

Translating Time-Series Foundation Models into AI-Enhanced Medication Inventory Control in a Large Urban Health System

Derek Long¹ Tianyu Wang² Charles Branas¹ Henry Lam¹ Jay Sethuraman¹ Donald U Apakama³ Sarah McCuskee³
Vikram Mulchandani³ Yosef Travis³ Alexis Zebrowski³ Rachael Piltch-Loeb³

¹Columbia University ²Harvard University ³Mount Sinai Hospital

Motivation

In 2024, the Mount Sinai Health System across the New York metropolitan area reported [3]:

7 hospital campuses	145,903 inpatient discharges	6,076,491 outpatient visits	372,959 emergency department visits
---------------------------	------------------------------------	-----------------------------------	---

Hence, **timely medication access is critical**, but **demand surges and heterogeneous medication needs make it difficult to maintain reliable medication availability** across all stations and departments, potentially leading to devastating consequences for patients.

Challenges:

- Low prediction accuracy from traditional forecasting methods in medication areas;
- Failure of current **contextual restocking policy for long-term planning**;
- Multi-objective optimization** in practical healthcare operations.



Key Contributions

We develop a **predict-then-optimize** framework that:

- Utilizes *distributional demand forecasts from time-series foundation models*;
- Generates an *interpretable replenishment policy* which balances holding cost, stockout-risk and refill workload;
- Achieves *lower average refill workload, holding cost, and stockouts* compared to the existing policy.

Methodology Overview

Inventory Dynamics

At the beginning of day t , for each medication-station pair $a \in \mathcal{A}$, we observe:

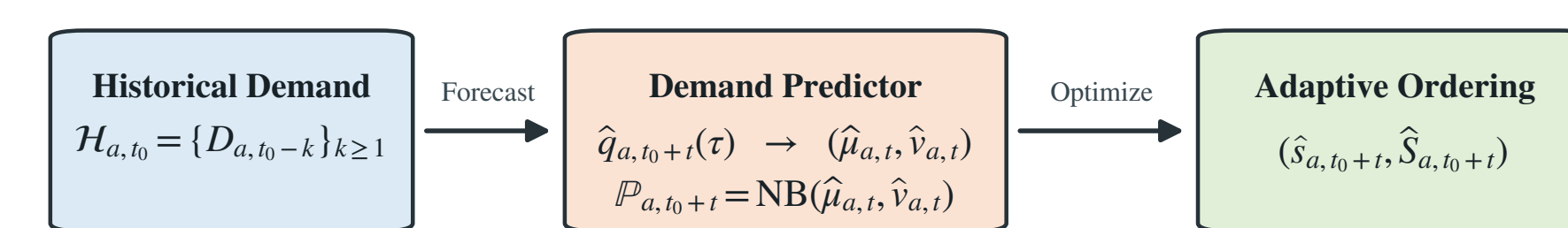
- $I_{a,t} \geq 0$: on-hand inventory;
- $x_{a,t}$: context (e.g., recent demand, patient-level information, and station-level information).

Restocking decision (s-S policy). The reorder threshold $s_{a,t}$ and order-up-to level $S_{a,t}$.

- When $I_{a,t} \leq s_{a,t}$, we refill the inventory to $S_{a,t}$ before demand $D_{a,t} \geq 0$ occurs;
- Otherwise, we do not change.

Predict-then-Optimize

The first known integration of time-series foundation models in healthcare optimizations



Distributional demand forecasts: Obtain accurate forecast of future demand distributions.

- Goal:** minimize forecast error.
- Output:** scenarios generated from the distributional model.



Constrained inventory optimization: Decide the inventory level over the future planning horizon.

- Goal:** minimize operational costs.

Demand Forecasting

Computational Procedure

- Step 1: Quantile generalization.** Generate quantile forecasts $\hat{q}_{a,t_0+t}(\tau)$ for a 14-day horizon at probability levels $\tau \in \mathcal{Q} = \{0.1, 0.2, \dots, 0.9\}$;
- Step 2: Tail Extrapolation.** Fit a negative-binomial (NB) marginal for every medication-station pair and forecast horizon, with mean μ and shape r solving:

$$(\hat{\mu}_{a,t_0+t}, \hat{r}_{a,t_0+t}) = \arg \min_{\mu, r > 0} \sum_{\tau \in \mathcal{Q}} \left[\frac{q^{\text{NB}}(\tau; \mu, r) - \hat{q}_{a,t_0+t}(\tau)}{1 + \hat{q}_{a,t_0+t}(\tau)} \right]^2.$$

Evaluation Pipeline

- Selected foundation model: **Chronos-Bolt** [2];
- In-context Learning with contexts and demand;

Empirical Validation

- Baseline methods:
 - Other time-series foundation models: **DeepAR** [4], **TimesFM-3** [1];
 - Classical ML methods: random forest, 28-day rolling mean.
- Chronos-Bolt** has the best performance overall.

Table 1. Demand forecast accuracy using different foundational models.

Metric	Chronos-Bolt	DeepAR	TimesFM-3	Random forest	rolling mean
MAE ↓	1.837	1.843	1.852	1.834	1.914
RMSE ↓	2.780	2.861	2.788	2.883	2.879
WAPE ↓	0.721	0.724	0.728	0.722	0.752
CRPS ↓	1.220	1.257	1.228	–	–
NLL ↓	1.801	1.761	1.775	–	–
Cov.90	0.891	0.882	0.906	–	–
PB.90 ↓	0.558	0.573	0.548	–	–

Inventory Optimization

Computational Procedure

Optimization Model. For each medication-station pair $a \in \mathcal{A}$, we compute the inventory decision indexed by $\{s_{a,t}, S_{a,t}\}_{t \in [14]}$ for the coming 14 days:

$$\begin{aligned} \min_{0 \leq s_{a,t} \leq S_{a,t}} \quad & \frac{1}{14} \sum_{t=1}^{14} \mathbb{E}[H_{a,t}] && \text{(minimize holding cost)} \\ \text{s.t.} \quad & \mathbb{P}(D_{a,t} > H_{a,t}) \leq \beta, \quad t = 1, \dots, 14 && \text{(low stockout risk)} \\ & \mathbb{E}[R_{a,t}] \leq B, \quad t = 1, \dots, 14 && \text{(low refill workload)} \end{aligned}$$

Extension of inventory decision: periodic restriction, borrowing behavior, etc.

Optimization Algorithm.

- Differential Rollout:** Smooth the non-differentiable stockout and refill indicators;
- Gradient Descent with Projected Dual Ascent:** Adam update + projected dual ascent.

Empirical Validation

Our learned policy reduces average holding cost by **3.3%**, average workload by **28.4%**, and stockout days by **21.2%** compared with the historical policy.

Table 2. Performance of the learned policy (L) and actual historical policy (H), L/H

Site-station code	Med.	Holding (Post-refill Inventory)	Workload (Daily Refills)	Stockout Days
Hosp-A-ED-1	1	10.64 /12.59	0.088 /0.108	0/0
Hosp-A-ED-2	1	14.06/ 12.47	0.126 /0.142	1/5
Hosp-B-ED-1	1	17.18 /21.91	0.140/ 0.126	4/4
Hosp-C-ED-1	1	52.00 /74.00	0.142/ 0.126	5/0
Hosp-D-ED-1	1	15.03/ 8.86	0.195 /0.358	6/9
Hosp-D-ED-2	1	14.30/ 8.47	0.149 /0.288	2/14
Hosp-E-ED-1	1	23.26/ 18.22	0.306 /0.389	6/18
Hosp-C-ED-1	2	22.50/ 19.45	0.095 /0.187	6/0
Hosp-E-ED-1	2	11.10/ 10.19	0.139 /0.206	11/2
Overall L/H (change)		20.01 /20.68 (−3.3%)	0.153 /0.214 (−28.4%)	41/52 (−21.2%)

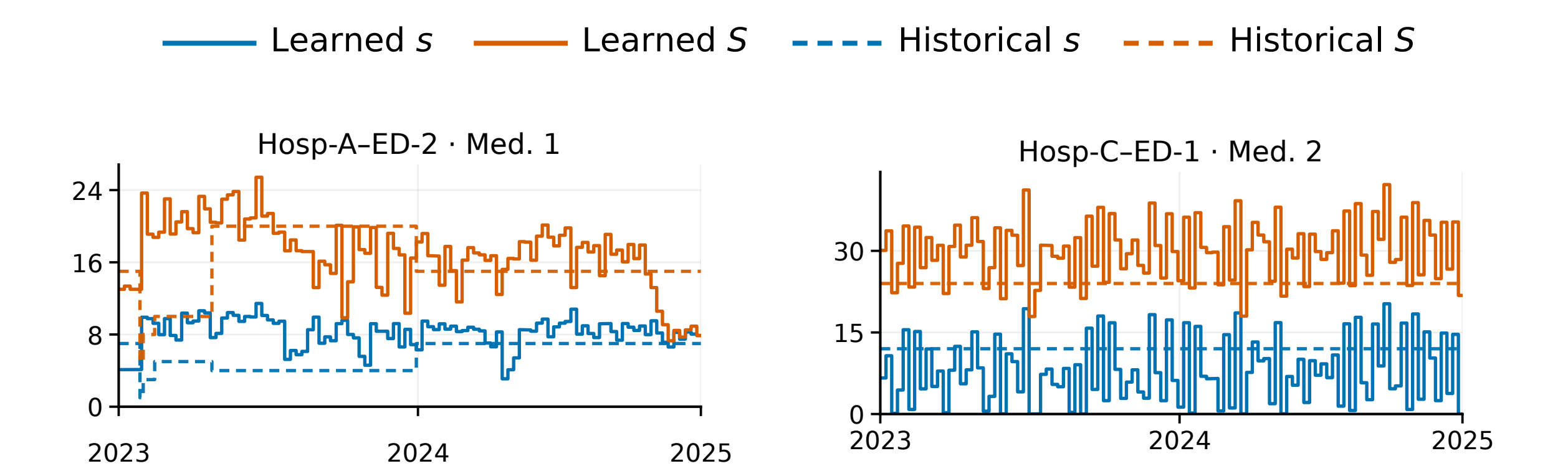


Figure 1. Learned and historical weekly (s, S) levels for representative stations.