Voice AI safety: A practical framework for healthcare leaders

The Risk Calculus:

Autonomous agents act without real-time human supervision, creating a fundamentally different exposure than passive suggestions in electronic health records (EHR).

Four-Layer Risk Matrix:

A consensus-based framework for evaluating agent performance, data safety, patient fit, and clinical integration.

Operational Imperatives:

Guidance on validation through pre-deployment simulation, defining autonomy tiers, and shifting to continuous monitoring of model drift.

Executive overview

Artificial intelligence is moving from passive, background decision support to active, autonomous patient interaction. AI voice agents are already placing outbound calls directly to patients, managing post-discharge recovery, closing chronic care gaps, and coordinating care across specialties. This is not yet happening at scale, and that is precisely the opportunity in front of us. Now, while adoption is still early, is the window to put guardrails in place and ensure this work is done with the well-being of patients, caregivers, and clinicians in mind.

That shift changes the risk calculus. These agents do not just recommend, they act, often without real-time human supervision or review before the action is taken. An error in one of these interactions can reach a patient before anyone has the chance to catch it. Human quality auditing after the agent interacts with the patient should, and will, always be part of any ethical deployment, but it is not a substitute for supervision at the point of action. That combination, autonomous action plus after-the-fact review, is a fundamentally different exposure than a flawed suggestion sitting in an electronic health record (EHR), waiting for a clinician to review it before anything happens.

Existing governance frameworks, including those from the Joint Commission and the Coalition for Health AI (CHAI), address AI broadly. None offer specific guidance for autonomous, patient-facing voice agents. This matters for every healthcare organization that interacts with patients directly, not just hospitals and health systems, but providers, payors, pharmacies, and pharmaceutical manufacturers as well. Deploying these systems without explicit human-supervision guardrails introduces real risk to patient safety, privacy, and equity, risk that most organizations are not yet equipped to measure, let alone manage. The framework in this paper is designed to slot into that emerging governance structure rather than compete with it. It operationalizes the Joint Commission/CHAI guidance and the NIST AI Risk Management Framework for the one category they do not yet reach directly: agents that speak with patients themselves.

This paper does two things. First, it lays out a four-layer framework for identifying where risk enters an autonomous voice interaction. Second, it shows how a panel of industry, academic, and clinical experts applied that framework to assess the likelihood and magnitude of those risks across five real-world use cases already in production today. That assessment was a formal two-round modified Delphi study: 14 experts spanning clinical medicine, informatics, health system leadership, payors, and industry rated 21 risks against each use case on an ISO 31000 likelihood and impact matrix, reaching consensus on all 105 risk-by-use-case combinations. The guidance that follows is grounded in that consensus, not in any single author's judgment.

The question this paper answers: how can healthcare organizations deploy voice AI safely and responsibly, at the speed the technology demands?

How a voice AI agent works, and where risk enters

Leaders do not need to become engineers, but they do need to understand the pipeline well enough to know where it can fail. Each stage below is also a point of potential breakdown, covered in the risk matrix later in this paper.



01 Voice interface

The direct audio touchpoint where the patient speaks and listens.



02 Speech-to-text (STT)

Converts the patient's spoken audio into text for the system to process. Transcription errors here propagate downstream.



03 Response generation

Processes the text using large language models (LLMs), standard operating procedures (SOPs), knowledge bases, and clinical rules to determine the appropriate response.



04 Orchestration and action

The agentic layer that carries out tasks beyond conversation — booking appointments, writing to the EHR, and triggering escalation pathways. This is where the agent's output reaches the real world, and where risks such as prompt injection and missed escalations concentrate.



05 Text-to-speech (TTS)

Converts the generated response back into audio, routed to the patient through the voice interface.

Five real-world use cases, and the risk hidden in each

These five are illustrative, not exhaustive. Workgroup members, drawn from a range of clinical disciplines and care settings, each brought forward use cases from their own experience. We selected these five because together they span enough variety, in acuity, population, and clinical context, to ground a general discussion without turning this paper into a long catalog of individual scenarios. The goal is not to enumerate every possible use case, but to surface the challenges and principles that recur across them.

Hypertension management

Agents prompt patients for home blood pressure readings and route data to care teams.

RISK

Failure to activate escalation protocols for dangerously abnormal values.

Post-surgical follow-up

Agents assess pain control, wound healing, and diet tolerance seventy-two hours post-discharge to triage urgent cases.

RISK

Misjudging acuity and routing a patient away from a necessary urgent callback.

Post-stroke care coordination

Agents confirm medication fills and screen for red-flag neurological symptoms.

RISK

Deployment in a population that frequently experiences aphasia or cognitive impairment, which can degrade both transcription accuracy and clinical signal.

HEDIS gap closure

Plan-focused agents contact members overdue for preventive screenings (HEDIS refers to the Healthcare Effectiveness Data and Information Set) to provide education and book appointments.

RISK

Failure to manage identity verification or fragmented underlying data at scale.

Pediatric diabetes management

Agents support insulin adherence in complex social environments, such as split custody.

RISK

Accidentally violating minor-consent laws or failing to navigate confidential medication protections.

A four-layer risk matrix for evaluating any agent

The pattern across these use cases is consistent: risk shows up at four levels, the agent itself, the data it touches, the patient it serves, and the clinician who has to trust it. Healthcare leaders, across provider, payor, pharmacy, and pharma organizations, have real opportunity to improve care efficiency and quality with agentic voice AI, but only if they evaluate it deliberately. Use this consensus framework to assess agent performance, data safety, patient fit, and clinical integration before procuring or expanding any patient-facing voice AI tool. Each layer also maps onto the seven elements of the Joint Commission/CHAI guidance (governance, privacy and transparency, data security, risk and bias assessment, quality monitoring, safety event reporting, and education) and the NIST AI Risk Management Framework, so adopting it extends the governance work your organization is likely already building rather than standing up a parallel process.

In practice, this matrix works as a due-diligence checklist. Before signing with any vendor, ask them to walk through each risk level and name a specific mitigation, not a general assurance about safety.

Agent level

RISK LEVELS AND KEY RISKS	POTENTIAL CONSEQUENCES	POSSIBLE MITIGATIONS
Transcription errors, hallucinations, omissions, social bias, missed escalations, misuse or prompt injection	Incorrect clinical data, fabricated recommendations, missed emergencies, disparate care across populations	Read-back confirmation, knowledge-base guardrails, SOP checklists, demographic bias testing, pre-deployment red-teaming

Data level

RISK LEVELS AND KEY RISKS	POTENTIAL CONSEQUENCES	POSSIBLE MITIGATIONS
Identity mis-verification, EHR data errors, fragmented records, model drift, data over-exposure	Privacy breaches, incorrect patient information relayed, AI mis-triage, gradual accuracy loss, HIPAA violations	Dual-factor identity verification, data quality audits, HL7 FHIR and TEFCA adoption, continuous monitoring, minimum-necessary data retrieval

Patient level

RISK LEVELS AND KEY RISKS	POTENTIAL CONSEQUENCES	POSSIBLE MITIGATIONS
Access and equity gaps, clinical inappropriateness, poor symptom recognition, over- or under-triage	Exclusion of vulnerable populations, missed diagnoses, patient harm from misjudged acuity	Multimodal and multilingual support, a priori eligibility criteria, clinically annotated NLP, calibrated triage thresholds

Clinician level

RISK LEVELS AND KEY RISKS	POTENTIAL CONSEQUENCES	POSSIBLE MITIGATIONS
Alert fatigue, automation complacency, liability ambiguity, workflow disruption, moral distress	Ignored critical alerts, erosion of clinical skills, unclear accountability, clinician workarounds, physician burnout	Tiered alert classification, explicit override authority, clear liability policies, EHR-integrated outputs, clinician co-design

Leader checklist: operational imperatives for safe deployment

The risk matrix tells you what to look for. This checklist tells you what to do about it before an agent ever speaks with a patient.

Allocate resources proportionately by use case:

Uniform resource allocation delivers uneven results. High-acuity deployments, such as post-stroke care, need intensive investment in escalation testing and transcription validation. Preventive outreach, such as HEDIS gap closure, should prioritize identity and data-fragmentation controls instead.

Validate through pre-deployment simulation:

Build the critical safety loops before launch. Use adversarial red-teaming simulations to validate how the voice agent handles rare, high-consequence clinical scenarios before it ever reaches a patient.

Define autonomy tiers before launch:

Treat all AI outputs as advisory, not autonomous decisions, with clinicians retaining clear, practical authority to override any agent recommendation. Within that, specify which actions the agent may complete on its own, such as appointment booking, data capture, and education, and which require human review before they reach the patient, such as triage dispositions, clinical advice, and escalation decisions, with tiers set by clinical acuity.

Disclose the AI and guarantee a human path:

Have the agent identify itself as AI at the start and end of every call, and transfer to a person on request. State law is already moving here: California's AB 3030 requires a disclaimer on generative AI-generated clinical communications and clear instructions for reaching a human provider, and several other states have similar bot-disclosure statutes.

Treat outbound calls as regulated communications:

The FCC ruled in 2024 that AI-generated voices are "artificial" voices under the Telephone Consumer Protection Act (TCPA), so consent, disclosure, and opt-out rules apply to any outbound call your organization initiates using one. This covers calls you place to patients; it does not reach calls patients place in to reach an agent. Call recording adds a separate layer of consent rules that vary by state, and two-party-consent states such as Illinois require the patient's agreement before recording.

Shift from periodic to continuous monitoring:

Model drift can quietly degrade performance over months. Automate real-time tracking of accuracy metrics, escalation rates, alert-to-action ratios, and demographic performance gaps. At minimum, track the missed-escalation rate, transcription word-error rate stratified by patient demographics, and the share of calls handed off to a human.

Give AI safety events a reporting home:

Route agent-related near-misses and harms through the same channels used for any other patient safety event, including incident reporting systems, Patient Safety Organizations, and sentinel event processes, and consider contributing de-identified reports to cross-industry registries such as CHAI's Health AI Registry.

Establish liability and accountability upfront:

Before signing vendor contracts or going live, define who bears responsibility when an autonomous agent contributes to patient harm, and encode that into institutional policy, vendor contracts, and documentation standards.

Design for equitable access from day one:

Build in multimodal support, multi-language capability, and caregiver proxy workflows so the tool serves every patient population safely, not just the easiest ones to reach.

Bottom line

Voice AI is already touching patients directly, and adoption is accelerating. The healthcare organizations, provider, payor, pharmacy, or pharma, that build safety into deployment now, while this window is still open, will be the ones able to expand these tools with confidence. The four-layer framework above turns a broad safety concern into a specific, answerable set of questions for every new voice agent your organization considers.

Contextualized from: Jayaraman S, Kumar A, Palm E, et al. "Safety of Patient-Facing Agentic AI: a Consensus Framework for Risk Assessment and Mitigation." JMIR, 2026. And: Kumar A, Saldanha A, Palm E, et al. "Safety Risk Prioritization for Patient-Facing Agentic Voice AI in Clinical Care: A Two-Round Modified Delphi Study." JMIR Preprints, 2026.

About the Healthcare Risk Management for Voice AI Taskforce

This paper is the first product of the Healthcare Risk Management for Voice AI Taskforce, an informal, multidisciplinary group formed to develop shared frameworks for the safe deployment of voice AI in healthcare. The taskforce brings together clinicians, technologists, and researchers from health systems, academic medicine, health plans, and industry to translate emerging voice AI safety concerns into practical, actionable guidance for healthcare leaders.

The group convened voluntarily, outside of any single institution, to pool expertise across clinical disciplines and care settings. This paper reflects that collective work and is intended as the first in an ongoing series.

Taskforce members

CONTRIBUTOR	INSTITUTION
Andrea Downing	The Light Collective, Boston, MA, USA
Melanie Fernando, MBA	Aetna Better Health of Illinois, Chicago, IL, USA
Sudha Jayaraman, MD, MSc, FACS	Suki AI, Redwood City, CA, USA; Departments of Surgery and Biomedical Engineering, University of Utah Health, Salt Lake City, UT, USA; Department of Surgery, University of Virginia, Charlottesville, VA, USA
Arooshi Kumar, MD	Department of Neurology, Rush University Medical Center, Chicago, IL, USA
Peter McCaffrey, MD	The University of Texas Medical Branch, Galveston, TX, USA
Erin Palm, MD, MBA, FACS	Infinitus Systems, Inc., San Francisco, CA, USA; Division of General Surgery, Department of Surgery, Stanford University School of Medicine, Stanford, CA, USA; Santa Clara Valley Medical Center, San Jose, CA, USA
Vishnu Ravi, MD	Stanford University School of Medicine, Stanford, CA, USA
Anil Saldanha, MS	Rush University System for Health, Chicago, IL, USA
Chethan Sarabu, MD	Cornell Tech Health Tech Hub, Cornell University, New York, NY, USA

Contributors

The following individuals contributed to this paper but are not core members of the taskforce.

CONTRIBUTOR	INSTITUTION
Paul Biondich, MD, MS	Regenstrief Institute and Indiana University School of Medicine, Indianapolis, IN, USA
Michael Choma, MD, PhD	Infinitus Systems, Inc., San Francisco, CA, USA
Bryon Frost, MD	McLeod Health, Florence, SC, USA
Chiraag Kulkarni, MD	Stanford University School of Medicine, Department of Medicine, Division of Gastroenterology
Raj Ratwani, PhD	MedStar Health National Center for Human Factors in Healthcare and Georgetown University School of Medicine, Washington, DC, USA
Maritza Suarez, MD	Department of Medicine, University of Miami Miller School of Medicine, Miami, FL, USA
Kevin Wang, MD	Suki AI, Redwood City, CA, USA

Contributors participated as individual experts rather than representatives of their respective institutions.