

Agentic Workflows III: The Patient Navigator

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 21 of "Healthcare Analytics and AI: Building Systems That Actually Work"

The essentials: what to know, why it matters at the bedside, and how models fail.

One-sentence thesis. A patient navigator is only as good as its worst-served user, so it must ground every answer in that patient's own record, escalate red flags with near-perfect recall, and never let the patient forget they are talking to a machine.

PART 1: The Navigation Gap After Discharge

Where the Preventable Harm Lives

Care does not fail only in the operating room. It fails in the unstructured space between leaving a clinical setting and returning to one, the handoff from cockpit to turbulence, where the patient goes from 24-hour monitoring by professionals to self-management in an environment the hospital cannot see. The numbers locate the danger precisely: the national 30-day readmission rate hovers near **14 percent**, each readmission averages about **\$16,300**, and the first 72 hours after discharge are the most dangerous window in a patient's trajectory. The Chapter 7 readmission models identify high-risk patients, but a risk score sitting in the EHR prevents nothing; the gap is not prediction, it is intervention.

Watch out. A fourteen-page discharge packet written at a twelfth-grade level, handed to a patient who reads at a sixth-grade level, is a failure waiting to happen even when the medicine was perfect.

Extend the Nurse, Do Not Replace Her

The fix is not to automate the nurse away but to let her reach everyone. In 2025 Universal Health Services deployed Hippocratic AI voice agents that call discharged patients, review instructions, confirm medication understanding, probe for symptoms, and escalate to human nurses on clinical concern, conversational LLM calls, not scripted IVR menus, and UHS reported **9.0 out of 10** patient satisfaction with full AI disclosure. The leverage is a triage net cast over everyone: a hospital can staff post-discharge calls to maybe 30 percent of patients, while the agent calls 100 percent within 24 hours and sorts them so nurses spend their time where it matters. The point is not replacing human judgment but concentrating it rather than smearing it thin across an undifferentiated call list, and this extend-don't-replace pattern recurs through the whole chapter.

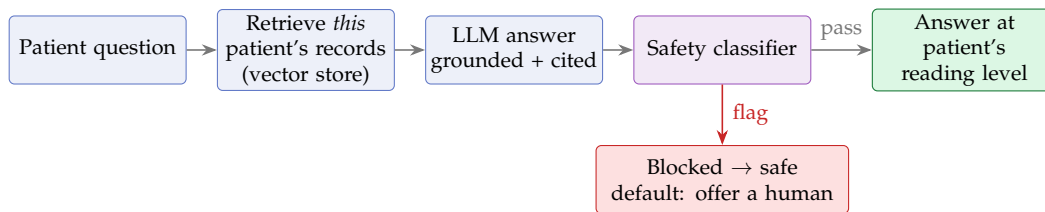
Watch out. The voice-agent stack needs the discharge summary, med list, and follow-up schedule; natural-language understanding to decode "I feel funny"; a protocol engine to decide when to escalate; and the reading-level calibration most clinical AI ignores.

RAG: Look It Up, Do Not Guess

“Can I take ibuprofen with my new blood thinner?” is a question about one specific patient’s medications and history, so a generic chatbot gives a generic, sometimes dangerous answer. The safer design looks it up instead of guessing: retrieval-augmented generation ingests the patient’s discharge summary, med list, and instructions into a vector store, retrieves the chunks relevant to the question, and generates an answer grounded in those documents.

In symbols, the answer is a function of the question conditioned on that patient’s retrieved records:

$$\text{answer} = f_{\text{LLM}}(\text{question} \mid \text{retrieved records for THIS patient})$$



Retrieval-augmented generation answers only from this patient’s own documents and cites the source; every response then passes a safety classifier that blocks hallucinated meds, contradicted allergies, or anything resembling a diagnosis, replacing a flagged answer with a safe default that offers a human.

The Health Literacy Constraint

The average discharge instruction is written for a reader who does not exist. **36 percent** of U.S. adults have basic or below-basic health literacy, yet most discharge instructions are written at a tenth- to twelfth-grade level, and LLMs can hit 100 percent readability compliance when told to rewrite at a sixth-grade level. The catch is that simplification is clinically necessary and technically treacherous: “avoid abrupt discontinuation” is code, “don’t stop suddenly” is plain and

ANALOGY

The difference between a student bluffing on a closed-book exam and one who opens the patient’s chart first, then answers.

A well-built RAG system answers only from retrieved context, cites the source document, communicates at the patient’s reading level, and refuses to diagnose or change a treatment plan, and every response then passes a safety classifier that blocks hallucinated meds, contradicted allergies, and anything resembling a diagnosis, replacing a flagged answer with a safe default that offers a human.

Watch out. That safety layer is not optional decoration; it is the last gate before a patient acts on the words. A confident hallucination dressed in hedging language (“you may be experiencing a mild reaction, consider contacting your doctor”) is worse than an honest “I don’t have that information.”

correct, and “never stop” is plain and wrong, because the drug can be tapered under guidance. The same rewrite that helps one patient can strip the urgency or invert the meaning for another, which is why human review remains a requirement for patient-facing RAG.

Watch out. Machine-translation scores (BLEU, METEOR) measure fluency, not clinical safety: “take with food” rendered as “eat before taking” is fluent and dangerous.

Classify Intent Before You Answer

The system's first move is not to answer but to figure out what kind of message just arrived, because the

category, not the words, determines the routing. Six core intents each carry their own risk tier, from a low-risk information request to a critical, act-immediately emergency.

Intent	Example	Routing	Risk
Information request	"What is metformin for?"	RAG to education content	Low
Symptom report	"My incision looks red."	Structured follow-up	Medium
Medication question	"I missed my morning dose."	Med-rec with dosing rules	Medium
Urgent concern	"Chest hurts, can't breathe."	Immediate 911; no clarifying	Critical
Administrative task	"Refill my prescription."	Staff or pharmacy handoff	Low
Emotional distress	"I can't do this anymore."	Validate; assess safety; offer human	High

Messages rarely stay in one box. "I'm so tired, is it the drug or the disease or am I depressed?" is symptom, medication, and emotional distress at once, and the rule is to route to the highest-clinical-risk category present, or straight to a human.

Escalation and the Recall That Cannot Miss

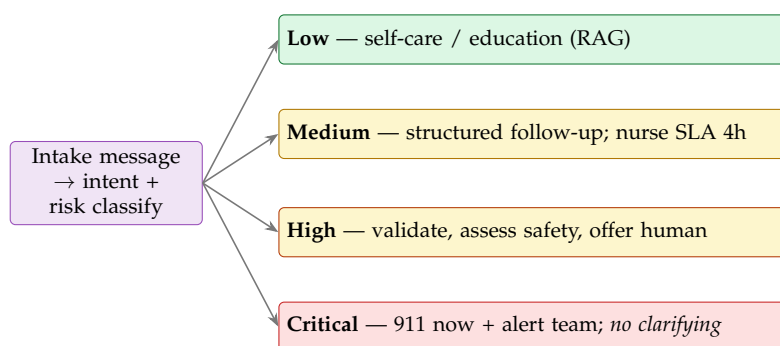
On the categories that can kill, one metric dominates: red-flag recall, the fraction of true emergencies the system actually escalates, which has to approach 1.

$$\text{recall}_{\text{red flag}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \rightarrow 1$$

Hard triggers (suicidal ideation, crushing or radiating chest pain, stroke signs, "I took two doses by mistake," anaphylaxis) always escalate with no follow-up questions and no clarification: "This could be an emergency. Call 911 now. I am also alerting your care

team." Soft triggers (repeated same-symptom messaging, escalating distress, plan non-adherence) flag for human review within a defined SLA. A smoke detector is judged by whether it ever misses a fire, not by how rarely it false-alarms, so on red flags the cost is asymmetric and a false negative can be fatal, the same asymmetric-loss logic as Chapter 7. SLAs make the net real, hard triggers to a human within one minute in business hours and soft triggers within four hours, because an escalation queue without SLAs gives the appearance of safety without the substance.

Watch out. State and memory are part of safety. If a patient reports ankle swelling Monday and the agent re-asks Wednesday as if new, it reads as incompetent, whereas "you mentioned swelling Monday, better, worse, or the same?" is clinically coherent, and that requires conversation state persisted across sessions, not a stateless loop.



Every message is classified before anything is generated, because the risk tier sets the routing. Red-flag (critical) messages escalate immediately with no clarifying questions; the metric that matters is red-flag recall approaching 1, since a false negative here can be fatal.

■ Multilingual Outreach as an Equity Lever

Language is one of the largest, most fixable gaps in post-discharge care. **26 million** U.S. adults have limited English proficiency (LEP), and LEP patients face **33 percent** higher odds of 7-day readmission; because LLMs are natively multilingual, the navigator becomes multilingual at near-zero marginal cost. The effect can be large: in a 2025 JMIR study, Spanish-speaking patients receiving AI outreach in their preferred language had a **2.6-fold** higher colorectal-screening (FIT) opt-in rate than English speakers, reversing the historical disparity.

Watch out. The emerging consensus is tiered: AI handles routine multilingual navigation while professional interpreters handle high-stakes conversations like consent and diagnosis disclosure. AI can be the interpreter in the hallway, never the only interpreter in the consent room, and validation in each target language by bilingual clinicians is the real bottleneck.

■ Trust, Disclosure, and False Intimacy

Trust is earned, not assumed, and the data shows how. **75 percent** of patients do not trust AI to make health-care decisions, but 65 percent say they would be more comfortable if their doctor explained how it is used,

so build trust incrementally: disclose the agent is AI, reveal what data it can see, keep a visible one-tap path to a human, and earn credibility on low-stakes tasks before clinical ones.

🔗 ANALOGY

Show your work. “I can see from your discharge summary you were prescribed lisinopril 10 mg, is that what you have?” builds trust; “Are you taking your lisinopril?” feels like surveillance.

Deception backfires, too: patients who discover undisclosed AI report lower trust, and the UHS 9.0 rating came with full disclosure. **Watch out.** False intimacy is a safety hazard, not a feature, and the Gavalas case (Chapter 18) showed its lethal end. Patients must never forget they are talking to a system, no terms of endearment, no first names without permission, no simulated caring, and an “I am not a doctor” disclaimer is necessary but insufficient; only interactive disclosure, where the patient demonstrates understanding or opts in, actually changes behavior. On channel and identity safety, verify identity before acting (Doctronic uses selfie plus photo ID) and never train on interaction data without explicit informed consent.

Doctronic and the Cases It Never Escalates

At the far end of the autonomy spectrum, the AI stops assisting and starts deciding. In January 2026 Utah authorized Doctronic, under a regulatory sandbox, to autonomously renew prescriptions from a **191-medication** formulary without a physician approving each renewal: a human writes the first prescription, the AI handles renewals via a structured questionnaire, red-flags to a physician when it sees new symptoms, interactions, or non-adherence, and issues the straightforward renewals itself. It earned that autonomy like a learner's permit, with physicians manually reviewing the first 250 prescriptions in each drug class before the model could operate solo, calibrating its judgment class by class. The exclusion list is where the safety actually lives: no controlled substances, no ADHD meds, no injectables, no narrow-therapeutic-index drugs needing lab monitoring.

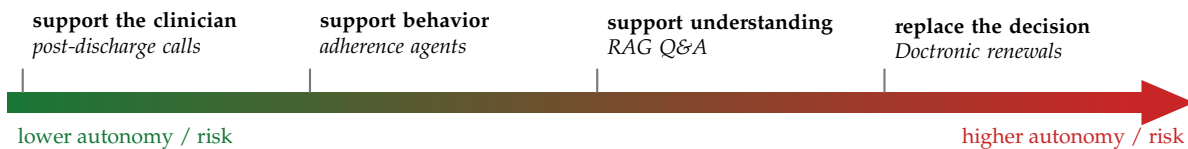
Utah's first public findings, in May 2026, are more informative than any vendor claim: the AI recommended renewal in **72 percent** of the cases it triaged and escalated the rest; reviewing physicians agreed with those renewal recommendations **91 percent** of the time (97 percent counting either of two reviewers) and judged **69 percent** of its escalations warranted. Read those numbers carefully, because they measure agreement on the cases the AI *surfaced*, not on the ones it quietly renewed. High concordance on easy renewals is expected; the danger lives with the patient whose "same as usual" masks a decline or a silent interaction and never gets escalated at all. Utah declined to disclose case counts, and Doctronic's au-

tonomy expands only if concordance stays above 98 percent, so the pilot has not yet run at a scale that could establish safety.

Watch out. Doctronic sits at the far, replace-the-decision end of a spectrum that runs from supporting the clinician (post-discharge calls) to supporting behavior (adherence agents) to supporting understanding (RAG Q&A), and each step is a different autonomy level, risk profile, and trust requirement. STAT News reported that the FDA has no clear authority over AI autonomous prescribers, so the sandbox is creating facts on the ground ahead of the law.

Should You Build This?

Build it when the failure you are targeting is navigation, the unsupported space between discharge and follow-up where the first 72 hours are the most dangerous, and the agent discloses that it is AI, starts with low-stakes administrative tasks, and earns trust incrementally, the pattern behind the 9.0-out-of-10 post-discharge deployment. **Think twice when** you cannot guarantee immediate escalation. "I can't breathe" is a hard trigger, not a dialogue turn, and a system that traps a patient in an AI loop after they ask for a human destroys the trust the entire design depends on; 75 percent of patients already distrust AI decisions, and one failed handoff confirms them. **Who bears the risk.** The patient, alone at home, with the least medical training and the fewest alternatives of any user in this book. The deploying system holds the liability, but the readmission, at an average of \$16,300, happens to a person.



The autonomy spectrum. The same technology spans from supporting the clinician to replacing the decision; each step up raises the risk profile and the trust it demands. Doctronic's autonomous renewals sit at the far end, where 91% physician agreement on the cases it escalated says nothing about the ones it never did.

Quick Self-Test

1. Why does identifying high-risk patients before discharge fail to prevent readmissions, and what closes that gap?
2. Why does a patient navigator need RAG instead of a general chatbot, and what does the safety classifier

block?

3. Give one example where simplifying a discharge instruction preserves the clinical intent and one where it destroys it.
4. On a hard escalation trigger, what must the agent NOT do, and why is red-flag recall the metric that matters?
5. What did the 2025 multilingual outreach study show, and where must AI still hand off to a human interpreter?
6. Utah reported that physicians agreed with 91 percent of Doctronic's renewal recommendations. Why is that number not the real safety question?

If You Remember One Thing

Build for the hardest user, not the easiest. An agent that shines for a tech-comfortable engineer but fails an 80-year-old Spanish-speaking patient with low health literacy does not narrow disparities, it widens them. Ground every answer in the patient's own record, escalate red flags with near-perfect recall, disclose that the system is a machine at every turn, and keep a human one tap away. The components exist today; the discipline of serving the worst-served user is what makes them safe.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552