MAYSCRIBE / TECHNICAL AND COMPLIANCE WHITE PAPER

Clinical Dictation Without the Cloud

Security, compliance, and real-time performance in hospital speech recognition, and why the safest place to run it is inside your own network.

July 2026 · MayScribe · Prepared for hospital IT, compliance, and clinical informatics leadership

1. Executive summary

Hospitals have spent two decades buying dictation software that lives in someone else's cloud. The pitch was always convenience: no servers to run, nothing to maintain, a subscription and a microphone. The costs of that convenience are now well documented. Recorded patient audio accumulates in vendor data centers. Transcription accuracy is good enough until it is not, and the errors that slip through tend to involve exactly the things that hurt patients: drug names, doses, units, and sidedness. When the vendor has an outage or a breach, every hospital on the platform goes down with it.

MayScribe takes a different position. Clinical dictation is infrastructure, and infrastructure that touches protected health information belongs inside the hospital's own security boundary. MayScribe is a self-hosted, real-time medical dictation system that runs entirely within the customer's cloud environment, on GPU instances the customer controls. Audio is processed in memory and discarded when the session ends. Nothing is retained. Between the speech model and the medical record sits a deterministic verification layer that checks doses, units, laterality, and sound-alike drug names against curated clinical vocabularies before a single word is committed to the note.

This paper covers four questions a hospital should ask of any dictation vendor. How is the audio and text protected, and what happens to it after the session? What is the compliance posture today, and what is the honest path to independent attestation? How does the system stay fast without cutting corners on safety? And what actually determines accuracy in clinical use, beyond a word error rate on a benchmark? Along the way it draws on published research and public incident history, because the case for doing dictation differently is not theoretical. It is sitting in the peer-reviewed literature and in vendors' own securities filings.

2. The documentation problem hospitals are actually trying to solve

Dictation software exists because clinical documentation consumes a staggering share of physician time. In a widely cited time and motion study across four specialties, Sinsky and colleagues found that ambulatory physicians spent close to two hours on electronic health record and desk work for every

hour of direct face time with patients, and only about a quarter of their day actually facing the people they were treating. Most of the physicians in that study then added another one to two hours of EHR work at night [1].

The picture in primary care is worse. Arndt and colleagues, using EHR event logs validated against direct observation, measured an average workday of 11.4 hours, with nearly six of those hours spent in the EHR. Two thirds of that computer time went to clerical and inbox tasks rather than clinical reasoning [2]. The downstream effects are not subtle. Burnout now affects more than half a million United States physicians and costs the health system roughly 5.6 billion dollars a year in 2023 dollars [3]. The relationship between after-hours documentation and exhaustion is measurable: in the Arch Collaborative's data across more than two hundred health organizations, physicians who kept home charting to five hours a week or less were about two and a half times more likely to report lower burnout than those charting six hours or more [4].

Dictation is one of the few interventions that attacks this problem directly. Speaking is several times faster than typing, and for surgeons, hospitalists, and emergency physicians it fits the rhythm of the work. The question was never whether hospitals should use speech recognition. The question is why the prevailing way of delivering it, a shared cloud service that records patients and holds their audio, was ever an acceptable default for this data.

3. What goes wrong with clinical dictation today

3.1 The errors are frequent, and the dangerous ones look ordinary

The most rigorous public error analysis of clinical speech recognition remains the 2018 study by Zhou and colleagues in JAMA Network Open, which examined 217 notes dictated by 144 physicians across two health systems. The raw speech recognition output contained 7.4 errors per hundred words, and 96.3 percent of the unedited notes contained at least one error. About one in six of those errors involved clinical information. Only after a professional transcriptionist reviewed the notes did the rate fall to 0.4 percent, and to 0.3 percent in the final signed versions [5]. The lesson is uncomfortable for anyone selling raw model output: the model got the words mostly right, and human review still caught a large volume of clinically meaningful mistakes.

Emergency medicine data tells the same story at the note level. Goss and colleagues found an average of 1.3 speech recognition errors per dictated emergency department note, with about 15 percent of those errors judged clinically significant [6]. A related analysis found that notes generated with speech recognition carried roughly four times the error rate of notes produced without it [7]. The Joint Commission considered the pattern serious enough to publish a patient safety advisory on speech recognition technology. Its examples are the kind that keep pharmacists up at night: a spoken order for 40 mg of Lasix captured as 400 mg, and a sentence about no episodes of unconsciousness transcribed with the negation dropped, reversing the clinical meaning [8].

Notice what these failure modes have in common. They are short, high-stakes spans inside otherwise fluent text. A dose. A unit. A negation. The word left. General-purpose accuracy metrics wash them out, because a system can be 97 percent accurate by word count and still corrupt the one token in the sentence that changes treatment.

3.2 Newer models added a new failure mode: confident fabrication

The current generation of transcription models introduced a problem the old systems did not have. In October 2024, an Associated Press investigation reported that OpenAI's Whisper model, the engine behind several medical transcription products, fabricates text that was never spoken, including invented medical treatments, and that it does so despite the vendor's own guidance against use in high-risk domains. At the time of the reporting, more than thirty thousand clinicians across roughly forty health systems were using a Whisper-based documentation tool [9]. Academic work quantified the behavior: Koenecke and colleagues found hallucinated content in about one percent of transcription samples, and judged that a large share of the fabricated passages could be harmful if taken at face value. The fabrications cluster around pauses and silences, which are common in clinical conversation [10].

One percent sounds small until you multiply it by the volume of a hospital's daily documentation and remember that the insertions are fluent, plausible, and invisible to a reader who was not in the room. A misrecognized word at least tends to look wrong. A hallucinated sentence looks like medicine.

3.3 Cloud concentration turns one vendor's bad day into every customer's outage

In June 2017 the NotPetya malware, spread through a compromised third-party software vendor, took down large parts of Nuance Communications, at the time the dominant supplier of hospital transcription. The company's own securities filings put the damage at roughly 68 million dollars in lost revenue, primarily in its healthcare segment, plus about 24 million dollars in remediation costs. Hospitals that depended on the platform lost dictation and transcription service for weeks; a month after the incident the company was still announcing restoration milestones, reporting that 75 percent of clients on its flagship transcription platform were back online [11][12]. None of those hospitals did anything wrong. They had simply concentrated a clinical workflow in someone else's infrastructure, and someone else's infrastructure failed.

The economics of a breach make the same point from a different angle. IBM's 2025 Cost of a Data Breach report puts the average healthcare breach at 7.42 million dollars, the highest of any industry for the fourteenth consecutive year, with an average of 279 days to identify and contain an incident [13]. The sector's largest single event, the Change Healthcare ransomware attack, ultimately affected roughly 192.7 million people according to federal breach reporting [14]. Industry surveys found ransomware hitting about two thirds of healthcare organizations in a single year, and more than a quarter of breached organizations reporting increased patient mortality following an incident [15].

Stored voice recordings sit squarely inside this risk. A cloud dictation archive is a warehouse of protected health information in its most intimate form: patients' own words, their clinicians'

assessments, spoken diagnoses, medication lists. Every month of retained audio adds to the pile an attacker can steal, a subpoena can reach, and a breach notification will eventually have to describe. The simplest way to shrink that exposure is also the most effective one. Do not keep the audio.

4. Security by architecture, not by policy

Most vendor security stories are stacks of policy: we promise not to look, we promise to delete, we promise our subprocessors promise the same. Policies matter, but they are commitments about behavior. MayScribe's position is that the strongest controls are structural, meaning the system is built so that the risky thing cannot happen, rather than promised not to.

4.1 The deployment boundary is the hospital's own cloud

MayScribe deploys into a virtual private cloud owned by the hospital, on private subnets with no internet gateway in the audio path. The speech models, the GPU inference servers, the verification services, and the audit store all run inside that boundary. Traffic between the clinician's workstation and the inference cluster moves over a site-to-site VPN with mutual TLS on top, so both ends of every connection authenticate each other with certificates rather than shared secrets. There is no MayScribe cloud in the path. Protected health information never transits infrastructure the hospital does not control, which collapses most of the third-party risk surface that the incidents in section 3 describe.

4.2 Zero audio retention: nothing stored means nothing to breach

Dictated audio streams into memory-backed buffers, is transcribed, and is discarded when the session closes. Buffers live on temporary in-memory filesystems, not disks, and are cleared at session end as a matter of system design. The platform does not write audio to persistent storage anywhere, in any mode, for any customer. This is not a retention setting with a short default. There is no retention machinery to configure.

The security consequence is easy to state. An attacker who fully compromises the environment finds no archive of patient voice recordings, because none exists. A records request or legal demand for stored dictation audio has nothing to attach to. The breach-cost arithmetic from IBM's data, hundreds of dollars per compromised record across hundreds of days of attacker dwell time, simply never accumulates on the audio side [13]. What remains is the text that was deliberately committed to the medical record, governed by the hospital's existing EHR controls, where it belongs.

4.3 An audit trail built for skeptical reviewers

Every session produces an append-only audit record: who dictated, when, into which system, what verification checks ran, which spans were flagged, and what the clinician did about them. Audit objects are written to storage with write-once-read-many object locking, so the record cannot be silently altered after the fact, including by an administrator. The point is not decoration for an auditor's binder. It is that when someone asks what the system did on a given Tuesday, the answer is a log

entry, not a recollection.

4.4 Model integrity and a stated threat model

The speech and verification models are self-hosted with hash-pinned weights, so the artifacts running in production are cryptographically the artifacts that were reviewed and approved. The system makes no external network calls at inference time. The threat model MayScribe designs against includes vendor compromise (removed by having no vendor cloud in the data path), stored-data theft (removed by retaining no audio), transport interception (addressed by mutual TLS inside a VPN), silent tampering with records (addressed by write-once audit storage), and model supply-chain drift (addressed by pinned weights). No architecture eliminates risk. This one removes entire categories of it rather than mitigating them.

5. A compliance posture a hospital can verify, and an honest path to attestation

5.1 HIPAA alignment today

HIPAA's Security Rule asks covered entities and their business associates for administrative, physical, and technical safeguards proportionate to risk. MayScribe maps cleanly onto it because the architecture was designed with the rule's categories in mind. Technical safeguards: encryption in transit via mutual TLS over a private VPN, no persistent audio at rest to encrypt, role-based access, and unique authentication for every user and service. Audit controls: the write-once session log described above. Administrative safeguards: documented risk analysis, access review, incident response, and workforce policies, all of which exist as living documents a hospital's compliance team can read rather than summaries on a trust page. MayScribe executes business associate agreements with its hospital customers, and the underlying cloud infrastructure runs under the provider's healthcare terms, so the BAA chain is complete from clinician to hardware.

Two properties do a large share of the compliance work. Because audio is never retained, the volume of PHI at rest is a small fraction of what a conventional dictation platform accumulates. And because everything runs inside the hospital's own environment, the hospital's existing perimeter, monitoring, and identity controls apply to the dictation system the same way they apply to any internal workload. Compliance reviewers are not asked to trust an opaque external service. They can inspect the network configuration, the storage policies, and the logs directly.

5.2 SOC 1 and SOC 2, stated plainly

Hospitals increasingly gate vendor selection on independent attestation, and they should. Here is MayScribe's position without varnish. No vendor can hand you a legitimate SOC 2 report on the first day of its first deployment, because a Type I attestation examines controls as designed and implemented at a point in time, and a Type II examines them operating over an observation window of months. Anyone offering a shortcut is describing something other than a SOC report.

MayScribe's controls are designed, documented, and operating now, mapped to the Trust Services Criteria for security, availability, and confidentiality. An independent SOC 2 Type I examination is targeted within the first year of production deployment. SOC 1, which addresses controls relevant to customers' financial reporting, and SOC 2 Type II, which requires the observation period, follow on a published roadmap. In the interim, MayScribe provides the control matrix, policies, architecture documentation, and evidence samples under NDA, so a hospital's risk team can perform substantially the same review an auditor will, against the same artifacts.

5.3 Questions worth asking any dictation vendor

The fastest way to evaluate this market is to ask every vendor the same six questions and compare the specificity of the answers. Where does the audio go, physically, and who holds root on those machines? How long is audio retained, under what setting, and who can change it? Is patient data used to train or tune models, and is the answer contractual or configurable? What happened to customers during your last significant outage or security incident? Which checks, if any, run on drug names, doses, and units before text enters the chart? And can our security team inspect the running system, not a diagram of it? MayScribe's answers are short: your VPC, zero retention, never, not applicable by architecture, deterministic checks on every commit, and yes.

6. How the system stays fast: real-time engineering without shortcuts

Clinicians abandon dictation tools that make them wait. The engineering problem is that the two things hospitals want, immediate text and verified text, pull in opposite directions. MayScribe resolves the tension with a two-pass design instead of a compromise.

6.1 Two passes: a streaming draft and a rescoring pass

The first pass is a streaming speech model tuned for latency. As the clinician speaks, partial text appears at the cursor within a fraction of a second, so the experience feels like the system is keeping up with speech rather than processing it. The second pass is a larger model that rescores each completed utterance in the background, correcting the places where streaming models typically stumble: word boundaries, rare terminology, and numbers. Published evaluations of transcription systems show why a single model is not enough; reported word error rates range from under nine percent in controlled dictation to far worse in conversational conditions, with specialized terminology and accented speech remaining persistent weak points across systems [16]. Running a fast model and an accurate model in tandem buys both properties instead of trading one for the other.

6.2 A latency budget, enforced end to end

The design target is a median of roughly 0.4 seconds from the end of speech to committed text at the cursor, with the full pipeline, capture, streaming inference, rescoring, verification, and insertion, engineered against that budget. (Figures describing MayScribe performance in this paper are design

targets, measured continuously in test environments and validated per deployment, not claims from a marketing benchmark.) Everything happens in memory on GPU hosts inside the hospital's VPC; there is no round trip to an external service to add jitter, and no disk write in the hot path. Capture is push-to-talk with voice activity detection, so the microphone is open only when the clinician deliberately opens it, which is both a privacy property and a latency one: the system is never chewing on ambient audio.

6.3 Insertion where clinicians actually work

Text lands at the cursor in the hospital's EHR, including in virtualized environments where the EHR runs as a published application rather than a local program. That detail matters more than it sounds. A large share of community hospitals run their EHR through virtual desktop infrastructure, a configuration many modern dictation products handle poorly or not at all. MayScribe treats cursor-level insertion in virtualized EHR sessions as a first-class requirement rather than an integration afterthought.

7. Accuracy is a system property, not a model score

Section 3 established the pattern: the errors that matter clinically are concentrated in a handful of token types, and the newest models add fluent fabrication to the list. A credible accuracy story therefore has to be about the system around the model, not the model alone. MayScribe's answer has three layers.

7.1 A clinical lexicon underneath everything

Verification runs against curated reference vocabularies: RxNorm for medications, SNOMED CT for clinical concepts, ICD-10-CM for diagnoses, and LOINC for laboratory terminology, together with the Institute for Safe Medication Practices list of look-alike, sound-alike drug pairs. When a clinician says a drug name, the system is not pattern-matching syllables. It is resolving the utterance against a formulary-aware vocabulary and checking whether the recognized drug is a known confusion risk for something that sounds similar. Sound-alike pairs on the ISMP list trigger review rather than silent commitment, because that list exists precisely because humans and machines both confuse those names.

7.2 Deterministic checks between the model and the chart

Before any span of text is committed, a rule layer validates the categories of content that the error literature identifies as dangerous. Doses are checked against plausible ranges for the resolved medication. Units are normalized deterministically, so a spoken milligram cannot silently become a microgram. Laterality terms are flagged when they lack support in the surrounding context. Negations are tracked so that a dropped word does not reverse a clinical statement, the exact failure the Joint Commission documented [8]. These checks are rules, not model opinions. They behave the same way every time, they are inspectable, and they do not hallucinate.

7.3 Confidence gating with a human on the residual

Recognition confidence from both model passes is fused with the rule layer's risk assessment to make a per-span decision: commit or hold. The design target is that clean, high-confidence dictation commits automatically and instantly, while roughly one span in every few dozen, the doses, units, sided terms, and negations the system is unsure about, is held for the clinician with alternatives ready for single-keystroke resolution. This is the inversion of the Zhou study's workflow, where humans reviewed everything and caught most errors at great cost in time [5]. Here the machine handles the volume, and human judgment is spent only where the risk actually is. The audio for a held span is retained in memory only for the seconds the hold exists, then discarded like everything else.

The honest limit of this design should be stated too. No verification layer catches an error it was never designed to see, and a clinician who accepts a flagged span without reading it defeats the mechanism. What the architecture guarantees is narrower and more useful: the categories of error with the worst safety record cannot pass silently from a speech model into a medical record.

8. Scaling on the hospital's terms

Cloud dictation is priced per seat per month, which means the cost of adopting it scales with exactly the thing a hospital wants to grow: the number of clinicians using it. Self-hosting inverts the model. MayScribe runs as containerized inference services on a small number of GPU instances inside the hospital's cloud account. The design target for a community hospital's full dictation load is two GPU nodes, with capacity added by adding nodes, not by renegotiating licenses. The hospital pays its cloud provider for compute it controls, under rates it already negotiates, and the marginal cost of the next physician who starts dictating is effectively zero.

The same structure answers the availability question that the 2017 vendor outage raised. A hospital running MayScribe is not sharing fate with a thousand other customers on a vendor's platform. Its dictation capacity is its own infrastructure, inside its own disaster recovery posture, and an incident elsewhere in the world does not reach it. Updates ship as versioned, hash-pinned artifacts the hospital applies on its own schedule, through the same change-management process it uses for any clinical system.

There is also a quieter operational benefit. Because the system is deployed per customer, tuning is per customer too: the formulary in the lexicon is the hospital's formulary, the specialty vocabulary reflects the hospital's case mix, and performance is measured against that hospital's real acoustic conditions rather than a global average.

9. Conclusion

The published record on clinical speech recognition points in one direction. Documentation burden is large enough to be a workforce crisis [1][2][3]. Raw model output contains errors at rates no hospital would accept in any other clinical instrument, and the dangerous errors hide in doses, units, sides,

and negations [5][6][8]. The newest models fabricate fluent text under exactly the acoustic conditions clinical conversation produces [9][10]. And the prevailing delivery model concentrates recorded patient audio in third-party clouds whose failures and breaches are a matter of public record [11][13][14].

MayScribe's design answers each finding with structure rather than promises. Run the models inside the hospital's own boundary. Keep no audio. Put deterministic clinical checks between recognition and the record. Spend human attention only on the spans that carry risk. Price the system like infrastructure, because that is what it is. Hospitals evaluating dictation vendors do not need to take any of this on faith; the architecture is inspectable in their own environment, and the roadmap to independent attestation is stated in plain terms. That is what it looks like when a vendor expects to be audited, and builds accordingly.

MayScribe provides a security whitepaper, control matrix, and architecture review under NDA to hospital security and compliance teams. Design-target figures cited in this document are validated per deployment during pilot.

References

[1] Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Annals of Internal Medicine. 2016;165(11):753-760. doi:10.7326/M16-0961

[2] Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Annals of Family Medicine. 2017;15(5):419-426. doi:10.1370/afm.2121

[3] Mayo Clinic Proceedings. Predicting primary care physician burnout from electronic health record use measures. 2024. Burnout affects more than 500,000 US physicians at an estimated annual cost of 5.6 billion dollars (2023 dollars). mayoclinicproceedings.org/article/S0025-6196(24)00037-5/fulltext

[4] Eiser AR, et al. Burnout related to electronic health record use in primary care. Journal of Primary Care and Community Health / Arch Collaborative data. 2023. pmc.ncbi.nlm.nih.gov/articles/PMC10134123

[5] Zhou L, Blackley SV, Kowalski L, et al. Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. JAMA Network Open. 2018;1(3):e180530. doi:10.1001/jamanetworkopen.2018.0530

[6] Goss FR, Zhou L, Weiner SG. Incidence of speech recognition errors in the emergency department. International Journal of Medical Informatics. 2016;93:70-73.

[7] Topaz M, Schaffer A, Lai KH, et al. Medical malpractice trends: errors in automated speech recognition. Journal of Medical Systems. 2018;42(8).

[8] The Joint Commission. Quick Safety: Speech recognition technology translates to patient risk. digitalassets.jointcommission.org/api/public/content/e6d87bd33b014b1790c2bad1d1bcbf39

[9] Associated Press. Researchers say AI transcription tool used in hospitals invents things no one ever said. October 2024. apnews.com/article/90020cdf5fa16c79ca2e5b6c4c9bbb14

[10] Koenecke A, et al. Careless Whisper: speech-to-text hallucination harms. ACM Conference on Fairness, Accountability, and Transparency (FAccT). 2024.

[11] KirkpatrickPrice. Million dollar malware losses: lessons learned from Nuance. Summarizing Nuance Communications SEC Form 10-Q disclosures on the June 2017 NotPetya incident: approximately 68 million dollars in lost revenue, primarily healthcare, and 24 million dollars in remediation costs.

kirkpatrickprice.com/blog/million-dollar-malware-losses-lessons-learned-from-nuance

[12] Imaging Technology News. Nuance restores service to majority of eScription clients following malware incident. July 28, 2017. itnonline.com/content/nuance-restores-service-majority-escription-clients-following-malware-incident

[13] IBM Security. Cost of a Data Breach Report 2025. Healthcare average breach cost 7.42 million dollars, highest of any industry for the 14th consecutive year; mean time to identify and contain 279 days.

[14] US Department of Health and Human Services, Office for Civil Rights breach reporting on the Change Healthcare ransomware incident, approximately 192.7 million individuals affected.

[15] Sophos, State of Ransomware in Healthcare 2024 (67 percent of healthcare organizations attacked); Ponemon Institute and Proofpoint, Cyber Insecurity in Healthcare (29 percent of affected organizations reporting increased patient mortality following incidents).

[16] Systematic review of AI-based transcription performance in clinical settings, 29 included studies; reported word error rates from 0.087 in controlled dictation to over 50 percent in conversational, multi-speaker scenarios. pmc.ncbi.nlm.nih.gov/articles/PMC12220090