EFFICIENT GENERATIVE PREDICTION FOR EHR FOUNDATION MODELS: THE SCOPE AND REACH ESTIMATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative inference for foundation models trained on tokenized electronic health records (EHR) estimates the risk of clinical outcomes through the autoregressive prediction of future patient timelines. Three practical limitations constrain the utility of current approaches for clinical generative prediction: sparse estimate distributions poorly differentiate patient risk levels, extreme computational requirements make inference both expensive and high-latency, and high sampling variance produces wildly different risk estimates purely by chance. We address these challenges by proposing two new estimators that leverage next-token probability distributions underutilized by standard Monte Carlo: the Sum of Conditional Outcome Probability Estimator (SCOPE) and Risk Estimation from Anticipated Conditional Hazards (REACH). We present formally verified proofs that both are unbiased, that REACH guarantees variance reduction over Monte Carlo for any model and outcome, and that REACH is a Rao-Blackwellization of any naive importance sampling scheme that preserves the non-outcome token distribution. Across 11 clinically important outcomes in MIMIC-IV and local EHR data, SCOPE and REACH match 100-sample Monte Carlo accuracy with a median of $2.5\times$ to $3.4\times$ fewer generated tokens and more than $80\times$ fewer for the rarest outcomes, with calibration preserved throughout. Because SCOPE reuses a single pool of sampled timelines across an arbitrary number of outcomes at no marginal generation cost while REACH provides a per-task guaranteed variance reduction when computed on the same number of timelines, the two estimators are complementary in deployment and meaningfully reduce the inference cost for generative EHR foundation model prediction, especially for the rare, high-impact outcomes common in healthcare.

1 INTRODUCTION

Foundation models pre-trained on sequences of tokenized electronic health records (EHR) data have shown performance and versatility across a range of downstream predictive tasks (1; 2; 3; 4; 5; 6). In this framework, the patient's interactions with healthcare are converted from their operational representation (tabular EHR databases) into a sequence of tokens that describe medical events in the patient's timeline, such as labs, vitals, and prescriptions (7; 8; 9). After being trained on the self-supervised task of predicting the next token in a patient's timeline, these models can autoregressively generate multiple potential future timelines for a patient. This pool of simulated timelines can then be used to predict an arbitrary number of downstream outcomes by using the proportion of simulated timelines in which the model generates the outcome of interest as a risk estimate (5; 10; 11; 6; 12; 13; 14; 15; 16, Figure 1).

Inference by simulating future timelines is typically substantially more computationally demanding than traditional supervised prediction methods, limiting the adoption of the technique. In additional experiments reproducing the pipeline of an existing paper (Appendix I), performing inference to generate future timelines required an estimated $1500+$ A100 GPU hours (11). This could be a major barrier to prospective clinical deployment, particularly in acute settings such as the intensive care unit (ICU) where risk estimates may need to be refreshed frequently and with low latency. Without

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.

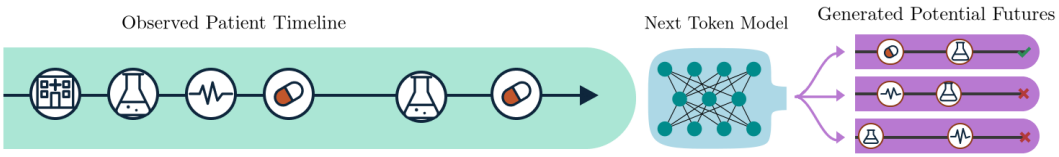


Figure 1: Visualization of the generative outcome prediction framework.

reducing inference-time requirements, the method may be limited to offline or low-frequency use rather than real-time decision support.

Due to this computational expense, current implementations are limited in the number of future timelines they can sample per patient. This is especially problematic because the granularity of this estimator is driven by the number of sampled timelines. Monte Carlo (MC) sampling risk resolves only to multiples of 1/n meaning it will estimate the risk of patients with very different true risks between the same two lowest possible risk scores. For an outcome with a prevalence of 1/10,000, a Monte Carlo estimator with 100 sampled futures will fail to differentiate between two patients where one has a 10-times greater risk over 90% of the time (see Appendix O). This arithmetic limitation is misaligned with clinical needs: the most consequential decisions are often driven by low-prevalence but high-impact outcomes, where even small absolute differences in risk can meaningfully change diagnostic and therapeutic actions (e.g. evaluating chest or back pain for aortic dissection (17)).

In response to these challenges, we propose two ways of converting generated patient futures into outcome probability estimates that decouple the granularity of the estimator from the number of sampled timelines, allowing for accurate prediction of rare outcomes using much smaller computational budgets: Sum of Conditional Outcome Probability Estimates (**SCOPE**) and Risk Estimation from Anticipated Conditional Hazards (**REACH**). In particular, SCOPE can be computed for any number of outcomes at exactly the same token generation cost as traditional MC estimation. REACH, on the other hand, requires a small task specific generation cost proportional to the prevalence of the outcome of interest. The task-marginal generation cost of REACH is therefore lowest for the rarest outcomes, where these methods provide the greatest increases in predictive accuracy. REACH is also demonstrated formally to reduce estimator variance over both the Monte Carlo and SCOPE methods, guaranteeing a more precise estimator given an equal number of samples.

We summarize our contributions as follows:

- **Continuous risk scoring** *(SCOPE & REACH).* We leverage next-token probability distributions to produce granular, continuous risk scores per timeline, resolving estimator dispersion and enabling accurate rare-outcome prediction with substantially fewer sampled futures.
- **Theoretical grounding** *(SCOPE & REACH).* We provide formal derivations, verified proofs of unbiasedness, and closed-form characterization of estimator variance.
- **Fixed-cost multi-outcome estimation** *(SCOPE).* A single pool of sampled futures suffices to compute SCOPE estimates for an arbitrary number of outcomes simultaneously, with no additional token generation cost beyond what is required for a standard MC estimate.
- **Guaranteed variance reduction** *(REACH).* With equal numbers of sampled timelines, REACH provably reduces variance relative to both MC and SCOPE across all model-task combinations, at a marginal inference cost proportional to outcome prevalence, with rarer outcomes incurring the least cost.

1.1 Problem Statement

In a standard workflow, electronic health records are represented as token sequences. Each sequence begins with demographic tokens, followed by chronologically ordered tokens for medical events such as: vitals, labs, medications, and procedures. Critically, these are **not** language models and all models presented are trained from scratch with random initializations. Each token describes a medical event, not letters, words, or subwords. A model $\mathbb{P}$ is then trained on these sequences using a standard language modeling objective (next-token prediction).

At inference time, given a partial sequence, the model conditions on all observed tokens up to a prediction time and autoregressively samples $n \geq 1$ possible continuation sequences $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$.

Let O denote a token corresponding to an event of interest. For a sequence X, define $T_O(X)$ as the first index at which O appears, or ∞ if it does not occur. Similarly, let $T_E(X)$ denote the first index at which the timeline is considered "terminated". This can occur for a number of reasons, an outcome time window has elapsed, a timeline ending token (hospital discharge or mortality) has been generated, or other conditions. We assume sequences are finite ($T_E(X) < \infty$) and that $T_E \neq T_O$; we prevent this by setting $T_E = T_O + 1$ in the event that they would otherwise coincide.

Our goal is to efficiently estimate $\mathbb{P}_{X\sim P}(T_O(X) < T_E(X))$: the probability that a given next token prediction model will sample the outcome of interest before sampling some terminating condition. Note that this probability is not the patient's true future risk of experiencing the outcome. Although we will use predictive performance as a metric to compare each method at different levels of sampling or computational budget, our empirical analysis is limited to analyzing how **efficient** each method is at converting inference to an estimate of the above probability. As a result, supervised outcome prediction methods, which instead are trained to predict the patient's true risk directly, will not be included in this analysis. See Appendix C for collected notation.

2 Description of SCOPE and REACH

SCOPE

$$\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{t=1}^{\min\{T_E(X^{(i)}),T_O(X^{(i)})\}} \mathbf{P}(X_t{=}O \mid X^{(i)}_{1:t-1})\right]$$

Algorithm 1 SCOPE $(\mathcal{O} = \text{⚠})$

$X_{1:i}$: Observed patient timeline
$p \leftarrow 0$
do
  ▷ + prob of outcome at next token
  $p \leftarrow p + \mathbb{P}(X_{i+1} = \mathcal{O} \mid X_{1:i})$
  ▷ Sample a next token from the distribution
  $X_{i+1} = \textbf{Sample}(\mathbb{P}(X_{i+1} \mid X_{1:i}))$
  $i \leftarrow i + 1$
▷ While no outcome and not terminated
while not isFinished$(X_{1:i}) \wedge (X_i \neq \mathcal{O})$
return p

Line	$X_{1:i}$	Next token distribution	p
1–4	🏥🩺	.20	$0 + .20 = \mathbf{0.20}$
5–6	🏥🩺⚠	.20	0.20
7–8	🏥🩺⚠	$X_i =$ ⚠: exit	**0.20**

Legend

- 🏥 Hospital admission
- 🩺 Observation
- 💊 Medical intervention
- ⚠ Adverse outcome $\mathcal{O}$
- 💀 Mortality
- Observed timeline
- $\mathbb{P}$(Outcome $\mathcal{O}$ = ⚠)
- Sampled token (🎲)
- Discharge

REACH

$$\mathcal{R} = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \prod_{t=1}^{T_E}(1 - \mathbf{P}(X_t{=}O \mid \hat{X}^{(i)}_{1:t-1}))\right]$$

Algorithm 2 REACH $(\mathcal{O} = \text{⚠})$

$X_{1:i}$: Observed patient timeline
  ▷ p starts out as the probability of **no** outcome
$p \leftarrow 1$
do
  ▷ × by prob of no outcome at next token
  $p \leftarrow p \cdot \mathbb{P}(X_{i+1} \neq \mathcal{O} \mid X_{1:i})$
  ▷ Sample from no outcome distribution
  $X_{i+1} = \textbf{Sample}(\mathbb{P}_{\varnothing}(X_{i+1} \mid X_{1:i}))$
  $i \leftarrow i + 1$
▷ While not terminated
while not isFinished$(X_{1:i})$
▷ Take complement to get prob of the outcome.
$p \leftarrow 1 - p$
return p

Line	$X_{1:i}$	Next token distribution	p
1–4	🏥🩺	$1 - .20 = .80$	$1 \times .80 = \mathbf{0.80}$
5–6	🏥🩺💊	0	0.80
7	🏥🩺💊	not finished	0.80
4	🏥🩺💊	$1 - .10 = .90$	$.80 \times .90 = \mathbf{0.72}$
5–6	🏥🩺💊🚪	0	0.72
7–9	🏥🩺💊🚪	🚪 finished: exit	$1 - .72 = \mathbf{0.28}$

Figure 2: Descriptions of SCOPE and REACH in equations and algorithms with step by step examples.

2.1 SUM OF CONDITIONAL OUTCOME PROBABILITY ESTIMATES (SCOPE)

The Sum of Conditional Outcome Probability Estimates method, referred to as $\mathcal{S}$ in equations, begins with the same process as MC: our model generates a pool of potential future timelines. However, where the above method only uses the next token distribution to sample, this method makes additional use of the distribution by taking a sum of the probabilities that the next token is the outcome of interest token to compute the estimator. This sum terminates when either the timeline ends or the outcome of interest token is sampled. By summing the next token probabilities of all relevant outcomes, an often negligible cost, this estimator can be computed using the same pool of sampled timelines as MC.

Since each sub-estimator is a sum of next token probabilities, the estimator produces probabilities with much higher resolution than those of MC. However, since the values for this estimator are a sum of raw probabilities, it is possible for the estimator to produce a value above 1. In situations where this is more probable, the variance of this estimator can exceed the variance of the MC method. If SCOPE predictions are not clipped to fall in the $[0, 1]$ range, they are not guaranteed to be valid probability estimates. On the other hand, if they are clipped to this range, SCOPE is no longer guaranteed to be an unbiased estimator. In our later experiments, we elected to clip the averaged estimator to the $[0, 1]$ range. With 100 sampled futures per observed timeline, this clip only activated on 0.016% of all observed estimates.

2.2 RISK ESTIMATION FROM ANTICIPATED CONDITIONAL HAZARDS (REACH)

Unlike the two above methods, Risk Estimation from Anticipated Conditional Hazards, which is referred to as $\mathcal{R}$ in equations, samples timelines where the outcome of interest token is **excluded** from the pool of next token candidates. These timelines are generated by setting the next token probability for the outcome to zero and re-normalizing the next token distribution, while simultaneously tracking what the probability of this outcome was before it was zeroed out. After this timeline has been generated, we perform a coin-flip at each newly generated token with probability of heads equal to the pre-zeroed outcome probability. It is shown in Appendix E that the probability of at least one coin-flip coming up heads is exactly equal to the probability of the patient experiencing the outcome of interest under the standard generation method. Critically, since we know the probabilities for each coin-flip, we can explicitly compute the probability of generating at least one head, integrating out all variance created by sampling whether or not this event occurs. REACH is, therefore, formally a Rao-Blackwellization in this alternative two-step probability space (Appendix P). See Figure 2 for a visual comparison of SCOPE and REACH.

As a result, REACH is an unbiased estimator with lower variance than both of the techniques described above. The primary limitation of this method is that, unlike MC and SCOPE, REACH requires some additional inference to be performed on a task by task basis. Since the outcome of interest is prohibited at each step in REACH timelines, starting from a pool of normally generated timelines, REACH will require all timelines that contain the outcome of interest to be re-completed starting just before the outcome of interest first appears, implying that the marginal cost of running REACH on additional outcomes is proportional to the probability of the outcome of interest. Note that this cost is also determined by the number of tokens required to terminate timelines in the outcome avoiding distribution, which can vary by outcome. Fortunately, this implies that the outcome-marginal computational cost of REACH is lowest for rarer outcomes, where REACH offers the greatest potential reductions in variance.

3 RELATED WORK

Existing statistical methods, namely importance sampling and splitting, have been used historically to address similar difficulties in rare event probability estimation (18). While potentially valuable for single outcome prediction, these approaches undermine the key appeal of generative foundation model based outcome prediction: they depend on the selection of an outcome of interest prior to future timeline sampling and therefore require sampling to occur separately for each outcome of interest. As a result, a key consideration for our problem is to ensure that our methods are able to re-use timelines or other information generated to make predictions across multiple outcomes of interest. Notably, SCOPE avoids these issues entirely by using the same pool of sampled future timelines to predict for an arbitrary number of outcomes. REACH, on the other hand, is able to

re-use all sampled tokens that occur prior to the generation of the outcome of interest, meaning the additional token generation required is proportional to the probability of the outcome of interest. Additionally, it is shown in Appendix Q that REACH is in fact a Rao-Blackwellization, and thus a variance reduction, of any importance sampling scheme that does not alter the non-outcome next token distribution.

Conditioning and other variance-reduction techniques have a rich history in Monte Carlo estimation. See Asmussen & Glynn (19) for an introduction. REACH can be viewed as importance sampling applied to SCOPE (see Appendix H) with proposal distribution $\hat{p}$. In the setting of Markov chains, McKeague and Wefelmeyer (20) introduced an estimator somewhat similar to SCOPE in spirit and showed that, for reversible Markov chains, it improves asymptotic variance. Previous research indicates that token sequences derived from EHRs tend to be highly non-Markovian, however, with long-range dependencies, interspersed repeating tokens, and increasing perplexity (8). It should also be mentioned that SCOPE is neither a form of conditional MC nor a Rao-Blackwellization, as there are cases in which it does not reduce variance (see Appendix J for one such example). In contemporaneous work, Amar et al. (21) introduce path-computing probabilities that score the probability of a target token along a sequence. Their approach is reminiscent of REACH; however, the authors do not appear to use an outcome-avoiding distribution or consider statistical properties of their estimator.

4 PROOFS

4.1 FORMAL VERIFICATION

The proofs that SCOPE and REACH are unbiased estimators, that REACH is a Rao-Blackwellization of MC, and consequently that REACH necessarily reduces variance compared to MC have been formally verified in Lean 4 and registered on Palomar (link redacted for anonymity).

4.2 DERIVATION AND UNBIASEDNESS OF $\mathcal{S}$ (SCOPE)

Theorem 1. *SCOPE is an unbiased estimator for the true probability that the outcome of interest is sampled.*

Proof. Note that the probability of the outcome of interest occurring is equal to the expectation of $\mathbb{1}_{\{T_O(X)<T_E(X)\}}$, which is the indicator random variable (r.v.) describing the event where the outcome of interest comes before the end of the timeline. We can rewrite this r.v. as a sum of random variables indexed over when the outcome of interest first appears: $\sum_{t=1}^{\infty} \mathbb{1}_{\{T_O(X)=t\}} \mathbb{1}_{\{T_E(X)>t-1\}}$. Note that for all terms where either the timeline has ended ($t > T_E(X)$) or the outcome has already appeared $t > T_O(X)$, the indicator variable is zero, allowing us to terminate our sum at the index $\min\{T_E(X), T_O(X)\}$. The expectation of this sum, which we recall is equal to the probability of the outcome of interest occurring, is then by the linearity of expectation:
$\sum_{t=1}^{\min\{T_E(X),T_O(X)\}} \mathbb{P}(X_t = O|X_{1:t-1})$, which is our exact equation for a SCOPE sub-estimator. Thus, we have shown that SCOPE is unbiased by deriving it from the exact probability we want to estimate. A more detailed rigorous proof is provided in Appendix D. □

4.3 CHARACTERIZATION OF THE VARIANCE OF $\mathcal{S}$

Theorem 2. *The variance of the SCOPE estimator minus the variance of the MC estimator is equal to* $\frac{1}{n}\mathbb{E}\big[S(S-1)\big]$ *(where S is a SCOPE sub-estimator)*

Proof. By the fact that each sample is i.i.d., $\mathrm{Var}(\mathcal{S}) = \mathrm{Var}(S)/n$. Therefore by the variance of a Monte Carlo estimator, defining $\mu = \mathbb{E}(\mathcal{S}) = \mathbb{E}(M_0)$:

$$\mathrm{Var}(\mathcal{S}) - \mathrm{Var}(M_0) = \frac{\mathrm{Var}(S) - \mu(1-\mu)}{n} = \frac{\mathbb{E}(S^2) - \mu^2 - \mu + \mu^2}{n} = \frac{1}{n}\mathbb{E}\big[S(S-1)\big]$$
$$\mathrm{Var}(\mathcal{S}) < \mathrm{Var}(M_0) \iff \mathbb{E}\big[S(S-1)\big] < 0$$

□

Note that if $S < 1$ with probability 1 then this statement is always true. However, for any timeline where there are sufficiently many tokens where the probability of the outcome of interest is sufficiently high, S will indeed exceed 1. This means that $\mathcal{S}$ is not necessarily a lower variance estimator than M_0 and should therefore be checked as an alternative for M_0 rather than a universal replacement. Additionally, since S, as an unbiased estimator for μ, is more likely to be large for more probable events, $\mathcal{S}$ is often a higher variance estimator than M_0 in contexts where the outcome of interest is more likely. This fact is reflected in our synthetic and empirical results. However, as is shown in Appendix J, $\mathrm{Var}(\mathcal{S}) > \mathrm{Var}(M_0)$ can hold for arbitrarily low mean. Fortunately, the additional computational cost required to compute both M_0 and $\mathcal{S}$ for a given outcome is limited to saving and summing the probabilities, which is negligible when compared to the computational cost of inference.

4.4 Derivation and Unbiasedness of $\mathcal{R}$ (REACH)

Theorem 3. *REACH is an unbiased estimator for the true probability that the outcome of interest is sampled.*

Proof. Let the event A be the subset of all timelines that, at some point before the end of the timeline, contain some outcome of interest token O. From this, we can define a Monte Carlo estimator for $P(A)$ by sampling m timelines and using the fraction of samples that include the outcome of interest token as an unbiased estimate of $P(A)$.

Define an outcome-avoiding probability space $\Omega_{\varnothing}$ as follows. Timelines are generated using the same model and time limit as in the previous probability space. Using the same next token prediction model as in the standard probability space, we define $p_t = P(X_t = O|X_{1:t-1})$. However, when sampling timelines in this probability space, we forbid the sampling of the outcome of interest token, renormalizing the next token distribution accordingly. Instead, alongside this outcome free timeline, at each token, we sample a Bernoulli trial (coin flip) with probability p_t. This means each sample in our space is defined by some outcome free timeline $X_{1:n}$ and a series of Bernoulli trials $c_{1:n}$. If at least one of these Bernoulli trials is a success, we label this sample as containing the outcome of interest and denote this subset of samples as B. Note that in order to show $P(A) = P(B)$, it is sufficient to show $P(A^C) = P(B^C)$. In other words, it is sufficient to show that the probability that the outcome of interest does not occur is equal between the two spaces.

Note that for each timeline in the original space that does not contain the outcome of interest there is exactly one sample in the second space that contains the exact same sequence of events and also does not contain the outcome of interest. There is also always only one such timeline in the original space for every outcome free timeline in the second space. Furthermore, by cancellation between the re-normalizing denominator and the Bernoulli trial failures, these samples have equal probability. Thus, by a probability preserving bijection, we conclude that $P(A^C) = P(B^C)$. This argument is provided in greater detail in Appendix E.

However, rather than fully sampling in the second space, REACH is achieved by sampling the outcome free timeline and then explicitly computing the probability that at least one Bernoulli trial succeeds given the outcome free timeline using the equation $1 - \prod_{t=1}^{T_E}(1 - P(X_t = O|\hat{X}_{1:t-1}))$. Note that this is exactly the sub-estimator of REACH defined above. Having derived the sub-estimator directly from our probability of interest, this reveals that our estimator $\mathcal{R}$ is an unbiased estimator of the probability that a sampled timeline will contain the outcome of interest token. □

4.5 Characterization of the Variance of $\mathcal{R}$

Theorem 4. *REACH is a variance reduction over the Monte Carlo estimator. This specific reduction is equal to:* $\mathrm{Var}(M_0) - \mathrm{Var}(\mathcal{R}) = \mathbb{E}(\mathrm{Var}(\mathbb{1}_B|\hat{X}_{1:n}))/n$.

Proof. Letting m_0 be the single timeline indicator r.v.s that determine the MC estimator M_0 and R be the sub-estimators of $\mathcal{R}$, by the law of total variance:

$$\begin{aligned}\mathrm{Var}(M_0) = \frac{1}{n}\mathrm{Var}(m_0) = \frac{1}{n}\mathrm{Var}(\mathbb{1}_A) = \frac{1}{n}\mathrm{Var}(\mathbb{1}_B) &= \frac{1}{n}\left(\mathrm{Var}(\mathbb{E}(\mathbb{1}_B|\hat{X}_{1:n})) + \mathbb{E}(\mathrm{Var}(\mathbb{1}_B|\hat{X}_{1:n}))\right) \\ &= \frac{1}{n}\left(\mathrm{Var}(R) + \mathbb{E}(\mathrm{Var}(\mathbb{1}_B|\hat{X}_{1:n}))\right) = \mathrm{Var}(\mathcal{R}) + \mathbb{E}(\mathrm{Var}(\mathbb{1}_B|\hat{X}_{1:n}))/n\end{aligned}$$

so that:

$$\mathrm{Var}(M_0) - \mathrm{Var}(\mathcal{R}) = \mathbb{E}(\mathrm{Var}(\mathbb{1}_B|\hat{X}_{1:n}))/n. \quad (1)$$

This also implies that $\mathrm{Var}(\mathcal{R}) \leq \mathrm{Var}(M_0)$. □

Intuitively, the above means that $\mathcal{R}$ presents a larger improvement to M_0 in scenarios in which an outcome-free timeline has a higher degree of uncertainty regarding whether or not the outcome of interest will actually occur, where the difference increases the closer p is to 0.5.

We call this statistic, $\mathbb{E}(\mathrm{Var}(\mathbb{1}_B|\hat{X}_{1:n}))$, "**spontaneity**." In Appendix S, it is shown that the efficiency gains presented by SCOPE and REACH can be predicted by a linear model using only prevalence and spontaneity as inputs with an R^2 of over 0.93 in MIMIC. The relationship between spontaneity and estimator variance is further explored in synthetic Markov chain experiments described in Appendix K. In addition to being a lower variance estimator than the MC method, $\mathcal{R}$ is also a lower variance estimator than $\mathcal{S}$ for all models and stopping conditions. This fact is proven in detail in Appendix H.

5 Empirical Study Methods

In addition to theoretical guarantees, it is important to study how these methods perform on real world patient data. We generated risk predictions for 11 clinically important outcomes across MIMIC-IV (22) and a local EHR data source, in which only 10 of the outcomes were available. Both datasets are converted to the Common Longitudinal ICU Format (CLIF: 23) version 2.1 standard and tokenized (Appendix R) using our 1385-token vocabulary. Timelines are split temporally, 70% for training, 10% for tuning, and 20% held out for evaluation. The training split is then used to train two next token prediction transformers, one per dataset, from scratch, implemented using a modified version of the Llama-3.2-1B architecture (24).

From the held out timelines, we remove stays in which the patient is discharged within the first 24 hours. Remaining timelines are truncated to tokens occurring in the first 24 hours of the stay and passed to the trained model for inference. We generate 100 potential future timelines per patient using SGLang's inference engine (25), with the model auto-regressively generating tokens until the patient is discharged or the maximum timeline length of 10,000 tokens is reached. SCOPE and REACH are computed iteratively from the next token distribution during inference. When a timeline generates an outcome of interest, the corresponding outcome-free completion required by REACH is generated in the same batch to avoid KV-cache misses. Confidence intervals are obtained by bootstrap subsampling 50% of the eligible held-out set 100 times. For each outcome, we restrict to the pool of patients who did not experience said outcome during the first 24 hours of their admission. Labels are defined as the outcome of interest any time after 24 hours and before discharge. See Appendix B for a breakdown of the number of filtered timelines, positives, and negatives for each outcome.

To compute a concrete inference cost reduction factor, we perform the following procedure. For each bootstrap and outcome, the 100-timeline Monte Carlo estimator is a baseline. Given a tolerance ϵ (set to 0.01 here), we find for each estimator the smallest sample size n such that AUROC remains within ϵ of this baseline for all sample sizes greater than n. To ensure fair comparison, this statistic is also computed for MC itself, and savings are reported as n_{MC}/n_{new}. For example, if MC remains within ϵ of 100-sample accuracy down to 60 samples while SCOPE does so down to 15, this is reported as a $4\times$ reduction. The same definition extends to token efficiency by letting n count generated tokens rather than sampled timelines.

The above framework does not capture how generative efficiency changes when multiple outcomes are estimated simultaneously. We therefore extend the per-task ϵ token cost to a cross-outcome total ϵ cost: the inference required to stay within ϵ of 100-sample MC on all tasks. For MC and SCOPE, cost is shared across tasks and equals the maximum individual task cost. REACH must still generate at least as many standard timelines as the most demanding outcome, but for less demanding outcomes it can subsample these timelines and perform completions only on the subsampled timelines containing the outcome of interest, slowing the rate at which its cost grows with the number of outcomes. To examine the total cost across all three methods, for each $k \in [11]$, we report the mean cost across all possible sets of k outcomes.

The necessary code to run these experiments, as well as the config files required to use the libraries to tokenize the dataset and train the initial model, will be made publicly available on acceptance.

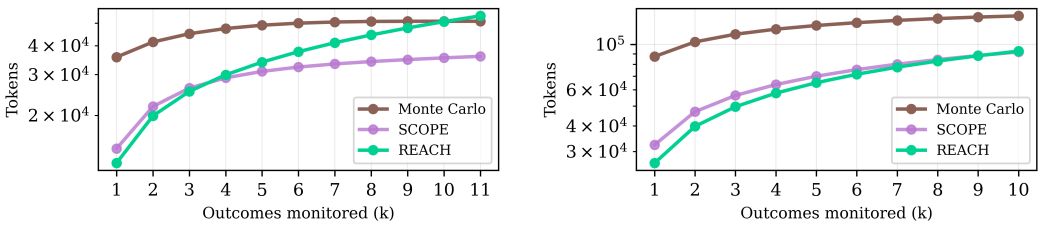
6 RESULTS

6.1 EXPERIMENTAL RESULTS

Across the two datasets both SCOPE and REACH match 100-sample Monte Carlo performance with substantially fewer tokens on nearly every outcome. The gap is largest for rare lab-defined outcomes (severe hypernatremia, profound hypokalemia, profound hyponatremia, severe hyperkalemia, hypoglycemia). Calibration is likewise preserved across estimators at reduced sample counts (per-outcome curves and Brier score analysis in Appendices M, N, and L). Table 1 quantifies these savings at $\epsilon = 0.01$. The most extreme reductions occur for severe hypernatremia and profound hypokalemia (average of 85.00× for SCOPE and 83.24× for REACH on MIMIC-IV); profound hypokalemia on local data shows comparable 75.00× and 72.89× reductions. Aggregate means are 20.97× (SCOPE) and 20.71× (REACH) on MIMIC, and 11.10× and 11.00× on local data, while the medians of 2.50× to 3.41× better reflect typical outcomes, since the means are pulled upward by the rare-event tail. The efficiency gain rank order is very similar across institutions, supporting external validity.

Table 1: Single task token efficiency of SCOPE and REACH vs. Monte Carlo baseline at $\varepsilon = 0.01$, on MIMIC-IV and Local Data. Values are mean token cost per patient. **Bold** indicates lowest token cost. (Full table with confidence intervals is available in Appendix T, AUC vs token count curves are available in Appendices F and G)

	MIMIC-IV					Local Data				
Outcome	T_{MC}	T_{SC}	T_{RE}	**MC/SC**	**MC/RE**	T_{MC}	T_{SC}	T_{RE}	**MC/SC**	**MC/RE**
Anemia	27k	12k	**11k**	2.25	2.42	55k	28k	**24k**	2.00	2.30
Hospital Mortality	33k	**5993**	6413	5.50	5.14	101k	**28k**	29k	3.67	3.54
Hypoglycemia	36k	8991	**6788**	4.00	5.30	92k	**9181**	10k	10.00	8.91
ICU Admission	**30k**	**30k**	31k	1.00	0.95	73k	55k	**53k**	1.33	1.39
IMV Event	33k	36k	**26k**	0.92	1.25	73k	92k	**66k**	0.80	1.11
Positive Confusion	27k	24k	**20k**	1.12	1.36	—	—	—	—	—
Profound Hypokalemia	51k	**599**	613	85.00	83.16	138k	**1836**	1889	75.00	72.89
Profound Hyponatremia	51k	**1199**	1240	42.50	41.09	119k	**28k**	29k	4.33	4.16
Severe Hyperkalemia	30k	12k	**11k**	2.50	2.68	73k	28k	**22k**	2.67	3.27
Severe Hypernatremia	51k	**599**	611	85.00	83.31	92k	**9181**	9670	10.00	9.49
Tachycardia	24k	27k	**21k**	0.89	1.12	55k	46k	**19k**	1.20	2.92
Aggregate:				**SC**	**RE**				**SC**	**RE**
Mean				20.97	20.71				11.10	11.00
Median				2.50	2.68				3.17	3.41



(a) MIMIC (b) Local Data

Figure 3: Mean cost per patient across all size k outcome subsets in tokens for predicting outcomes within tolerance $\epsilon = 0.01$ of 100-sample accuracy for both datasets.

Figure 3 shows cumulative token cost as the number of jointly monitored outcomes grows from 1 to 11 (MIMIC) or 10 (local data). Because SCOPE reuses a single sampled pool across all outcomes at no marginal generation cost, its cumulative cost flattens almost immediately and remains well below MC for every k. REACH grows more steeply, since each outcome requires its own bank of outcome-free completions, but remains at or below MC through 10 outcomes, only slightly exceeding MC for the MIMIC-IV dataset with all 11 outcomes. This pattern supports a hybrid deployment

strategy: SCOPE as the default for arbitrary multi-outcome panels, with REACH reserved for the highest-priority outcomes where its variance guarantee is most valuable.

These results demonstrate that SCOPE and REACH deliver consistent reductions in the amount of required inference for generative outcome prediction, especially for rare outcomes. Of our 11 outcomes, only on the task of ICU Admission do both SCOPE and REACH fail to deliver reductions in required inference. This failure mode is consistent with our spontaneity-based account: ICU admission is a clinician-initiated action that is typically signaled in advance by tokens already present in the timeline, leaving little residual outcome uncertainty for REACH to integrate out and giving SCOPE little dispersion to resolve. We recommend empirical screening with the spontaneity statistic before assuming gains on a new outcome.

7 Conclusion

SCOPE and REACH address three important limitations of MC generative outcome prediction identified in the introduction: estimator dispersion, computational cost, and sampling variance. Both estimators are unbiased, REACH is provably lower variance than either MC or SCOPE for any model and outcome, and across 11 clinically important outcomes in two health systems they match 100-sample MC accuracy with median reductions of $2.5\times$ to $3.4\times$ in tokens generated and reductions $> 80\times$ for the rarest outcomes. The estimators are complementary in deployment: SCOPE is the natural default for multi-outcome panels, since a single sampled pool yields estimates for an arbitrary number of outcomes at no marginal generation cost, while REACH is preferred when a small number of high-priority outcomes justify task-specific completions in exchange for its variance guarantee. Furthermore, the fact that a linear model in $\log$-prevalence and $\log$-spontaneity explains $> 93\%$ of the variance in efficiency gains across outcomes (Appendix S) demonstrates that the efficiency gains from these methods are substantial while highly predictable and well understood.

Addressing the key limitations of the MC sampling also eases the deployment constraints that were inherent to this method. Lower inference cost makes more frequent, lower-latency risk updates feasible in acute settings such as the ICU. Additionally, lower compute requirements for inference reduce energy usage for this existing class of models. Continuous risk estimates replace stepwise MC proportions with smoother trajectories, separating physiologic change from sampling noise and supporting the construction of foundation-model-derived survival curves (Section 7.1). Together, these contributions move generative EHR foundation models closer to real-time feasibility.

7.1 Limitations and Future Work

Three limitations bound these results. First, SCOPE lacks a guarantee of variance reduction and the estimator can exceed 1.0. Although these factors did not prevent SCOPE from delivering efficiency improvements across nearly all outcomes, future work must keep this property in mind. Second, as illustrated in Figure 3, the marginal cost of computing REACH completions adds up. In applied settings, it may be prudent to only compute REACH completions for the least prevalent outcomes. Finally, both estimators require the outcome to be captured by the appearance of a single token, requiring pre-processing to inject computed label tokens where multi-token logic is required.

Since these estimators provide continuous risk scores for patient timelines, rather than binary labels of whether or not the outcome of interest token was observed, these methods could be used to improve clinician understanding of generative event models by identifying common events or features of high risk timelines. Furthermore, not only are these risk scores continuous at the timeline level, but for REACH in particular, the risk scores can easily be computed over time, which could provide clinicians with clearer intuition regarding the evolution of a patient's risk with respect to specific outcomes. This potential for the construction of these foundation model derived survival curves has a large number of potentially valuable applications.

Additionally, while these methods improve the computational cost of generative risk prediction by resolving the dispersion of estimates and reducing variance of estimators, future work could further address this problem by varying the number of sampled timelines per patient. This would allow this technique to expend fewer computational resources computing the risk in situations where the outcome of interest is either impossible or guaranteed, allowing for compute time to be focused on risk stratification in more ambiguous cases.

REFERENCES

[1] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Sci. Rep.*, 10, 2020.

[2] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.*, 4(1):86, 2021.

[3] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digit. Med.*, 6(1):135, 2023.

[4] Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Alan Fries, and Nigam Shah. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. In *Neurips Datasets and Benchmarks Track*, volume 36, pages 67125–67137, 2023.

[5] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T Teo, and Richard J B Dobson. Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit. Health*, 6(4), 2024.

[6] Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative medical event models improve with scale, 2025.

[7] Matthew McDermott, Bret Nestor, Peniel Argaw, and Isaac S Kohane. Event stream GPT: A data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. In *Adv. Neural Inf. Process. Syst.*, volume 36, pages 24322–24334, 2023.

[8] Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, E Steinberg, J Fries, Christopher Ré, Oluwasanmi Koyejo, and Nigam H Shah. Context clues: Evaluating long context models for clinical prediction tasks on EHRs. In *ICLR*, 2025.

[9] Michael C. Burkhart, Bashar Ramadan, Luke Solo, William F. Parker, and Brett K. Beaulieu-Jones. Quantifying surprise in clinical care: Detecting highly informative events in electronic health records with foundation models. In *Pacific Symposium on Biocomputing*, volume 31, pages 173–188, 2026.

[10] Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *npj Digit. Med.*, 7(1):256, 2024.

[11] Pawel Renc, Michal K Grzeszczyk, Nassim Oufattole, Deirdre Goode, Yugang Jia, Szymon Bieganski, Matthew B A McDermott, Jaroslaw Was, Anthony E Samir, Jonathan W Cunningham, David W Bates, and Arkadiusz Sitek. Foundation model of electronic medical records for adaptive risk estimation. *GigaScience*, 14, 2025.

[12] Pavel Blinov and Vladimir Kokh. Medical profile model: Scientific and practical applications in healthcare. *IEEE Journal of Biomedical and Health Informatics*, 28(1):450–458, 2024.

[13] Yu Akagi, Tomohisa Seki, Yoshimasa Kawazoe, Toru Takiguchi, and Kazuhiko Ohe. A generative foundation model for structured patient trajectory data. In *AMIA Annu Symp Proc.*, 2025.

[14] Chao Pang, Jiheum Park, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Shalmali Joshi, Noémie Elhadad, and Karthik Natarajan. Cehr-xgpt: A scalable multi-task foundation model for electronic health records, 2025.

[15] Haresh Rengaraj Rajamohan, Xiang Gao, Weicheng Zhu, Shih-Lun Huang, Long Chen, Kyunghyun Cho, Cem M Deniz, and Narges Razavian. Foundation models for clinical records at health system scale. In *1st ICML Workshop on Foundation Models for Structured Data*, 2025.

[16] Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume PMLR 259, pages 291–307, 2025.

[17] Renee Y. Hsia, Zachariah Hale, and Jeffrey A. Tabas. A National Study of the Prevalence of Life-Threatening Diagnoses in Patients With Chest Pain. *JAMA Internal Medicine*, 176(7):1029–1032, 2016.

[18] Jack P. C. Kleijnen, Ad Ridder, and Reuven Rubinstein. Variance reduction techniques in monte carlo methods. *Available at SSRN*, November 2010.

[19] Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York, NY, 2007.

[20] Ian W. McKeague and Wolfgang Wefelmeyer. Markov chain monte carlo and Rao–Blackwellization. *J. Stat. Plann. Inference*, 85(1):171–182, 2000.

[21] Jonathan Amar, Edward Liu, Alessandra Breschi, Liangliang Zhang, Pouya Kheradpour, Sylvia Li, Lisa Soleymani Lehmann, Alessandro Giulianelli, Matt Edwards, Yugang Jia, David Nola, Raghav Mani, Pankaj Vats, Jesse Tetreault, T. J. Chen, and Cory Y. McLean. Integrating Genomics into Multimodal EHR Foundation Models, 2025.

[22] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

[23] Juan C. Rojas, Patrick G. Lyons, Kaveri Chhikara, Vaishvik Chaudhari, Sivasubramanium V. Bhavani, Muna Nour, Kevin G. Buell, Kevin D. Smith, Catherine A. Gao, Saki Amagai, Chengsheng Mao, Yuan Luo, Anna K. Barker, Mark Nuppnau, Michael Hermsen, Jay L. Koyner, Haley Beck, Rachel Baccile, Zewei Liao, Kyle A. Carey, Brenna Park-Egan, Xuan Han, Alexander C. Ortiz, Benjamin E. Schmid, Gary E. Weissman, Chad H. Hochberg, Nicholas E. Ingraham, and William F. Parker. A common longitudinal intensive care unit data format (CLIF) for critical illness research. *Intensive Care Med.*, 2025.

[24] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models, 2024.

[25] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. *arXiv preprint arXiv:2312.07104*, 2024.

[26] Michael C. Burkhart, Bashar Ramadan, Zewei Liao, Kaveri Chhikara, Juan C. Rojas, William F. Parker, and Brett K. Beaulieu-Jones. Foundation models for electronic health records: representation dynamics and transferability, 2025.

[27] Inhyeok Lee, Luke Solo, Michael C. Burkhart, Bashar Ramadan, Sahil Sethi, Sarah Jabbour, William F. Parker, and Brett K. Beaulieu-Jones. Representation before training: A practical benchmark for generative medical event model tokenization, 2026.

A AI USE DISCLOSURE

In the process of creating this paper, generative AI tools were used for:

- Revision of the paper to improve readability and fix typographical errors
- Generating, reviewing, and testing code
- Exploring the relevant literature
- Auto-formalization of our core proofs using prove2.me

Despite being used for the auto-formalization, generative AI was not used to create the initial concepts for the proposed methods, create the original proofs, create the proof sketches featured in the body, or discover the primary theoretical facts established in this paper.

All AI-assisted work was carefully reviewed (and tested where applicable) by multiple authors at all steps. All references discovered using AI were subsequently directly reviewed. All code generated by AI has been double checked for correctness and consistency with our experimental design.

B OUTCOME COHORTS

Outcome	Timelines	Excluded (past)	Positives	Negatives	Prevalence (%)
ICU Admission	83485	13187	3567	79918	4.27
IMV Event	92413	4259	2567	89846	2.78
Hospital Mortality	96672	0	2048	94624	2.12
Profound Hyponatremia	96087	585	612	95475	0.64
Severe Hypernatremia	96588	84	107	96481	0.11
Profound Hypokalemia	96525	147	213	96312	0.22
Severe Hyperkalemia	88641	8031	8274	80367	9.33
Anemia	90300	6372	8181	82119	9.06
Hypoglycemia	94745	1927	4949	89796	5.22
Tachycardia	90585	6087	4616	85969	5.10
Positive Confusion Test	94939	1733	2600	92339	2.74

Table 2: Label statistics by outcome in MIMIC-IV. Total timelines in cohort: 96,672.

Outcome	Timelines	Excluded (past)	Positives	Negatives	Prevalence (%)
ICU Admission	3157	843	250	2907	7.92
IMV Event	3622	378	176	3446	4.86
Hospital Mortality	4000	0	163	3837	4.08
Profound Hyponatremia	3942	58	35	3907	0.89
Severe Hypernatremia	3986	14	14	3972	0.35
Profound Hypokalemia	3965	35	27	3938	0.68
Severe Hyperkalemia	3478	522	420	3058	12.08
Anemia	3575	425	584	2991	16.34
Hypoglycemia	3632	368	300	3332	8.26
Tachycardia	1958	2042	661	1297	33.76

Table 3: Label statistics by outcome on the local dataset. Total timelines in cohort: 4,000.

C NOMENCLATURE

M_0 Equation shorthand for the Monte Carlo estimator.

$\mathcal{S}$ Equation shorthand for the SCOPE estimator.

$\mathcal{R}$ Equation shorthand for the REACH estimator.

T_O The random variable equal to the index of the first outcome of interest token.

T_E The random variable equal to the index of the terminating token of a sequence or the token coming after the first outcome of interest token.

A The event describing the subset of timelines in the first probability space that contain the outcome of interest token.

B The event describing the subset of timelines that have a successful Bernoulli trial under the second probability space.

V The vocabulary of tokens.

O The outcome of interest token, an element of V.

$X_{1:n}$ A tokenized timeline of length n.

$\hat{X}_{1:n}$ A tokenized timeline of length n that excludes O from the set of possible next tokens.

Ω_{M_τ} The probability space of timelines sampled under some model M and time limit τ.

$\Omega_{\not\mathcal{O}}$ The probability space of outcome-free timelines and Bernoulli trials at each token.

$\mathbf{c}_{1:n}$ The result of n Bernoulli trials for some outcome-free tokenized timeline of length n.

D DERIVATION OF SCOPE

$$\mathbb{P}(T_O < T_E) = \mathbb{E}_{X\sim P}[\mathbb{1}_{T_O(X)<T_E(X)}] = \mathbb{E}_{X\sim P}\big[\sum_{t=1}^{\infty} \mathbb{1}_{\{T_O(X)=t\}}\mathbb{1}_{\{T_E(X)>t-1\}}\big]$$

Equivalent to the product of the indicator r.v.s in the interior of the term above, to simplify the notation, we define a new function f over partial sequences:

$$g(x_{1:t-1}) = \mathbb{1}_{\{O\notin x_{1:t-1}\}}\mathbb{1}_{\{T_E(x_{1:t-1})>t\}}, \quad f(x_{1:t}) = \mathbb{1}_{\{x_t=O\}}g(x_{1:t-1})$$

Note that for any sequences exceeding the time horizon or containing O, the above function will map to zero. So that, by the linearity of expectation and the law of total expectation:

$$\begin{aligned}\mathbb{P}(T_O < T_E) &= \textstyle\sum_{t=1}^{\infty} \mathbb{E}_{X\sim P}\big[f(X_{1:t})\big] = \sum_{t=1}^{\infty} \mathbb{E}_{X_{1:t-1}\sim P}\big[\mathbb{E}[f(X_{1:t})|X_{1:t-1}]\big] \\ &= \textstyle\sum_{t=1}^{\infty} \mathbb{E}_{X_{1:t-1}\sim P}\big[g(X_{1:t-1})\mathbb{E}[\mathbb{1}_{\{X_t=O\}}|X_{1:t-1}]\big] \\ &= \textstyle\sum_{t=1}^{\infty} \mathbb{E}_{X_{1:t-1}\sim P}\big[g(X_{1:t-1})\mathbb{P}(X_t = O|X_{1:t-1})\big]\end{aligned}$$

Next by the law of total expectation

$$\begin{aligned}\mathbb{E}_X(g(X_{1:t-1})\mathbb{P}(X_t = O|X_{1:t-1})|X_{1:t-1})) &= \mathbb{E}_{X_{1:t-1}}[\mathbb{E}[g(X_{1:t-1})\mathbb{P}(X_t = O|X_{1:t-1})|X_{1:t-1}]] \\ &= \mathbb{E}_{X_{1:t-1}}[g(X_{1:t-1})\mathbb{P}(X_t = O|X_{1:t-1})]\end{aligned}$$

Substituting this into the above

$$\begin{aligned}P(T_O < T_E) &= \textstyle\sum_{t=1}^{\infty} \mathbb{E}_{X\sim P}\big[g(X_{1:t-1})\mathbb{P}(X_t = O|X_{1:t-1})\big] = \mathbb{E}_{X\sim P}\big[\sum_{t=1}^{\infty} g(X_{1:t-1})\mathbb{P}(X_t = O|X_{1:t-1})\big] \\ &= \mathbb{E}_{X\sim P}\big[\sum_{t=1}^{\min\{T_E(X),T_O(X)\}} \mathbb{P}(X_t = O|X_{1:t-1})\big]\end{aligned}$$

The inside of this expression, which we will denote S_i, can then be sampled to generate an unbiased estimator, $\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n} S_i$.

E PROOF OF EQUAL PROBABILITIES

Consider a sequence generation model M that generates a probability distribution for the next token from some vocabulary V, which contains our outcome of interest token O. Let $Y_{1:n}$ denote the sequence of tokens that the model has processed before generating any new tokens. Note that each token that is associated with an increasing time stamp $\mathcal{T}_i$. Let $P_M(Y_0 = t_j)$ denote the probability that the next token generated by the model is t_j. Likewise, if the model has already generated $Y_{1:i}$, $P_M(Y_{i+1} = t_j | Y_{1:i})$ denotes the probability that the next token generated will be t_j.

Before introducing our improved sampling method, we will introduce the standard Monte Carlo method. Define probability space Ω_{M_τ}:

$$\Omega_{M_\tau} := \left\{ X_{1:n} |\ \forall i \in [n],\ X_i \in V,\ \ \mathcal{T}_i < \tau \right\} \tag{2}$$

$$P_{\Omega_{M_\tau}}(X_{1:n}) = P_M(Y_0 = X_0) \prod_{i=1}^{n} P_M(Y_i = X_i | Y_{1:i-1} = X_{1:i-1}) \tag{3}$$

Intuitively, the samples in this probability space correspond to the sequences of tokens generated by by sampling the tokens directly at the rate dictated by the model. Within this distribution, define the event A:

$$A := \{ X_{1:n} \in \Omega_{M_\tau} |\ \exists i \in [n] \text{ s.t. } X_i = O \} \tag{4}$$

Next, consider a second probability space $\Omega_{\not O}$:

$$\Omega_{\not O} := \left\{ (X_{1:n}, \mathbf{c}_{1:n}) |\ \forall i \in [n],\ X_i \in V \setminus O,\ \ \mathcal{T}_i < \tau,\ \mathbf{c}_i \in \{0, 1\}] \right\} \tag{5}$$

$$P_{\Omega_{\not O}}(X_{1:n}) = \left(\frac{P_M(Y_0 = X_0)}{P_M(Y_0 \neq O)} \prod_{i=1}^{n} \frac{P_M(Y_i = X_i | Y_{1:i-1} = X_{1:i-1})}{P_M(Y_i \neq O | Y_{1:i-1} = X_{1:i-1})} \right) \tag{6}$$

Intuitively, this can be understood as the probability of generating a given sequence $X_{1:n}$ by sampling next tokens at the rates suggested by the model but this time **excluding** the outcome token.

Define $h_i(X_{1:n}) = P_M(Y_i = O | X_{1:i-1})$

(Note that in order to keep notation simple $X_{1:n}$ may be omitted)

$$P_{\Omega_{\not O}}(X_{1:n}, \mathbf{c}_{1:n}) = P_{\Omega_{\not O}}(X_{1:n}) \prod_{i=1}^{n} \Big((1 - \mathbf{c}_i)(1 - h_i) + \mathbf{c}_i h_i \Big) \tag{7}$$

Each $\mathbf{c}_i$ can be understood as an independent Bernoulli trial with probability h_i. The purpose of these variables is to simulate the tokens in the timeline being randomly "flipped" to the outcome token.

Now define the event corresponding to all outcomes where at least one token in the timeline is randomly flipped to the outcome of interest as B.

Consider the complements of our two events described above: $A^{\complement}$, which corresponds to all sampled timelines that **do not** contain the outcome of interest token, and $B^{\complement}$, which corresponds to all sampled timelines where **none** of the Bernoulli trials succeed and flip a token to the outcome of interest.

Since the elements of both probability spaces are determined by the same next token distribution model, note that there exists a natural bijection $\phi : A^{\complement} \to B^{\complement}$:

$$\phi(X_{1:n}) := (X_{1:n}, (0, ..., 0)) \tag{8}$$

Furthermore, from our definition of our probability spaces:

$$P_{\Omega_{\not O}}(\phi(X_{1:n})) = P_{\Omega_{\not O}}((X_{1:n}, (0, ...,)) \tag{9}$$

Looking at the definitions once more, it is clear that the Bernoulli trial product terms from (7) cancel with the denominator of (6) to yield exactly (3). Therefore:

$$P_{\Omega_{\not O}}(\phi(X_{1:n})) = P_{\Omega_{M_\tau}}(X_{1:n}) \tag{10}$$

By the definition of the probability of an event, this allows for the substitution:

$$P_{\Omega_\varnothing}(B^C) = \sum_{(X_{1:n},(0,...,0))\in B^C} P_{\Omega_\varnothing}((X_{1:n},(0,...,0))) \tag{11}$$

$$= \sum_{X_{1:n}\in A^C} P_{\Omega_\varnothing}(\phi(X_{1:n})) = \sum_{X_{1:n}\in A^C} P_{\Omega_{M_\tau}}(X_{1:n}) = P_{\Omega_{M_\tau}}(A^C) \tag{12}$$

So, by the law of total probability, we have shown that $P(A) = P(B)$.

F MIMIC-IV AUC vs. Tokens Curves

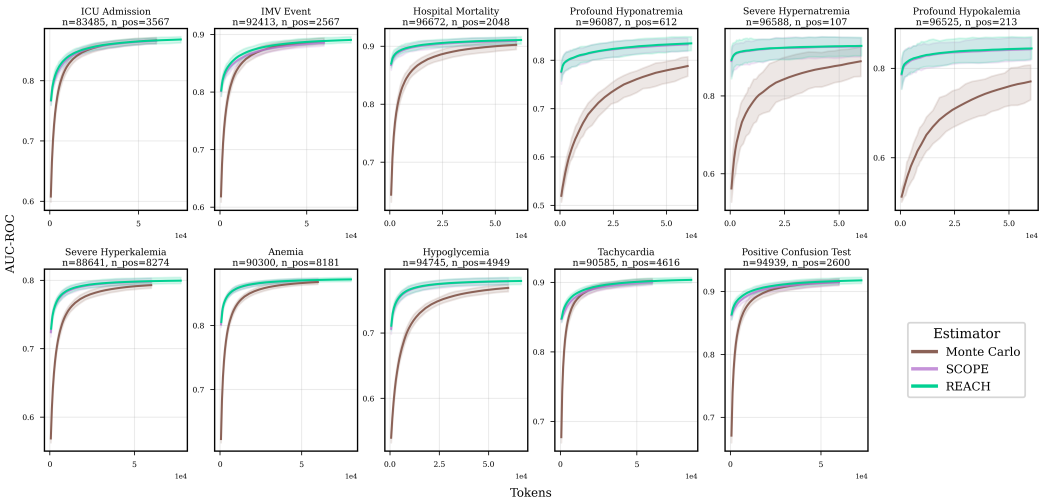


Figure 4: Prediction accuracy comparison across MIMIC-IV patients on all 11 outcomes. The x-axis tracks the number of tokens generated per patient to achieve the predictive accuracy reflected on the y-axis. Local data curves can be found in Appendix G.

Figure 4 plots AUROC against tokens generated per patient for all 11 MIMIC-IV outcomes.

G Local Data AUC Curves

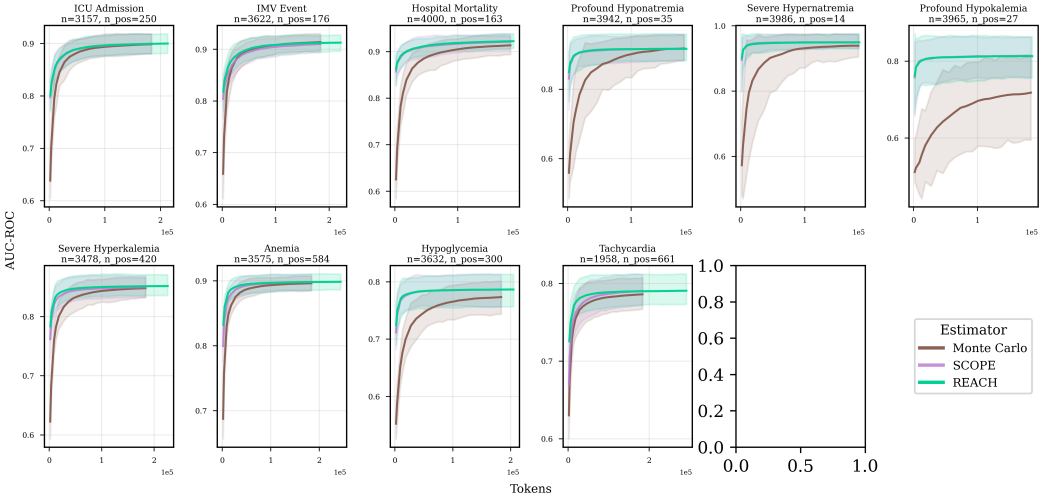


Figure 5: AUC vs Tokens Generated for Local Data

H PROOF OF $\mathrm{Var}(\mathcal{R}) \leq \mathrm{Var}(\mathcal{S})$

From the body of the paper, recall that we can define our SCOPE estimator as

$$\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{t=1}^{\min\{T_E(X^{(i)}),T_O(X^{(i)})\}} P(X_t = O|X^{(i)}_{1:t-1})\Big] \tag{13}$$

If we then define

$$f_t(x) = P(X_T = O|X_{1:t-1} = x_{1:t-1})\mathbb{1}_{\{T_O(x_{1:t-1}),T_E(x)>t-1\}} \tag{14}$$

We can instead express $\mathcal{S}$ as:

$$\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t\geq 1} f_t(X^{(i)}) \tag{15}$$

As an alternative to our previous derivation of the REACH estimator, let $\hat{P}$ denote the restriction of P to $V \setminus \{O\}$ and apply importance sampling to (13) to obtain:

$$\mathcal{R} = \frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{t=1}^{T_E(\hat{X}^{(i)})} P(X_t = O|\hat{X}^{(i)}_{1:t-1})\frac{P(\hat{X}^{(i)_{1:t-1}})}{\hat{P}(\hat{X}^{(i)_{1:t-1}})}\Big] \tag{16}$$

where $\forall i \in [n],\ X^{(i)} \sim^{\text{i.i.d.}} \hat{P}$. For any sequence x in the vocabulary $V \setminus \{O\}$, we have

$$\hat{P}(X_{1:t-1} = x_{1:t-1}) = \prod_{j=1}^{t-1}\hat{P}(X_j = x_j|X_{1:j-1} = x_{1:j-1})$$

$$= \prod_{j=1}^{t-1}\frac{P(X_j = x_j|X_{1:j-1} = x_{1:j-1})}{1 - P(X_j = O|X_{1:j-1} = x_{1:j-1})}$$

so that

$$\frac{P(\hat{X}^{(i)}_{1:t-1})}{\hat{P}(\hat{X}^{(i)}_{1:t-1})} = \prod_{j=1}^{t-1}(1 - P(X_j = O|X_{1:j-1} = x_{1:j-1})) \tag{17}$$

which, when substituted into (16), yields our original expression for $\mathcal{R}$ provided in the body.

Define

$$h^0_t(x) = \prod_{j=1}^{t-1}(1 - P(X_j = O|X_{1:j=1} = x_{1:j-1})) \tag{18}$$

$$h_t(x) = P(X_t = O|X_{1:t-1} = x_{1:t-1})\mathbb{1}_{\{T_E(x)>t-1\}}h^0_t(x) \tag{19}$$

Using these definitions, $\mathcal{R}$ can be simplified as:

$$\mathcal{R} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t\geq 1} h_t(\hat{X}^{(i)}) \tag{20}$$

Next, using our result from (17):

$$\mathbb{E}[h_t(\hat{X})^2] = \int h_t(\hat{x}_{1:t})\hat{p}(\hat{x}_{1:t})d\hat{x}_{1:t} \tag{21}$$

$$= \int h_t(\hat{x}_{1:t})\frac{\hat{p}(\hat{x}_{1:t})}{p(x_{1:t})}p(x_{1:t})d\hat{x}_{1:t} \tag{22}$$

$$= \int f_t(\hat{x}_{1:t})^2 h^0_t(x_{1:t})^2\frac{1}{h^0_t(\hat{x}_{1:t})}p(x_{1:t})d\hat{x}_{1:t} = \mathbb{E}[f_t(X)^2 h^0_t(X)] \tag{23}$$

Since $0 \leq h_s(x) \leq 1$, we can conclude that:

$$\mathbb{E}[h_t(\hat{X})^2] \leq \mathbb{E}[f_t(X)^2] \tag{24}$$

Since $\mathcal{S}$ and $\mathcal{R}$ have the same expectation as unbiased estimators of the same probability, (24) demonstrates that:

$$\mathrm{Var}(h_t(\hat{X})) \leq \mathrm{Var}(f_t(\hat{X})) \tag{25}$$

Next, for $s < t$, we have

$$\mathrm{Cov}(h_s(\hat{X}), h_t(\hat{X})) = \mathbb{E}[h_s(\hat{X})h_t(\hat{X})] - \mathbb{E}[h_s(\hat{X})]\mathbb{E}[h_t(\hat{X})] \tag{26}$$

By a nearly identical argument to the one presented in (21) to (23), it can be shown that

$$\mathbb{E}[h_s(\hat{X})h_t(\hat{X})] - \mathbb{E}[h_s(\hat{X})]\mathbb{E}[h_t(\hat{X})] = \mathbb{E}[f_s(X)f_t(X)h_s^0(X)] - \mathbb{E}[f_s(X)]\mathbb{E}[f_t(X)]$$

$$\mathrm{Cov}(h_s(\hat{X}), h_t(\hat{X})) \leq \mathrm{Cov}(f_s(X), f_t(X)) \tag{27}$$

And, by (15) and (20), we reach our final conclusion that:

$$\mathrm{Var}(\mathcal{R}) \leq \mathrm{Var}(\mathcal{S})$$

I ETHOS-ARES: TOTAL INFERENCE TIME COMPUTATION

In order to estimate the time required by the standard ETHOS-ARES library, the inference code was left to run with 8xA100 GPUS for 1 hour. This generated 22395 samples out of the required 4304700. This implies a little over 192 hours will be required to complete inference using $8\times$ A100s. Multiplying by 8 to determine the GPU hours required, we get over $\boxed{1537}$ GPU hours.

J COUNTEREXAMPLE DEMONSTRATING THE IMPOSSIBILITY OF A POSITIVE BOUND M SUCH THAT IF $P(A) < M$ THEN $\mathrm{Var}(M_0) > \mathrm{Var}(\mathcal{S})$ FOR ALL MODELS AND STOPPING CONDITIONS

In the body, it was demonstrated that $\mathrm{Var}(\mathcal{S})$ is not necessarily less than $\mathrm{Var}(M_0)$. Specifically, timelines for which the $\mathcal{S}$ sub-estimator predicts an estimate for the value of the probability above 1 contribute to the inequality $\mathrm{Var}(M_0) < \mathrm{Var}(\mathcal{S})$. Given this fact, it seemed plausible that there could exist some positive lower bound M for the probability of the outcome of interest, $P(A)$, for which the following could hold: for any model and stopping condition, if $P(A) < M$ then $\mathrm{Var}(\mathcal{S}) < \mathrm{Var}(M_0)$. Such a bound would be quite useful as it would allow one to determine that $\mathcal{S}$ would be a lower variance estimator than M_0 based on the probability of the outcome alone.

However, this fact is not true and is shown to be false by the following counterexample.

Consider the following model and stopping condition. At first, the model has two possible token outputs A and B. Let it generate token A with arbitrary probability p, and B otherwise. Neither contribute toward the time limit or are the outcome of interest. The model then operates according to the following rules:

- If A was generated at the first step:
- Immediately terminate the timeline without generating any further tokens.
- If B was generated at the first step:
- Generate new tokens with equal probability of it being token H or token T. Terminate either after 3 tokens are generated or after the appearance of the first heads.

For this counterexample, we will treat H as the outcome of interest token. Naturally:

$$P(H \in X) = \frac{7}{8}(1-p) \tag{28}$$

Since $\mathcal{S}$ and M_0 are unbiased estimators for this probability, this likewise is the expectation of both of these random variables. In order to compute the difference in variance between $\mathcal{S}$ and M_0, we can proceed directly from the definition of variance and the unbiasedness of our estimators:

$$\mathrm{Var}(\mathcal{S}) - \mathrm{Var}(M_0) = \mathbb{E}(\mathcal{S}^2) - \mathbb{E}(M_0^2) \tag{29}$$

From the table:

Outcome	Probability	M_0^2 contribution	$\mathcal{S}^2$ contribution
A	p	0	0
BH	$\frac{1}{2}(1-p)$	1	$\frac{1}{4}$
BTH	$\frac{1}{4}(1-p)$	1	1
BTTH	$\frac{1}{8}(1-p)$	1	$\frac{9}{4}$
BTTT	$\frac{1}{8}(1-p)$	0	$\frac{9}{4}$

$$\mathbb{E}(M_0^2) = \frac{7}{8}(1-p)$$

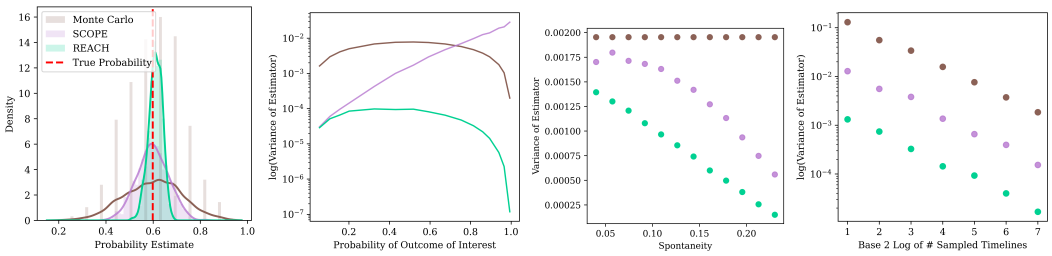
$$\mathbb{E}(\mathcal{S}^2) = \frac{15}{16}(1-p)$$

$$\mathrm{Var}(\mathcal{S}) - \mathrm{Var}(M_0) = \frac{1}{16}(1-p) \tag{30}$$

Note that if $p < 1$, (30) implies $\mathrm{Var}(\mathcal{S}) - \mathrm{Var}(M_0) > 0$. Further, if we allow p get arbitrarily close to 1, $P(H \in X) \to 0$, while the variance of $\mathcal{S}$ remains strictly worse than the variance of M_0. This shows that there cannot exist a positive value M that guarantees that if $P(H \in X) < M$ then $\mathrm{Var}(\mathcal{S}) < \mathrm{Var}(M_0)$.

K MARKOV CHAIN SYNTHETIC EXPERIMENTS

In order to measure how the theoretical guarantees of these techniques translate to actual improvements in the accuracy of generative event model derived predictors, we first implement a suite of Markov Chain synthetic experiments that allow us to analyze the properties of these new estimators in an easily controlled and computationally cheap manner. Under this framework, the next token distribution relies only on the most recent token, rather than the entire sequence of tokens. For these experiments, we reduce the generative event model timeline simulation paradigm of "Generate next tokens until a certain time limit or terminal condition" to "Progress in the discrete time Markov chain until a specified number of steps have been taken." And, standing in for a clinically significant token, we choose one state in the chain to act as the "outcome of interest" and use our two new methods and the traditional Monte Carlo method to estimate the probability of reaching this "outcome of interest" state. Note: All figures presented in this section were generated using the default parameters of 4



(a) (b) (c) (d)

Figure 6: **(a)** Distribution of estimators vs. true probability. **(b)** Variance of estimator vs. probability. **(c)** Variance vs. spontaneity. **(d)** Variance vs. number of sampled timelines.

experiments in the Markov chain notebook.

Figure 6(a) visually demonstrates the way in which both SCOPE and REACH resolve the dispersion of estimates problem, which can be seen in the fully separated spikes of the Monte Carlo estimator's distribution. REACH and SCOPE also visibly reduce variance over the Monte Carlo estimator.

In Figure 6(b), the relationship between the variance of SCOPE and REACH and the probability of the outcome of interest is clearly shown. For lower true probabilities, the variance of SCOPE is extremely low and is not visibly different from that of REACH. However, as the true probability grows, and with it the probability of $\mathcal{S}$ producing sampled values that exceed 1, the variance of SCOPE grows rapidly and exceeds that of the MC estimator when the probability of the outcome of interest is above 0.8.

Next, Figure 6(c) reflects how the difference in variance between REACH and the MC estimator depends on the uncertainty of the non-outcome timelines. The Markov chains for this figure are uniquely determined by fixing both the probability of reaching the outcome of interest at 0.5 and the spontaneity of the chain at various values between 0 and 0.25. Since spontaneity is equal to the difference in variance between the Monte Carlo and REACH methods, the linear decrease in variance of the REACH estimator exactly matches this theoretical result. However, although not directly implied by our earlier work, a similar relationship seems to hold between SCOPE and the MC estimator. Finally, Figure 6(d) clearly shows all 3 estimators decreasing in variance as the number of sampled timelines increases. Note that the x axis is in log base 2 scaling.

L Low Sample Calibration Skill Score Results

Below, we reproduce the core concept of our AUC experiments with Brier Skill Score (BSS) as our new object of study. In these experiments, we treat 100 sample MC as the baseline estimator and compute the BSS at all lower sample counts.

The table is computed by setting n for each method equal to the smallest number of samples required for each method to achieve BSS > 0 on 50% of bootstrap samples.

These results demonstrate that, when a calibration metric such as the Brier score replaces a rank based prediction metric like AUROC, these models provide comparable gains in the number of required samples to produce accurate estimates.

Table 4: Token efficiency of SCOPE and REACH vs. Monte Carlo baseline under the Brier skill score criterion on MIMIC-IV. Values are the smallest per-patient token budget at which more than 50% of 100 bootstrap replications achieve BSS ≥ 0 relative to MC at $n = 100$ [95% bootstrap CI]. REACH cost is $M_1 + p \times M_2$. **Bold** indicates lowest token cost. Ratios > 1 indicate fewer tokens needed than MC. †SCOPE never matched MC at $n = 100$; the baseline is instead SCOPE at $n = 100$, and the ratio is the MC budget needed to match it divided by SCOPE's budget at $n = 100$ (values < 1 indicate SCOPE is less efficient than MC). Omitted entries reflect estimators that do not have a n that satisfies the given condition.

Outcome	T_{MC}	T_{SCOPE}	T_{REACH}	**MC/SCOPE**	**MC/REACH**
ICU Admission	**48k [48k, 48k]**	60k [60k, 60k]	59k [59k, 59k]	0.80 [0.80, 0.80]	0.81 [0.81, 0.81]
IMV Event	**48k [48k, 51k]**	—	53k [49k, 53k]	0.75 [0.75, 0.80]†	0.91 [0.91, 0.98]
Hospital Mortality	48k [48k, 48k]	54k [51k, 54k]	**35k [32k, 35k]**	0.89 [0.89, 0.94]	1.36 [1.36, 1.50]
Profound Hyponatremia	48k [45k, 48k]	18k [18k, 21k]	**15k [15k, 15k]**	2.67 [2.29, 2.67]	3.09 [2.90, 3.09]
Severe Hypernatremia	54k [54k, 57k]	11k [9589, 12k]	**9172 [8560, 9783]**	5.00 [4.74, 5.94]	5.88 [5.51, 6.65]
Profound Hypokalemia	54k [51k, 54k]	8390 [8390, 8989]	**7350 [7350, 7963]**	6.43 [6.00, 6.43]	7.34 [6.40, 7.34]
Severe Hyperkalemia	42k [42k, 42k]	24k [24k, 24k]	**19k [16k, 19k]**	1.75 [1.75, 1.75]	2.25 [2.25, 2.56]
Anemia	42k [42k, 45k]	33k [30k, 33k]	**22k [22k, 22k]**	1.27 [1.27, 1.40]	1.88 [1.88, 2.01]
Hypoglycemia	45k [45k, 45k]	13k [13k, 15k]	**12k [11k, 12k]**	3.41 [3.00, 3.41]	3.68 [3.68, 4.14]
Tachycardia	**48k [45k, 48k]**	—	51k [51k, 51k]	0.70 [0.70, 0.70]†	0.94 [0.88, 0.94]
Positive Confusion	**48k [48k, 48k]**	—	52k [52k, 52k]	0.75 [0.75, 0.80]†	0.93 [0.93, 0.93]
Aggregate:				**SCOPE**	**REACH**
Mean				2.22 [2.14, 2.29]	2.64 [2.57, 2.73]
Median				1.27 [1.27, 1.40]	1.88 [1.88, 2.01]

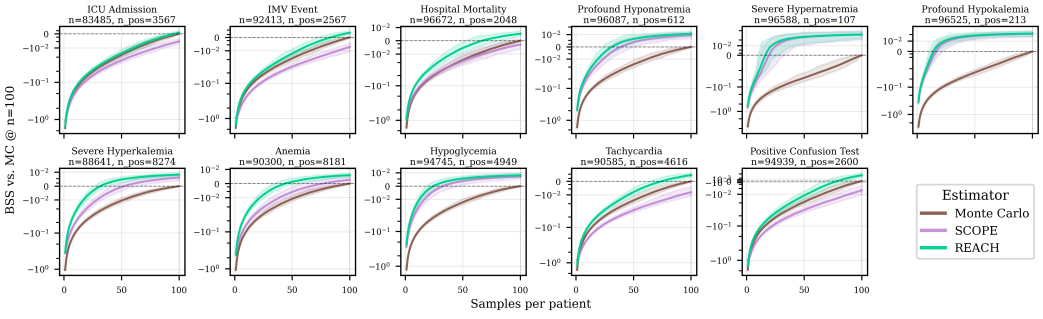


Figure 7: Brier skill scores with 100 sample Monte Carlo as the baseline on MIMIC-IV

Table 5: Token efficiency of SCOPE and REACH vs. Monte Carlo baseline under the Brier skill score criterion on the local dataset. Values are the smallest per-patient token budget at which more than 50% of 100 bootstrap replications achieve BSS ≥ 0 relative to MC at $n = 100$ [95% bootstrap CI]. **Bold** indicates lowest token cost. Ratios > 1 indicate fewer tokens needed than MC.

Outcome	T_{MC}	T_{SCOPE}	T_{REACH}	**MC/SCOPE**	**MC/REACH**
ICU Admission	138k [129k, 147k]	92k [83k, 110k]	**74k [74k, 84k]**	1.50 [1.25, 1.78]	1.87 [1.53, 1.99]
IMV Event	156k [138k, 156k]	156k [147k, 174k]	**133k [122k, 144k]**	1.00 [0.83, 1.06]	1.17 [1.04, 1.28]
Hospital Mortality	129k [119k, 147k]	64k [55k, 73k]	**57k [48k, 57k]**	2.00 [1.62, 2.50]	2.25 [2.09, 2.90]
Profound Hyponatremia	156k [147k, 156k]	28k [28k, 37k]	**19k [13k, 29k]**	5.67 [4.00, 5.67]	8.17 [5.44, 10.98]
Severe Hypernatremia	174k [165k, 174k]	28k [28k, 37k]	**19k [19k, 29k]**	6.33 [4.75, 6.33]	9.02 [6.01, 9.02]
Profound Hypokalemia	156k [147k, 165k]	**3673 [3673, 3673]**	3779 [3779, 3779]	42.50 [40.00, 45.00]	41.30 [38.87, 43.73]
Severe Hyperkalemia	138k [129k, 147k]	110k [101k, 138k]	**45k [45k, 56k]**	1.25 [1.00, 1.36]	3.07 [2.45, 3.27]
Anemia	138k [129k, 147k]	92k [73k, 101k]	**60k [48k, 60k]**	1.50 [1.27, 2.00]	2.30 [2.15, 3.07]
Hypoglycemia	129k [119k, 138k]	18k [18k, 18k]	**14k [10k, 14k]**	7.00 [6.50, 7.50]	8.91 [8.27, 13.36]
Tachycardia	129k [119k, 138k]	73k [73k, 92k]	**40k [40k, 54k]**	1.75 [1.40, 1.88]	3.18 [2.39, 3.41]
Aggregate:				**SCOPE**	**REACH**
Mean				7.05 [6.76, 7.37]	8.12 [7.77, 8.82]
Median				1.88 [1.70, 2.19]	3.13 [2.73, 3.24]

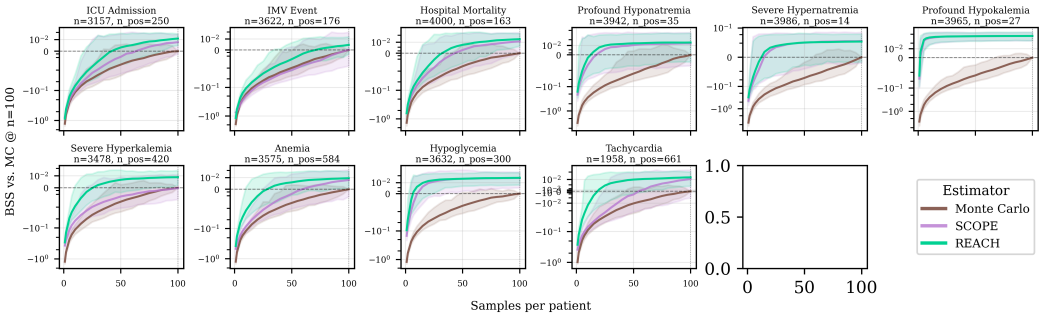


Figure 8: Brier skill scores with 100 sample Monte Carlo as the baseline on the local dataset

M MIMIC-IV CALIBRATION RESULTS

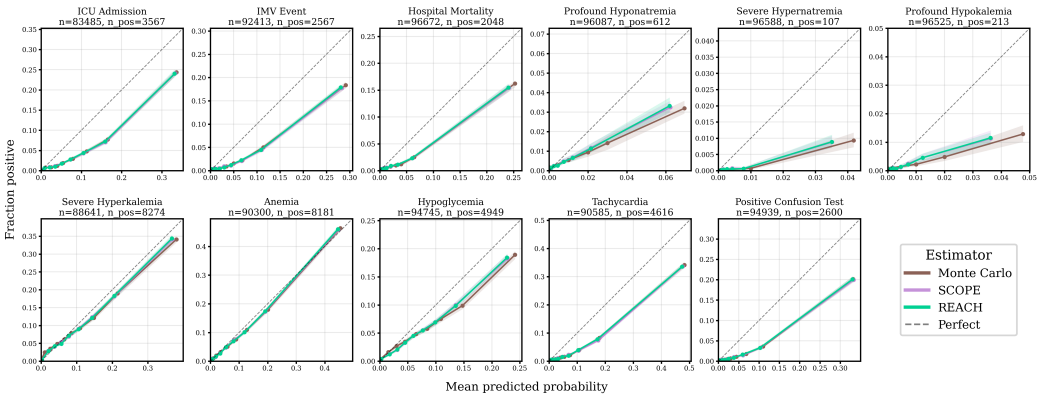


Figure 9: Calibration curves of MIMIC-IV prediction tasks

N LOCAL DATA CALIBRATION RESULTS

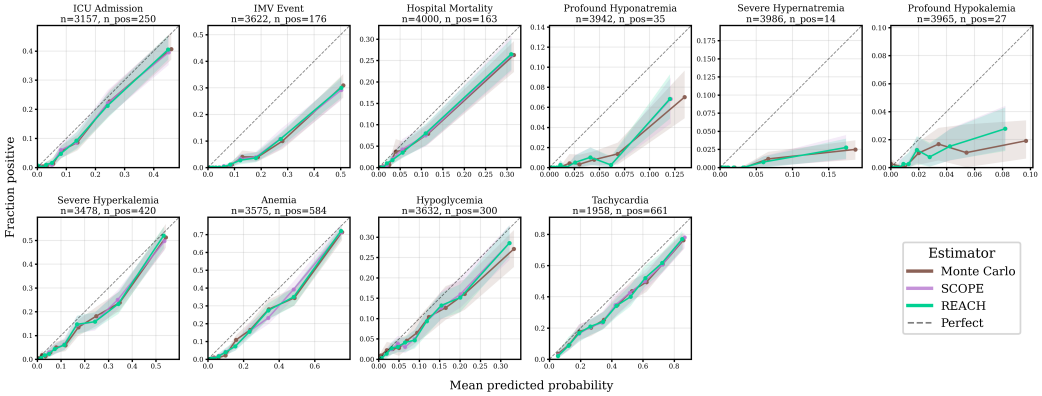


Figure 10: Calibration curves of Local Data prediction tasks

O DISPERSION OF ESTIMATES COMPUTATION

Consider some outcome with general prevalence of $1/10000$. Next, we want to consider the probability that a patient with $10\times$ the general population risk is assigned a risk score by the Monte Carlo estimator that exceeds the risk score of a patient with risk equal to the prevalence.

The PMF for the Binomial distribution is

$$f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

In order to compute the probability that the higher risk patient is ranked higher, we sum over all possible outcomes for the average risk patient and multiply each by the probability that the higher risk patient generates more samples that indicate the outcome of interest:

$$\sum_{k=0}^{100} f\big(k; 100, \tfrac{1}{10000}\big) \cdot \sum_{j=k+1}^{100} f\big(j; 100, \tfrac{1}{1000}\big) = 0.0943.$$

Thus there is only a 9.43% chance that a patient with risk $10\times$ baseline is given a higher risk score than a patient with average risk.

P REACH IS A RAO–BLACKWELLIZATION

Let $(\hat{X}_{1:T}, C_{1:T})$ be a pair of random variables corresponding to a backbone, so that $\hat{X}_{1:T} \sim \hat{p}$ is drawn from the outcome-avoiding distribution, $T = T_E(\hat{X})$, and

$$C_t \sim \text{Bernoulli}(h(\hat{X}_{1:t-1}))$$

where $h(\hat{x}_{1:t-1}) = P(X_t = O | X_{1:t-1} = \hat{x}_{1:t-1})$ for all $1 \leq t \leq T$. Note that each C_t is independent after conditioning on $\hat{X}_{1:T}$. From the argument that $P(A) = P(B)$, it follows that

$$P(T_O < T_E) = \mathbb{E}[\mathbb{1}_{\{\exists t:\, C_t = 1\}}].$$

Clearly, $\hat{X}_{1:T}$ is a sufficient statistic for each C_t and thus all $C_{1:T}$, so that Rao–Blackwell's theorem implies that the following estimator has reduced variance:

$$\begin{aligned}
\mathcal{R}_1 &= \mathbb{E}[\mathbb{1}_{\{\exists t:\, C_t = 1\}} | \hat{X}_{1:T}] \\
&= 1 - \mathbb{E}[\mathbb{1}_{\{C_t \neq 1\, \forall t\}} | \hat{X}_{1:T}] \\
&= 1 - \textstyle\prod_{t=1}^{T} \mathbb{E}[\mathbb{1}_{\{C_t \neq 1\}} | \hat{X}_{1:t-1}] \\
&= 1 - \textstyle\prod_{t=1}^{T} (1 - h(\hat{X}_{1:t-1}))
\end{aligned}$$

where conditional independence and the fact that

$$\mathbb{E}[\mathbb{1}_{\{C_t \neq 1\}} | \hat{X}_{1:T}] = \mathbb{E}[\mathbb{1}_{\{C_t \neq 1\}} | \hat{X}_{1:t-1}]$$

justifies the third equality. REACH is the sampled version of $\mathcal{R}_1$.

Q REACH IS RAO-BLACKWELLIZATION OF ARBITRARY RARE-EVENT IS

Consider some next token generation model that generates a probability distribution for the next token conditioned on the sequence up to this point. Given some arbitrary stopping condition, we want to estimate $P(T_O < T_E)$.

If this outcome of interest token appearing before the end of the timeline is relatively rare event, one straightforward way to decrease the variance of a Monte Carlo sampling estimate is to perform importance sampling in order to increase the probability that the next token is the outcome of interest token during sampling.

Let f denote the ordinary distribution of timelines. Assume that we have some arbitrary proposal distribution g with the following restriction:

$$P_g(X_n \,|X_{1:n-1} \cap O \notin X_{1:n}) = P_f(X_n \,|X_{1:n-1} \cap O \notin X_{1:n})$$

In other words, this class of importance sampling techniques only scales the probability that the outcome of interest token is generated. This is a reasonable restriction since there is no trivial choice of distribution that changes the probability of the non-outcome of interest tokens in a medical setting.

Given this restriction, consider the alternative probability space $(\hat{X}_{1:n}, C_{1:n})$. $\hat{X}_{1:t}$ is drawn from the outcome avoiding distribution and $C_t \sim \text{Bernoulli}(P_g(X_t = O|\hat{X}_{1:t-1}))$. Note that the estimator now becomes: $\frac{1}{m}\mathbb{E}[\frac{f((\hat{X}_{1:T},C_{1:T}))}{g((\hat{X}_{1:T},C_{1:T}))}\mathbb{1}_{C_i=1}]$

By our above restriction, $\frac{f((\hat{X}_{1:T},C_{1:T}))}{g((\hat{X}_{1:T},C_{1:T}))} = \frac{f(C_{1:T})}{g(C_{1:T})}$.

Let Ω_1 be the subset of the original sample space where the outcome of interest is sampled at some point. Let Ω_2 be the subset of the second space where $C_i = 1$ for some i. Let $h : \Omega_1 \to \mathcal{P}(\Omega_2)$ be the following mapping between Ω_1 and the subsets of Ω_2.

(Note that for this argument we are assuming sequences terminate at T_O) For each sequence $Y_{1:T_0}$ in the original sample space, map it to the set of all $(\hat{X}_{1:T}, C_{1:T})$ such that $Y_{1:T_0-1} = \hat{X}_{1:T_0-1}$ and $C_{1:T_O-1} = 0$ but $C_{T_0} = 1$. Note that the probability of $X_{1:T_0}$ and the probability of the event $h(X_{1:T_0})$ are equal:

$$\begin{aligned}
P_g(h(X_{1:T_0})) &= \frac{P_g(Y_1 = X_1)}{1 - P_g(Y_1 = O)} \\
&\quad \cdot \left(\prod_{t=2}^{T_O-1} \frac{P_g(Y_t = X_t|Y_{1:t-1} = X_{1:t-1})}{1 - P_g(Y_t = O|Y_{1:t-1} = X_{1:t-1})} \right) P_g(Y_{T_0} = O|Y_{1:T_0-1} = X_{1:T_0-1}) \\
&\quad \cdot (1 - P_g(Y_1 = O) \prod_{t=2}^{T_O-1} (1 - P_g(Y_t = O|Y_{1:t-1} = X_{1:t-1})) \\
&= P_g(Y_1 = X_1) \left(\prod_{t=2}^{T_O-1} P_g(Y_t = X_t|Y_{1:t-1} = X_{1:t-1}) \right) P_g(Y_{T_0} = O|Y_{1:T_0-1} = X_{1:T_0-1}) \\
&= P_g(X_{1:T_O})
\end{aligned}$$

Next, note that by our restriction:

$$\begin{aligned}
\frac{f(h(X_{1:T_O}))}{g(h(X_{1:T_O}))} &= \frac{P_f(X_1 \neq O)(\prod_{t=2}^{T_O-1} P_f(X_t \neq O|X_{1:t-1}))(P_f(X_{T_O} = O|X_{1:t-1}))}{P_g(X_1 \neq O)(\prod_{t=2}^{T_O-1} P_g(X_t \neq O|X_{1:t-1}))(P_g(X_{T_O} = O|X_{1:t-1}))} \\
&= \frac{f(X_{1:T_O})}{g(X_{1:T_O})}
\end{aligned}$$

Thus, we have shown that the this mapping maintains both probabilities and sample weights. Thus, we can conclude that the variance of the two estimators are equal.

Next, consider the Rao-Blackwellization using the outcome-free backbone as the sufficient statistic. By our earlier work, it is simple to show that this yields:

$$\hat{p} = \sum_{t=1}^{T} \frac{P_f(C_{1:t-1} = 0)P_f(C_t = 1)}{P_g(C_{1:t-1} = 0)P_g(C_t = 1)} P_g(C_{1:t-1} = 0) \cdot P_g(C_t = 1)$$

Which is exactly equal to the non-importance sampling REACH estimator. Since this was a Rao-Blackwellizaiton of an estimator with equal variance to the importance sampling estimator in the original space, this implies that REACH is a lower variance estimator for all possible choices of g that do not alter the distribution of timelines conditioned on the non-appearance of the outcome of interest.

R TOKENIZATION

Tokenization proceeded as follows. Both MIMIC and local data were converted to the Common Longitudinal ICU Format (CLIF: 23) version 2.1 standard that harmonizes critical care data across institutions by mapping site-specific laboratory and vital codes to a minimum set of common ICU data elements. Each hospitalization was assigned a sequence of tokens, starting with a `BOS` beginning-of-sequence token. Tokens for age, sex, race, ethnicity, and admission type were inserted in this order at the time of admission. Interdepartmental transfers within the hospital received one token each for ingress and egress, at their respective times. Medications with dosages, labs with quantitative results, and vitals with measured values each received category-value tokenization. For each category (standardized kind medication, type of lab, type of vital), all numerical values reported within the training set corresponding to that category were used to determine decile cutoffs. All values for that category (in training, tuning, and held-out data) were then binned into `Q0`, `Q1`, ..., `Q9` according to those deciles (cf. 26). For each category-value pair, a new "fused" token (27) was created (for example, `LAB-RES//so2_arterial_Q0` corresponds to a lab result for arterial SO_2 in the lowest decile). Tokens for medications and labs were inserted at the time of administration or measurement. For labs, a token corresponding to the lab was inserted at the time the lab was ordered, and two tokens correspond to both the lab and its binned numerical value were inserted at the time the result was available. Changes in code status were tokenized similarly. Continuous renal replacement therapy with blood flow rate, patient assessments with numerical value where appropriate, and respiratory support with fio2, peep, and tidal volume values also received category-value tokenization. Changes in code status, the practice of proning a patient, and discretized measurements of sofa score (as provided by the clifpy package) were also tokenized. All sequences terminate with an `EOS` end-of-sequence token. See Table 6 for further details on the tokens.

Table 6: A breakdown of the 1385-token vocabulary by token type

category	description	example token(s)	count
ADMN	admission type	`ADMN//direct`, `ADMN//elective`	3
AGE	age	`AGE//age_Q0`, `AGE//age_Q5`, `AGE//age_Q9`	10
ASMT	assessment	`ASMT//braden_activity_Q6`, `ASMT//rass_Q4`	75
CODE	code status	`CODE//dnr`, `CODE//full`	4
CRRT	dialysis	`CRRT//cvvh_Q0`, `CRRT//cvvhd_Q2`	23
DSCG	discharge	`DSCG//acute_care_hospital`, `DSCG//home`	12
ETHN	ethnicity	`ETHN//hispanic`, `ETHN//unknown`	3
LAB-ORD	lab order	`LAB-ORD//albumin`, `LAB-ORD//alt`, `LAB-ORD//wbc`	45
LAB-RES	lab result	`LAB-RES//albumin_Q0`, `LAB-RES//alt_Q2`	473
MED-CTS	continuous medication	`MED-CTS//dopamine_Q8`	388
MED-INT	intermittent medication	`MED-INT//adenosine_Q7`	155
POSN	proning	`POSN//prone`	1
RACE	race	`RACE//asian`, `RACE//unknown`, `RACE//white`	7
RESP	respiratory support	`RESP//fio2_set_Q1`	35
SEX	sex	`SEX//female`, `SEX//male`	2
SOFA	sofa score	`SOFA//cns-0`, `SOFA//coag-0`, `SOFA//resp-4`	29
TIME	time-spacing	`TIME//12h-1d`, `TIME//1w-2w`, `TIME//6mt+`	13
VTL	vitals	`VTL//dbp_Q0`, `VTL//dbp_Q5`, `VTL//weight_kg_Q9`	88
XFR-IN	transfer in	`XFR-IN//ed`, `XFR-IN//psych`, `XFR-IN//ward`	8
XFR-OUT	transfer out	`XFR-OUT//ed`, `XFR-OUT//psych`, `XFR-OUT//ward`	8
special	miscellaneous	`BOS`, `EOS`, `UNK`	3

S SPONTANEITY AND PREVALENCE FIT

MC / SCOPE multivariate $R^2 = 0.936$

$$\log(\textbf{efficiency}) = -1.436\log(\textbf{prev}) + 1.022\log(\textbf{spont}) + 0.406$$

MC / REACH multivariate $R^2 = 0.943$

$$\log(\textbf{efficiency}) = -1.384\log(\textbf{prev}) + 1.017\log(\textbf{spont}) + 0.526$$

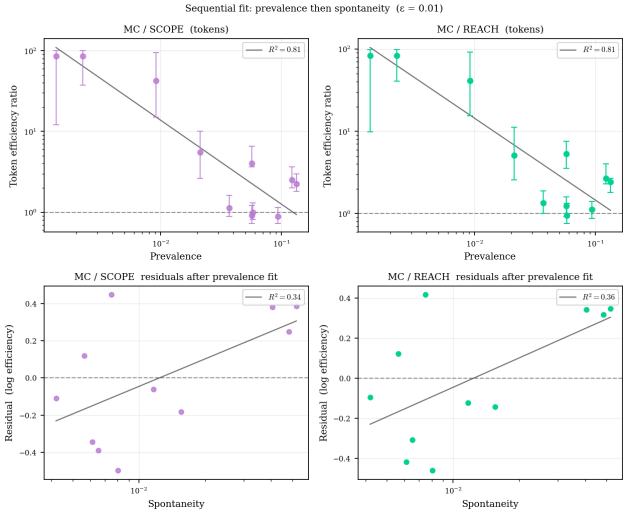


Figure 11: MIMIC-IV Sequential Fits

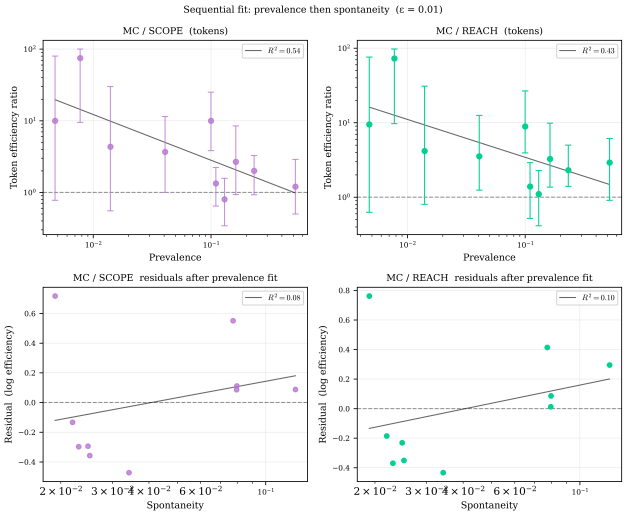


Figure 12: Local data sequential fits

T FULL TABLE 1

Table 7: Token efficiency of SCOPE and REACH vs. Monte Carlo baseline at $\varepsilon = 0.01$ on MIMIC-IV. Values are median token cost per patient [95% bootstrap CI] across 100 replications. REACH cost is $M_1 + p \times M_2$. **Bold** indicates lowest token cost (where M_1 is mean standard generation length and M_2 is mean task specific REACH completion rate and p is predicted prevalence). Ratios > 1 indicate fewer tokens needed than MC.

Outcome	T_{MC}	T_{SCOPE}	T_{REACH}	MC/SCOPE	MC/REACH
ICU Admission	30k [27k, 33k]	30k [24k, 36k]	31k [24k, 35k]	1.00 [0.82, 1.31]	0.95 [0.76, 1.34]
IMV Event	33k [27k, 36k]	36k [27k, 42k]	**26k [23k, 34k]**	0.92 [0.73, 1.22]	1.25 [0.93, 1.59]
Hospital Mortality	33k [27k, 42k]	**5993 [3566, 12k]**	6413 [3207, 13k]	5.50 [2.62, 10.00]	5.14 [2.57, 11.21]
Profound Hyponatremia	51k [42k, 57k]	**1199 [599, 2996]**	1240 [620, 3099]	42.50 [15.00, 95.00]	41.09 [15.47, 91.85]
Severe Hypernatremia	51k [30k, 60k]	**599 [599, 4195]**	611 [611, 5243]	85.00 [12.04, 100.00]	83.31 [9.80, 98.01]
Profound Hypokalemia	51k [39k, 60k]	**599 [599, 1199]**	613 [613, 1225]	85.00 [37.50, 100.00]	83.16 [40.30, 97.84]
Severe Hyperkalemia	30k [27k, 33k]	12k [8989, 15k]	**11k [7457, 11k]**	2.50 [2.00, 3.67]	2.68 [2.31, 4.02]
Anemia	27k [24k, 30k]	12k [8989, 15k]	**11k [11k, 15k]**	2.25 [1.80, 3.00]	2.42 [1.81, 2.68]
Hypoglycemia	36k [33k, 42k]	8989 [5993, 8989]	**6788 [4752, 10k]**	4.00 [3.67, 6.50]	5.30 [3.53, 7.57]
Tachycardia	24k [21k, 27k]	27k [21k, 33k]	**21k [17k, 26k]**	0.89 [0.73, 1.14]	1.12 [0.87, 1.40]
Positive Confusion	27k [24k, 33k]	24k [18k, 30k]	**20k [16k, 28k]**	1.12 [0.89, 1.62]	1.36 [1.01, 1.89]
Aggregate:				**SCOPE**	**REACH**
Mean				20.97 [12.91, 26.98]	20.71 [12.74, 26.68]
Median				2.50 [2.25, 3.51]	2.68 [2.42, 4.02]

Table 8: Token efficiency of SCOPE and REACH vs. Monte Carlo baseline at $\varepsilon = 0.01$ on the local dataset. Values are median token cost per patient [95% bootstrap CI] across 100 replications. REACH cost is $M_1 + p \times M_2$. **Bold** indicates lowest token cost. Ratios > 1 indicate fewer tokens needed than MC.

Outcome	T_{MC}	T_{SCOPE}	T_{REACH}	MC/SCOPE	MC/REACH
ICU Admission	73k [37k, 119k]	**55k [28k, 92k]**	53k [32k, 84k]	1.33 [0.64, 2.23]	1.39 [0.52, 2.91]
IMV Event	73k [37k, 115k]	92k [50k, 170k]	**66k [38k, 133k]**	0.80 [0.34, 1.57]	1.11 [0.41, 2.28]
Hospital Mortality	101k [55k, 156k]	**28k [9181, 83k]**	29k [9513, 72k]	3.67 [1.00, 11.43]	3.54 [1.24, 12.55]
Profound Hyponatremia	119k [64k, 174k]	**28k [3673, 174k]**	29k [3822, 162k]	4.33 [0.55, 30.00]	4.16 [0.80, 30.87]
Severe Hypernatremia	92k [15k, 174k]	**9181 [1836, 125k]**	9670 [1934, 125k]	10.00 [0.78, 80.00]	9.49 [0.63, 75.96]
Profound Hypokalemia	138k [28k, 184k]	**1836 [1836, 9181]**	1889 [1889, 9447]	75.00 [9.47, 100.00]	72.89 [9.72, 97.18]
Severe Hyperkalemia	73k [46k, 119k]	28k [13k, 65k]	**22k [11k, 45k]**	2.67 [0.93, 8.46]	3.27 [1.36, 9.85]
Anemia	55k [37k, 73k]	28k [18k, 55k]	**24k [12k, 36k]**	2.00 [0.93, 3.26]	2.30 [1.40, 5.01]
Hypoglycemia	92k [50k, 143k]	**9181 [3673, 23k]**	10k [4122, 18k]	10.00 [3.83, 25.00]	8.91 [3.93, 26.73]
Tachycardia	55k [28k, 101k]	46k [28k, 83k]	**19k [13k, 40k]**	1.20 [0.50, 2.88]	2.92 [0.91, 6.14]
Aggregate:				**SCOPE**	**REACH**
Mean				11.10 [4.34, 19.05]	11.00 [4.81, 19.03]
Median				3.17 [1.45, 5.50]	3.41 [1.89, 6.03]