Healthcare Patient-Flow Optimization Agent

AI Decision-Support System for Near-Term Hospital Congestion Risk Screening

Author: Taash Chikosi

Target roles: AI Consultant / AI Product Manager / AI Solutions Architect

Target employers: Top-tier consulting firms and large enterprises adopting AI at scale

Value proposition: Design and deliver end-to-end AI agent systems that solve real business problems, quantify ROI, and integrate human-in-the-loop decision-making

Strengths: Business problem framing, AI workflow design, ML-driven predictions, RAG, executive dashboards

Outcome focus: Measurable efficiency gains, cost reduction, or revenue impact

Table of Contents

1. Executive Summary

This engagement delivers a live AI decision-support system designed to help hospital leaders **anticipate near-term congestion risk** and prioritize operational attention over the next 24 hours.

The system forecasts likely **maximum occupancy pressure** and **average patient wait times** based on the hospital's current operational state, recent demand patterns, and time-of-day effects. It then translates those forecasts into a **simple, conservative risk signal** supported by explicit assumptions, limitations, and governance controls.

The system is intentionally designed for **screening and early warning**, not for automated staffing, clinical decisions, or real-time control. Its purpose is to help decision-makers answer a single operationally critical question early enough to matter:

"Is this hospital at meaningful risk of congestion in the next 24 hours, and should human review be triggered now?"

The deployed solution combines:

- a machine-learning forecasting layer trained on realistic hospital-like operational data, and
- a retrieval-augmented reasoning (RAG) layer that explains outputs, limitations, and appropriate responses using a curated knowledge base.

2. Decision Context & Objectives

Decision context

Hospitals operate under constant uncertainty. Demand fluctuates by hour, day, and season; staffing levels vary; and congestion often emerges **before** it is visible in headline metrics.

While hospitals collect vast amounts of operational data, leaders often lack a **forward-looking, decision-grade signal** that integrates current conditions into a near-term risk outlook.

The challenge is not perfect prediction, but **earlier awareness**.

Business question

Based on the hospital's current operational state, what is the likelihood of congestion over the next 24 hours, and does that risk justify proactive human review or intervention?

Intended users

- **Hospital operations leaders** — early warning and prioritization
- **Patient-flow and capacity teams** — situational awareness and escalation
- **Executive leadership** — governance-safe summaries of operational risk
- **Public-sector and health-system planners** — resilience and surge preparedness

Success criteria

- Forward-looking forecasts tied to current operational conditions
- Simple, interpretable outputs suitable for rapid decision-making
- Explicit signaling of risk rather than false precision
- Clear boundaries on appropriate and inappropriate use
- Transparent, auditable logic suitable for governance review

3. How the System Works (High Level)

The system follows a deliberately simple and defensible workflow:

Observe → Forecast → Screen → Explain → Decide

Inputs (current operational state only)

Users describe the hospital's **current state**, not future assumptions:

- Bed capacity
- Current occupancy (%)
- Recent arrivals pressure (low / normal / high)
- Staffing level (relative)

- Current average wait time
- Date and hour (to capture time-of-day and weekday effects)

Recent 24-hour rolling indicators (arrivals, occupancy, wait times) are **approximated conservatively** to reflect realistic recent history without requiring full historical feeds.

Outputs

For each scenario, the system produces:

- **Predicted maximum occupancy ratio (next 24h)**
- **Predicted mean patient wait time (next 24h)**
- **Screening-level congestion risk flag (LOW / HIGH)**
- Evidence-backed explanations and governance guidance via the RAG assistant

Outputs are framed to support **human judgment**, not automated control.

4. What the Results Show

For a representative hospital scenario, the system produces:

- A quantified forecast of near-term occupancy pressure
- A predicted average wait time for the next 24 hours
- A conservative risk classification indicating whether congestion risk is elevated

The deployed model achieves strong screening-grade performance, with low absolute error relative to realistic operational variability. Importantly, outputs are intentionally **directional rather than deterministic**, answering:

“Should we be paying closer attention right now?”

The system does not prescribe staffing levels, bed closures, or clinical actions.

5. Why the Results Are Trustworthy

This system was designed to be **decision-grade**, not merely predictive. Trustworthiness is established through realistic data design, disciplined modeling, validation rigor, and explicit governance controls.

Data strategy and realism

Access to real hospital operational data is limited and highly sensitive. The dataset was therefore engineered to **behave like real hospital operations**, not idealized simulations.

Key principles included:

- Separation of structural capacity, demand dynamics, and service performance
- Explicit modeling of diurnal, weekly, and seasonal patterns
- Inclusion of volatility, noise, and demand surges
- Avoidance of best-case or steady-state bias

Synthetic data generation and ML readiness

Synthetic data was generated via a reproducible, code-driven pipeline that preserved causal relationships (e.g., demand → occupancy → waits) while injecting uncertainty and measurement error.

The dataset was designed to be:

- Sufficient for supervised learning
- Representative of operational decision contexts
- Suitable for time-aware validation

Feature engineering and unit of analysis

The unit of analysis is **one hospital state at one point in time**, forecasting outcomes over the next 24 hours.

All features are derived from:

- current observable conditions, or
- recent historical summaries available at decision time

This ensures:

- No look-ahead bias
- Deployment realism
- Alignment with real operational workflows

Training, validation, and performance criteria

Models were trained using a **chronological split**, ensuring that training data always precedes test data in time. This mirrors real deployment, where forecasts are always made for the future.

Performance evaluation prioritized:

- Mean Absolute Error (MAE) over abstract fit metrics
- Stability across scenarios
- Robustness under noisy inputs

A Random Forest model was selected for its balance of performance, stability, and interpretability under operational uncertainty.

Governance, confidence, and human oversight

Risk classification is deliberately conservative and rule-based. Governance controls include:

- Explicit separation between forecasts and decisions
- Deterministic risk-flag logic
- Mandatory human review under high-risk signals
- Clear language constraints to prevent over-claiming
- Logged inputs and outputs for auditability

Together, these mechanisms ensure the system **supports human judgment rather than replacing it**.

6. What Is Live & What Comes Next

What is live today

- Deployed Streamlit web application
- Version-controlled ML inference artifacts
- Persisted RAG vector store for healthcare-specific knowledge
- Secure secrets management and reproducible deployment

Users can interactively adjust operational conditions and observe how near-term risk responds.

What the deployment proves

- End-to-end operability
- Stable, low-latency inference
- Interpretable outputs for non-technical stakeholders
- Practical governance integration

Scope boundaries

The system **is designed for**:

- Early warning and screening
- Operational awareness
- Decision support and escalation

The system **is not designed for**:

- Automated staffing decisions
- Clinical decision-making
- Real-time control systems
- Regulatory or compliance certification

What comes next

Potential enhancements include:

- Integration with live hospital data feeds
- Expanded risk tiers (e.g., LOW / MODERATE / HIGH)
- Longer-horizon forecasting
- Cross-hospital benchmarking
- Integration with surge and contingency planning workflows

Closing Summary

This Healthcare Patient-Flow Optimization Agent demonstrates how AI can be applied responsibly to **improve operational resilience in complex, high-stakes environments**.

By combining realistic data design, disciplined forecasting, explicit governance, and evidence-backed explanation, the system delivers **early, defensible signals** that help hospital leaders act sooner—without overstating certainty or automating judgment.