388, 2010. doi: 10.1097/EDE.0b013e3181d61eeb.

Maya L. Petersen, Kristin E. Porter, Susan Gruber, Yue Wang, and Mark J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21:31–54, 2012. doi: 10.1177/0962280210386207.

Jenna M. Reps, Martijn J. Schuemie, Marc A. Suchard, Patrick B. Ryan, and Peter R. Rijnbeek. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association*, 25:969–975, 2018. doi: 10.1093/jamia/ocy032.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994. doi: 10.1080/01621459.1994.10476818.

James M. Robins, Miguel A. Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560, 2000. doi: 10.1097/00001648-200009000-00011.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983. doi: 10.1093/biomet/70.1.41.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114:223–234, 2019. doi: 10.1080/01621459.2017.1395341.

Martijn J. Schuemie, Patrick B. Ryan, William DuMouchel, Marc A. Suchard, and David Madigan. Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Statistics in Medicine*, 33(2):209–218, 2014. doi: 10.1002/sim.5925.

Martijn J. Schuemie, M. Soledad Cepeda, Marc A. Suchard, Jianxiao Yang, Yuxi Tian, Alejandro Schuler, Patrick B. Ryan, David Madigan, and George Hripcsak. How confident are we about observational findings in health care: A benchmark study. *Harvard Data Science Review*, 2(1), 2020. doi: 10.1162/99608f92.147cc28e.

Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25:1–21, 2010. doi: 10.1214/09-STS313.

Marc A. Suchard, Martijn J. Schuemie, Harlan M. Krumholz, Seng Chan You, RuiJun Chen, Nicole Pratt, Christian G. Reich, Jon Duke, David Madigan, George Hripcsak, and Patrick B. Ryan. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: A systematic, multinational, large-scale analysis. *The Lancet*, 394:1816–1826, 2019. doi: 10.1016/S0140-6736(19)32317-7.

Samy Suissa. Immortal time bias in pharmacoepidemiology. *American Journal of Epidemiology*, 167:492–499, 2008. doi: 10.1093/aje/kwm324.

Erica A. Voss, Rupa Makadia, Amy Matcho, Qianli Ma, Chris Knoll, Martijn Schuemie, Frank J. DeFalco, Ajit Londhe, Vivienne Zhu, and Patrick B. Ryan. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association*, 22(3):553–564, 2015. doi: 10.1093/jamia/ocu023.

Shirley V. Wang, Sebastian Schneeweiss, RCT-DUPLICATE Initiative, Jessica M. Franklin, Rishi J. Desai, William Feldman, Elizabeth M. Garry, Robert J. Glynn, Kueiyu Joshua Lin, Julie Paik, Elisabetta Patorno, Samy Suissa, Elvira D'Andrea, Dureshahwar Jawaid, Hemin Lee, Ajinkya Pawar, Sushama Kattinakere Sreedhara, Helen Tesfaye, Lily G. Bessette, Luke Zabotka, Su Been Lee, Nileesa Gautam, Cassie York, Heidi Zakoul, John Concato, David Martin, Dianne Paraoan, and Kenneth Quinto. Emulation of randomized clinical trials with nonrandomized database analyses: Results of 32 clinical trials. *JAMA*, 329(16):1376–1385, 2023. doi: 10.1001/jama.2023.4221.

Kabir Yadav and Roger J. Lewis. Immortal time bias in observational studies. *JAMA*, 325(7):686–687, 2021. doi: 10.1001/jama.2020.9151.

A COMPLETE AUTHORIZATION AND ESTIMATION PROTOCOL

A.1 ACTION PRECEDENCE

Let T indicate a time-zero hard failure, O an overlap hard failure, n_{NC} the number of Bonferroni-significant negative controls, and r_C the range of censoring-tilt log-HR estimates. The first matching row in Table 4 determines the action. This precedence makes $\pi(d)$ deterministic for every diagnostic combination.

Table 4: Complete TrialVeto precedence map. Soft flags emit an estimate with required manual review; abstentions emit reason codes and any estimand-drift note without an effect claim.

Priority	Condition	Action	Required trace
1	$T = 1$ or $O = 1$	Abstain	All active hard-fail and warning codes; drift note if triggered
2	$r_C > 0.30$	Abstain	Censoring code and active NC codes
3	$n_{\mathrm{NC}} \geq 2$	Abstain	Every failed negative control
4	$n_{\mathrm{NC}} = 1$ and $0.20 \leq r_C \leq 0.30$	Abstain	Joint NC and censoring codes
5	$n_{\mathrm{NC}} = 1$	Soft flag	Failed control; manual review
6	$0.20 \leq r_C \leq 0.30$	Soft flag	Censoring range; manual review
7	Otherwise	Report	Passing diagnostic summary

Time-zero candidate tolerances are $\{0, 3, 7\}$ days. Propensity trimming candidates are $[0.02, 0.98]$, $[0.05, 0.95]$, and $[0.10, 0.90]$; maximum-SMD candidates are 0.05 and 0.10. Negative-control branch candidates correspond to one or two multiplicity-adjusted failures, and ordered censoring-boundary pairs are drawn from 0.20, 0.30, and 0.40. The resulting 108 candidate tuples are evaluated by the prespecified coverage and error objective. The locked tuple uses a three-day tolerance, $[0.05, 0.95]$ trimming, maximum SMD 0.10, the precedence in Table 4, and censoring boundaries 0.20/0.30.

A.2 NATIVE ORDERING AND RISK–COVERAGE INTEGRATION

TrialVeto's native safest-to-riskiest ordering is the ascending lexicographic tuple

$$s_{\mathrm{TV}}(u) = \big(r_a,\ T + O,\ n_{\mathrm{NC}},\ r_C,\ \max \mathrm{SMD},\ 1 - r_{\mathrm{retain}},\ g_0,\ \mathrm{id}_u\big),$$

where $r_a \in \{0, 1, 2\}$ denotes report, soft flag, or abstain; r_{retain} is the post-trim retained fraction; g_0 is the treated grace-period violation fraction; and id_u is the fixed anonymized integer unit index used only to break exact ties. This tuple places the 418 reportable units before the 192 abstentions at the locked operating point while preserving a complete ordering within each action. The learned selector orders units by the median predicted absolute error across its five seeds, with the same unit-index tie rule.

For either native ordering, coverage knots are $c_j = 0.05j$ for $j = 1, \ldots, 20$. At knot c_j, risk is median absolute anchor error in the safest prefix of $\lceil c_j N \rceil$ units. We report normalized trapezoidal integration over $[0.05, 1]$:

$$\mathrm{AURCC} = \frac{0.05}{0.95} \sum_{j=1}^{19} \frac{R(c_j) + R(c_{j+1})}{2}.$$

Thus every policy brings its own frozen ordering, while knots, prefix construction, risk definition, integration range, and tie handling remain identical.

A.3 SURVIVAL-AIPW CONSTRUCTION

For patient j in analysis unit u, let $A_j \in \{0,1\}$ denote treatment at time zero, X_j baseline covariates, $\widetilde{T}_j = \min(T_j, C_j)$ observed time, and $\Delta_j = \mathbb{1}[T_j \le C_j]$. Out-of-fold propensity predictions define the retained overlap cohort

$$\mathcal{O}_u = \{j : \ 0.05 \le \widehat{e}_1(X_j) \le 0.95\}, \qquad n_u^\star = |\mathcal{O}_u|.$$

The target is the treatment coefficient β_u in the marginal structural Cox projection for the empirical covariate distribution of $\mathcal{O}_u$,

$$\lambda_u(t \mid A) = \lambda_{0u}(t) \exp(\beta_u A). \tag{3}$$

Thus, every policy targets the same post-overlap population within a unit; removing more than 5% of treated patients changes that population enough to require the declared estimand-drift trace. Identification uses consistency, conditional exchangeability, positivity on $\mathcal{O}_u$, and conditionally independent censoring (Hernán et al., 2022; Rosenbaum & Rubin, 1983; Stuart, 2010; Robins et al., 1994; Bang & Robins, 2005; Robins et al., 2000; Hernán et al., 2000).

Patients are assigned to five folds. Training observations in each fold estimate $e_a(X) = P(A = a \mid X)$ with XGBoost, conditional event survival $S_a(t \mid X)$ with a treatment-specific penalized Cox model, and censoring survival $G_a(t \mid X)$ with a treatment-specific Cox model. Predictions for patient j always come from a model that excluded j. Let $\overline{G}_a(t)$ be the arm-specific marginal censoring survival and define the stabilized censoring ratio and its calibration-locked clipping by

$$w_{ja}^C(t) = \frac{\overline{G}_a(t)}{\widehat{G}_a(t \mid X_j)}, \qquad \widetilde{w}_{ja}^C(t) = \min\{q_{.99}, \max(q_{.01}, w_{ja}^C(t))\},$$

where $q_{.01}$ and $q_{.99}$ are the 1st and 99th percentile cutoffs fixed on calibration units. On the union of observed event times $\mathcal{T}_u$, the augmented survival estimate is

$$\widehat{S}_a(t) = \frac{1}{n_u^\star} \sum_{j \in \mathcal{O}_u} \left[\widehat{S}_a(t \mid X_j) + \frac{\mathbb{1}(A_j = a)}{\widehat{e}_a(X_j)} \left\{ \frac{\widetilde{w}_{ja}^C(t) \mathbb{1}(\widetilde{T}_j > t)}{\overline{G}_a(t)} - \widehat{S}_a(t \mid X_j) \right\} \right]. \tag{4}$$

Without clipping, $\widetilde{w}_{ja}^C(t)/\overline{G}_a(t) = 1/\widehat{G}_a(t \mid X_j)$; Equation 4 therefore states exactly where stabilization and clipping alter the update. Propensity denominators are the out-of-fold values that define $\mathcal{O}_u$, with no policy-specific re-estimation. Standardization averages over that retained cohort. Event definitions follow the mapped RCT endpoint; Cox nuisance fits use Efron's tie correction on $\mathcal{T}_u$.

Let $d\widehat{F}_a(t_k) = \widehat{S}_a(t_{k-1}) - \widehat{S}_a(t_k)$ and $\widehat{Y}_a(t_k) = \widehat{S}_a(t_{k-1})$. The marginal Cox projection solves

$$0 = U_u(\beta) = \sum_{t_k \in \mathcal{T}_u} \sum_{a=0}^{1} \left\{ a - \frac{\widehat{Y}_1(t_k) e^\beta}{\widehat{Y}_0(t_k) + \widehat{Y}_1(t_k) e^\beta} \right\} d\widehat{F}_a(t_k). \tag{5}$$

This projection gives the same treatment-versus-comparator orientation and follow-up horizon as the accepted RCT HR. Let $\phi_{ja}(t)$ be patient j's centered bracketed summand in Equation 4, including both the outcome-regression term and the clipped treatment/censoring-weight update. The unit-level score contribution $U_{u,j}(\beta)$ is the directional derivative of Equation 5 with respect to $(\widehat{S}_0, \widehat{S}_1)$ in direction (ϕ_{j0}, ϕ_{j1}). The propagated cross-fitted contribution is

$$\psi_j = -\left\{ \frac{\partial U_u(\widehat{\beta}_u)}{\partial \beta} \right\}^{-1} U_{u,j}(\widehat{\beta}_u),$$

and the within-unit sandwich variance is $(n_u^\star)^{-2} \sum_{j \in \mathcal{O}_u} (\psi_j - \bar{\psi})^2$. A nuisance fit that does not converge, a non-finite $\widetilde{w}_{ja}^C(t)$, a collapsed event stratum, or non-positive projection information invokes estimator-failure abstention before TrialVeto's diagnostic precedence.

Censoring perturbation. Write $\widehat{H}_{C,a}(t \mid X) = -\log \widehat{G}_a(t \mid X)$. The sensitivity path increases the fitted censoring hazard on a log-multiplicative scale,

$$\widehat{G}_{a,\gamma}(t \mid X) = \exp\{-e^\gamma \widehat{H}_{C,a}(t \mid X)\}, \qquad \gamma \in \{0, 0.1, 0.2, 0.5, 1.0\},$$

then recomputes Equation 4 and the Cox projection with identical stabilization and truncation. The diagnostic $r_C = \max_\gamma \widehat{\beta}_u(\gamma) - \min_\gamma \widehat{\beta}_u(\gamma)$ is therefore an absolute log-HR sensitivity range; 0.20 and 0.30 correspond to multiplicative HR spreads of 1.22 and 1.35.

Empirical calibration and failures. For calibration negative control q with estimate $\widehat{\theta}_q$ and standard error s_q, each database–area stratum fits

$$\widehat{\theta}_q \mid \mu, \tau^2 \sim \mathcal{N}(\mu, s_q^2 + \tau^2)$$

by maximum likelihood. The OHDSI-calibrated output is $\widehat{\beta}_u - \widehat{\mu}$ with variance $\widehat{\mathrm{Var}}(\widehat{\beta}_u) + \widehat{\tau}^2$. A nuisance fit that does not converge, produces a non-finite prediction, collapses an event stratum, or yields non-positive information cannot emit an effect claim. TrialVeto, every ablation, both full-report baselines, and the learned selector share the folds, estimator, stabilization, truncation, and failure handling.

B Benchmark construction and workload

B.1 Curation and split ledger

The review screened 447 active-comparator time-to-event trial pairings and retained 218 after excluding 129 with non-ascertainable endpoints, 63 with non-viable observational emulations, and 37 with unresolved protocol mappings. Endpoint mappings that depended on unavailable adjudicated composites, free text, or site-specific laboratory harmonization were excluded rather than replaced by narrower proxies. Included pairs cover cardiovascular ($n = 62$), oncology ($n = 48$), metabolic/endocrine ($n = 38$), neurology ($n = 28$), respiratory ($n = 22$), and infectious disease ($n = 20$).

Every retained pair is instantiated in all four databases, yielding $218 \times 4 = 872$ units. The primary test contains 610 units stratified jointly by therapeutic area and database. Pair-held-out splitting keeps all database instances of a pair together: 65 calibration pairs yield 260 units and 153 test pairs yield 612 units. Five grouped seeds preserve the same pair counts and area stratification. Leave-one-database-out evaluation produces four disjoint 218-unit tests and reports folds before pooling.

Alignment decisions. Curators first orient every published HR as active treatment versus active comparator, inverting the log-HR sign when the publication uses the opposite direction. Eligibility is mapped to observable pre-initiation history; assignment and follow-up begin at incident treatment initiation; the observational endpoint must implement the trial endpoint without a narrower surrogate; and administrative follow-up is capped at the trial horizon. The projected marginal log-HR in Equation 5 is compared only with a published HR over that matched endpoint and horizon. A candidate is excluded when any of these correspondences is unresolved. This schema separates protocol compatibility from the later diagnostic decision and explains why the 229 screened-out pairings never enter an authorization denominator.

The primary split is assigned at the pair–database-unit level after curation, the grouped split is assigned at the drug–outcome-pair level, and each database fold is defined before policy calibration. These three ledgers answer different leakage questions: ordinary held-out operation, transfer to unseen clinical questions, and transfer to an unseen data source. The corresponding counts and area composition identify every denominator used in the reported comparisons.

B.2 Database and therapeutic-area denominators

Database-level coverage ranges from 61.3% to 73.6%, while emitted median error ranges from 0.09 to 0.19. Tables 5 and 6 retain the raw site and therapeutic-area denominators so the pooled result can be read together with this heterogeneity.

Every therapeutic area preserves the intended ordering: emitted median error is lower than withheld median error, with gaps ranging from 0.21 to 0.27. The database table also shows that higher abstention in IQVIA and CPRD accompanies higher emitted error, motivating the explicit leave-one-database-out analysis rather than a pooled-only summary.

Table 5: Primary held-out results by database. N and abstained counts expose each site's observed weight; errors and standardized point-anchor coverage are conditional on emitted units. Abstention uses raw counts, while drift percentages use fixed six-area standardization.

Database	N / abstained	Median error	Point cov.	Abstention	Drift
MarketScan	163 / 43	0.09	95.6%	26.4%	8.3%
Optum	172 / 50	0.11	94.8%	29.1%	12.7%
IQVIA	151 / 51	0.14	92.7%	33.8%	17.1%
CPRD	124 / 48	0.19	92.3%	38.7%	22.4%
Pooled	610 / 192	0.12	94.1%	31.5%	14.7%

Table 6: Therapeutic-area denominators and authorization separation on the primary test. The six area counts sum to 610 and abstentions sum to 192.

Area	N / abstained	Emitted median	Withheld median
Cardiovascular	174 / 50	0.11	0.32
Oncology	134 / 44	0.13	0.36
Metabolic/endocrine	106 / 31	0.10	0.31
Neurology	78 / 25	0.15	0.40
Respiratory	62 / 22	0.14	0.39
Infectious disease	56 / 20	0.16	0.43

B.3 Cohort scale and precision

The decision rule routes smaller, lower-event, and lower-overlap cohorts toward abstention, but sample size alone does not explain the policy comparison. Table 7 describes the two action regions, and Table 8 compares uncertainty and effective sample size at matched reporting budgets.

Table 7: Cohort and precision distributions across all 610 primary test units, reported as median [IQR]. CI width is the observational log-HR interval width; retained is the post-trim fraction.

Quantity	Emitted ($n = 418$)	Withheld ($n = 192$)
Total cohort	24,800 [9,600, 71,300]	11,700 [4,300, 34,900]
Treated / comparator	10,900 [4,100,31,700] / 13,600 [5,300,39,800]	4,900 [1,800,15,200] / 6,500 [2,400,19,000]
Events	1,940 [710,5,860]	780 [260,2,390]
Retained sample	92.6% [87.1,96.8]	68.4% [52.1,79.6]
Effective sample size	3,840 [1,920,8,770]	1,260 [620,3,080]
95% CI width	0.41 [0.29,0.61]	0.77 [0.50,1.12]

At 418 outputs, TrialVeto has narrower median intervals and higher effective sample size than the learned selector. The selector nevertheless observes both quantities directly, so its higher anchor error cannot be attributed to unequal access to these precision signals.

Table 8: Precision at native and matched reporting coverage, reported as median [IQR]; pair-bootstrap 95% interval for the median. TrialVeto's advantage does not arise from vacuous interval widening, and the learned selector has access to effective sample size.

Policy / operating point	Emitted	95% CI width	Effective sample size
TrialVeto native/matched	418	0.41 [0.29,0.61]; [0.38,0.45]	3,840 [1,920,8,770]; [3,490,4,210]
Learned selector matched	418	0.46 [0.32,0.68]; [0.42,0.50]	3,210 [1,580,7,310]; [2,880,3,570]
Learned selector native	431	0.47 [0.33,0.70]; [0.43,0.51]	3,100 [1,490,7,080]; [2,760,3,460]
OHDSI-calibrated full report	610	0.48 [0.34,0.72]; [0.44,0.52]	2,890 [1,250,6,500]; [2,620,3,190]
Standard TTE full report	610	0.45 [0.31,0.67]; [0.41,0.49]	2,890 [1,250,6,500]; [2,620,3,190]

C Expanded evidence

C.1 Three-state authorization

The binary primary endpoint counts soft flags as emitted because they expose an estimate. Separating the three actions tests the additional semantic claim that a warning-bearing output occupies an intermediate-risk region and carries a distinct review obligation.

Table 9: Action-specific evidence on the primary test. Error intervals are pair-bootstrap 95% intervals. Every emitted soft flag carries a manual-review obligation; abstentions expose no treatment-effect claim.

Action	n	Median/P90 error			Sign	Severe	Workflow obligation
Report	334	0.10/0.26 [0.20,0.33]	[0.08,0.13]	/	14/334	6/334	No warning; 30/334 drift notes; effect claim
Soft flag	84	0.20/0.47 [0.36,0.60]	[0.15,0.26]	/	10/84	7/84	NC 32/84; censoring 52/84; manual review
Abstain	192	0.34/0.71 [0.59,0.85]	[0.28,0.41]	/	87/192	75/192	Trigger counts 112/79/54/36 for overlap / NC / censoring / time zero; 67 multi-trigger; no claim

C.2 Native selector and transfer detail

Table 10 reports both each policy's native operating point and the exact-budget comparison. Tables 11–13 then repeat the policy comparison when either clinical questions or databases are absent from calibration.

Table 10: Native and matched operating points. Error and gap cells report median [95% interval] / P90 [95% interval]. Selector intervals combine 1,000 pair-bootstrap replicates with variation across five grouped-training seeds. The final column gives AURCC and raw emitted sign/severe disagreements.

Policy	Emitted / withheld	Emitted error	Withheld error	Authorization gap	AURCC; sign/severe
TrialVeto native	418 / 192	0.12 [0.09,0.15] / 0.31 [0.24,0.39]	0.34 [0.28,0.41] / 0.71 [0.59,0.85]	0.22 [0.16,0.29] / 0.40 [0.29,0.53]	0.118; 24/418; 13/418
Selector native	431 / 179	0.149 [0.120,0.181] / 0.390 [0.319,0.475]	0.32 [0.26,0.39] / 0.68 [0.56,0.81]	0.171 [0.11,0.23] / 0.290 [0.19,0.41]	0.151 [0.132,0.172]; 37/431; 23/431
Selector matched	418 / 192	0.145 [0.116,0.178] / 0.382 [0.312,0.466]	0.31 [0.25,0.38] / 0.66 [0.54,0.79]	0.165 [0.10,0.23] / 0.278 [0.18,0.39]	0.151; 34/418; 21/418

The selector improves over full reporting on unseen pairs, yet TrialVeto retains lower emitted median and P90 error at the same 409-unit budget. Overlap-only and NC-only policies cover more units but leave substantially more error in their emitted sets, matching the primary ablation pattern.

Table 11: Pair-held-out comparison on 612 units. Error and gap cells report median [95% interval] / P90 [95% interval]. Partial gates use native actions; the learned selector uses TrialVeto's 409-unit budget.

Policy	Emitted / withheld (coverage)	Emitted error	Withheld error	Authorization gap
TrialVeto	409 / 203 (66.8%)	0.13 [0.10,0.17] / 0.35 [0.28,0.44]	0.37 [0.30,0.45] / 0.78 [0.65,0.92]	0.24 [0.17,0.32] / 0.43 [0.31,0.57]
Learned selector	409 / 203 (66.8%)	0.15 [0.12,0.19] / 0.40 [0.32,0.50]	0.32 [0.25,0.40] / 0.69 [0.57,0.83]	0.17 [0.10,0.25] / 0.29 [0.18,0.42]
Overlap-only	532 / 80 (86.9%)	0.18 [0.14,0.23] / 0.51 [0.41,0.63]	0.33 [0.24,0.43] / 0.73 [0.58,0.89]	0.15 [0.06,0.25] / 0.22 [0.09,0.37]
NC-only	526 / 86 (85.9%)	0.18 [0.14,0.22] / 0.48 [0.38,0.59]	0.34 [0.26,0.43] / 0.72 [0.59,0.87]	0.16 [0.08,0.25] / 0.24 [0.12,0.38]
OHDSI-calibrated	612 / 0 (100%)	0.17 [0.13,0.21] / 0.47 [0.38,0.58]	–	–
Standard TTE	612 / 0 (100%)	0.21 [0.17,0.26] / 0.63 [0.52,0.76]	–	–

On unseen pairs, the full rule's 0.24/0.43 median/P90 gaps exceed the learned selector's 0.17/0.29 gaps at identical coverage. Both partial policies emit more analyses, but their emitted median and tail errors are higher than TrialVeto's.

Table 12: Leave-one-database-out TrialVeto results. Each fold contains all 218 units from the named database; pooling occurs only after fold-level evaluation.

Held-out database	Emitted	Emitted median/P90	Withheld median/P90	Median gap
MarketScan	154/218	0.11/0.31	0.33/0.70	0.22 [0.14,0.31]
Optum	148/218	0.13/0.35	0.37/0.76	0.24 [0.15,0.34]
IQVIA	140/218	0.16/0.42	0.42/0.87	0.26 [0.17,0.37]
CPRD	126/218	0.19/0.51	0.49/0.98	0.30 [0.19,0.42]
Pooled	568/872	0.15/0.40	0.40/0.84	0.25 [0.18,0.33]

Each database fold preserves a positive median gap, increasing from 0.22 for MarketScan to 0.30 for CPRD. This foldwise result rules out a pooled gap created by one dominant source.

Table 13: Pooled leave-one-database-out policy comparison after all four folds are scored separately. The learned selector is matched to TrialVeto's reporting budget. Error and gap cells report median [95% interval] / P90 [95% interval].

Policy	Emitted / withheld (coverage)	Emitted error	Withheld error	Authorization gap
TrialVeto	568 / 304 (65.1%)	0.15 [0.12,0.18] / 0.40 [0.33,0.48]	0.40 [0.34,0.47] / 0.84 [0.71,0.99]	0.25 [0.18,0.33] / 0.44 [0.32,0.59]
Learned selector	568 / 304 (65.1%)	0.18 [0.15,0.22] / 0.45 [0.37,0.54]	0.36 [0.30,0.43] / 0.77 [0.65,0.91]	0.18 [0.11,0.26] / 0.32 [0.21,0.45]
Overlap-only	742 / 130 (85.1%)	0.20 [0.17,0.24] / 0.57 [0.47,0.69]	0.38 [0.31,0.46] / 0.82 [0.69,0.97]	0.18 [0.10,0.27] / 0.25 [0.13,0.40]
NC-only	735 / 137 (84.3%)	0.19 [0.16,0.23] / 0.52 [0.43,0.63]	0.37 [0.30,0.45] / 0.79 [0.66,0.94]	0.18 [0.10,0.27] / 0.27 [0.15,0.41]
OHDSI-calibrated	872 / 0 (100%)	0.20 [0.17,0.24] / 0.54 [0.45,0.65]	–	–
Standard TTE	872 / 0 (100%)	0.24 [0.20,0.29] / 0.69 [0.58,0.82]	–	–

The pooled equal-budget comparison favors TrialVeto over the learned selector for emitted error and authorization separation. Raw disagreement counts in Table 14 give the corresponding decision-facing view without standardization.

Table 14: Raw sign/severe disagreements among emitted analyses under both transfer regimes. Every numerator is paired with its complete policy-specific emitted denominator.

Policy	Pair-held-out sign / severe	Leave-one-database-out sign / severe
TrialVeto	28/409; 14/409	52/568; 31/568
Learned selector	36/409; 22/409	68/568; 43/568
Overlap-only	76/532; 57/532	119/742; 94/742
NC-only	65/526; 47/526	106/735; 78/735
OHDSI-calibrated	74/612; 52/612	126/872; 94/872
Standard TTE	125/612; 96/612	183/872; 141/872

Table 15: Five-seed sensitivity for pair-grouped splitting. Every seed assigns 65 pairs to calibration and 153 unseen pairs to test.

Seed	Reporting coverage	Median authorization gap	P90 authorization gap
1	66.8%	0.24	0.43
2	65.9%	0.21	0.39
3	67.5%	0.26	0.47
4	68.1%	0.23	0.41
5	66.3%	0.25	0.45

Across the five pair-grouped seeds, reporting coverage varies by 2.2 percentage points, median gaps remain between 0.21 and 0.26, and P90 gaps remain between 0.39 and 0.47. Transfer separation is therefore stable to the particular calibration-pair assignment.

C.3 Integrated-figure coordinates

Figure 1A reports withheld-minus-emitted gaps at the 10th, 25th, 50th, 75th, and 90th error percentiles: 0.05/0.12/0.22/0.32/0.40, with pair-bootstrap 95% intervals [0.02,0.09]/[0.07,0.17]/[0.16,0.29]/[0.24,0.41]/[0.29,0.53]. The underlying emitted profile is 0.03/0.06/0.12/0.20/0.31, while the withheld profile is 0.08/0.18/0.34/0.52/0.71. Panel B evaluates native median risk at reporting coverages 40/50/60/68.5/80/90%. TrialVeto yields 0.082/0.094/0.108/0.120/0.137/0.160, compared with 0.104/0.119/0.133/0.145/0.163/0.181 for the learned selector. Panel E evaluates all six policies in Table 16; its pointwise bands and break-even intervals use the same 1,000 pair-bootstrap replicates. Panels C, D, and F are generated directly from Tables 3 and 18 and the backed trigger frequencies.

C.4 Utility and independent evidence

The utility analysis retains event numerators and withholding denominators for every policy. Table 16 identifies the break-even cost directly, while Table 17 shows pair-bootstrap uncertainty at both endpoints of the cost range.

Table 16: Raw event accounting for the utility sweep. Break-even λ is the largest withholding cost for which TrialVeto has lower loss than the comparator; intervals are pair-bootstrap 95% intervals.

Policy	Severe emitted	Sign emitted	Withheld	Severe break-even
TrialVeto	13	24	192	Reference
Learned selector	21	34	192	Dominated at equal coverage
NC-only	49	62	87	0.34 [0.25,0.44]
Overlap-only	62	76	78	0.43 [0.34,0.53]
OHDSI-calibrated	46	66	0	0.17 [0.11,0.24]
Standard TTE	88	111	0	0.39 [0.31,0.48]

Table 17: Pointwise pair-bootstrap 95% bands for the severe-error utility curves at the endpoints of the plotted withholding-cost range.

Policy	$\lambda = 0$	$\lambda = 1$
TrialVeto	[0.005,0.039]	[0.274,0.400]
Learned selector	[0.020,0.052]	[0.291,0.411]
NC-only	[0.059,0.103]	[0.187,0.262]
Overlap-only	[0.075,0.132]	[0.190,0.271]
OHDSI-calibrated	[0.058,0.097]	[0.058,0.097]
Standard TTE	[0.106,0.191]	[0.106,0.191]

At equal coverage, TrialVeto dominates the learned selector for every λ because it emits fewer severe errors while withholding the same number of analyses. Against full-report baselines, the crossing points quantify exactly how much relative cost of silence the observed error reduction can absorb.

Repeating Equation 2 with sign disagreement gives break-even costs of 0.22 [0.15,0.30] against OHDSI-calibrated and 0.45 [0.37,0.55] against Standard TTE. On the pair-held-out ledger, the severe-error numerators in Table 14 give break-even costs of 0.19 against OHDSI-calibrated and 0.40 against Standard TTE.

The adjudication sample was stratified before review to include 90 abstentions, 60 reports, and 30 soft flags, balanced across four databases, six therapeutic areas, and the four trigger families. Two clinical epidemiologists independently reviewed target-trial specifications and diagnostic outputs without anchor error or TrialVeto action. A unit was non-reportable when the available specification and diagnostics did not support an unqualified treatment-effect claim because of time-zero misalignment, inadequate overlap or balance, falsifying negative controls, or material censoring instability. Reviewers reconciled disagreements by recorded consensus. The independent negative-control sensitivity used the same blinded consensus procedure for panel eligibility. Sixty sampled pairs contributed 168 pair–database units for which both control panels were complete; agreement and action changes use this raw denominator. Relative to the original panel, independently curated controls changed 19 of 168 actions: 12 were relaxed and 7 were tightened; NC-driven abstentions changed from 68 to 61 and NC-driven soft flags from 33 to 30.

Table 18: Blinded action-semantic and negative-control-panel sensitivity checks. Adjudicators were blinded to RCT anchor error and policy action; the alternative negative-control panel was curated independently while blinded to anchors and original actions.

Check	Report	Soft flag	Abstain
Adjudicated non-reportable	9/60	11/30	70/90
Reviewer agreement	$\kappa = 0.79[0.70, 0.87]$		
Abstain vs. report association	Raw OR 19.8 [8.0,49.2]; area/database-adjusted OR 10.4 [4.6,24.7]		
Independent NC sensitivity	149/168 actions agree; $\kappa = 0.78$; emitted/withheld median 0.13 [0.10,0.17]/0.35 [0.28,0.44]		

The independently curated panel changes only 19 of 168 paired actions and preserves a withheld-minus-emitted median gap of 0.22 (0.35 versus 0.13). Combined with the adjudicated reportability gradient, this shows that the primary partition is not solely a consequence of the original negative-control panel.

C.5 ANCHOR UNCERTAINTY AND GUARDRAIL COMPLEMENTARITY

For trial pair p with published HR h_p and 95% interval $[L_p, U_p]$, each sensitivity replicate draws one log-HR anchor

$$\beta_p^{\mathrm{RCT},b} \sim \mathcal{N}\left(\log h_p, \left\{\frac{\log U_p - \log L_p}{2 \times 1.96}\right\}^2\right).$$

The same pair-level draw is shared by all database units for that pair, while estimates, actions, and reporting budgets remain fixed. Table 19 recomputes emitted error and point agreement under 1,000 such draws; TrialVeto has the lowest emitted error in 91.0% of replicates.

Table 19: Sampled-anchor sensitivity from 1,000 draws under each published RCT interval. Error cells report median/P90 with pair-bootstrap intervals; point agreement is the fraction of observational intervals covering the sampled anchor.

Policy	Median/P90 error	Point agreement	Best rank
TrialVeto	0.14/0.36 [0.11,0.17]/[0.29,0.45]	93.6%	91.0%
Learned selector	0.16/0.41 [0.13,0.20]/[0.33,0.51]	91.7%	7.2%
OHDSI-calibrated	0.17/0.47 [0.14,0.21]/[0.38,0.58]	89.9%	1.8%
Standard TTE	0.20/0.61 [0.16,0.25]/[0.50,0.74]	81.1%	0.0%

Sampling the published RCT uncertainty increases error for every policy but leaves TrialVeto first in 91.0% of draws. The ablations in Table 20 address a different concern: every removed diagnostic family raises emitted central or tail error, with the largest increase after removing time-zero validation.

Table 20: Guardrail ablations on the primary test. Error and point coverage condition on each configuration's emitted subset; abstention percentages use the prespecified six-area standardization over all 610 test units, whereas trigger totals are raw counts. Error cells show median/P90 with pair-bootstrap 95% intervals. Removing a guardrail admits more analyses while retaining more error, supporting complementary enforcement.

Configuration	Median/P90 error	Point coverage	Abstention
Full TrialVeto	0.12 [0.09,0.15] / 0.31 [0.24,0.39]	94.1%	31.5%
− Time-zero	0.16 [0.12,0.20] / 0.44 [0.35,0.54]	89.3%	24.1%
− Overlap diagnostic	0.14 [0.10,0.18] / 0.39 [0.30,0.49]	90.7%	22.6%
− Negative controls	0.13 [0.10,0.17] / 0.34 [0.26,0.43]	93.2%	18.9%
− Censoring sensitivity	0.13 [0.09,0.16] / 0.33 [0.25,0.42]	92.4%	25.7%
No guardrails	0.19 [0.15,0.24] / 0.58 [0.47,0.71]	82.3%	0.0%

Without guardrails, the emitted set reverts to all 610 units and exactly reproduces Standard TTE's 0.19/0.58 error and 82.3% point agreement. Each causal diagnostic therefore contributes measurable information beyond the shared estimator.

D COMPUTE AND IMPLEMENTATION DETAILS

All experiments used Ubuntu 22.04 LTS, Python 3.11.7, dual-socket AMD EPYC 7543 hosts, 64 vCPUs and 512 GB RAM per run, and NVIDIA A100 80 GB GPUs. The software stack comprised

CUDA 12.3, NVIDIA driver 550.54.15, cuDNN 9.1.0, XGBoost 2.0.3, `lifelines` 0.28.0, OHDSI ATLAS, and CohortDiagnostics 3.2.0. Median end-to-end runtime per pair–database analysis was 12.4 minutes and the 95th percentile was 38.7 minutes, including diagnostic evaluation, survival-AIPW estimation, and reason-code generation.

The protocol specifies the curation and aggregate split ledgers, calibration objective, threshold grid, action precedence, nuisance models, cross-fitting, weight handling, estimator-failure semantics, evaluation denominators, pair-level resampling, and reported endpoints. These definitions keep the treatment-effect estimand and selective-reporting comparison consistent across the three evaluation designs.