ZERO-SHOT EXTERNAL-CONTROL CONSTRUCTION FOR SINGLE-ARM CLINICAL TRIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Oncology single-arm clinical trials (SCTs) evaluate treatments without a control group. External controls constructed from real-world data (RWD) enable comparative evaluation, but require trial-specific elements governing patient inclusion, follow-up initiation, and historical borrowing. Existing methods largely leave these elements to manual specification. Additionally, the absence of a dedicated benchmark has restricted prior evaluations to individual trials or single cancer settings, leaving unclear whether the methods can be transferred to unseen trials. To address these gaps, we introduce SCT-Bench, the first benchmark for this setting, comprising 9,163 clinical trials organized into seven complementary tasks that evaluate protocol extraction, feature ascertainability, control-arm similarity, outcome recovery, survival-curve recovery, treatment-effect recovery, and drug-efficacy recovery. We formulate the SCT evaluation task as *zero-shot external-control construction*: selecting construction elements for an unseen SCT using only observable evidence. Withheld randomized control arms provide reference outcomes and populations for evaluating both trial-specific element determination and outcome recovery. Such zero-shot construction requires transferring prior methodological knowledge into executable trial-specific actions. Building on this idea, we propose Task-Adaptive External-Control Construction (TAEC), an LLM-based knowledge-to-action framework. Skill-guided planning converts such prior knowledge, together with the target protocol and RWD context, into ten structured construction elements. Data-grounded construction identifies candidate donors and assigns weights through partial optimal transport. Outcome-free refinement verifies the construction without target-trial outcomes, revises unsupported elements and their supporting skills, and updates the skill library within the current target trial. Across SCT-Bench evaluations with two RWD sources, TAEC consistently improves external-control construction and downstream outcome recovery over existing baselines. Further experiments show that transferring TAEC's selected time-zero choice into otherwise unchanged baseline pipelines yields large performance gains. These results identify upstream task instantiation as a major source of generalization beyond estimator choice.

1 INTRODUCTION

Oncology single-arm clinical trials (SCTs) substantially reduce trial size and resource requirements by enrolling only an investigational treatment arm without a control arm, accelerating drug development (Agrawal et al., 2023). However, comparative efficacy must instead be assessed against an external control constructed from real-world data (RWD) (Liu et al., 2023). Given the treatment arm and an RWD source, external-control construction requires several trial-specific elements to be specified, including the eligible population and matching procedure. These elements determine which RWD patients form the external-control cohort, and they depend on both the protocol and the source. The resulting cohort provides control outcomes, such as median overall survival and the survival curve, which are compared with the treatment arm to estimate treatment effects and assess drug efficacy. A highly related historical trial can inform these trial-specific elements (Pocock, 1976), but such a trial is often unavailable (Burcu et al., 2020). For a new SCT, these elements must therefore be determined for an unseen protocol without a directly reusable historical control (Hashmi et al., 2021).

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.

Existing methods mainly automate downstream patient matching, while upstream trial-specific elements, such as time origin, are still determined manually by experts for each new trial (Rippin et al., 2022; Signorovitch et al., 2012; Ren et al., 2023). We reframe this unresolved problem as zero-shot task-adaptive analysis planning for external-control construction: adapting a construction plan for an unseen trial. Although the plan must be task adaptive, its underlying elements, such as temporal alignment and feature selection, follow recurring methodological principles across studies. This leads to our central question: can methodological knowledge accumulated from prior trials and literature be transferred into an executable construction plan for an unseen trial?

Answering this question requires evaluating both upstream element determination and downstream outcomes. However, most existing methods are demonstrated on individual SCTs or cancer settings where the true control is unobserved (Rippin et al., 2022), and no public benchmark currently enables both stages to be evaluated across trials. To address this limitation, we introduce SCT-Bench, to our knowledge the first benchmark for zero-shot external-control construction from unseen trial protocols. SCT-Bench spans 9,163 clinical trials with heterogeneous patient- and trial-level information, pairing historical evidence for construction with withheld randomized control arms as evaluation references. Following FDA considerations and the external-control construction pipeline in Figure 1, SCT-Bench defines seven tasks spanning both upstream and downstream stages: Protocol Extraction to extract trial requirements, Feature Ascertainability to determine whether required features are available, Control-Arm Similarity to compare control populations, Outcome Recovery to recover the control-arm median outcome, Survival-Curve Recovery to compare survival trajectories, Treatment-Effect Recovery to recover treatment effects, and Drug-Efficacy Recovery to determine drug efficacy.

Building on this formulation, we propose **T**ask-**A**daptive **E**xternal-**C**ontrol Construction (TAEC), an LLM-based knowledge-to-action framework that automatically generates adaptive construction plans for unseen trials. Here, knowledge-to-action means transferring reusable construction knowledge from historical evidence and methodological literature into an executable construction plan. A skill library organizes this knowledge, which an LLM planner uses to adapt ten structured trial-specific elements to the target protocol. The resulting plan is executed with partial optimal transport, which allows unsupported patients to remain unmatched and preserves the distinction between missing and satisfied evidence. Because target-trial outcomes are unavailable during construction, an outcome-free verifier checks each element and revises unsupported elements before estimation.

On SCT-Bench, TAEC consistently outperforms existing baselines across two RWD sources, achieving the best performance on nearly all five downstream tasks from control-outcome recovery to drug-efficacy assessment. This broad advantage indicates that TAEC improves not only a single endpoint, but the overall quality of the constructed external control. To understand where these gains come from, we further transfer time-zero choice selected by TAEC into otherwise unchanged classical statistical pipelines. This transfer substantially reduces APE for nearly all outcome-free classical pipelines, even though patient selection, donor weights, and downstream estimators remain unchanged. These results show that a substantial portion of the performance gain comes from how the trial-specific construction elements are determined, rather than from replacing the downstream estimator. In other words, TAEC improves generalization by adapting the upstream construction plan itself. Both the model code and benchmark are publicly available at GitHub[1].

Our contributions are summarized as follows:

- We formulate external-control construction as *zero-shot task-adaptive analysis planning*. To our knowledge, we introduce SCT-Bench, the first benchmark for this setting, with seven tasks spanning upstream element determination and downstream trial outcomes.
- We introduce TAEC, an LLM-based knowledge-to-action framework that transfers reusable methodological knowledge into trial-adaptive construction elements and executes them through fixed statistical procedures for unseen trials.
- We show that upstream task instantiation is a major source of generalization: adapting trial-specific construction elements yields large performance gains even when downstream statistical estimators remain unchanged.

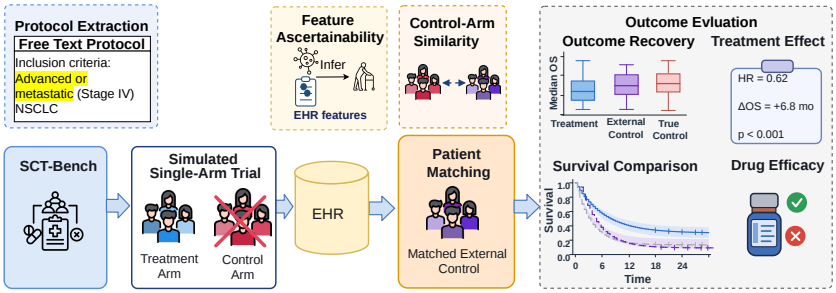


Figure 1: Overview of SCT-Bench and the external-control construction pipeline.

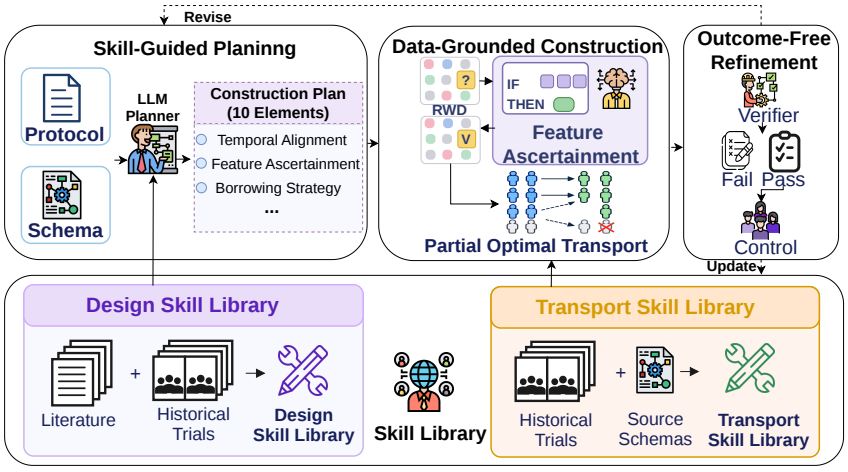


Figure 2: Architecture of TAEC. The left side shows how the LLM planner uses the protocol, schema, and retrieved design skills to construct trial-specific elements. In the middle, these elements are grounded in RWD through feature ascertainment and partial optimal transport using transport skills. On the right, an outcome-free verifier evaluates the constructed control and feeds failures back for iterative refinement.

2 METHODS

2.1 PROBLEM FORMULATION

We consider an unseen single-arm clinical trial with protocol P, observed treatment cohort $\mathcal{T}$, and access to real-world data $\mathcal{D}$. The randomized control arm $\mathcal{C}^*$ is withheld and used only for retrospective evaluation. Rather than applying a fixed external-control procedure across trials, we study zero-shot task-adaptive external-control construction: for each target trial, the framework constructs a trial-specific plan π, consisting of the elements needed to construct and analyze an external control, and executes it on $(P, \mathcal{T}, \mathcal{D})$:

$$(\widehat{\mathcal{C}}, \widehat{\theta}_C) = \text{Execute}(\pi; P, \mathcal{T}, \mathcal{D}), \tag{1}$$

where $\widehat{\mathcal{C}}$ denotes the constructed, weighted external-control cohort and $\widehat{\theta}_C$ denotes its estimated control outcome under the selected endpoint and follow-up definition.

The framework targets unseen trials and constructs a trial-specific analysis plan without target-outcome feedback. The key challenge is knowledge-to-action: transferring reusable construction knowledge to a new trial, grounding the selected choices in the available RWD, and revising unsupported decisions without access to the missing control. The plan may use reusable knowledge distilled before target-trial evaluation, but never $\mathcal{C}^*$ or its outcome.

[1] https://anonymous.4open.science/r/TAEC/

2.2 OVERVIEW

To address these challenges, we propose TAEC, a zero-shot task-adaptive framework that translates reusable external-control knowledge into executable trial-specific constructions. As illustrated in Figure 2, TAEC consists of three modules. Skill-Guided Planning retrieves reusable construction knowledge and composes a trial-specific construction plan. Data-Grounded Construction operationalizes the plan on the target RWD source and constructs the external-control cohort. Outcome-Free Refinement verifies the construction and iteratively revises unsupported decisions.

2.3 SKILL-GUIDED PLANNING

For an unseen trial, the first challenge is to determine which construction choices are appropriate without target-control feedback. The protocol and RWD provide trial-specific evidence, but many of these choices also depend on reusable external-control knowledge. We therefore translate such knowledge into a trial-specific construction plan. Trial-specific elements determine which real-world patients can serve as controls and how they are aligned and matched to the trial. TAEC represents ten construction elements (e.g., time zero and feature ascertainability), each as a selection over predefined construction operators. For each element, the system retrieves skills according to the target disease, enrollment setting, protocol characteristics, available time-zero evidence, and feature ascertainment in the RWD source. Each skill links these trial conditions to a ranked set of applicable operators (e.g., treatment-anchored versus diagnosis-anchored time zero), together with the evidence required to support them and conditions that indicate failure or abstention.

We maintain a reusable skill library $\mathcal{S}$, initialized before target-trial evaluation by distilling construction knowledge from methodological literature and historical trials. Let $\mathcal{R}_t \subseteq \mathcal{S}$ denote the skills retrieved for target trial t, with

$$\mathcal{R}_t = \mathrm{Retrieve}(P_t, \mathcal{D}, \mathcal{S}).$$

The LLM planner uses $\mathcal{R}_t$ together with the protocol and RWD context to select an operator for each element. During outcome-free refinement, the skill library may be updated for the current target trial; these updates are reset before the next evaluation trial.

$$\pi_t = F_\theta(P_t, \mathcal{D}, \mathcal{R}_t) = (e_t^1, \ldots, e_t^{10}), \tag{2}$$

where e_t^k denotes the selected operator for the k-th element. The resulting plan π_t specifies, for example, temporal alignment, control-source selection, feature ascertainment, borrowing strategy, and matching configuration. Because each element is instantiated from a predefined operator space, the plan can be directly executed on the RWD and subsequently checked by the verifier.

2.4 DATA-GROUNDED CONSTRUCTION

Given the plan π_t, the construction module applies population, control-source, treatment-history, and time-zero elements to the RWD $\mathcal{D}$, yielding the candidate donor pool $\mathcal{E}_t = \{i \in \mathcal{D} : g(i; \pi_t) = 1\}$, where g indicates whether a patient satisfies the requirements of these elements.

To determine each candidate's contribution to the external control, the module assesses how well donors match the target population on clinical features. This requires checking whether the features needed for comparison can be obtained from both sources. Feature ascertainment determines whether each feature is observed, inferable, or unavailable in each source. Schema alignment harmonizes the ascertainable features, and covariate alignment selects those used for matching, yielding

$$\mathbf{x}_{i,t} = \phi_t(r_i^{(s)}, s), \qquad s \in \{\mathrm{trial}, \mathrm{RWD}\}, \tag{3}$$

where ϕ_t maps source-specific records $r_i^{(s)}$ into the shared feature space.

Based on this representation, partial optimal transport assigns donor weights $w_{i,t} \geq 0$ according to the borrowing and matching elements in π_t, with zero weight for unmatched candidates. The resulting external control

$$\widehat{\mathcal{C}}_t = \{(i, w_{i,t}) : i \in \mathcal{E}_t,\ w_{i,t} > 0\} \tag{4}$$

is used for downstream outcome estimation.

2.5 OUTCOME-FREE REFINEMENT

The outcome-free verifier examines the plan π_t, cohort $\widehat{\mathcal{C}}_t$, retrieved skills $\mathcal{R}_t$, and execution diagnostics to identify elements that violate protocol requirements or lack evidence. It returns element-specific failure signals, such as unsupported time-zero choices, unavailable required features, or insufficient donor support, without using target-trial outcomes or downstream outcome estimates.

For each failed element, the signal specifies why the selected operator is unsupported under the current trial and RWD context. The planner uses this feedback to retrieve an alternative skill, adapt an existing one, or create a new skill when no applicable alternative exists. The corresponding construction element is then updated, while supported elements remain unchanged. Revised skills are written back to the persistent library $\mathcal{S}$ for reuse within the current trial; if no supported operator can be identified, the element abstains. The revised plan is then executed and verified again:

$$(\mathcal{R}_t^{(r)}, \pi_t^{(r)}) \to \widehat{\mathcal{C}}_t^{(r)} \to V_t^{(r)} \to (\mathcal{R}_t^{(r+1)}, \pi_t^{(r+1)}), \tag{5}$$

where r denotes the refinement round and $V_t^{(r)}$ denotes verifier feedback. Refinement stops when verification passes or the revision budget is exhausted.

New or revised skills can be incorporated into the skill library during refinement:

$$\mathcal{S}^{(r+1)} = \mathrm{Update}\big(\mathcal{S}^{(r)}, \Delta\mathcal{S}_r\big), \tag{6}$$

where r indexes refinement iterations within the current target trial, and $\Delta\mathcal{S}_r$ contains skill updates supported by outcome-free verification. These updates are used only within the current trial and are reset before the next evaluation trial.

3 SCT-BENCH

SCT-Bench contains 9,163 oncology clinical trials, including 41 with patient-level data: 11 randomized trials with individual time-to-event records for end-to-end evaluation and 30 data-sharing releases for feature ascertainability. The remaining 9,122 trials provide trial-level evidence from ClinicalTrials.gov and associated records, including protocols, eligibility criteria, treatment arms, baseline summaries, and reported outcomes. Dataset descriptions are detailed in section C.

Following FDA guidance, SCT-Bench evaluates external-control construction at both patient and trial levels. For randomized trials replayed as SCTs, the control arm and treatment-arm outcomes are withheld during construction and used only for evaluation. Task availability depends on the reference evidence available for each trial.

3.1 TASK DESIGN

3.1.1 UPSTREAM CAPABILITIES

Protocol Extraction. Given trial documents, the system extracts construction-relevant fields, including treatment setting, prior treatment lines, disease extent, and control regimen. Predictions are compared with reference fields from the primary results publication using exact matching for categorical and integer fields and normalized drug-set matching for control regimens. For treatment setting, we additionally report results on cases where independent reference signals agree.

Feature Ascertainability. Using 30 trials with released individual participant data, we mask an observed feature and assess whether it can be recovered from the remaining available features, using the held-out value as ground truth.

3.1.2 EXTERNAL-CONTROL RECOVERY

Control-Arm Similarity. Constructed controls are compared with true controls using absolute standardized mean differences (SMDs) in mean age and female proportion. References come from randomized control-arm records or registry-reported control-arm baseline summaries. We report mean absolute SMD and the proportions meeting prespecified similarity thresholds.

Outcome Recovery. This is the primary task in SCT-Bench, as it directly tests whether the constructed external control can recover the outcome of the withheld randomized control arm, which is

the central objective of external-control construction and the basis for downstream treatment-effect and efficacy estimation. The system estimates the control-arm median outcome, evaluated using absolute percentage error (APE) against the reported control median. We report the median APE across trials, supplemented by error-threshold and confidence-interval-based assessments.

Survival-Curve Recovery. This task is evaluated on trials with released patient-level data, where we assess recovery of the randomized-control survival curve using absolute error in 24-month restricted mean survival time (RMST), with survival-probability errors at 6, 12, and 24 months as additional measures.

Treatment-Effect Recovery. We compare the treatment effect obtained using the constructed external control with the corresponding reference estimate obtained using the randomized control arm.

Drug-Efficacy Recovery. We evaluate whether the treatment conclusion derived using the constructed external control agrees with that obtained using the randomized control arm. We assess conclusion-level agreement using chance-corrected three-class κ, treatment-effect ordering using Spearman's ρ, and estimate agreement against the randomized-reference confidence interval.

4 Experiments

4.1 Experimental Setup

We use MarketScan administrative claims (Merative, 2022) as the primary RWD source for external-control construction and MIMIC-IV (Johnson et al., 2024) as a complementary source with greater clinical granularity but substantially smaller scale and more limited longitudinal coverage. Outcome read-outs follow the endpoint available in each source; when mortality information is unavailable, follow-up duration is used as a source-specific event-time proxy and is not interpreted as overall survival. Data-source construction, available variables, endpoint definitions, and preprocessing are detailed in subsection D.1. The curation of SCT-Bench is detailed in section C.

4.1.1 Baseline

Throughout the LLM-based experiments, we evaluate Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B (Qwen Team, 2026a), Qwen3.8-27B (Qwen Team, 2026b), and GPT-5.5 (OpenAI, 2026) as backbones, covering multiple model scales and generations across both open-weight and proprietary model families. Unless otherwise specified, TAEC uses Qwen3.5-9B as the default backbone because it provides a favorable performance–efficiency trade-off, with performance broadly comparable to larger backbones across our evaluations.

For Protocol Extraction, we compare direct prompting across multiple LLM backbones. For Feature Ascertainability, we compare alternative inference interfaces under the same LLM backbones: our rule-based LLM interface (TAEC), zero-shot decision-tree induction (Knauer et al., 2024), and TabLLM-style direct prediction (Hegselmann et al., 2023). TAEC generates reusable feature-ascertainment rules without fitting to target-trial labels. For external-control construction, we compare TAEC with an unweighted donor pool, TSC, and classical cohort-construction methods, including *weighting-based methods*—IPTW (Robins et al., 2000), MAIC (Signorovitch et al., 2010), and entropy balancing (Hainmueller, 2012); *matching*—propensity-score matching (PSM) (Rosenbaum & Rubin, 1983); and *transport-based methods*—Sinkhorn and partial optimal transport (OT) (Cuturi, 2013; Chapel et al., 2020).

4.2 Main Results

Protocol Extraction. As shown in Table 1, within Qwen3.5, scaling most clearly improves treatment-setting extraction, likely because setting is often stated through direct trial-level cues. Prior lines and disease extent rely more on dispersed eligibility evidence and show less consistent scaling. Comparing Qwen3.5-27B with Qwen3.8-27B further suggests that newer backbones shift performance across field types rather than uniformly improving extraction.

Feature Ascertainability. Performance depends jointly on the backbone and inference interface (Table 2). Our rule-based interface is generally stronger than zero-shot trees, particularly for smaller Qwen backbones, while TabLLM becomes competitive or better for continuous features with larger

Table 1: Protocol Extraction performance across different LLM backbones and protocol-field types, measured by accuracy (%). Setting (Agreement) evaluates the subset with consistent reference annotations. Higher is better.

Backbone / Method	Setting ↑	Setting (Agreement) ↑	Prior lines ↑	Extent ↑
Qwen3.5-4B	47.5	63.1	57.9	56.2
Qwen3.5-9B	54.9	76.9	62.8	**71.4**
Qwen3.5-27B	**59.6**	**82.9**	59.3	65.4
Qwen3.8-27B	57.7	76.9	**66.4**	68.5
GPT-5.5	55.1	76.6	62.2	69.1

Table 2: Zero-shot Feature Ascertainability across LLM backbones and inference interfaces. TAEC uses the LLM directly to generate and apply feature-ascertainment rules. Binary features are evaluated using AUC, while continuous and ordinal features use Spearman's ρ; higher is better for all metrics. No target-trial labels are used for fitting or demonstrations.

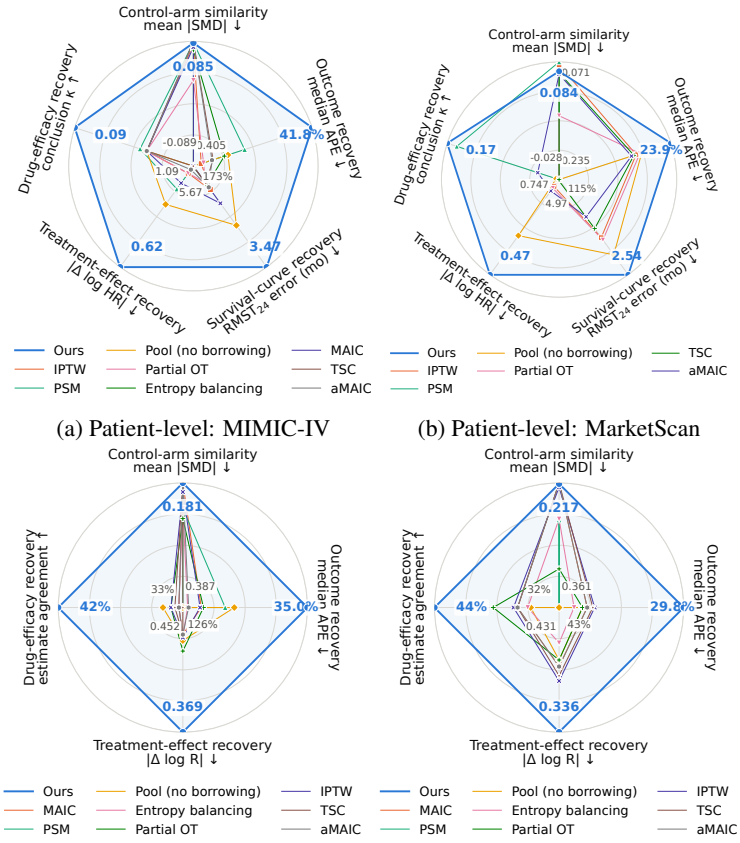
Backbone	Inference interface	Binary AUC ↑	Continuous ρ ↑	Ordinal ρ ↑
Qwen3.5-4B	TAEC	.554	.131	.082
	Zero-shot tree	.489	.001	.001
	TabLLM-style direct	.535	.122	.038
Qwen3.5-9B	TAEC	.529	.118	.073
	Zero-shot tree	.497	.059	.062
	TabLLM	.519	.109	.050
Qwen3.5-27B	TAEC	.565	.185	.066
	Zero-shot tree	.555	.143	.084
	TabLLM	.545	.252	.142
Qwen3.8-27B	TAEC	.540	.242	.087
	Zero-shot tree	.558	.215	.072
	TabLLM	.550	.293	.086
GPT-5.5	TAEC	**.610**	**.299**	**.197**
	Zero-shot tree	.592	.277	.177
	TabLLM	.590	.334	.069

models. Backbone scaling most clearly improves continuous-feature recovery, whereas binary and ordinal performance is less monotonic. GPT-5.5 achieves the strongest binary and ordinal results with TAEC, while the best continuous correlation is obtained by TabLLM with GPT-5.5.

Outcome Recovery. As shown in Figure 3, TAEC achieves the lowest control-outcome error across all four patient- and trial-level evaluations. At the trial level, median APE is 35.0% on the validated-death MIMIC-IV endpoint and 29.8% on MarketScan, compared with 88.2% and 40.6% for the unweighted pools, respectively. The advantage on MIMIC-IV is particularly important because it persists when outcomes are evaluated using observed deaths and censoring-aware Kaplan–Meier estimation rather than the follow-up proxy. At the patient level, TAEC obtains 41.8% APE on MIMIC-IV and 23.9% on claims, remaining substantially below the classical baselines. These results show that the trial-specific construction decisions improve control-outcome recovery across data granularities, RWD sources, and both proxy and validated mortality endpoints. Detailed results are provided in section E.1.

Survival-Curve Recovery. TAEC also achieves the lowest survival-curve error on both patient-level donor sources. It obtains an RMST_{24} error of 3.47 months on MIMIC-IV and 2.54 months on claims, improving over all classical cohort-construction baselines. This extends the recovery gains beyond a single control-arm median to integrated survival over the first 24 months. The improvement on the validated-death MIMIC-IV evaluation further shows that the advantage is not restricted to the claims follow-up proxy. Survival-curve evaluation is restricted to the 11 trials with individual time-to-event data; detailed results are provided in section E.1.

Treatment-Effect Recovery. The gains in control reconstruction consistently translate into more accurate treatment-effect estimates. At the trial level, TAEC achieves the lowest absolute log-ratio



(a) Patient-level: MIMIC-IV (b) Patient-level: MarketScan

(c) Trial-level: MIMIC-IV (d) Trial-level: MarketScan

Figure 3: External-control construction performance on SCT-Bench across patient-level and trial-level evaluations using MIMIC-IV and MarketScan. Each radar axis represents one evaluation metric, with values normalized so that farther from the center indicates better performance.

error on both MIMIC-IV and MarketScan, with errors of 0.369 and 0.336, respectively. At the patient level, TAEC likewise achieves the lowest error on both sources, with $|\Delta \log HR|$ of 0.623 on MIMIC-IV and 0.470 on claims. Thus, improved external-control construction propagates beyond recovery of the control outcome to the downstream comparison between treatment and control arms. Detailed results are provided in section E.1.

Drug-Efficacy Recovery. TAEC achieves the strongest drug-efficacy recovery across four evaluations. At the trial level, estimate agreement reaches 42% on MIMIC-IV and 44% on MarketScan. At the patient level, TAEC attains the highest chance-corrected three-class agreement, with $\kappa = 0.089$ on MIMIC-IV and $\kappa = 0.171$ on claims. These results show that gains in control reconstruction propagate to downstream treatment conclusions. Results are provided in section E.1.

4.3 Additional Analyses

In this section, we examine the validity and mechanisms of our framework. We first audit whether comparable historical controls are sufficiently available to support direct reuse, then use component ablations and time-zero transfer to identify where the recovery gains arise. We further examine the theoretical mechanisms and subgroup-level behavior of the resulting controls.

Availability of Historical Controls. Our zero-shot setting addresses trials for which construction decisions cannot simply be copied from a comparable historical study. We audit historical candidates for 12,467 single-arm oncology trials. Although 99.8% have an earlier-registered candidate with the same tumour type and histology, only 7.0% retain a candidate after the first four operationalized conditions of Pocock (1976), falling to 0.72% after the fifth. Missing baseline reports account for 85.6% of exclusions at the fourth condition. Thus, finding a related trial rarely establishes that its

control is transferable, motivating construction from reusable knowledge when comparability cannot be verified. Details are provided in subsection E.2.

Ablation Study. We separately evaluate the contributions of retrieved skills, structured verification, and iterative refinement within the construction loop, before the final balancing post-step. Removing the skill library, replacing verifier feedback with self-reflection, and giving a one-shot planner the candidate time-zero landmarks upfront all degrade recovery. On claims, median APE increases from 24.2% for the full loop to 25.9–29.7% across these variants; on validated-death MIMIC-IV, it increases from 41.8% to 72.4–183.1%. These results show that the gains arise from combining reusable construction knowledge with verifier-guided refinement rather than from evidence access alone. Full within-backbone results are provided in subsection E.3.

Time-Zero Selection. To isolate the value of the time-zero element, we keep each baseline's selected patients, donor weights, and outcome estimator unchanged, and modify only the start of follow-up using the trial-compatible time zero identified by TAEC. On the nine trials where this alternative follow-up start can be constructed from the available RWD, the change substantially reduces APE for nearly all outcome-free baselines, with particularly large gains in later-line settings. The improvement also persists across 40 clinical strata, showing that time-zero selection is a transferable construction decision rather than a benefit specific to our downstream estimator. Implementation details and complete results are provided in subsection E.5.

Theoretical Analysis. We characterize how errors in temporal alignment affect outcome recovery, when correcting the start of follow-up can reduce this error, and how partial optimal transport constrains donor-weight concentration. Empirical diagnostics further examine whether the observed results are consistent with these mechanisms. These analyses clarify why the selected construction elements can affect recovery even when the downstream estimator is unchanged. Details are provided in Appendix F.

Stratified Evaluation. Whole-trial recovery can conceal errors within clinically distinct populations. We therefore evaluate baseline-defined strata by sex, age, and performance status when subgroup-level randomized controls are available. Performance varies across subgroup types and data sources, with larger errors when clinically important characteristics cannot be represented directly in the RWD source. In particular, performance-status strata remain challenging because the external data provide only an indirect utilisation-derived representation rather than the trial's clinical performance-status measure. These results show that strong trial-level recovery does not guarantee subgroup fidelity when key characteristics are incompletely represented. Detailed results are provided in subsection E.6.

5 Conclusion

We formulate external-control construction for oncology single-arm trials as a zero-shot, task-adaptive analysis-planning problem: given an unseen trial, the system must translate reusable methodological knowledge into executable construction choices without access to the missing control or its outcomes. We introduce SCT-Bench, a benchmark of 9,163 oncology trials spanning seven complementary tasks, and TAEC, a knowledge-to-action framework that combines planning over a reusable skill library, data-grounded construction, and outcome-free refinement with trial-local skill updates. Across patient- and trial-level evaluations on MIMIC-IV and MarketScan, TAEC achieves the lowest control-outcome, survival-curve, and treatment-effect errors while maintaining near-best control-arm similarity, including median APEs of 35.0% and 29.8% at the trial level and 41.8% and 23.9% at the patient level. Transferring its selected time-zero choice to classical pipelines also substantially reduces their reconstruction error, showing that gains arise not only from downstream estimation but from selecting appropriate trial-specific construction elements. Together, these results establish external-control construction as a knowledge-to-action problem in which upstream analysis planning is central to reliable recovery of missing controls.

AI USE STATEMENT

Generative AI tools were used in this work to assist with writing and language polishing, retrieval and discovery of related literature, research ideation and execution, and drafting portions of the manuscript. In particular, these tools were used to brainstorm and refine methodological and experimental ideas, assist with research code and analysis workflows, identify potentially relevant literature, prototype figures, and improve the organization, clarity, and presentation of the paper. All AI-assisted research decisions, implementations, references, analyses, and manuscript content were reviewed and verified by the authors. References suggested through AI-assisted retrieval were independently checked against the original sources, and all reported experimental results were obtained from the underlying data and validated by the authors. The authors take full responsibility for the final content of the work.

REFERENCES

M. Abdallah, S. Nakken, M. Georges, et al. Trialmatchai: an end-to-end ai-powered clinical trial recommendation system to streamline patient-to-trial matching. *Nature Communications*, 17: 4472, 2026.

Sundeep Agrawal, Shaily Arora, Laleh Amiri-Kordestani, R Angelo de Claro, Lola Fashoyin-Aje, Nicole Gormley, Tamy Kim, Steven Lemery, Gautam U Mehta, Emma C Scott, et al. Use of single-arm trials for us food and drug administration drug approval in oncology, 2002-2021. *JAMA oncology*, 9(2):266–272, 2023.

Isabelle Rose I Alberto et al. The impact of commercial health datasets on medical research and health-care algorithms. *The Lancet Digital Health*, 5(5):e288–e294, 2023. doi: 10.1016/S2589-7500(23)00025-0.

Sanchita Bhattacharya, Patrick Dunn, Cristel G. Thomas, Barry Smith, Henry Schaefer, Jieming Chen, Zicheng Hu, Kelly A. Zalocusky, Ravi D. Shankar, Shai S. Shen-Orr, Elizabeth Thomson, Jeffrey Wiser, and Atul J. Butte. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific Data*, 5:180015, 2018. doi: 10.1038/sdata.2018.15.

Barbara E. Bierer, Rebecca Li, Mark Barnes, and Ida Sim. A global, neutral platform for sharing trial data. *New England Journal of Medicine*, 374(25):2411–2413, 2016. doi: 10.1056/NEJMp1605348.

Mehmet Burcu, Nancy A Dreyer, Jessica M Franklin, Michael D Blum, Cathy W Critchlow, Eleanor M Perfetto, and Wei Zhou. Real-world evidence to support regulatory decision-making for medicines: considerations for external control arms. *Pharmacoepidemiology and drug safety*, 29(10):1228–1235, 2020.

Harlan Campbell and Antonio Remiro-Azócar. Doubly robust augmented weighting estimators for the analysis of externally controlled single-arm trials and unanchored indirect treatment comparisons. *Research Synthesis Methods*, pp. 1–34, 2026. doi: 10.1017/rsm.2026.10106.

Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E. Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J. Byrne, Michael L. Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P. Goldberg, Chris Sander, and Nikolaus Schultz. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012. doi: 10.1158/2159-8290.CD-12-0095.

Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2903–2913. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1e6e25d952a0d639b676ee20d0519ee2-Paper.pdf.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The

cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013. doi: 10.1007/s10278-013-9622-7.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269):pl1, 2013. doi: 10.1126/scisignal.2004088.

Carol A. Giffen, Leslie E. Carroll, John T. Adams, Sean P. Brennan, Sean A. Coady, and Elizabeth L. Wagner. Providing contemporary access to historical biospecimen collections: Development of the NHLBI biologic specimen and data repository information coordinating center (BioLINCC). *Biopreservation and Biobanking*, 13(4):271–279, 2015. doi: 10.1089/bio.2014.0050.

Angela K. Green, Katherine E. Reeder-Hayes, Robert W. Corty, Ethan Basch, Mathew I. Milowsky, Stacie B. Dusetzina, Antonia V. Bennett, and William A. Wood. The project data sphere initiative: Accelerating cancer research by sharing data. *The Oncologist*, 20(5):464–e20, 2015. doi: 10.1634/theoncologist.2014-0431.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012. doi: 10.1093/pan/mpr025. URL https://doi.org/10.1093/pan/mpr025.

Mustafa Hashmi, Jeremy Rassen, and Sebastian Schneeweiss. Single-arm oncology trials and the nature of external controls arms. *Journal of Comparative Effectiveness Research*, 10(12):1053–1066, 2021.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5549–5581. PMLR, 2023.

Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. Matching patients to clinical trials with large language models. *Nature communications*, 15(1):9074, 2024.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. https://archive.ics.uci.edu.

Ricardo Knauer, Mario Koddenbrock, Raphael Wallsberger, Nicholas M. Brisson, Georg N. Duda, Deborah Falla, David W. Evans, and Erik Rodner. "Oh LLM, I'm Asking Thee, Please Give Me a Decision Tree": Zero-Shot Decision Tree Induction and Embedding with Large Language Models. *arXiv preprint arXiv:2409.18594*, 2024.

Jiali Liu, Minghong Yao, Mingqi Wang, et al. Design, conduct, and analysis of externally controlled trials. *JAMA Network Open*, 8(9):e2530277, 2025. doi: 10.1001/jamanetworkopen.2025.30277.

Lu Liu, Jungyeon Choi, Jammbe Z. Musoro, et al. Single-arm studies involving patient-reported outcome data in oncology: A literature review on current practice. *The Lancet Oncology*, 24(5):e197–e206, 2023.

Matthew D. Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, 39(10):1181–1186, 2007. doi: 10.1038/ng1007-1181.

Merative. Marketscan research databases. https://www.merative.com/documents/merative-marketscan-research-databases, 2022. Accessed: 2026-08-03.

OpenAI. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/, April 2026.

J. Park, Y. Fang, C. Ta, et al. Criteria2query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. *Journal of Biomedical Informatics*, 154:104649, 2024. doi: 10.1016/j.jbi.2024.104649.

Stuart J. Pocock. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29(3):175–188, 1976. doi: 10.1016/0021-9681(76)90044-8.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, August 2026. ISSN 3005-0693. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026a.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. https://qwen.ai/blog?id=qwen3.8, August 2026b.

Rifaquat Rahman, Steffen Ventz, Jon McDunn, Bill Louv, Irmarie Reyes-Rivera, Mei-Yin C Polley, Fahar Merchant, Lauren E Abrey, Joshua E Allen, Laura K Aguilar, et al. Leveraging external data in the design and analysis of clinical trials in neuro-oncology. *The lancet oncology*, 22(10): e456–e465, 2021.

J. Ren, P. Cislo, J. C. Cappelleri, P. Hlavacek, and M. DiBonaventura. Comparing g-computation, propensity score-based weighting, and targeted maximum likelihood estimation for analyzing externally controlled trials with both measured and unmeasured confounders: a simulation study. *BMC Medical Research Methodology*, 2023.

G. Rippin, N. Ballarini, H. Sanz, et al. A review of causal inference for external comparator arm studies. *Drug Safety*, 2022.

James M. Robins, Miguel Ángel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000. doi: 10.1097/00001648-200009000-00011. URL https://pubmed.ncbi.nlm.nih.gov/10955408/.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

Joseph S. Ross, Joanne Waldstreicher, Stephen Bamford, Jesse A. Berlin, Karla Childers, Nihar R. Desai, Ginger Gamble, Cary P. Gross, Richard Kuntz, Richard Lehman, Peter Lins, Sandra A. Morris, Jessica D. Ritchie, and Harlan M. Krumholz. Overview and experience of the YODA project with clinical trial data sharing after 5 years. *Scientific Data*, 5:180268, 2018. doi: 10.1038/sdata.2018.268.

Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, and Charles Fisher. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*, 18(2):329–356, 2022. doi: 10.1515/ijb-2021-0072.

James E. Signorovitch, Eric Q. Wu, Andrew P. Yu, Charles M. Gerrits, Evan Kantor, Yanjun Bao, Shiraz R. Gupta, and Parvez M. Mulani. Comparative effectiveness without head-to-head trials: A method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *PharmacoEconomics*, 28(10):935–945, 2010. doi: 10.2165/11538370-000000000-00000. URL https://doi.org/10.2165/11538370-000000000-00000.

James E. Signorovitch, Vanja Sikirica, M. H. Erder, et al. Matching-adjusted indirect comparisons: A new tool for timely comparative effectiveness research. *Value in Health*, 2012.

U.S. Food and Drug Administration. Considerations for the design and conduct of externally controlled trials for drug and biological products: Draft guidance for industry. Guidance document, February 2023. URL https://www.fda.gov/media/164960/download. Draft guidance, not for implementation.

Yuxin Wang, Dennis Frauen, Emil Javurek, Konstantin Hess, Yuchen Ma, and Stefan Feuerriegel. Targeted synthetic control method, 2026. URL https://arxiv.org/abs/2602.04611.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

A Related Work

External controls for single-arm trials. Existing methods estimate a treatment effect once the comparison has been specified. Propensity-score adjustment, outcome regression, doubly robust estimation, and matching-adjusted indirect comparison all take a population, a set of measurements, and a time origin as given and balance covariates within them (Rippin et al., 2022; Signorovitch et al., 2012; Ren et al., 2023). Their validity rests on those specifications, which are made by hand for each study (Rahman et al., 2021; Liu et al., 2025). Digital twins take a different route, training a prognostic model on historical data to adjust a randomized trial (Schuler et al., 2022), but presume a randomized control still exists. We study how to make the specification for an unseen trial without feedback from its missing control.

Language models for clinical-trial retrieval and matching. Recent work has applied LLMs to several components of clinical-trial matching. Criteria2Query translates free-text eligibility criteria into executable database queries (Park et al., 2024). TrialGPT retrieves candidate trials, performs criterion-level eligibility assessment, and ranks trials for individual patients (Jin et al., 2024). TrialMatchAI similarly combines retrieval, re-ranking, and LLM-based eligibility classification over heterogeneous patient information (Abdallah et al., 2026). These systems address whether a patient satisfies a trial's eligibility criteria or which trial is most appropriate for a patient. Our setting instead assumes the target trial is already given and asks how its missing comparison arm should be constructed from RWD. The LLM therefore does not directly predict eligibility or outcomes; it selects trial-specific construction elements, whose numerical realization is delegated to fixed execution code.

B Implementation Details of TAEC

B.1 Inputs and Information Access

For target trial t, the framework receives the protocol P_t, the observed treatment population $\mathcal{T}_t$, and the candidate RWD source $\mathcal{D}$. Treatment-population information includes released baseline characteristics. Consistent with FDA recommendations to prespecify external-control selection and analysis before trial initiation, we withhold treatment-arm outcomes during construction to prevent outcome-informed design decisions (U.S. Food and Drug Administration, 2023). The randomized control arm $\mathcal{C}_t^*$, including its baseline characteristics and outcomes, is reserved for retrospective evaluation.

The RWD context supplied to the planner describes the source schema, available clinical variables, treatment records, and recorded clinical events that may define time zero, such as diagnosis or treatment initiation. Using this context, the planner maps protocol requirements to source variables and checks whether their clinical definitions and measurement times support the intended eligibility criteria, baseline features, and follow-up origin.

Historical retrieval excludes target-trial materials. Retrieved knowledge therefore supports construction decisions without exposing the withheld randomized reference. The planner and verifier do not use target-trial outcomes or downstream outcome estimates to select, revise, or accept a plan. Donor outcome records are used by the downstream estimator after construction; these read-outs are not fed back into refinement.

B.2 Construction Elements and Executable Plans

The construction plan comprises ten elements:

$$\pi_t = (e_t^1, \ldots, e_t^{10}), \qquad e_t^k \in \mathcal{O}_k,$$

where $\mathcal{O}_k$ is the predefined operator space for element k. These elements configure temporal alignment, outcome read-out, donor-source selection, feature alignment, borrowing, and transport diagnostics. Table 3 lists the implemented operator spaces and their defaults. The executor translates each selected operator into its associated parameters and checks. Depending on the element, execution changes the feature costs, transported mass, time-origin scale, outcome quantile, or verification settings. The elements are evaluated jointly because their feasibility depends on shared source information. A time-zero choice requires a named source event compatible with the enrollment setting

Table 3: Trial-specific construction elements and their admissible choices. Bold denotes the default choice.

Element	Admissible Choices
Time Zero	Treatment anchored, **Diagnosis anchored**, Procedure anchored, Landmark anchored, Unresolved
Outcome Readout	**Median**, Time-zero scaled, Upper quartile, Keyword-informed quantile, Public-prior quantile
Control Source	**Strict pool**, Line-of-care pool, Unrestricted pool
Feature Alignment	Demographics only, **Protocol conditioned**, Full schema
Feature Ascertainability	Directly observed only, **Allow supported inference**, Allow unsupported inference, Abstain
Bilateral Observability	**Require both sides**, Allow unilateral unknown
Borrowing	Full, **Partial**, Restricted, No borrowing
Matching Diagnostics	**Full checks**, Matching only, None
Schema Alignment	Raw coding, **Harmonize known coding**, Abstain on mismatch
Treatment-History Alignment	Ignore, **Line matched**, Prior therapy required

and sufficient patient support. Feature alignment depends on feature ascertainability, bilateral observability, and schema compatibility. Borrowing is additionally constrained by the effective sample size of the resulting weighted cohort.

Temporal alignment and outcome read-out. The time-zero operator selects the source event from which follow-up should begin so that the external control uses a clock compatible with the target trial. When the selected event occurs after the source's default index, we estimate a two-clock ratio, $r_{\text{two_clock}}$, defined as the ratio between follow-up measured from the selected event and follow-up measured from the default source index among patients for whom both clocks are observable. This ratio is used to calibrate the outcome read-out to the selected time origin. If the required event cannot be identified in the source data, the time-zero choice remains unresolved and no temporal correction is applied; the verifier treats this as missing evidence rather than evidence of valid alignment.

The outcome-read-out operator determines how the weighted donor outcome distribution is summarized. Standard choices use the median or upper quartile, while time-zero-adjusted read-out retains the median quantile and applies the estimated temporal scale. When an external literature estimate M_{pub} is available, the public-prior operator maps it to its corresponding quantile in the source-pool outcome distribution, $F_{\text{pool}}(M_{\text{pub}})$. The verifier checks that the selected read-out is compatible with the trial setting, time origin, and available source support.

Feature representation and admissibility. Feature-alignment operators determine which patient characteristics are used to compare the trial and external-control populations, ranging from demographics alone to protocol-conditioned or broader source features. Feature ascertainability determines whether a clinically relevant factor can be used directly, inferred from supported evidence, or should be excluded when it cannot be represented reliably. We summarize this with a recoverability score that reflects how faithfully a trial-side concept can be reconstructed from the available source information; poorly supported factors are omitted from the matching representation rather than treated as observed.

Bilateral observability further checks whether the information required for comparison is available on both the trial and external-data sides. Missing information is treated as uncertainty rather than positive evidence, while explicit contradictions can make a candidate inadmissible. Schema alignment harmonizes compatible representations across sources, such as different categorical encodings, and treatment-history alignment determines whether prior treatment and line-of-therapy information is ignored, incorporated into matching, or required for eligibility. The verifier additionally tracks how much of the constructed cohort depends on incompletely observed information and flags designs that rely too heavily on such evidence.

Borrowing and transport diagnostics. The borrowing operator controls how much of the external candidate population may contribute to the constructed control, ranging from broad borrowing to restricted borrowing or abstention when adequate support is unavailable. The selected borrowing level is executed through partial transport, allowing unsupported patients to remain unmatched rather than forcing the entire donor pool into the comparison.

Transport diagnostics evaluate whether the resulting weighted cohort is sufficiently supported and stable. The verifier checks effective sample size, residual population imbalance, sensitivity to perturbations of the transport solution, and other indicators of unstable or overly concentrated borrowing. Designs that fail essential support requirements are rejected, while weaker diagnostic concerns are returned as signals for revision.

B.3 Skill Representation and Planning

The skill library stores reusable construction knowledge derived from methodological literature and historical trials. Design skills guide the selection of trial-specific construction elements, whereas transport skills describe how those choices can be operationalized under the available trial and source information. Each skill links an applicability context to candidate operators, the evidence required to support them, and conditions under which the recommendation should be rejected or the system should abstain. We represent a skill abstractly as

$$s = (c_s, \mathcal{O}_s, q_s, f_s),$$

where c_s defines its applicability context, $\mathcal{O}_s$ contains candidate operators, q_s specifies the supporting evidence, and f_s records failure or abstention conditions.

For each target trial, relevant skills are retrieved using the protocol, treatment setting, disease context, and information available in the external source. The planner combines these skills with the current trial and source context to select construction elements. Retrieved recommendations are treated as conditional knowledge rather than fixed rules: an operator is used only when its supporting requirements are satisfied. During refinement, the planner may retrieve alternative skills, combine compatible recommendations, specialize general knowledge to the current trial, or revise unsupported choices. These operations alter the construction plan using only observable evidence and never expose the withheld target control outcome.

B.4 Donor Eligibility and Temporal Alignment

Population, control-source, treatment-history, and time-zero choices jointly determine which source patients can serve as candidate controls. Given a construction plan π_t, the eligible donor pool is

$$\mathcal{E}_t = \{i \in \mathcal{D} : g(i; \pi_t) = 1\},$$

where g applies the selected eligibility requirements to the available source records. Requirements that cannot be determined from the source are treated as unknown rather than as satisfied or violated.

Temporal alignment then determines the event from which donor follow-up begins. The selected time origin must both match the target trial's enrollment setting and be identifiable in the external data. These are separate requirements: selecting an appropriate time-zero operator does not guarantee that the corresponding event can actually be observed for a given source or patient. When the required time origin is unavailable, the framework does not substitute a different clock as if it were equivalent. Instead, the alignment remains unresolved and the missing support is passed to the verifier for revision or abstention.

B.5 Feature Ascertainability and Alignment

For each clinical feature required by the construction plan, the framework assesses its ascertainability separately in the trial and the RWD source:

$$a_{j,t}^{(s)} \in \{\text{observed}, \text{inferable}, \text{unavailable}\}, \qquad s \in \{\text{trial}, \text{RWD}\}.$$

An observed feature is directly represented in the available records. An inferable feature is not directly recorded but can be reconstructed from available information using a supported rule. An unavailable feature lacks a reliable representation in that source. For inferable features, the LLM generates a reusable rule that is subsequently executed across patient records:

$$\widehat{x}_{ij,t}^{(s)} = h_{j,t}^{(s)}\left(r_i^{(s)}\right).$$

The rule is generated without access to target-trial feature labels and is shared across records rather than prompting the LLM separately for each patient. This makes the transformation explicit and reproducible while preserving the distinction between directly observed and inferred information.

Schema alignment then maps ascertainable features into compatible clinical definitions and encodings across sources, while covariate alignment determines which aligned features are used for patient comparison. A feature can contribute to cross-source matching only when it has a supported representation on both sides. The resulting patient representation is

$$\mathbf{x}_{i,t} = \phi_t(r_i^{(s)}, s),$$

where ϕ_t combines the selected ascertainment rules, schema mappings, and matching features. The ascertainability status of each feature is retained for downstream verification even after its representation has been aligned.

B.6 Partial Optimal Transport and Donor Weights

Partial optimal transport compares the aligned external-source and target-trial representations while allowing unsupported candidates to remain unmatched. Let $C_{ij,t}$ denote the cost of aligning donor unit i with target unit j, and let μ_t denote the amount of mass selected for borrowing. The feasible transport set is

$$\mathcal{U}_{\mu_t}(a,b) = \left\{\Gamma \geq 0 : \Gamma\mathbf{1} \leq a,\ \Gamma^\top\mathbf{1} \leq b,\ \mathbf{1}^\top\Gamma\mathbf{1} = \mu_t\right\}.$$

The transport plan is obtained by minimizing the alignment cost,

$$\Gamma_t^\star \in \arg\min_{\Gamma\in\mathcal{U}_{\mu_t}(a,b)} \left[\langle C_t, \Gamma\rangle + \varepsilon_t\Omega(\Gamma)\right],$$

where Ω denotes the regularization used by the selected transport operator.

The transported mass assigned to each donor induces its external-control weight,

$$w_{i,t} = \frac{\sum_j \Gamma^\star_{ij,t}}{\mu_t}, \qquad \sum_i w_{i,t} = 1.$$

Donors receiving no transported mass remain unmatched. The resulting weighted donor set forms the constructed external control,

$$\widehat{\mathcal{C}}_t = \{(i, w_{i,t}) : i \in \mathcal{E}_t,\ w_{i,t} > 0\}.$$

Execution diagnostics characterize whether the selected transport is feasible, sufficiently supported, and overly concentrated. We summarize weight concentration using the effective sample size,

$$\mathrm{ESS}_t = \frac{1}{\sum_i w_{i,t}^2}.$$

All transport construction and diagnostics depend only on observable trial and source information and do not use the withheld control outcome.

B.7 Outcome-Free Verification and Refinement

The verifier evaluates the current construction plan after execution. It receives the selected elements, retrieved skills, constructed cohort, and execution diagnostics, and checks whether the plan is consistent with the protocol, supported by the available evidence, and feasible under the current source.

Verifier feedback identifies both the affected construction element and the reason for failure. This distinction separates semantic incompatibility from source limitations and execution failures. For example, a time-zero choice may be appropriate for the target setting but unusable because the corresponding source event cannot be identified. Similarly, an inferred feature may be executable but insufficiently supported as a representation of the required clinical concept.

Refinement uses these structured failures to revise the construction plan and, when needed, retrieve or adapt its supporting skills:

$$(\mathcal{R}_t^{(r)}, \pi_t^{(r)}) \to \widehat{\mathcal{C}}_t^{(r)} \to V_t^{(r)} \to (\mathcal{R}_t^{(r+1)}, \pi_t^{(r+1)}).$$

Each revised plan is re-executed before verification is repeated. The loop terminates when no actionable failure remains or the refinement budget is exhausted.

Across visited plans, the verifier assigns a construction score that prioritizes hard incompatibilities and otherwise aggregates softer evidence, support, and stability concerns. The framework returns the best-supported visited plan rather than selecting according to target outcomes or downstream reconstruction accuracy. Consequently, verification certifies only that a construction is better supported by the observable protocol and source evidence; it does not guarantee accurate recovery of the withheld control outcome.

B.8 Skill Updates and Evaluation Isolation

The framework allows skills to be revised or extended during refinement, but benchmark evaluation keeps these updates trial-local. Every target trial starts from the same initial skill library,

$$\mathcal{S}_{\mathrm{start},i} = \mathcal{S}_0 \qquad \text{for all evaluation trials } i.$$

Skills retrieved, specialized, or created while processing one trial may influence that trial's refinement, but all trial-specific updates are reset before the next evaluation trial. This prevents later trials from benefiting from information accumulated on earlier benchmark targets.

All skill updates rely only on the protocol, RWD context, and outcome-free verification signals; withheld control outcomes and downstream reconstruction errors are never used to create, select, or update skills.

B.9 Outcome Read-Out

After external-control construction, the selected donor weights are applied to the endpoint available in each source. Outcome estimation is strictly downstream of construction: these estimates are used only for evaluation and are never fed back to select or revise the construction plan.

The resulting estimand is source dependent. When mortality information is available, the read-out summarizes the corresponding survival endpoint. MarketScan, however, does not provide death dates. For this source, we therefore use duration from the selected time origin to the end of observed follow-up, treating follow-up termination as the read-out event. This quantity measures alignment of the observed follow-up-time distribution and should not be interpreted as overall survival or a mortality-based survival distribution.

Accordingly, we distinguish mortality-based survival recovery from follow-up-duration recovery throughout the evaluation, and use the broader term outcome recovery when results from both endpoint types are summarized together.

C SCT-Bench Data and Evaluation Details

Trial Curation. For SCT-Bench, we screened 11 repositories containing patient-level clinical trial data: Project Data Sphere (Green et al., 2015), Vivli (Bierer et al., 2016), the Yale Open Data Access Project (Ross et al., 2018), ImmPort (Bhattacharya et al., 2018), dbGaP (Mailman et al., 2007), The Cancer Imaging Archive (Clark et al., 2013), BioLINCC (Giffen et al., 2015), the UCI Machine Learning Repository (Kelly et al.), cBioPortal (Cerami et al., 2012; Gao et al., 2013), the AIDS Clinical Trials Group (ACTG),[2] and the NCI National Clinical Trials Network / NCI Community Oncology Research Program (NCTN/NCORP) Data Archive.[3]

Three repositories have relatively low access barriers in practice, requiring registration and a standard data-use agreement without project-specific sponsor or committee approval for each request: Project Data Sphere, UCI, and cBioPortal. Vivli instead follows a data-request process in which the proposal is reviewed by the data contributor and by an independent review panel before release. The three low-barrier repositories provide directly downloadable patient-level trial datasets: Project Data Sphere, UCI, and cBioPortal. Across the screened resources, 41 trials contribute patient-level data to SCT-Bench. Of these, 30 are used for Feature Ascertainability, while 11 two-arm randomized trials provide both treatment and control patient records and are replayed as single-arm trials by withholding the randomized control arm.

[2] https://actgnetwork.org

[3] https://nctn-data-archive.nci.nih.gov

Table 4: Task coverage in SCT-Bench. Trial-level and patient-level evaluations use different subsets depending on the reference evidence required by each task.

Task	Trial-level	Patient-level
Protocol Extraction	2,975 trials with publication-grounded evidence. Scored fields: Setting 2,046 (933 in the two-source agreement subset), Prior Lines 1,164, Extent 2,247.	–
Feature Ascertainability	–	374 masked-feature instances from 30 released PDS trials: 104 binary, 215 continuous, and 55 ordinal.
Control-Arm Similarity	Evidence corpus: 8,303 MIMIC / 1,729 MarketScan trials (2,760 / 695 with a hidden second arm); registry: 822 / 354 trials with a reported control row.	11 randomized trials (trial 201 reports sex and stage proxies but no age).
Outcome Recovery	858 MIMIC trials (validated death) and 366 MarketScan trials (follow-up proxy).	11 randomized trials; the control-median APE is evaluable for 10 because trial 201 does not reach the median.
Survival-Curve Recovery	–	11 randomized trials evaluated with RMST_{24}.
Treatment-Effect Recovery	783 MIMIC / 333 MarketScan trials whose treated and control arms report the same outcome measure.	11 randomized trials.
Drug-Efficacy Recovery	783 / 333 trials; 662 / 294 additionally post confidence intervals for estimate agreement.	11 randomized trials.

We further construct the trial-level portion of SCT-Bench from ClinicalTrials.gov,[4] yielding 9,122 trials with trial-level information after curation. These records provide structured trial characteristics, arm information, and reported outcomes for the trial-level benchmark tasks. Together, the patient-level and trial-level collections comprise 9,163 trial instances.

C.1 Data Sources and Benchmark Split

SCT-Bench links trial protocols and primary results publications with released individual participant data (IPD), registry records, and arm-level baseline and outcome summaries. Evaluation is organized at two levels. The patient-level collection contains released IPD used for Feature Ascertainability and randomized-control replay, while the trial-level collection uses registry- and publication-derived trial information. For outcome-related tasks, control-arm information and treatment-arm outcomes used as references are withheld during construction and accessed only for final scoring.

C.2 Task Coverage and Eligibility

The seven tasks operate on different subsets of SCT-Bench according to the evidence and reference information required for evaluation. Table 4 summarizes the available evaluation instances.

For Protocol Extraction, an instance is included when the target field can be supported by an identifiable reference in the associated trial record or primary results document.

For Feature Ascertainability, all 30 released IPD trials are considered. Within each trial, every eligible released variable is treated in turn as the masked target and all remaining variables are made available as predictors. A target is retained when it has at least 25 non-missing observations. For binary targets, the minority class must additionally account for at least 3% of observations.

Control-Arm Similarity requires a control-arm reference for the evaluated baseline characteristic. Patient-level randomized trials use the withheld control records directly, whereas trial-level studies use reported randomized-control baseline summaries.

[4] https://clinicaltrials.gov

Outcome Recovery requires an observed randomized-control median outcome under the corresponding source-specific endpoint. Trials for which the control median is not reported or is not reached are excluded from median-survival metrics.

Survival-Curve Recovery is restricted to the 11 randomized trials with individual time-to-event data. Trials represented only at the trial level are excluded because full survival trajectories cannot be reconstructed from fixed-time survival summaries.

Treatment-Effect Recovery requires sufficient treatment- and control-arm information to define the same effect measure for both the constructed and randomized controls.

Drug-Efficacy Recovery further requires that a reference treatment conclusion can be derived from the randomized comparison.

C.3 Evaluation Metrics

Protocol Extraction. Let z_{ik} denote the reference value of protocol field k for trial i and $\hat{z}_{ik}$ the extracted value. For categorical and integer-valued fields, we report exact-match accuracy,

$$\mathrm{Acc}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{1}[\hat{z}_{ik} = z_{ik}] .$$

For treatment setting, we additionally report accuracy on the subset for which the independent reference sources agree. Control regimens are normalized to canonical drug sets before exact-set comparison,

$$\mathbb{1}\Big[\mathcal{N}(\widehat{D}_i) = \mathcal{N}(D_i)\Big] ,$$

where $\mathcal{N}(\cdot)$ denotes regimen normalization. Accuracy is reported separately for each protocol field; higher is better.

Feature Ascertainability. For each eligible feature in a released IPD trial, the target feature is masked and inferred from the remaining observed features without using target-feature labels for model fitting or demonstrations. Metrics are computed within each masked-feature instance and aggregated by feature type.

For binary targets, we report area under the receiver-operating-characteristic curve (AUC). For continuous targets, we use Spearman rank correlation,

$$\rho_s = \mathrm{corr}\left(\mathrm{rank}(y), \mathrm{rank}(\hat{y})\right) .$$

Ordinal targets are evaluated using the same rank-correlation metric. Higher values indicate better ascertainment for all three feature types.

Control-Arm Similarity. We evaluate whether the constructed external control reproduces observable baseline characteristics of the reference control arm, the two characteristics with clinically compatible representations on both sides. For each characteristic k, similarity is measured using the absolute standardized mean difference,

$$\mathrm{SMD}_k = \frac{\left|\widehat{\mu}_k - \mu_k^{(0)}\right|}{s_k},$$

where $\widehat{\mu}_k$ and $\mu_k^{(0)}$ denote the corresponding summaries in the constructed and reference control arms, and s_k is the appropriate standardizing dispersion for the characteristic. We report absolute SMD and the proportion of trials meeting prespecified similarity thresholds; lower absolute SMD and higher threshold-attainment proportions indicate better similarity.

Outcome Recovery. Let M_0 denote the reference control-arm median for the source-specific outcome and $\widehat{M_0}$ the corresponding median obtained from the constructed external control. We measure recovery using absolute percentage error,

$$\mathrm{APE} = 100 \times \frac{\left|\widehat{M_0} - M_0\right|}{M_0}.$$

We report median APE across eligible trials; lower is better.

The interpretation of M_0 depends on the endpoint available for the corresponding evaluation. When mortality information is available, it represents median survival. When only a source-specific follow-up endpoint is available, it represents median follow-up duration rather than mortality-based survival. We therefore use Outcome Recovery as the umbrella task name across these settings.

Survival-Curve Recovery. Survival-Curve Recovery is evaluated only on randomized trials with individual time-to-event data. Let $\widehat{S}(t)$ denote the survival curve reconstructed from the external control and $S_0(t)$ the Kaplan–Meier curve of the withheld randomized control arm. Restricted mean survival time through horizon τ is

$$\mathrm{RMST}_\tau(S) = \int_0^\tau S(t)\,dt.$$

The primary metric is absolute RMST error at 24 months,

$$E_{\mathrm{RMST}_{24}} = \left|\mathrm{RMST}_{24}(\widehat{S}) - \mathrm{RMST}_{24}(S_0)\right|.$$

We additionally report relative RMST_{24} error, absolute RMST_{12} error, and survival-probability error at selected horizons,

$$E_S(t) = \left|\widehat{S}(t) - S_0(t)\right|, \qquad t \in \{6, 12, 24\} \text{ months}.$$

Lower values indicate better curve recovery.

Treatment-Effect Recovery. Let ψ_0 denote the treatment effect obtained using the randomized or reported control reference and $\widehat{\psi}$ the corresponding effect obtained after replacing that control with the constructed external control. Recovery is measured by

$$E_{\mathrm{TE}} = \left|\widehat{\psi} - \psi_0\right|,$$

with lower values indicating better recovery.

For trials with individual time-to-event data,

$$\psi_0 = \log HR(T, C_0), \qquad \widehat{\psi} = \log HR(T, \widehat{C}),$$

where both hazard ratios are estimated using the same procedure and differ only in whether the comparator is the randomized control C_0 or the constructed external control $\widehat{C}$.

For trial-level evaluations, treatment and control arms must report the same outcome measure. The corresponding effect is defined using the ratio of their reported median outcomes,

$$\psi_0 = \log \frac{M_T}{M_0}, \qquad \widehat{\psi} = \log \frac{M_T}{\widehat{M}_0}.$$

Thus, the trial-level metric does not assume that every reported median corresponds to overall survival.

Drug-Efficacy Recovery. Drug-Efficacy Recovery evaluates whether replacing the reference control with the constructed external control preserves the resulting treatment conclusion. Each comparison is mapped by the same prespecified rule to one of three efficacy categories,

$$y_i, \hat{y}_i \in \{\text{benefit}, \text{null}, \text{harm}\}.$$

Because these categories can be substantially imbalanced, raw accuracy may be high even for a constant predictor. We therefore use three-class Cohen's κ as the primary conclusion-level metric,

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where

$$p_o = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = y_i]$$

is the observed agreement and

$$p_e = \sum_c p_c \widehat{p}_c$$

is the agreement expected from the empirical reference and predicted class frequencies. Thus, a constant predictor receives $\kappa = 0$ rather than benefiting from the prevalence of the majority class. Higher κ indicates better discrimination among benefit, null, and harm. We additionally report raw conclusion agreement for interpretability.

For patient-level trials with individual time-to-event data, we further evaluate whether the reconstructed treatment effects preserve the ordering of treatment effects across trials. Let

$$\psi_i = \log HR_{\mathrm{RCT},i}, \qquad \widehat{\psi}_i = \log HR_{\mathrm{pred},i}.$$

We report Spearman rank correlation,

$$\rho_{\mathrm{effect}} = \mathrm{corr}\left(\mathrm{rank}(\psi), \mathrm{rank}(\widehat{\psi})\right),$$

together with its confidence interval. This threshold-free metric measures whether the reconstructed controls preserve relative treatment-effect magnitude independently of the class distribution or efficacy thresholds.

We additionally report confidence-interval-based estimate agreement where suitable reference intervals are available. At the patient level, the reconstructed treatment-effect estimate is considered concordant when it falls within the 95% confidence interval of the randomized treatment effect,

$$\mathbb{1}[|\log HR_{\mathrm{pred}} - \log HR_{\mathrm{RCT}}| \le 1.96\, SE(\log HR_{\mathrm{RCT}})].$$

The hazard ratios are estimated using Cox models, with the reference standard error obtained from the randomized trial data.

At the trial level, published confidence intervals are available for the control-arm median rather than for the reconstructed treatment-effect measure. We therefore evaluate whether the external-control estimate of the control-arm median falls within the published 95% confidence interval,

$$\mathbb{1}\left[L_i^{(0)} \le \widehat{M}_{0,i} \le U_i^{(0)}\right],$$

where $[L_i^{(0)}, U_i^{(0)}]$ is the reported confidence interval for the control-arm median. Confidence-interval agreement is therefore defined at the treatment-effect level for patient-level trials and at the control-outcome level for trial-level trials.

C.4 Metric Implementation

Patient-level cohort metrics are computed using the final external-control weights produced by each evaluated method. The same weighted cohort is used for baseline similarity, outcome recovery, survival-curve estimation, and downstream treatment-effect estimation. For randomized evaluations, the treatment arm is held fixed when comparing the randomized and constructed controls, so differences in downstream metrics arise solely from replacing the comparator. Unless otherwise specified, metrics are computed separately for each trial and then aggregated across trials, preventing larger trials from dominating benchmark-level results. Trials missing the reference quantity required for a metric are excluded only from that metric and remain eligible for the remaining tasks.

D Experimental Setup

D.1 RWD Data Sources

MIMIC-IV. MIMIC-IV v3.1 (Johnson et al., 2024; Pollard et al., 2026) is a publicly available, de-identified EHR resource. It contains approximately 365,000 unique individuals, more than 546,000 hospitalizations, and over 94,000 ICU stays. Its main advantages are clinical depth and reproducibility: it provides diagnoses, procedures, medication orders and administrations, laboratory measurements, microbiology results, and detailed hospital and ICU encounters in a widely used open research dataset.

Because observation is concentrated around hospital encounters, however, outpatient oncology treatment and longitudinal treatment histories are incompletely represented. Later-line treatment landmarks consequently have limited support, and relatively few patients satisfy the full eligibility criteria of a given oncology trial. Candidate pools contain 634–3,000 patients per trial. We therefore use MIMIC-IV as a complementary source for evaluating the framework in a clinically rich, publicly reproducible, but substantially smaller and less longitudinal setting.

MarketScan. MarketScan (Merative, 2022) is one of the largest longitudinal administrative claims resources in the United States. It has included more than 273 million unique individuals since 1995 (Alberto et al., 2023), with data contributed by more than 350 large employers and health plans and coverage spanning all 50 U.S. states. Records link inpatient admissions, outpatient services, and outpatient pharmacy claims and include ICD diagnosis codes, CPT/HCPCS procedure codes, NDC drug codes, enrollment information, and basic demographics. Chemotherapy administrations are captured across inpatient and outpatient settings, enabling reconstruction of treatment histories and lines of therapy.

The data used in this study contain neither laboratory measurements nor dates of death. For outcome estimation, we use duration from the selected time origin to the end of observed enrollment as an event-time proxy, with follow-up termination treated as the read-out event. Consequently, MarketScan outcome estimates characterize weighted follow-up duration rather than directly observed survival and are not interpreted as overall survival. Each trial is associated with a candidate pool of 3,000 patients represented by 47 covariates.

Primary RWD source. MarketScan serves as our primary RWD source because its substantially larger population and broader longitudinal coverage better match the external-control construction setting. Trial-specific eligibility criteria often leave only a small fraction of real-world patients as plausible controls, making source scale particularly important. MarketScan therefore provides a larger and more longitudinally complete candidate pool from which eligible controls can be identified.

D.2 Baselines

We compare against baselines matched to the output required by each task. Cohort-construction baselines produce patient weights or matched external-control cohorts and can therefore be evaluated on baseline similarity and downstream recovery tasks. Outcome-informed baselines additionally access observed donor outcomes and are reported separately from our outcome-free framework. Direct-prediction baselines do not construct an external-control cohort and are evaluated only on downstream quantities that they can predict directly.

Protocol Extraction. Protocol Extraction compares direct prompting across the LLM backbones used in the corresponding experiment, including Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B (Qwen Team, 2026a), Qwen3.8-27B (Qwen Team, 2026b), and GPT-5.5 (OpenAI, 2026). All models receive the same protocol evidence and extraction schema and are evaluated without task-specific fine-tuning. The manually curated reference fields are used only as evaluation ground truth and are not treated as a baseline.

Feature Ascertainability. Feature Ascertainability compares three inference interfaces under the same LLM backbone. For zero-shot decision-tree induction (Knauer et al., 2024), the LLM constructs a decision tree from the target and available feature definitions using prior knowledge without labeled training rows. For TabLLM-style direct prediction (Hegselmann et al., 2023), the observed features of each patient are serialized into text and the LLM directly predicts the masked feature. These interfaces are compared under the same backbone wherever results are reported.

Cohort-construction baselines. The following baselines construct an external-control cohort or patient weights and are used for Control-Arm Similarity and the applicable downstream recovery tasks.

Unweighted donor pool. All eligible candidate controls receive equal weight.

IPTW. Trial-membership propensity scores are estimated from the available baseline covariates using logistic regression, and inverse-probability weights are constructed following the propensity-score framework of Rosenbaum and Rubin (Rosenbaum & Rubin, 1983).

MAIC. Matching-adjusted indirect comparison reweights candidate controls so that their weighted baseline covariate moments align with those of the target population (Signorovitch et al., 2010).

Entropy balancing. Nonnegative donor weights are estimated subject to baseline covariate-balance constraints using entropy balancing (Hainmueller, 2012).

Propensity-score matching. Candidate controls are matched to the target population using nearest-neighbor matching on estimated trial-membership propensity scores.

Sinkhorn OT. Entropy-regularized optimal transport is applied to the aligned baseline representations, with transported donor mass defining external-control weights (Cuturi, 2013).

Fixed-mass partial OT. Partial optimal transport is evaluated using prespecified transported masses rather than adapting the borrowing level to the target trial (Chapel et al., 2020).

Outcome-informed baselines. For Outcome Recovery and downstream evaluations supported by their outputs, we additionally compare against methods that access observed donor outcomes. These methods are reported separately because the corresponding information is unavailable to our outcome-free construction procedure.

Prognostic model / matching. A prognostic model is fitted from baseline covariates to observed donor outcomes. Depending on the evaluation, its predictions are used either to construct prognostically similar controls or to estimate the target control outcome.

TSC. We adapt Targeted Synthetic Control (Wang et al., 2026) to the external-control setting, using observed donor outcomes when estimating the target counterfactual.

Augmented MAIC (aMAIC). Augmented MAIC combines MAIC-style covariate balancing with an outcome-regression correction fitted from observed donor outcomes (Campbell & Remiro-Azócar, 2026).

Direct LLM baselines. We additionally test whether an LLM can predict downstream quantities without constructing an external control. In **Direct LLM prediction**, the model is prompted to return the requested prediction directly. In **Chain-of-Thought prediction**, the model is prompted to reason through the available trial information before producing its final prediction (Wei et al., 2022). Neither baseline receives hidden target-control outcomes. Because these methods do not produce an external-control cohort, they are not evaluated on Control-Arm Similarity or Survival-Curve Recovery.

Task-specific downstream evaluation. Control-Arm Similarity is evaluated only for methods that produce an explicit matched or weighted external-control cohort. Outcome Recovery admits both cohort-construction methods and methods that directly estimate the control outcome. Treatment-Effect Recovery applies the corresponding reconstructed control estimate to the treatment-effect definition of each evaluation level, while Drug-Efficacy Recovery applies the same prespecified efficacy rule to the resulting comparison. Thus, Treatment-Effect and Drug-Efficacy Recovery reuse the applicable baseline outputs rather than introducing separate baseline families.

Survival-Curve Recovery. Survival-Curve Recovery is restricted to the 11 randomized trials with individual time-to-event data. We evaluate methods that produce patient-level external-control cohorts, including the unweighted donor pool, IPTW, MAIC, entropy balancing, propensity-score matching, Sinkhorn OT, and fixed-mass partial OT. Prognostic matching is additionally reported as an outcome-informed comparison. Methods that predict only a scalar control outcome, including GBM-CF and direct LLM prediction, cannot recover a patient-level survival trajectory and are therefore not included in this task.

Table 5: Control-arm similarity across trial- and patient-level evaluations. Values are mean absolute standardized mean differences (age and sex); lower is better. Sinkhorn returns the uniform pool weights.

Method	Trial MIMIC-IV ↓	Trial MarketScan ↓	Patient Claims ↓	Patient MIMIC-IV ↓
TSC	.182	**.216**	.086	**.081**
aMAIC	**.181**	**.216**	.088	.083
Pool	.387	.361	.235	.405
IPTW	.196	.221	**.072**	.089
MAIC	**.181**	**.216**	.088	.096
Entropy balancing	.223	.258	.086	.103
Sinkhorn	.387	.361	.235	.405
Partial OT	.240	.316	.195	.182
PSM	.223	.259	.124	.084
TAEC	**.181**	.217	.084	.085

E Additional Experiments

This section provides additional analyses of the empirical setting and the mechanisms underlying TAEC. We examine the availability of suitable historical controls, component ablations, transfer of the selected time origin to classical estimators, subgroup-level recovery, backbone sensitivity, and empirical diagnostics of the temporal-alignment and transport mechanisms.

E.1 Detailed Recovery Results

We report detailed external-control reconstruction results at both trial and patient levels. Trial-level TSC and aMAIC use criteria-only construction with external prior moments as the target under the same observable-information setting as the other baselines. Patient-level classical baselines on claims are refit after harmonizing sex coding between the treatment and candidate cohorts.

Control-Arm Similarity. Table 5 compares the constructed and randomized control populations using age and sex. At the trial level, TAEC ties the best similarity on MIMIC-IV ($|\mathrm{SMD}| = 0.181$) and remains close to the best result on MarketScan (0.217 versus 0.216). At the patient level, IPTW achieves the lowest SMD on claims (0.072), while TSC achieves the lowest value on MIMIC-IV (0.081). TAEC remains competitive on both sources, with SMDs of 0.084 and 0.085, respectively.

Outcome Recovery. TAEC achieves the lowest control-outcome error across all four evaluations. At the trial level, median APE is 35.0% on the validated-death MIMIC-IV endpoint and 29.8% on MarketScan. TSC and aMAIC achieve strong population similarity but substantially larger outcome error on MIMIC-IV, with APEs of 125.6% and 121.4%, respectively, compared with 88.2% for the unweighted pool. On MarketScan, TSC and aMAIC obtain 39.5% and 40.1%, close to the pool at 40.6%.

At the patient level, TAEC obtains 23.9% APE on claims and 41.8% on MIMIC-IV. After harmonizing sex coding in the claims baselines, the classical methods remain substantially less accurate, with APEs ranging from 53.3% to 88.9%. Thus, the advantage of TAEC persists after correcting the baseline implementation.

Survival-Curve Recovery. Survival-curve recovery is restricted to the 11 randomized trials with individual time-to-event data. TAEC achieves the lowest RMST_{24} error on both donor sources, with 2.54 months on claims and 3.47 months on MIMIC-IV. This extends control recovery beyond a single median to integrated survival over the first 24 months. Trial-level studies are not evaluated because summary outcomes do not provide complete survival trajectories.

Treatment-Effect Recovery. Improved control reconstruction translates into more accurate treatment-effect estimates. At the trial level, TAEC achieves the lowest absolute log-ratio error on both MIMIC-IV and MarketScan, with errors of 0.369 and 0.336, respectively. TSC and aMAIC

Table 6: Control-outcome recovery. Values are median absolute percentage error (APE; %, lower is better). Patient-level medians are evaluable for 10 of the 11 randomized trials. Claims baselines use the harmonized sex encoding.

Method	Trial MIMIC-IV ↓	Trial MarketScan ↓	Patient Claims ↓	Patient MIMIC-IV ↓
TSC	125.6	39.5	54.2	151.0
aMAIC	121.4	40.1	56.2	152.2
Pool	88.2	40.6	47.8	134.6
IPTW	113.1	39.3	53.3	164.1
MAIC	112.2	40.1	56.2	172.7
Entropy balancing	122.1	41.5	61.0	138.6
Sinkhorn	86.3	40.9	48.0	134.6
Partial OT	110.6	40.6	54.1	161.5
PSM	94.8	43.1	88.9	116.2
TAEC	**35.0**	**29.8**	**23.9**	**41.8**

Table 7: Patient-level survival-curve recovery on 11 randomized trials. Values are absolute RMST_{24} error in months; lower is better. Claims baseline values use the harmonized sex encoding.

Method	Claims ↓	MIMIC-IV ↓
TSC	3.72	5.30
aMAIC	4.02	5.21
Pool	3.42	4.38
IPTW	3.54	5.16
MAIC	4.02	4.86
Entropy balancing	4.24	5.26
Sinkhorn	3.42	4.38
Partial OT	3.73	5.31
PSM	4.05	5.67
TAEC	**2.54**	**3.47**

obtain errors of 0.434 on MIMIC-IV and 0.379 and 0.386 on MarketScan, respectively. At the patient level, TAEC achieves $|\Delta \log HR| = 0.470$ on claims and 0.623 on MIMIC-IV, again providing the lowest error across the evaluated methods.

Drug-Efficacy Recovery. TAEC achieves the strongest drug-efficacy recovery across the four evaluations. At the trial level, estimate agreement reaches 0.420 on MIMIC-IV and 0.439 on MarketScan, with corresponding κ values of 0.093 and 0.133. At the patient level, estimate agreement reaches 0.515 on claims and 0.242 on MIMIC-IV, with $\kappa = 0.171$ and 0.089, respectively. These results show that gains in control reconstruction are preserved through downstream treatment conclusions.

Summary. Across the downstream reconstruction tasks, strong covariate balance does not by itself guarantee accurate recovery of the missing control. TSC, aMAIC, and several classical weighting methods achieve highly competitive control-arm similarity, yet their outcome and treatment-effect errors remain substantially larger on several evaluations. TAEC maintains competitive population similarity while achieving the lowest control-outcome, survival-curve, and treatment-effect errors and the strongest efficacy recovery across the patient- and trial-level evaluations.

E.2 Availability of Historical Controls

A central assumption of our zero-shot setting is that a new target trial rarely has a historical study from which its trial-specific external-control construction can be directly reused. We therefore screen 12,467 single-arm oncology trials using progressively stricter comparability requirements derived from Pocock (1976). The results are summarized in Table 10.

Although 99.8% of target trials have at least one earlier trial in the same tumor setting, this high retrieval rate decreases sharply once comparability requirements are imposed. Only 7.0% retain a

Table 8: Treatment-effect recovery. Trial-level columns report absolute log-ratio error, while patient-level columns report $|\Delta \log HR|$; lower is better. Claims baseline values use the harmonized sex encoding.

Method	Trial MIMIC-IV ↓	Trial MarketScan ↓	Patient Claims ↓	Patient MIMIC-IV ↓
TSC	.434	.379	.745	1.088
aMAIC	.434	.386	.714	1.072
Pool	.430	.391	.726	.912
IPTW	.432	.375	.740	1.055
MAIC	.434	.386	.714	1.010
Entropy balancing	.433	.404	.833	1.087
Sinkhorn	.430	.387	.726	.912
Partial OT	.423	.391	.754	1.055
PSM	.452	.431	.830	.983
TAEC	**.369**	**.336**	**.470**	**.623**

Table 9: Drug-efficacy recovery across trial- and patient-level evaluations. EA denotes estimate agreement with the randomized reference, while κ denotes chance-corrected three-class agreement; higher is better. Claims baseline values use the harmonized sex encoding.

	Trial MIMIC-IV		Trial MarketScan		Patient Claims		Patient MIMIC-IV	
Method	EA ↑	κ ↑	EA ↑	κ ↑	EA ↑	κ ↑	EA ↑	κ ↑
TSC	.332	.048	.357	.124	.091	-.028	.091	-.019
aMAIC	.329	.049	.357	.124	.182	.010	.091	-.019
Pool	.341	.041	.344	.117	.091	-.028	.091	-.019
IPTW	.335	.042	.361	.121	.091	-.028	.091	-.019
MAIC	.329	.049	.357	.124	.182	.048	.091	-.019
Entropy balancing	.329	.045	.347	.116	.273	.064	.182	-.089
Sinkhorn	.341	.041	.344	.117	.091	-.028	.091	-.019
Partial OT	.335	.047	.381	.126	.091	-.028	.091	-.019
PSM	.326	.057	.316	.093	.364	.000	.000	-.009
TAEC	**.420**	**.093**	**.439**	**.133**	**.515**	**.171**	**.242**	**.089**

historical candidate after the first four operationalized conditions, and only 0.72% remain after the fifth. Thus, the presence of a disease-matched historical trial is substantially more common than the presence of a trial whose construction can plausibly be transferred to the target setting.

The largest loss occurs when population comparability must be established. Missing baseline reports account for 85.6% of exclusions at the fourth condition, showing that historical evidence often lacks the information required to determine whether two trial populations are sufficiently comparable. This distinction is important for our formulation: historical trials can still provide reusable knowledge about appropriate construction strategies even when they cannot provide a target-specific construction to copy.

These results therefore support the zero-shot setting in a stronger sense than simply showing that historical controls are scarce. In most cases, prior studies exist, but they do not provide a verified trial-specific solution for the unseen target. Construction must instead be inferred from the target protocol, the currently available RWD, and reusable knowledge accumulated across previous studies, without target-control feedback.

E.3 Component Ablations on Claims

We first isolate the contributions of retrieved skills, structured verification, and iterative revision on the patient-level claims evaluation. All variants use Qwen3.5-9B and share the same candidate RWD pool, execution procedure, outcome estimator, and scorer, so that only the planning and refinement components are changed.

We consider three variants. The first removes the skill library while retaining structured verification and iterative revision. The second replaces the structured verifier with LLM self-reflection. The

Table 10: Targets with at least one acceptable historical control, applying Pocock (1976)'s conditions cumulatively over 12,467 single-arm oncology trials registered 2010–2025. Columns restrict candidates to those registered within the stated window.

Conditions satisfied	Any time	$\leq$10 yr	$\leq$5 yr	$\leq$3 yr
1 Same standard treatment	8,205 (65.8%)	8,205 (65.8%)	8,205 (65.8%)	8,205 (65.8%)
+ 2 Recent, same eligibility	5,352 (42.9%)	5,186 (41.6%)	4,731 (38.0%)	4,250 (34.1%)
+ 3 Same outcome assessment	2,648 (21.2%)	2,492 (20.0%)	2,114 (17.0%)	1,742 (14.0%)
+ 4 Comparable patient characteristics	875 (7.0%)	618 (5.0%)	147 (1.2%)	10 (0.08%)
+ 5 Same institution and investigators	90 (0.72%)	73 (0.59%)	13 (0.10%)	0

Table 11: Component ablations on the patient-level claims evaluation. All variants use Qwen3.5-9B and the same execution and scoring procedure. Lower is better for APE, RMST_{24} error, and $|\Delta \log HR|$; higher is better for estimate agreement (EA).

Variant	Skills	Structured Verifier	Time-Zero Evidence	Iteration	APE (%) ↓	RMST_{24} ↓	$\lvert\Delta \log HR\rvert$ ↓	EA ↑	Abstain
Unweighted pool					47.8	3.42	.726	.091	.000
One-shot + time-zero evidence	✓		✓		29.7	3.05	.621	.242	.000
Self-reflection	✓		✓	✓	27.9	2.71	.565	.364	.091
No skill library		✓	✓	✓	25.9	2.58	.537	.455	.000
Full planning + verification	✓	✓	✓	✓	**24.2**	**2.54**	**.471**	**.485**	.091

third uses a one-shot planner that receives the candidate time-zero evidence upfront but performs no verifier-guided revision. This last comparison tests whether the benefit of refinement can be reproduced simply by providing the planner with additional evidence.

Retrieved skills. Removing the skill library degrades all four recovery metrics. Median APE increases from 24.2% to 25.9%, RMST_{24} error from 2.54 to 2.58 months, and treatment-effect error from 0.471 to 0.537, while estimate agreement decreases from 0.485 to 0.455. Because the no-skill variant retains structured verification and iterative revision, the difference isolates the value of reusable construction knowledge supplied to the planner.

Structured verification. Replacing structured verification with self-reflection produces a larger degradation. APE increases to 27.9%, RMST_{24} error to 2.71 months, and $|\Delta \log HR|$ to 0.565, while estimate agreement falls to 0.364. Both configurations allow repeated revision, but only structured verification identifies failures of specific construction elements using protocol compatibility, RWD support, and execution diagnostics. Explicit checks therefore provide more useful revision signals than asking the LLM to reconsider its previous decision generically.

Evidence access versus iterative revision. Providing candidate time-zero evidence directly to a one-shot planner does not recover the performance of iterative refinement. The one-shot configuration obtains 29.7% APE, 3.05 months RMST_{24} error, and 0.621 treatment-effect error, with estimate agreement decreasing to 0.242. Thus, the gain does not come simply from exposing the planner to additional temporal information. It depends on using that evidence to identify unsupported construction elements and revise the plan.

The results are consistent across the four downstream measures: the complete planning and verification procedure performs best, followed by removal of the skill library, self-reflection, and one-shot planning. Retrieved construction knowledge, explicit verification, and iterative revision therefore provide complementary contributions rather than interchangeable sources of improvement.

E.4 Component Ablations on Validated-Death MIMIC-IV

We repeat the component analysis on MIMIC-IV using observed deaths and censoring-aware Kaplan–Meier estimation. The same three changes are evaluated: removing the skill library, replacing structured verification with self-reflection, and replacing iterative refinement with a one-shot planner that receives the candidate time-zero evidence directly.

The component effects are substantially larger on validated-death MIMIC-IV. Removing the skill library increases APE from 41.8% to 72.4%. Replacing structured verification with self-reflection

Table 12: Component ablations on the patient-level MIMIC-IV evaluation using the validated-death endpoint. APE is macro-averaged across the 10 trials with a reference value. Lower is better for APE, RMST_{24} error, and $|\Delta \log HR|$; higher is better for estimate agreement (EA).

Variant	Skills	Structured Verifier	Time-Zero Evidence	Iteration	APE (%) ↓	RMST_{24} ↓	$\lvert\Delta \log HR\rvert$ ↓	EA ↑	Abstain
Unweighted pool					134.6	4.38	.912	.091	.000
One-shot + time-zero evidence	✓		✓		138.4	3.88	.738	**.273**	.061
Self-reflection	✓		✓	✓	183.1	4.12	.842	.182	.030
No skill library		✓	✓	✓	72.4	3.97	.837	.212	.121
Full planning + verification	✓	✓	✓	✓	**41.8**	**3.47**	**.623**	.242	.000

Table 13: Transferring the trial-compatible time-zero adjustment to existing patient-level claims pipelines. Patient selection and donor weights are unchanged. Δ denotes the paired change in macro APE over the nine trials for which the alternative start of follow-up can be constructed; negative values indicate improvement.

Pipeline	Macro-10 Original	Macro-9 Original	Macro-10 + Time Zero	Macro-9 + Time Zero	Δ Macro-9	95% CI	Later-line: Original → + Time Zero
Pool	47.8	43.8	25.6	26.8	-17.0	$[-41.4, 8.3]$	71.2 → 26.7
PSM	115.4	124.5	53.2	57.2	-67.3	$[-182.4, 0.0]$	176.9 → 52.6
IPTW	52.4	50.3	22.4	24.2	-26.2	$[-61.9, 2.8]$	82.1 → 22.0
MAIC	73.2	77.7	40.9	43.5	-34.1	$[-79.3, 0.0]$	97.4 → 32.9
Entropy balancing	51.7	48.0	21.0	21.7	**-26.3**	$[\mathbf{-50.4}, \mathbf{-5.5}]$	82.0 → 20.7
Sinkhorn OT	48.0	44.0	25.5	26.8	-17.2	$[-42.0, 8.2]$	71.7 → 26.8
Partial OT	49.3	45.5	24.2	25.3	-20.2	$[-44.3, 1.4]$	76.4 → 26.1
Prognostic matching†	29.8	29.5	30.4	31.8	+2.3	$[-14.9, 19.9]$	38.0 → 39.2

increases it to 183.1%, while one-shot planning with the same time-zero evidence reaches 138.4%. The same pattern extends beyond the control median. The complete planning and verification procedure obtains an RMST_{24} error of 3.47 months and $|\Delta \log HR| = 0.623$, compared with 3.97 and 0.837 without skills, 4.12 and 0.842 with self-reflection, and 3.88 and 0.738 with one-shot planning.

Estimate agreement is less discriminative on the 11-trial set: the complete procedure obtains 0.242 while the one-shot planner obtains 0.273, despite large differences in their continuous recovery errors. Together, the MIMIC-IV results reinforce the claims ablation: reusable construction knowledge is most effective when combined with explicit verification and iterative revision, and providing the relevant evidence without this refinement process is insufficient.

E.5 Time-Zero Transfer

We next test whether the time-zero choice contributes independently of the downstream estimator. For each classical pipeline, we keep its selected patients and donor weights unchanged and modify only the start of follow-up used for outcome read-out.

In the claims data, follow-up is originally represented relative to a default RWD start point. For trials whose protocol requires follow-up from a later treatment stage, we construct the corresponding trial-compatible start from the available treatment history. We denote the resulting adjustment by $r^\star$, defined as the ratio between median follow-up from the trial-compatible start and median follow-up from the original RWD start among the same patients. Applying $r^\star$ therefore changes the temporal alignment of the outcome while leaving patient selection and weighting unchanged.

Among the 10 trials with an evaluable control median, this alternative follow-up start can be constructed for nine. We report results on all 10 trials where possible and separately report the paired comparison over these nine trials.

Changing only the start of follow-up substantially improves nearly every outcome-free pipeline. On the nine paired trials, APE decreases from 50.3% to 24.2% for IPTW, from 48.0% to 21.7% for entropy balancing, from 45.5% to 25.3% for partial OT, and from 43.8% to 26.8% for the unweighted pool. PSM and MAIC also show large reductions. The paired confidence interval excludes zero for entropy balancing and reaches zero for PSM and MAIC, while the remaining whole-trial comparisons have wider intervals because only nine paired trials are available.

The effect is especially large in later-line trials. APE decreases from 82.1% to 22.0% for IPTW, from 82.0% to 20.7% for entropy balancing, and from 76.4% to 26.1% for partial OT. These trials begin follow-up at a later treatment stage, making correct selection of the follow-up start particularly important.

Table 14: Time-zero transfer across 40 patient-level claims strata. Patient selection and donor weights are unchanged; only the start of follow-up is modified. Negative Δ indicates lower error after applying the trial-compatible time zero.

Pipeline	Original	+ Time Zero	Δ	95% CI	Better / Worse
Pool	45.5	29.9	-15.6	$[-33.3, -0.7]$	14 / 7
Sinkhorn OT	45.9	29.9	-16.0	$[-33.8, -1.0]$	13 / 7
Partial OT	51.6	27.9	-23.7	$[-38.2, -11.2]$	14 / 6
IPTW	52.6	27.2	-25.4	$[-44.4, -8.5]$	14 / 6
PSM	75.9	40.9	-35.0	$[-81.8, -3.6]$	11 / 9
MAIC	88.4	44.6	-43.8	$[-90.4, -11.1]$	12 / 8
Entropy balancing	94.7	39.2	**-55.4**	$[\mathbf{-92.9}, \mathbf{-24.1}]$	16 / 4
Prognostic matching†	48.6	40.3	-8.3	$[-19.1, 1.7]$	12 / 8

Prognostic matching behaves differently: its macro-9 APE changes from 29.5% to 31.8%. Unlike the other pipelines, this baseline directly models donor outcomes, so changing the follow-up start does not provide the same improvement.

Because patient selection, donor weights, and downstream estimators are fixed, this experiment isolates the effect of the time-zero element. The same trial-compatible follow-up start improves pipelines based on propensity weighting, matching, moment balancing, and optimal transport, showing that time-zero selection contributes independently of the choice of weighting method.

E.5.1 Time-Zero Transfer Across Clinical Strata

We further test the same time-zero intervention within 40 clinical strata. For each pipeline and subgroup, the original selected patients and donor weights are retained; only the start of follow-up is changed using the trial-compatible time-zero adjustment defined above. We report macro APE over the available strata together with paired bootstrap confidence intervals.

The subgroup results strengthen the whole-trial finding. Applying the trial-compatible time zero reduces macro APE for every outcome-free classical pipeline, and all corresponding paired confidence intervals exclude zero. The reduction ranges from 15.6 percentage points for the unweighted pool to 55.4 points for entropy balancing. Partial OT, IPTW, PSM, and MAIC improve by 23.7, 25.4, 35.0, and 43.8 points, respectively.

The improvement also occurs across many individual strata rather than being driven by a small number of extreme cases. Entropy balancing improves 16 strata and worsens four, while partial OT and IPTW each improve 14 and worsen six. Thus, the effect of time-zero selection persists when recovery is evaluated within clinically distinct subgroups.

These results separate temporal alignment from population adjustment. Methods based on propensity weighting, matching, moment balancing, and optimal transport produce substantially different donor weights, yet all improve when the same trial-compatible follow-up start is applied. Time-zero selection therefore captures a source of reconstruction error that cannot be addressed by reweighting patients alone.

The outcome-informed prognostic baseline again shows a smaller change, from 48.6% to 40.3%, with a confidence interval that includes zero. This is consistent with its different use of donor outcomes in the estimation step.

E.6 Stratified Evaluation

Whole-trial outcome recovery may hide systematic errors within clinically distinct subpopulations. We therefore repeat the evaluation within available strata defined by sex, age, and performance status, asking whether a construction that recovers the overall control outcome also preserves recovery within clinically relevant subgroups.

On claims, 40 evaluable strata are available. The Qwen3.5-9B TAEC loop obtains a macro APE of 33.4%, improving over the static planner (46.9%) and pool fallback (45.7%). The deterministic rule

Table 15: Stratified control-outcome recovery using Qwen3.5-9B. Values are macro APE across evaluable strata (claims: 40; MIMIC: 41); lower is better. Abstention denotes the fraction of strata for which the TAEC loop returns no supported estimate.

Method	Claims ↓	MIMIC ↓	Abstention (claims / MIMIC)
TAEC loop	33.4	**32.8**	0.19 / 0.15
Deterministic rule	**28.3**	40.9	–
Static planner	46.9	40.9	–
Pool median (fallback)	45.7	44.6	–

performs best at 28.3%, indicating that adaptive refinement is not uniformly superior once evaluation is restricted to smaller clinical strata.

On MIMIC, across 41 evaluable strata, the TAEC loop achieves the lowest macro APE at 32.8%, compared with 40.9% for both the deterministic rule and static planner and 44.6% for the pool fallback. The reversal in relative performance across the two RWD sources suggests that subgroup recovery depends not only on the construction strategy but also on how well the source represents the characteristics defining each subgroup.

The loop abstains on 19% of claims strata and 15% of MIMIC strata. Finer stratification reduces the number of compatible RWD patients, making some trial-specific constructions unsupported under the same requirements used for the full trial. Performance-status strata are particularly challenging because the corresponding clinical measurement is not directly observed in the RWD and must instead be represented indirectly.

Overall, the stratified evaluation provides a stricter test than whole-trial recovery: TAEC improves substantially over static construction and the unweighted fallback, but aggregate recovery does not guarantee uniformly accurate reconstruction across all clinical subgroups.

F Theoretical Analysis

Setting. For a target trial, let Y_i denote candidate i's follow-up measured from the source time origin, and let R_i denote follow-up measured from the trial-compatible enrolment landmark. These quantities describe the same endpoint under different time origins; re-indexing follow-up does not establish that the endpoint itself is a valid measure of survival.

Let $w_i \geq 0$ be normalized donor weights with $\sum_i w_i = 1$, and let $Q_q(Y; w)$ denote the weighted empirical q-quantile:

$$Q_q(Y; w) = \inf \left\{ u : \sum_i w_i \mathbf{1}\{Y_i \leq u\} \geq q \right\}, \qquad q \in (0, 1).$$

We analyze the estimator

$$\widehat{\theta} = s\, Q_{1/2}(Y; w), \qquad s > 0,$$

where $\theta^* > 0$ is the randomized-control reference median. All medians appearing in ratios below are assumed positive. The analysis applies to this multiplicatively calibrated median estimator, rather than to every possible outcome estimation procedure.

For a landmark cohort $\mathcal{P}_\ell$ on which both time origins can be represented, define

$$r_\ell = \frac{Q_{1/2}(R; \mathcal{P}_\ell)}{Q_{1/2}(Y; \mathcal{P}_\ell)}. \tag{7}$$

Setting $s = r_\ell$ aligns the two medians within this landmark cohort. It does not automatically align medians under different donor weights or in another population.

Quantile scaling.

Lemma 1 (Positive scaling of quantiles)**.** *For a random variable Y and a constant $a > 0$,*

$$Q_q(aY) = aQ_q(Y).$$

For empirical observations, the corresponding identity

$$Q_q(aY; w) = aQ_q(Y; w)$$

holds when each observation is multiplied by a and the weights are unchanged.

Proof. For the population distribution, $F_{aY}(u) = F_Y(u/a)$. Substituting this identity into the definition of the quantile gives

$$\inf\{u : F_Y(u/a) \geq q\} = a \inf\{v : F_Y(v) \geq q\}.$$

The weighted empirical distribution satisfies the same identity because $\mathbf{1}\{aY_i \leq u\} = \mathbf{1}\{Y_i \leq u/a\}$. □

Error decomposition. The estimation ratio admits the exact factorization

$$\frac{\widehat{\theta}}{\theta^*} = \underbrace{\frac{s\, Q_{1/2}(Y; w)}{Q_{1/2}(R; w)}}_{\kappa_{\text{clock}}} \cdot \underbrace{\frac{Q_{1/2}(R; w)}{\theta^*}}_{\kappa_{\text{pop}}}. \tag{8}$$

Consequently,

$$\log \frac{\widehat{\theta}}{\theta^*} = \log \kappa_{\text{clock}} + \log \kappa_{\text{pop}}.$$

The first factor captures residual time-origin mismatch under the selected weights and scale. The second captures the remaining discrepancy between the re-indexed donor median and the target reference. We call it the population factor, although it may also contain endpoint-measurement differences and other residual discrepancies. Both factors can depend on w.

The decomposition does not imply that absolute total error is bounded below by the absolute clock error: the signed log terms may offset one another. A conditional lower bound requires control of the second term.

Proposition 1 (Conditional lower bound)**.** *If $|\log \kappa_{\text{pop}}| \leq \epsilon$, then*

$$\left|\log \frac{\widehat{\theta}}{\theta^*}\right| \geq \max\left\{0,\ |\log \kappa_{\text{clock}}| - \epsilon\right\}.$$

Proof. Apply the reverse triangle inequality to Equation 8 on the logarithmic scale. □

Thus, when the re-indexed donor median is already close to the target reference, a substantial residual time-origin error cannot be removed through cancellation by a small population term. This is a conditional statement, not a weight-independent lower bound.

Bounded time-origin distortion. Time-origin changes need not be multiplicative: a landmark shift can instead subtract a patient-specific interval. The following result therefore requires an explicit approximation assumption.

Proposition 2 (Clock-factor bound)**.** *Suppose $Y_i \geq 0$ and there exist $r > 0$ and $\delta \in [0, 1)$ such that*

$$r(1 - \delta)Y_i \leq R_i \leq r(1 + \delta)Y_i$$

for every candidate with positive weight. Then

$$\frac{s}{r(1 + \delta)} \leq \kappa_{\text{clock}} \leq \frac{s}{r(1 - \delta)}.$$

In particular, choosing $s = r$ gives

$$|\log \kappa_{\text{clock}}| \leq -\log(1 - \delta).$$

Proof. Weighted quantiles preserve pointwise ordering. Together with Lemma 1, this yields

$$r(1 - \delta)Q_{1/2}(Y; w) \leq Q_{1/2}(R; w) \leq r(1 + \delta)Q_{1/2}(Y; w).$$

Dividing $sQ_{1/2}(Y; w)$ by these positive bounds proves the first claim. When $s = r$, the log factor lies between $-\log(1 + \delta)$ and $-\log(1 - \delta)$. The latter has the larger absolute magnitude. □

More generally,

$$|\log \kappa_{\text{clock}}| \le \left|\log \frac{s}{r}\right| - \log(1-\delta).$$

This separates scale-selection error from departures from the multiplicative approximation. Combined with $|\log \kappa_{\text{pop}}| \le \epsilon$, it gives

$$\left|\log \frac{\widehat{\theta}}{\theta^*}\right| \le \left|\log \frac{s}{r}\right| - \log(1-\delta) + \epsilon.$$

The empirical landmark ratio r_ℓ is a candidate calibration scale. Applying this bound with $r = r_\ell$ requires the distortion assumption to hold on the positively weighted candidates. Its definition as a ratio of landmark-cohort medians does not establish that assumption.

Predicting a calibration scale.

Proposition 3 (Optimal prediction under absolute log-scale loss)**.** *Let $r > 0$ be a random calibration ratio with $\mathbb{E}|\log r| < \infty$, and let Z denote information available at construction time. If $m(Z)$ is a conditional median of $\log r$ given Z, then*

$$f^*(Z) = \exp(m(Z))$$

minimizes expected absolute log-scale loss among positive measurable predictors. In particular, for any such f,

$$\mathbb{E}\,|\log r - \log f(Z)| \ge \mathbb{E}\,|\log r - m(Z)|\,.$$

Proof. For each $Z = z$, a conditional median minimizes $\mathbb{E}[|\log r - c| \mid Z = z]$ over constants c. Apply this fact to $c = \log f(z)$ and then take expectations over Z. □

This proposition characterizes an optimal predictor under a specified loss. It does not establish that the empirical landmark ratio is that predictor. Within $\mathcal{P}_\ell$, choosing $s = r_\ell$ gives the algebraic identity

$$s\, Q_{1/2}(Y; \mathcal{P}_\ell) = Q_{1/2}(R; \mathcal{P}_\ell).$$

Generalizing this calibration beyond the landmark cohort requires additional assumptions about population and time-origin compatibility.

Weight concentration under partial transport.

Lemma 2 (Effective sample size under uniform donor capacities)**.** *Consider a feasible partial-transport plan over m candidate donors with total transported mass $\mu \in (0, 1]$. Suppose donor j receives mass a_j, where*

$$0 \le a_j \le \frac{1}{m}, \qquad \sum_{j=1}^{m} a_j = \mu,$$

and define normalized weights $w_j = a_j/\mu$. Then

$$\text{ESS}(w) = \frac{1}{\sum_j w_j^2} \ge \mu m.$$

Equality holds if and only if every positively weighted donor receives its full capacity $1/m$.

Proof. The capacity constraint implies $w_j \le 1/(\mu m)$. Therefore,

$$\sum_j w_j^2 \le \frac{1}{\mu m} \sum_j w_j = \frac{1}{\mu m}.$$

Taking reciprocals proves the bound. Equality requires every positive w_j to equal $1/(\mu m)$, equivalently $a_j = 1/m$. □

This guarantee depends on uniform donor capacities, fixed transported mass, and the stated normalization. It does not directly extend to subsequent weight transformations that violate these constraints. The bound limits weight concentration; it does not guarantee covariate balance, source validity, or accurate outcome recovery.

Table 16: Empirical checks for the theoretical analysis using the Qwen3.5-9B mainline configuration on the claims source. $r^\star$ denotes the protocol-matched time-origin ratio, with $r^\star = 1$ for first-line and adjuvant trials in this diagnostic. Correlations use signed log errors. Hidden reference outcomes are used only for retrospective evaluation. Partial-OT diagnostics are backbone-independent. These results assess consistency with the proposed mechanisms but do not verify the theoretical assumptions.

Diagnostic	Statistic	Value
Time-origin association	Corr. of pool log error and $-\log r^\star$ (10 trials)	0.64
	Corr. of IPTW / Sinkhorn / partial OT with pool error	.99 / 1.00 / .99
	Mean $\lvert \log \kappa_{\mathrm{pop}} \rvert$	0.31
Scale diagnostics	Numeric interface: mean / median $\lvert \log s - \log r^\star \rvert$	0.60 / 0.50
	Later-line trials with $s > 1$	0.87
	Across-definition MAD of $\log r$	0.36
Transport constraints	Executed designs	583
	Min. / median $\mathrm{ESS}/(\mu m)$	.999 / 1.001
	Max. selected-candidate count / (μm)	1.065
Refinement ordering	Pairs violating hard-constraint ordering	0 / 313
	Maximum soft component of J	7.3
	Hard-free design visited $\rightarrow$ hard-free returned	109 / 109
	$J(d^\star) \leq J(d_0)$; strictly lower	129/129; 93/129
Time-axis diagnostics	Hard time-axis failure | clock error $> 1.5\times$	1.00
	No time-axis failure | clock error $\leq 1.5\times$	0.76
	Clock error $> 1.5\times$: one-shot $\rightarrow$ returned	.38 $\rightarrow$ .07