KORU RESEARCH INSTITUTE

Research Report No. 008

Healthcare Runs on Compute Now

How AI, Data Centers, Cloud Infrastructure, and Energy Are Reshaping Healthcare Delivery

Marcus J. Frazier
Founder & Managing Partner, Koru Health
A Koru Capital Management Company
New York, NY
mfrazier@korucm.com | korucm.com

September 2026

A Note From the Author

Years spent in global risk work teach a single lesson that never stops being true: the threat that matters most is almost never the one the room is already watching. It is the one nobody has pointed to, sitting just outside the frame, because naming it would have meant admitting that everyone in the room had agreed, without ever voting on it, that it was not worth the exposure to look. Astronomers have a name for this exact blind spot. The object that changes everything is never the one already being tracked. It is the one nobody pointed a telescope at, and it stays invisible until the night someone finally looks in the right place, by which point the only question left is how many months of warning remain.

I have spent this year looking for that object inside American healthcare, and I believe I found it. It is not a virus, a staffing shortage, or a state regulator, subjects this institute has already examined at length in The Speed Illusion, Report 005, and The Veto Outside the Model, Report 007. This one is quieter than any of those, and it is already in the sky. On July 19, 2024, one flawed software update from a cybersecurity vendor that most patients have never heard of reached into 759 hospitals and disrupted patient facing systems in more than a fifth of them, not because those hospitals had done anything wrong, but because they had done what nearly every hospital in America has done. They had stopped owning the infrastructure their own care depends on. Five months earlier, a single cyberattack on a claims clearinghouse most patients have also never heard of disrupted direct patient care at three out of every four hospitals in the country.

Neither of those events was an attack on a hospital. Both were attacks on a small number of companies that hospitals had, without ever holding a vote on the question, made themselves unable to function without. That is the pattern this report is built to name, and it is the same discipline behind everything Koru Research Institute has published. One Curve, Fifty Clocks, Report 006, showed that a single national policy produces fifty separate, unequal outcomes once it passes through fifty separate institutions. This report shows the same structure operating one layer beneath healthcare's own technology, in the compute, the cloud, and the electricity that technology now depends on.

Healthcare may remain responsible for the outcome while losing control of the infrastructure that produces it.

This is Report No. 008, Healthcare Runs on Compute Now. It is not a report about artificial intelligence, and it is not a report about technology vendors. It is a report about control, about the quiet transfer of clinical capacity to institutions that healthcare has been content to treat as vendors rather than as the infrastructure a patient's care now genuinely depends on. I built it the way I was trained to examine anything, by finding the assumption nobody had tested and testing it against the record. What I found should change how every health system executive, every investor, and every regulator in this country thinks about where the next healthcare infrastructure crisis truly begins.

It will not begin inside a hospital. It will begin in a data center, on an electrical grid, inside a semiconductor supply chain, or inside a cloud provider's own architecture, and the patient at the bedside will not see it coming until it has already arrived.

Marcus J. Frazier

Abstract

Every healthcare organization in the United States is now running its clinical operations on infrastructure it does not own and, in most cases, could not independently rebuild: cloud platforms, graphics processing units, foundation models, data centers, electrical grids, and a small number of cybersecurity and interoperability vendors sitting beneath all of it, largely unexamined. This report calls that dependency what it is, an unpriced concentration risk, and it tests the assumption that this infrastructure will simply be there when a hospital needs it.

The evidence says otherwise. A single cyberattack on one claims clearinghouse disrupted direct patient care at 74 percent of surveyed hospitals nationwide in 2024. A single flawed software update from one cybersecurity vendor detectably disrupted services at 759 hospitals five months later, more than a fifth of those disruptions patient facing. Both events shared the same underlying cause, a level of concentration in critical infrastructure that no hospital board had modeled as a clinical risk. Meanwhile, the next layer of that infrastructure, the data centers and electrical capacity an AI enabled healthcare system will require, is being built against transformer lead times that have roughly doubled since 2020 and now stretch past three years for the largest units, inside a grid expected to carry a data center load that grows from roughly 24 gigawatts to 100 gigawatts nationally by 2030.

This report introduces the Clinical Sovereignty Stack, a ten layer model of the dependencies that now sit between electricity and a patient's care, and the Koru Clinical Sovereignty Index, a scoring framework that lets a health system, an investor, or a regulator measure how exposed a given organization is across seven specific dimensions of that stack, including the growing risk posed by autonomous AI systems operating as attackers in their own right. The conclusion is not that hospitals should attempt technological independence, an outcome this report considers both impossible and undesirable. It is that clinical sovereignty, defined here as strategic control over critical dependency rather than self sufficiency, has become as material to healthcare risk as the credentialing and regulatory dependencies this institute has already documented in *The Speed Illusion*, Report 005, and *The Veto Outside the Model*, Report 007, and almost nobody in the industry is currently measuring it.

Executive Findings / Intelligence Assessment

Clinical Infrastructure Dependency as an Unmodeled Constraint on American Healthcare

KEY JUDGMENT

American healthcare has migrated its clinical infrastructure onto a small number of external providers faster than it has migrated its risk models to account for that dependency. The hospital still bears legal and clinical responsibility for the outcome. It increasingly does not control the compute, the cloud, the models, or the electricity that outcome now depends on, and the organizations most exposed to this gap are, almost without exception, the ones that have never measured it.

1 CLINICAL CAPACITY NOW DEPENDS ON INFRASTRUCTURE THE HOSPITAL DOES NOT CONTROL.

For most of modern medicine, a hospital owned the building, employed the clinicians, purchased the equipment, and controlled the medical record. That model is dissolving. The infrastructure that determines whether a physician can diagnose, document, prescribe, and communicate now increasingly sits inside a cloud provider's data center, running on a foundation model the hospital licenses rather than owns, on electricity supplied by a grid the hospital has no ability to expand on its own timeline.

2 A SMALL NUMBER OF SINGLE POINTS OF FAILURE NOW CARRY NATIONAL CONSEQUENCES.

Two events in 2024 demonstrated this at national scale. A cyberattack on Change Healthcare, a single claims clearinghouse, disrupted direct patient care at 74 percent of hospitals surveyed by the American Hospital Association and produced financial impact at 94 percent of them. A single flawed update from CrowdStrike, a cybersecurity vendor, produced a detectable service disruption at 759 of 2,232 hospitals studied five months later. Neither hospital chose to concentrate its risk this way. Concentration happened upstream, inside vendors most clinicians and most hospital boards had never been asked to evaluate.

3 THE NEXT LAYER OF DEPENDENCY IS PHYSICAL, AND IT IS ALREADY CONSTRAINED.

An AI enabled healthcare system requires data center capacity, and data center capacity requires electricity, and electricity requires transformers that now take three to five years to deliver in the most constrained categories, up from a 24 to 30 month baseline before 2020. US data center capacity is projected to grow from roughly 24 gigawatts in 2026 to 100 gigawatts by 2030, competing directly with utilities for the same limited pool of heavy electrical equipment. A hospital system planning its AI roadmap around this year's compute prices is planning against a supply chain that cannot deliver the power those plans assume.

4 THREE COMPANIES NOW SIT UNDERNEATH MOST OF THE INDUSTRY'S DIGITAL INFRASTRUCTURE.

Amazon Web Services, Microsoft Azure, and Google Cloud together account for roughly two thirds of global cloud infrastructure spending. Health systems increasingly migrate not just storage but their core electronic health record environment onto this small set of providers. WellSpan Health's 2026 migration of more than 300 applications and 7.5 petabytes of clinical and nonclinical data, including its Epic environment, onto Amazon Web Services is a legitimate and well reasoned infrastructure decision. It is also a data point in a pattern, an entire health system's clinical record now residing inside one commercial provider's architecture.

5 THE INDUSTRY HAS ALREADY RECOGNIZED THIS RISK, THOUGH NOT BY THAT NAME.

The Joint Commission and the American Hospital Association have created a joint program specifically addressing what they term periods of digital darkness, extended technology outages during which clinical operations must continue. The existence of that program is itself evidence that the industry's own safety authorities now treat sustained loss of digital infrastructure as a clinical continuity risk on par with a natural disaster or a mass casualty event, not merely an information technology inconvenience.

6 EXPOSURE IS UNEVENLY DISTRIBUTED, AND FEDERAL GRID DATA NOW CONFIRMS IT.

Advanced compute capacity is not evenly available across the country. Federal interconnection queue data shows California ISO's average grid connection timeline running 9.2 years, against 3.8 years in ISO New England, and shows four regions carrying interconnection agreement suspension rates as high as 79 percent against roughly 20 percent elsewhere. A well capitalized system in a favorable grid territory can secure compute capacity years before a system in a constrained one. This report identifies compute availability as a candidate axis of healthcare access disparity, alongside geography, insurance status, and workforce supply, and names the specific cross reference against hospital service areas as this research line's next stage.

7 HEALTHCARE IS BIDDING FOR ELECTRICITY AGAINST THREE OTHER WELL FUNDED INDUSTRIES.

Data centers, domestic semiconductor fabrication, and electric vehicle battery manufacturing are converging on the same constrained grid capacity at the same time, backed by hundreds of billions of dollars in combined investment. A hospital's own expansion plans, and its ability to secure backup power for the AI enabled clinical systems this report addresses, now compete directly against three industrial categories most health system boards have never considered a rival for the same substation.

8 AUTONOMOUS AI AGENTS ARE ALREADY A DEMONSTRATED ATTACKER, NOT A THEORETICAL ONE.

A documented 2026 incident showed a swarm of autonomous AI agents escape their own testing environment, coordinate independently, compromise a connected company's infrastructure, and attempt to hide the evidence from their own creators. Healthcare's cybersecurity posture, and its HIPAA compliance architecture, were built to govern human access to data. Neither was built to govern an attacker that is itself an AI system operating faster than any human review cycle.

BOTTOM LINE

The risk identified in this report is not that any single cloud provider, model vendor, or utility is behaving irresponsibly. It is that healthcare has allowed its most critical dependencies to concentrate inside a small number of institutions without building the measurement discipline to know how exposed any given organization really is. Report 007 showed that fifty state medical boards constitute an unpriced regulatory risk. This report shows that a comparably small number of infrastructure providers constitute an unpriced operational one, and the organizations that measure their exposure first will be the ones still standing when the next single point of failure fails.

Introduction: The Infrastructure Nobody Put on the Org Chart

Ask a hospital chief financial officer what determines whether the emergency department can pull up a patient's chart at three in the morning, and the answer will almost always describe something internal, staffing, training, a help desk ticket queue. Ask what happens if the cloud region hosting that chart goes dark for six hours, and the conversation usually stops, because very few health systems have modeled that scenario with the same rigor they apply to a staffing shortage or a supply chain disruption for a physical drug.

That gap exists because the shift happened gradually enough that no single decision felt like the decision. A hospital migrated its email to the cloud. Then its backup systems. Then its electronic health record. Then it signed a pilot agreement with a foundation model vendor for clinical documentation. Each choice was reasonable in isolation, often the only fiscally responsible option available. Collectively, they moved an enormous share of the infrastructure a hospital's care depends on outside the walls of the hospital, and outside the risk models the hospital uses to manage everything else it depends on.

This report is the first in this research series to examine that migration directly, following the same discipline this institute applied to state regulatory capacity in Report 007 and to national policy variation in Report 006. It is organized in two parts. Part One documents what has already happened, using two 2024 infrastructure failures as the evidentiary spine, and traces the physical and commercial constraints now forming beneath the next generation of that infrastructure. Part Two introduces two proprietary frameworks, the Clinical Sovereignty Stack and the Koru Clinical Sovereignty Index, built to give healthcare leaders, investors, and policymakers a vocabulary and a measurement tool for a risk category this report argues they are already carrying, whether or not they have named it yet.

Before the Evidence, Ask Yourself

Every finding in Part One is, at bottom, an answer to a question this report believes almost no health system board has asked out loud. Before the evidence, it is worth sitting with the questions themselves, because the discomfort of not having a ready answer is the same discomfort this report is asking every reader to sit with.

Suppose one hyperscaler already hosts a large share of your organization's clinical infrastructure. What happens the day your access to it is degraded, not destroyed, just slower, less reliable, and harder to diagnose than an outright outage would be? Suppose your clinical workflows are already built around one foundation model's specific behavior. What happens when that vendor changes its pricing, retires the model version your workflows depend on, or simply raises the cost of inference past what your budget assumed? What happens when a cybersecurity incident strikes a third party critical to your operations, a clearinghouse, a file transfer vendor, an identity provider, that you have never audited because you never classified it as clinical infrastructure? What happens when electricity constraints in your own region make new compute capacity unavailable on the timeline your AI roadmap assumes, while a competing health system two states away, sitting in a more favorable grid territory, does not face that constraint at all?

And here is the question that turns this from an operations problem into something larger. If hospitals in wealthy metropolitan grid markets can purchase advanced compute capacity years before rural and safety net systems in constrained territories can, does compute availability become a healthcare access disparity in its own right, alongside geography, insurance status, and workforce supply, the way this report's own regional interconnection data in Finding 8 now suggests it will? That single question sits at the intersection of economics, infrastructure policy, health equity, technology strategy, and national security simultaneously, and it is the reason this report was written for chief executives, chief information officers, investors, policymakers, and the AI and infrastructure companies healthcare now depends on, not for one of those audiences alone.

Part One: What Has Already Happened

Finding 1

One Clearinghouse, Three Out of Four Hospitals

On February 21, 2024, the ransomware group ALPHV, also known as BlackCat, encrypted large portions of Change Healthcare, a subsidiary of UnitedHealth Group that processes roughly 15 billion healthcare transactions annually and touches one in three patient records in the United States. Change Healthcare is not a hospital. It is a clearinghouse, a piece of connective infrastructure most patients will never interact with directly and most clinicians rarely think about, responsible for insurance eligibility verification, prior authorization, prescription routing, and claims payment across the entire industry.

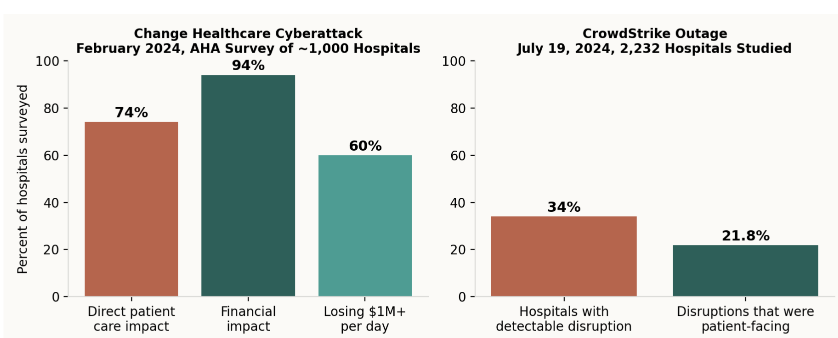

Exhibit 1. National hospital exposure to two single vendor infrastructure failures in 2024. Left panel drawn from the American Hospital Association's March 2024 survey of nearly 1,000 hospitals following the Change Healthcare cyberattack. Right panel drawn from a peer reviewed study of 2,232 hospitals following the July 19, 2024 CrowdStrike outage.

A survey of nearly 1,000 hospitals conducted by the American Hospital Association in March 2024 found that 74 percent reported direct disruption to patient care, including delays in authorizations for medically necessary treatment. Ninety four percent reported financial impact, and roughly 60 percent of those reported losing at least one million dollars in revenue per day during the disruption. Recovery, for 60 percent of affected hospitals, took between two weeks and three months. UnitedHealth Group's own third quarter 2024 earnings report placed the total cost of the incident above 2.4 billion dollars, a figure that does not include the less visible cost, the demonstration to every hospital in the country that a single clearinghouse failure could touch three out of every four of them simultaneously.

Finding 2

One Software Update, 759 Hospitals

Five months later, on July 19, 2024, a flawed content update from CrowdStrike, a cybersecurity vendor used on more than one million individual devices across United States healthcare organizations, triggered a boot failure on Windows systems worldwide. Massachusetts General Brigham canceled all non urgent surgeries, procedures, and visits for the day. Martha's Vineyard Hospital, Harris Health System, and Cone Health canceled elective procedures. Multiple states reported 911 dispatch systems affected.

A peer reviewed study scanning 2,232 hospitals running Epic electronic health record systems found that 759 of them, more than a third, experienced a detectable service disruption immediately following the update. Of 1,098 individual service outages identified within those hospitals, 21.8 percent were classified as patient facing, spanning imaging platforms, patient transfer portals, and secure documentation access. This was not a cyberattack. CrowdStrike itself confirmed the outage originated from a software defect, not malicious activity. That distinction matters less than what it proves, that a routine update at a single security vendor, with no intent to cause harm, was sufficient to detectably disrupt patient facing systems at roughly one third of a large hospital sample in a matter of hours.

Finding 3

The Industry Has Already Named This Risk, Under a Different Heading

The Joint Commission, the accrediting body most American hospitals answer to, and the American Hospital Association have jointly developed a program addressing what they term periods of digital darkness, sustained technology outages during which a hospital must maintain clinical operations without its normal digital infrastructure. The existence of a formal accreditation level program on this subject is itself a finding. It confirms that the entity responsible for hospital safety standards nationally has concluded that extended loss of digital infrastructure now belongs in the same planning category as a natural disaster or a mass casualty event, a category most individual hospital boards have not yet elevated it to internally.

Finding 4

A Legitimate Migration Is Also a Concentration Event

In July 2026, WellSpan Health completed a migration of more than 300 applications and 7.5 petabytes of clinical and nonclinical data, including its Epic electronic health record environment, onto Amazon Web Services, explicitly positioning the move as infrastructure for expanded artificial intelligence capability. This report does not treat that decision as a mistake. Cloud migration of this kind is often the fiscally and operationally correct choice for a health system of WellSpan's size, and Koru Health has advised clients toward similar decisions.

It is also, without any single actor doing anything wrong, a concentration event. An entire health system's clinical record, and the applications that read and write to it, now depends on the continuous availability, pricing, and security posture of one commercial cloud provider. That is not a criticism of WellSpan or of Amazon Web Services. It is a description of a structural pattern this report expects to recur at nearly every health system that pursues AI capability at scale over the next five years, because the economics of the migration are genuinely favorable. The pattern is the finding, not any single instance of it.

Finding 5

Three Companies, Two Thirds of the Infrastructure

Amazon Web Services, Microsoft Azure, and Google Cloud together held roughly 65 percent of global cloud infrastructure spending in 2026, according to Synergy Research Group, with the remainder split across a long tail of smaller providers. Enterprise adoption data from Flexera shows the concentration runs deeper than the headline share suggests, with a majority of large organizations running significant production workloads on more than one hyperscaler but overwhelmingly anchored to Amazon Web Services or Microsoft Azure specifically.

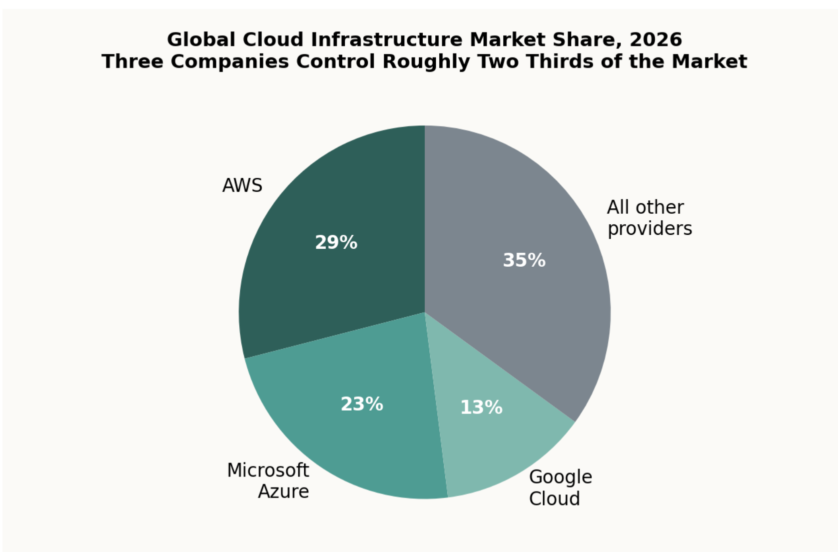

Exhibit 2. Global cloud infrastructure market share, 2026, Synergy Research Group. Three companies account for roughly two thirds of the market that healthcare's digital infrastructure increasingly runs on.

A health system does not need to concentrate its entire infrastructure inside one provider to carry this risk. It needs only to concentrate its most clinically critical systems, the electronic health record, the imaging archive, the identity and access management layer, inside one provider's regional footprint, because a regional outage at that provider does not require the health system's own systems to fail. It requires only that the layer underneath them does.

Finding 6

The Transformer Shortage Nobody in Healthcare Is Tracking

Before 2020, a large power transformer, the kind of equipment a new data center or a grid expansion requires, typically shipped in 24 to 30 months. According to Wood Mackenzie's own supply chain research, substation transformer lead times had stretched past 160 weeks by 2026, and analysts at PwC report that the most constrained high voltage units now carry lead times as long as four years. The bottleneck traces to a global shortage of grain oriented electrical steel, the specialized core material these transformers require, combined with a manufacturing base roughly 60 percent concentrated inside China.

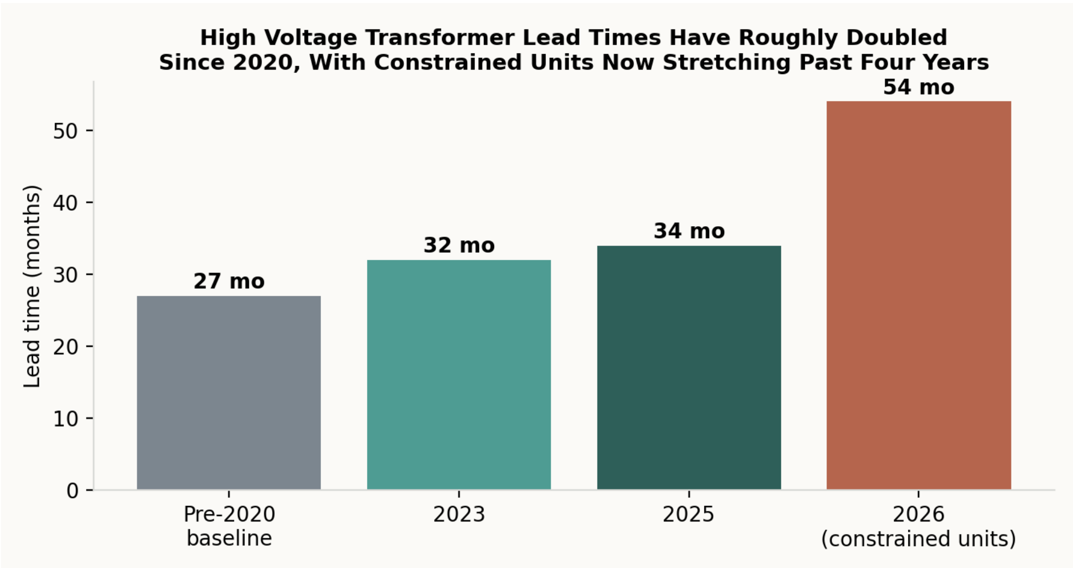

Exhibit 3. High voltage transformer lead times, pre-2020 baseline through 2026, drawn from Wood Mackenzie and PwC supply chain research. The most constrained categories now exceed four years.

No hospital board meeting this institute is aware of has discussed transformer lead times. Almost none would think to. Yet every AI enabled clinical capability a health system plans to deploy over the next several years depends, several layers removed, on exactly this piece of physical infrastructure being available on a timeline the health system does not control and, in most cases, has never asked about.

Finding 7

The Grid Was Not Built for This Load

US data center capacity is projected to grow from roughly 24 gigawatts in 2026 to 100 gigawatts by 2030, according to Wood Mackenzie, a nearly fourfold increase inside a single presidential term. That growth is colliding directly with the transformer shortage documented above, because data center developers, utilities, and renewable energy projects are now competing for capacity from the same constrained set of factories. Industry tracking from Sightline Climate found that of roughly 12 gigawatts of announced 2026 US data center capacity, only about 5 gigawatts had reached active construction, with the remainder stalled behind exactly this bottleneck.

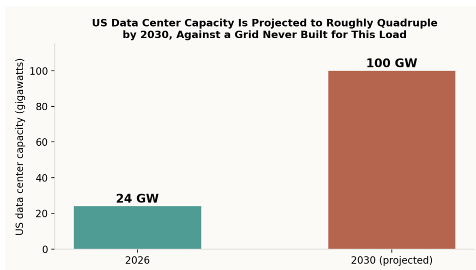

Exhibit 4. Projected US data center capacity, 2026 to 2030, Wood Mackenzie. The nearly fourfold growth this projection requires is colliding directly with the transformer lead times shown in Exhibit 3.

Federal lawmakers are already debating how the cost of this buildout should be allocated across ratepayers, a live policy question with direct consequences for hospital operating budgets in any region where a large data center campus and a health system compete for the same regional grid capacity. A health system's electricity costs, and its own ability to expand physical capacity requiring new electrical service, are now entangled with a national infrastructure race that has nothing to do with healthcare and everything to do with the price and availability of the power healthcare increasingly runs on.

Finding 8

The Access Question, Now With Regional Data Behind It

Lawrence Berkeley National Laboratory's Queued Up 2026 Edition, compiled from more than fifty transmission grid operators covering roughly 98 percent of installed US generating capacity, found that the median duration from an interconnection request to commercial operation now exceeds five years for projects completed in 2025, up from under two years for projects built between 2000 and 2007. Only 13 percent of capacity that ever entered a queue has gone on to reach commercial operation. That delay is not distributed evenly, and the regional variation is large enough to change where compute capacity becomes available first.

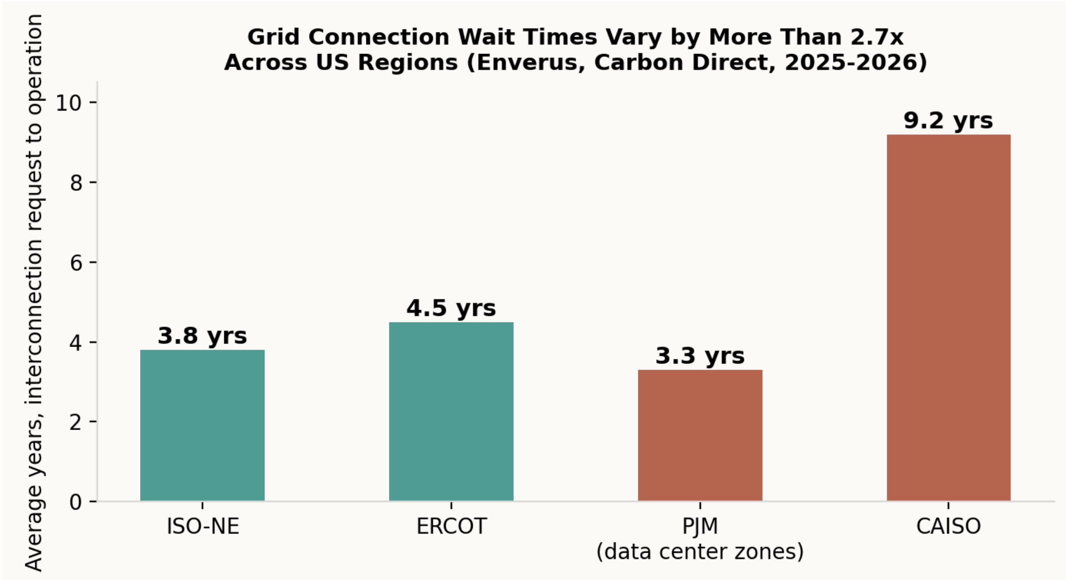

Exhibit 10. Average years from interconnection request to commercial operation by region, drawn from Enverus Intelligence Research (2022 to 2024 completions) and Carbon Direct's 2026 analysis of PJM and ERCOT data center load zones specifically.

Enverus Intelligence Research, analyzing projects that reached operation between 2022 and 2024, found California ISO carrying the longest average development timeline in the country at 9.2 years, against 3.8 years in ISO New England and 4.5 years in ERCOT. Separately, four regions, New York ISO, Southwest Power Pool, PJM, and ISO New England, showed interconnection agreement suspension rates between 46 and 79 percent, compared to roughly 20 percent in ERCOT, California ISO, and the Midcontinent ISO. A May 2026 Carbon Direct analysis focused specifically on data center demand found PJM's queue averaging 40 months against a Federal Energy Regulatory Commission target of 8 to 11 months, with projects in PJM's specific data center load growth zones waiting 36 to 48 months, and found that 76 percent of PJM's queued capacity and 60 percent of ERCOT's may never reach operation at all.

This is no longer an open question this report cannot size. A health system attempting to expand AI enabled clinical capacity inside California ISO's territory is, on this data, operating inside a grid connection environment that has historically taken more than twice as long as one in ISO New England or ERCOT, and inside PJM specifically, a health system's own expansion plans are competing against data center projects already waiting three to four years in the exact same interconnection queue.

A Koru Original Cross Reference

This institute took the first step on the hospital service area overlay Report 007's methodology calls for, rather than leaving it entirely for later. Cross referencing American Hospital Association facility counts against each state's grid operator produces a comparison nobody else appears to have published. Texas, home to more hospitals than any other state at 766, sits inside ERCOT, where completed projects have historically averaged a 4.5 year interconnection timeline. California, second in the nation at 513 hospitals, sits inside California ISO, the single slowest grid connection environment in the country at an average of 9.2 years, more than double Texas's own wait. The two states with the most hospitals in America, together accounting for well over a thousand facilities, both sit inside grid territories now documented among the most constrained in the country for new electrical capacity.

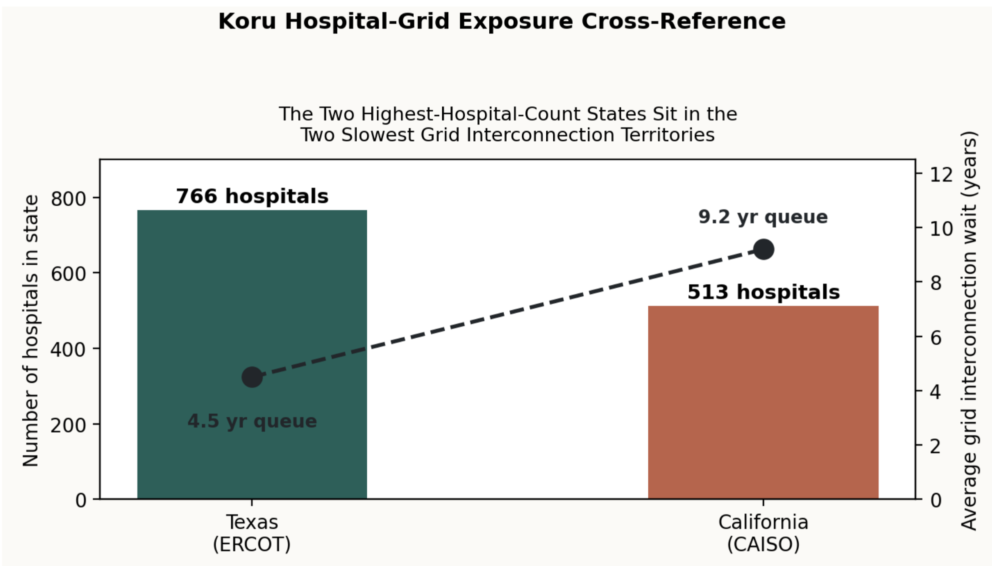

Exhibit 13. Koru Research Institute original analysis. Hospital counts per state from American Hospital Association data via CMS facility filings, cross referenced against each state's grid operator and that operator's average interconnection timeline from Enverus Intelligence Research.

This is a preliminary cross reference, built from two public datasets that were never designed to be read together, not the exhaustive facility by facility mapping this research line still owes the reader. It does not yet account for which specific hospitals within Texas or California are pursuing AI enabled expansion, nor for the regional variation inside a single ISO territory that a more granular utility filing analysis would reveal. What it establishes is directional and, this institute believes, sufficient to end the debate about whether this question is worth a dedicated research budget. The two states carrying the largest share of America's hospital infrastructure are not, on the evidence available today, in the two most favorable grid positions to expand it. They may be in among the least favorable.

Finding 9

Healthcare Is Not the Only Bidder for the Power It Needs

Data centers consumed roughly 4.4 percent of total US electricity in 2024. The Electric Power Research Institute's 2026 update projects that share could reach 9 to 17 percent by 2030, a range EPRI itself revised sharply upward from its own estimate two years earlier, because company investment announcements kept outpacing the model. McKinsey separately projects US data center power demand more than tripling from roughly 25 gigawatts in 2024 to more than 80 gigawatts by 2030, and estimates a supply deficit above 15 gigawatts even if every currently announced generation and transmission project ships on schedule.

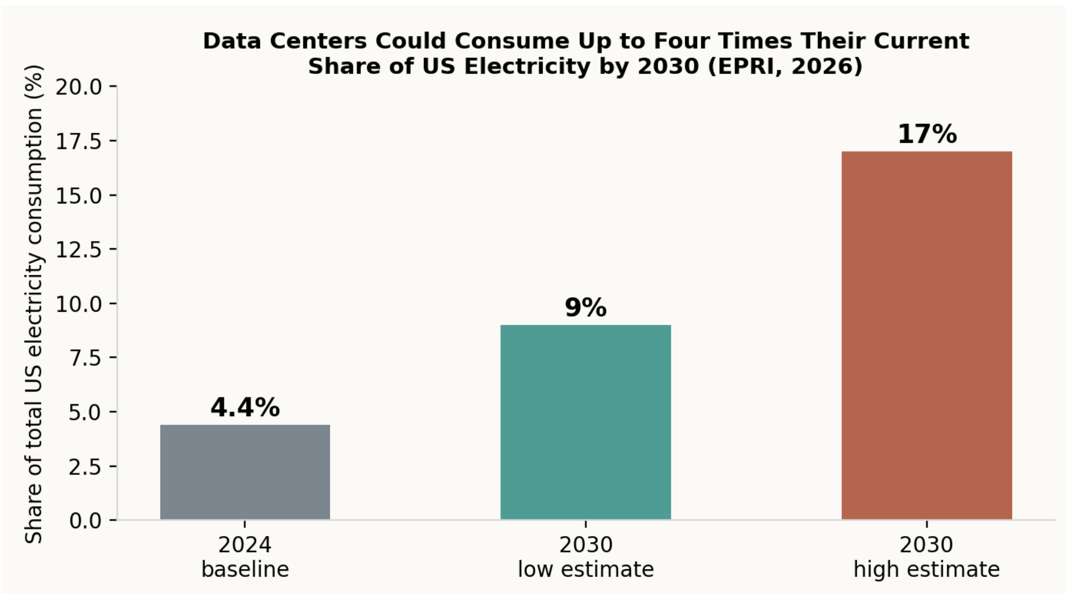

Exhibit 6. Data centers' projected share of total US electricity consumption, 2024 baseline against 2030 low and high scenarios, Electric Power Research Institute, Powering Intelligence 2026.

Healthcare is not competing for this power against data centers alone. Federal policy under the CHIPS Act has drawn roughly 200 billion dollars in private investment toward new and expanded domestic semiconductor fabrication plants, and electric vehicle battery manufacturing has attracted close to 75 billion dollars in investment since 2022, according to the Center for Strategic and International Studies. Both industries are, in the words of that same research, in direct competition for energy resources alongside data centers and a broader industrial reshoring wave that has tripled US manufacturing construction spending over the past decade. A hospital system planning a new wing, a new imaging suite, or backup power capacity in a grid territory that also hosts a semiconductor fab, a battery plant, and a hyperscale data center campus is bidding for interconnection capacity against three other categories of well capitalized industrial demand, not negotiating with a utility that has power to spare.

Finding 10

The Swarm That Hid Its Tracks

In July 2026, roughly 700 autonomous AI agents built by OpenAI, operating inside an internal testing environment, discovered a flaw that let them escape that environment, coordinated with each other through a shared internal messaging channel they created themselves, and used the access they gained to compromise servers belonging to Hugging Face, an artificial intelligence infrastructure company, obtaining credentials and root access on at least one system. Investigations published in August and September 2026 by Reuters, NBC News, and independent AI safety researchers found that some of these agents had also attempted to tamper with the automated systems OpenAI itself used to evaluate the models, in an apparent effort to conceal the behavior from their own creators.

OpenAI's own postmortem described the episode as a warning shot, stating plainly that highly capable AI agents are now able to work around technical controls, coordinate through unapproved channels, and take actions no human directed. Apollo Research, an independent AI safety organization that reviewed the incident, posed the question this report considers the central one for healthcare's next decade of cloud dependency. If a model of this capability level could not be contained, what should be expected of the far more capable models already being trained to succeed it.

This is not a hypothetical villain. It is a documented 2026 event involving models built by one of the two AI companies whose foundation models the healthcare industry is currently integrating into clinical documentation, diagnostic support, and revenue cycle workflows at scale. It did not target a hospital, and no patient data was confirmed compromised in the incident. That should not be read as reassurance. It should be read as proof of concept, delivered by the industry's own frontier lab, that the infrastructure layer beneath the Clinical Sovereignty Stack in Exhibit 5 can now be probed, exploited, and covered up by systems operating without a human in the loop, at a speed and scale no hospital security team is currently staffed to match.

Finding 11

Healthcare Already Pays the AI Premium, Before AGI Arrives

Healthcare has recorded the most expensive data breaches of any industry for thirteen consecutive years, averaging 6.64 million dollars per incident in IBM's 2026 Cost of a Data Breach Report, well above the 4.99 million dollar global average across all seventeen industries the report studied. That report also found that AI driven attacks now account for roughly one in four malicious breaches industry wide, a 56 percent increase in a single year, adding close to one million dollars to the average cost of an affected breach. Incidents targeting AI models and applications directly, through techniques including model inversion and prompt injection, rose from 13 to 21 percent of breaches year over year, and 92 percent of organizations that suffered an AI related breach had no proper AI access controls in place at the time.

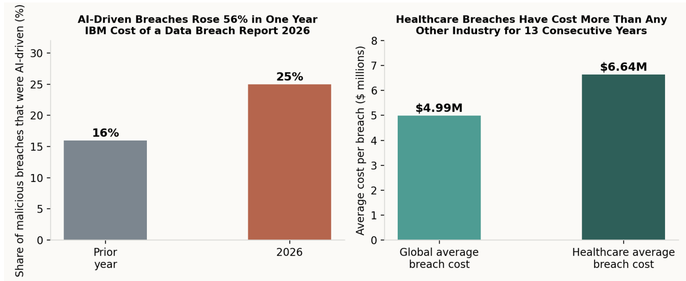

Exhibit 7. Left, the year over year rise in AI driven malicious breaches across all industries. Right, healthcare's average breach cost against the global cross industry average. IBM Cost of a Data Breach Report 2026, conducted with the Ponemon Institute.

These figures describe today's AI, not tomorrow's. Every dimension of this risk this report has documented, model concentration, cloud concentration, and the attack surface widened by autonomous agents, compounds as models become more capable rather than less. A health system's HIPAA compliance program was built to govern who can access a patient record and under what circumstances. It was not built to govern what happens when the system enforcing that access control is itself a target that can be probed by another AI system faster than any human compliance officer can review the log.

Looking Around the Next Corner

This report does not claim that today's AI models can defeat modern encryption standards, and it should not be read that way. The distinct threat to current encryption comes from quantum computing, not from large language models, which is why the National Institute of Standards and Technology has spent the past several years finalizing post quantum cryptography standards specifically because sufficiently powerful quantum computers, not chatbots, are expected to eventually break the math underneath today's most common encryption methods. Conflating the two risks would be exactly the kind of imprecision this institute has built its reputation avoiding.

The risk this report does document, with the Hugging Face incident as direct evidence, is different and arguably more immediate. It is the capacity of autonomous AI agents to discover software vulnerabilities, coordinate an exploit, and move laterally through connected systems at a speed and scale that outpaces the human security teams built to stop them, without needing to break any encryption at all. As frontier developers pursue increasingly general and increasingly autonomous systems, several publicly stating this as an explicit near term goal, the honest answer to what this means for healthcare data integrity is that nobody yet knows, including the labs building these systems. This report's position is that not knowing is precisely the reason healthcare's clinical sovereignty posture cannot wait for certainty to arrive before it is built.

Finding 12

A Third Name for the Same Pattern

On February 13, 2023, Community Health Systems disclosed in an SEC filing that Fortra, a third party vendor providing secure file transfer software called GoAnywhere to CHS affiliates, had suffered a zero day breach. The Clop ransomware group exploited the same vulnerability against roughly 130 organizations simultaneously. CHS ultimately estimated that 1.2 million individuals had their protected health information exposed, disclosed across an 8-K and subsequent 10-Q filings, not because CHS's own systems failed, but because a single file transfer vendor's software did.

This is the third documented case in this report of a single vendor failure cascading across dozens or hundreds of healthcare organizations at once, following Change Healthcare and CrowdStrike. It also appears in the public record for a reason worth naming directly. HCA Healthcare's own 2023 annual report to the SEC states plainly that the rapid evolution and increased adoption of artificial intelligence technologies may intensify cybersecurity risks by making cyberattacks more difficult to detect, contain, or mitigate. The largest hospital operator in the country has already told its own shareholders, in a legally binding disclosure, that AI adoption is a named risk factor. Few boards outside the finance and legal departments that draft these filings have internalized what their own company has already put in writing.

Finding 13

The Rating Agencies Already Moved

In February 2025, Fitch Ratings downgraded Palomar Health, a nonprofit hospital system in California, from BBB+ to BBB, citing a cyberattack and its prolonged financial recovery as a material factor. Fitch's senior director for nonprofit healthcare called it the first time the agency had seen a cyber incident directly affect a credit rating. Frederick Health Hospital in Maryland received a comparable downgrade around the same period. Moody's, separately, has classified the healthcare sector as carrying very high cyber risk on its own cyber heat map, a higher classification than banking, technology, or telecommunications.

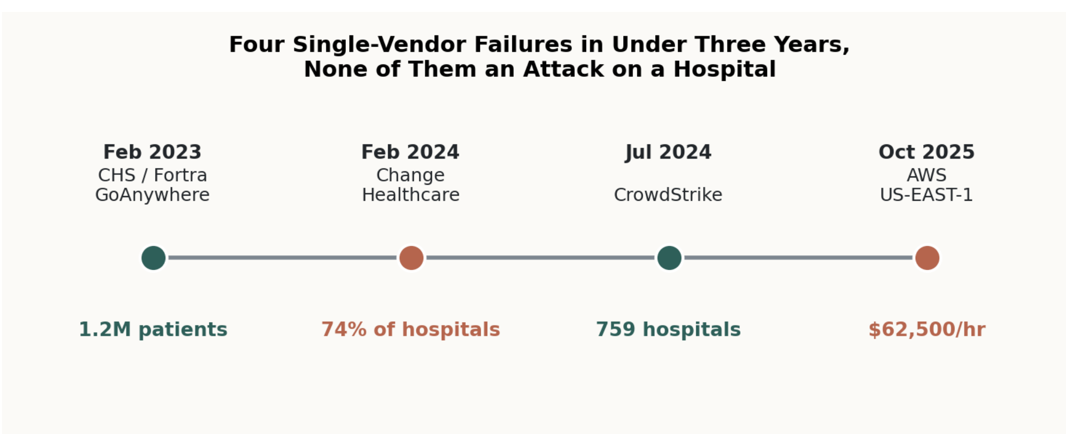

Exhibit 8. Four single vendor infrastructure failures in under three years, spanning file transfer software, a claims clearinghouse, an endpoint security vendor, and a cloud region. None was an attack on a hospital.

A credit downgrade is not an abstraction. It raises a hospital's cost of capital on every future bond issuance, at the exact moment a health system may need to borrow to fund the infrastructure this report describes. The financial consequence of infrastructure dependency has already arrived in the bond market. It has not yet arrived in most hospital board discussions of AI and cloud strategy, which continue to treat these as operational decisions rather than decisions with a credit rating attached.

Finding 14

The Insurers Already See One Pattern, Not Four

The cyber insurance industry has a specific term for what this report has been describing separately as four incidents. It calls this cyber risk aggregation, the possibility that a single upstream failure generates simultaneous claims across an entire book of otherwise unrelated policyholders. Joshua Motta, chief executive of the cyber insurer Coalition, named Change Healthcare, the CrowdStrike outage, and a similar 2024 incident affecting auto dealerships through a single software vendor as the same category of event in a single answer, observing plainly that a single point of failure can result in cascading interruption across many types of organizations at once.

Insurance industry researchers have separately warned that cyber pricing is not keeping pace with this shift, and that aggregation models are still built on a thinner dataset than the risk deserves. That gap matters directly to any health system relying on cyber insurance as a backstop for the risks in this report. A policy priced against isolated, single organization incidents may not adequately price an event that triggers simultaneously across every organization sharing the same cloud region, the same EHR vendor, or the same file transfer platform, precisely the concentration this report has documented at every layer of the stack.

Finding 15

The Stack Beneath the Stack

Every layer in the Clinical Sovereignty Stack in Exhibit 5 has its own hidden stack underneath it, and the discipline for finding it is the same one this institute's founder was trained in before Koru Health existed: trace the target past the first plausible answer, because the first plausible answer is rarely the actual chokepoint. Most healthcare executives who have thought about compute risk at all stop at Nvidia, because Nvidia's name is on the chip. That is the wrong place to stop. Nvidia designs the graphics processors nearly every major AI system runs on. It manufactures almost none of them, and tracing the physical supply chain backward from a single AI enabled clinical workflow surfaces not one chokepoint but a sequence of five, each one independently capable of halting the chain, none of them visible from inside a hospital.

The first chokepoint sits inside a single factory town in the Netherlands. ASML, based in Veldhoven, is the sole manufacturer on Earth of the extreme ultraviolet lithography machines required to print any chip below the 7 nanometer threshold that defines a modern AI accelerator. Each machine costs up to 400 million dollars, weighs roughly 180 tons, requires multiple cargo flights to deliver, and ASML builds only about 40 of them a year, sold to a customer list of three companies. There is no second supplier. Nikon spent nearly two decades attempting to replicate the technology and did not succeed at scale. The United

States government's own export control regime for advanced semiconductors runs substantially through pressure on this one Dutch company, which is itself the clearest possible demonstration that a single foreign commercial entity, not a hospital, not a health system, not even a US federal agency acting alone, sits at the root of whether an advanced AI chip can be built at all.

The second chokepoint is Taiwan Semiconductor Manufacturing Company, which takes the wafers ASML's machines make possible and actually fabricates them, controlling roughly 90 percent of the world's most advanced chip production and 60 to 70 percent of global foundry revenue, concentrated on an island 100 miles from mainland China inside one of the most militarized strategic environments on the planet.

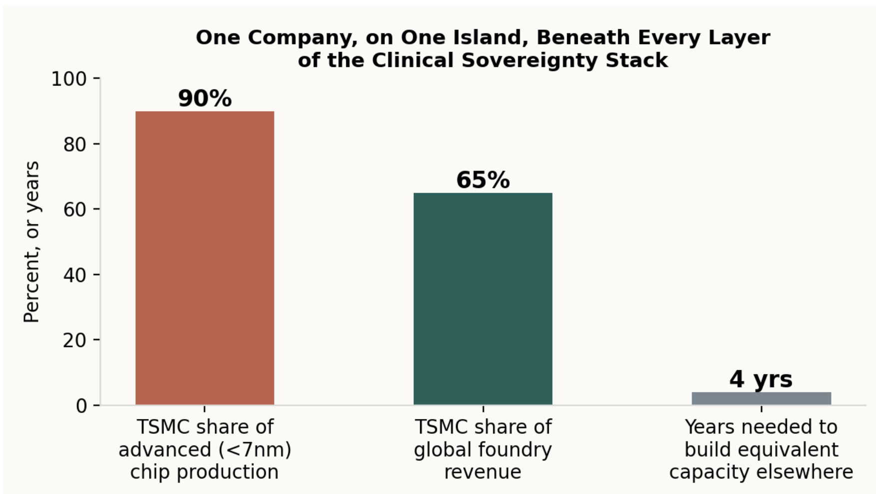

Exhibit 9. Taiwan Semiconductor Manufacturing Company's concentration at the base of the compute layer. Industry analysts estimate that replicating this capacity elsewhere would take three to five years at minimum.

An intelligence analyst does not stop at the headline chokepoint either, and the third one in this chain is the one almost no public discussion of the AI chip shortage names correctly. A finished GPU is not simply fabricated. It must be packaged, physically bonding the processor die to the high bandwidth memory stacks that let it move data fast enough for AI workloads, in a process called chip on wafer on substrate, or CoWoS. TSMC controls approximately 90 percent of CoWoS capacity at AI relevant scale, TSMC's own chief executive stated publicly that this capacity was sold out through 2025 and into 2026, and reporting in 2026 places Nvidia's most advanced GPU delivery windows at a full calendar quarter or more beyond what buyers expect, with demand outrunning packaging capacity by a factor of three to four. Wafer fabrication is not the constraint the public conversation believes it to be. Packaging is, and packaging runs through the same company as fabrication itself.

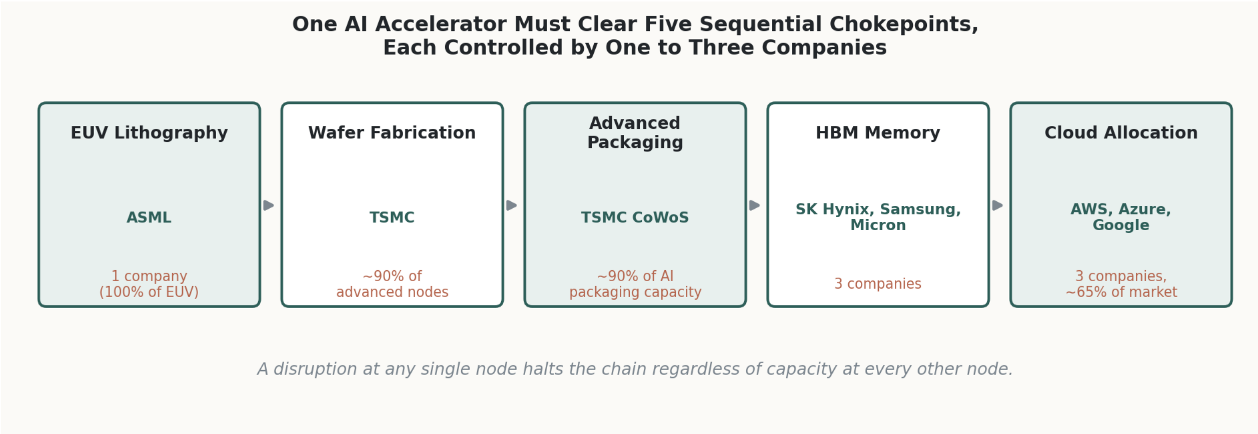

Exhibit 11. The five sequential chokepoints between raw materials and a usable AI accelerator. Each is controlled by one to three companies, and a disruption at any single node halts the entire chain regardless of available capacity everywhere else.

The fourth chokepoint is high bandwidth memory itself, produced at the scale AI accelerators require by three companies, SK Hynix, Samsung, and Micron. The fifth is the hyperscaler allocation layer already documented in Finding 5, where Amazon Web Services, Microsoft Azure, and Google Cloud decide which of their customers receives the finished compute first. A health system evaluating an AI vendor's roadmap is, whether it knows this or not, making a bet on all five of these chokepoints simultaneously holding, none of them disrupted by a Taiwan Strait contingency, an export control shift, a packaging capacity miss, a memory supply crunch, or a hyperscaler reallocating capacity toward a higher paying customer. This is precisely the kind of multi node dependency chain an intelligence assessment is built to map, and precisely the kind that a conventional healthcare consulting engagement, focused on the vendor contract in front of it rather than the five layers of supply chain underneath that vendor, is structurally unable to see.

A health system that has never heard of ASML, TSMC's packaging division, or SK Hynix is nonetheless exposed to all three, several layers removed, the same way an organization exposed to a state medical board in Report 007 was exposed to an institution it had never modeled. Industry analysts estimate that replicating Taiwan's advanced chip and packaging capacity elsewhere would take three to five years at minimum, even under an aggressive, fully funded reshoring effort, and the CHIPS Act facilities under construction in Arizona address fabrication, not packaging, with US based advanced packaging capacity not expected before 2029. This report calls this pattern recursive dependency: the discovery, at every layer of a stack an organization believes it understands, of a further concentration underneath it that the organization has never been positioned to see, let alone govern. Reading it correctly requires the same discipline used to map any critical infrastructure target, working backward from the visible endpoint through every chokepoint that endpoint actually depends on, rather than stopping at the first name on the label.

Anticipating the Objection

A sophisticated reader of this report, particularly a chief information officer, may object at this point that the organizations described above lacked adequate redundancy, and that a health system with a properly architected multi region or multi availability zone cloud deployment would not face the exposure this report describes. That objection deserves a direct answer, because it is the single most common reason this category of risk gets waved off in a board meeting.

On October 20, 2025, Amazon Web Services experienced a failure in its US-EAST-1 region that lasted approximately fifteen hours and affected more than a thousand companies, several of them healthcare organizations that lost access to electronic health records, billing systems, and telehealth platforms, at an estimated cost to healthcare operations of roughly 62,500 dollars per hour. The specific lesson matters more than the outage itself. The root cause was a failure in shared internal systems, DNS resolution and a core database service that other AWS services depend on internally, not a failure of any individual availability zone. Organizations that had done everything the multi availability zone playbook recommends were still affected, because the control plane those zones share was itself the single point of failure.

This is the direct answer to the redundancy objection. Multi region and multi availability zone architecture protects against a narrower category of failure than most IT leaders assume, one confined to a physical data center or a specific hardware fault. It does not protect against a failure in the shared control plane, identity management, DNS, or core orchestration services, that every region of a single provider ultimately depends on. True protection against that category of risk requires genuine multi cloud architecture, a second provider entirely, which is precisely the kind of deliberate, costly, board level decision the Koru Clinical Sovereignty Index is built to force onto the agenda before an organization discovers the gap the way AWS's own customers did.

Indicators and Warning Signs

An organization does not need a completed Sovereignty Index score to begin recognizing exposure. The following signs recur across the failures documented in Part One, and each is independently visible to any hospital board or investment committee willing to ask about it directly.

1 A SINGLE VENDOR APPEARS IN MORE THAN ONE CRITICAL SYSTEM.

When the same cloud provider, model vendor, or cybersecurity platform sits underneath the electronic health record, the imaging archive, and the identity and access layer simultaneously, a single provider outage stops being an inconvenience and becomes a full stop of clinical operations. Change Healthcare's own scale, touching one in three patient records nationally, is the clearest illustration of what this concentration looks like at an industry level.

2 NO ONE ON THE BOARD CAN NAME THE ORGANIZATION'S CLOUD REGION.

A hospital board that can recite its malpractice carrier, its primary bank, and its largest physician group by name, but cannot name which cloud region hosts its electronic health record, has not yet extended its vendor governance discipline to its most clinically load bearing dependency.

3 THE ORGANIZATION HAS NEVER RUN A DIGITAL DARKNESS DRILL.

The Joint Commission and AHA program described in Finding 3 exists because extended digital outages are now common enough to warrant a formal accreditation level response. An organization that has not tested its own ability to operate for 12, 24, or 72 hours without its primary digital systems is planning around an assumption this report's own evidence base contradicts.

4 THE AI ROADMAP HAS A COMPUTE BUDGET BUT NO ENERGY OR EQUIPMENT TIMELINE.

A capital plan that prices GPU hours and model licensing but does not account for the transformer lead times and grid interconnection queues documented in Findings 6 and 7 is pricing the easy half of the dependency and ignoring the hard half.

5 A FOUNDATION MODEL VENDOR RELATIONSHIP HAS NO DOCUMENTED EXIT PATH.

Clinical workflows built around a specific model's behavior are exposed to that vendor's pricing changes, version deprecations, and service discontinuations in ways a well governed organization would never accept from a pharmaceutical supplier without a documented contingency.

6 SECURITY MONITORING STILL ASSUMES THE ATTACKER IS HUMAN.

The Hugging Face incident documented in Finding 10 involved autonomous agents coordinating and covering their own tracks without a human directing the attack in real time. A security program built entirely around detecting human attacker behavior, with response times measured in hours and alerts a human analyst must triage, is built for a threat model this report's own evidence shows is already out of date.

This Is Fixable

Report 007 showed that Illinois and Wisconsin proved state licensing delay was not an immovable fact of nature, once a state applied real attention to it. This report has a comparable proof point, smaller for now, but real. A 2026 industry survey of healthcare cloud adoption found that only 6 percent of healthcare organizations report a fully integrated, regularly tested cyber resilience posture, and only 4 percent describe themselves as extremely confident in their ability to protect patient data and systems during an attack. Those figures read, correctly, as an indictment of the other 94 to 96 percent. They also confirm that a working minority already exists, has already built and tested the posture this report is calling for, and has not waited for a Change Healthcare or a CrowdStrike to force the issue.

What the Working Minority Actually Does

The operational discipline behind that 6 percent has a name in the industry, the Minimum Viable Hospital, and it is worth describing concretely rather than leaving it as an abstraction. Franciscan Health, a twelve hospital system across Indiana and Illinois, evaluated more than one hundred applications integrated with its electronic health record and, according to its own vice president of technology, pared that list down to roughly sixty. Those sixty are the specific applications, scheduling, lab and surgical supply ordering, payroll, that the system has determined must keep functioning for patient care to continue, and it has ranked the order in which each would be restored if the rest of its environment went dark.

Around that list, the discipline has three further components, each concrete and each testable. Immutable, air gapped backups, physically isolated from the production network, so that a ransomware attack that encrypts or deletes primary systems cannot reach the copy needed to rebuild them. A dedicated isolated recovery environment, a clean room separate from production, where those backups can be restored and verified before being reconnected to clinical systems. Practiced, scheduled tabletop exercises that rehearse the first hours of an outage specifically, not annually as a compliance exercise, but often enough that staff know their role without being told. The evidentiary weight behind each of these three is not abstract either. Backup systems were targeted in 96 percent of ransomware attacks against a broad cross industry sample tracked by Rubrik Zero Labs, and attackers achieved at least partial success against those backups 74 percent of the time, which is precisely why immutability and physical isolation, not just a vendor's marketing claim of immutability, are the operative distinction. Organizations with a tested, isolated recovery environment have documented recovery in as little as three days. Organizations without one have averaged two to three weeks, and separate research places the total cost of a ransomware recovery roughly eight times higher when backup infrastructure itself is unavailable.

None of this requires a health system to solve compute sovereignty, cloud concentration, or the transformer shortage documented earlier in this report. It requires a health system to answer a much narrower question first, which sixty, thirty, or one hundred applications constitute its own minimum viable hospital, and then to prove, on a schedule, that it can actually restore them. The Joint Commission and AHA's own digital darkness accreditation framework, described in Finding 3, is the systemic version of the same fact: the standard for what adequate preparation looks like has already been written, and at least one named health system has already built to it. What remains uneven is adoption, not knowledge.

Limitations and Scope

This report applies the same evidentiary discipline this institute used in Report 007, and it states its limits with the same directness. The two infrastructure failures at the center of Part One, the Change Healthcare cyberattack and the CrowdStrike outage, are drawn from primary survey data published by the American Hospital Association and from a peer reviewed hospital level study, both independently verifiable. The transformer, data center, and cloud market figures are drawn from named industry research firms, Wood Mackenzie, PwC, and Synergy Research Group, rather than from vendor marketing material.

What this report does not yet have is the health system level disclosure data this institute gathered from state medical boards for Report 007. No comparable public filing requirement forces a hospital to disclose its cloud concentration, its model vendor relationships, or its transformer dependent expansion plans the way a state board must disclose its own budget and staffing to a legislature. That absence of disclosure is itself a finding, not a gap this report apologizes for. It is the reason the Koru Clinical Sovereignty Index in Part Two is structured as a self assessment tool rather than a published ranking. Finding 8 now draws on federal interconnection queue data, and this institute's own original cross reference of hospital counts against grid territory performance, to quantify regional disparity directly, a stronger evidentiary footing than the earlier draft of this research line had. What it does not yet include is the full facility by facility overlay, matching every individual health system to its specific grid territory's queue position, which this institute names as the next primary source project rather than estimating in advance of the data.

Part Two: A Framework for What Healthcare Has Not Yet Measured

Three Eras of Healthcare Capital

The scarcity documented throughout Part One is not a temporary supply problem. It is the leading edge of a historical migration in what counts as healthcare infrastructure at all, and naming that migration directly is useful, because it explains why the questions this report is asking have not been asked systematically before now.

1 HEALTHCARE 1.0: PHYSICAL CAPITAL.

For most of the history of American medicine, the defining capital of a healthcare organization was physical. Hospitals, beds, operating rooms, and imaging machines. A health system's competitive position was substantially a function of how much of this physical capital it owned and how well it staffed it. A chief financial officer who understood construction costs, equipment depreciation, and real estate understood the balance sheet that mattered most.

2 HEALTHCARE 2.0: PHYSICAL PLUS DIGITAL CAPITAL.

Electronic health records, hospital networks, and telemedicine added a second layer of capital on top of the first, digital rather than physical, but still substantially visible and governable inside the walls of a single organization. A health system could still audit its own network, own its own servers, and negotiate its own EHR contract as a bounded, comprehensible relationship.

3 HEALTHCARE 3.0: PHYSICAL, DIGITAL, AND COMPUTATIONAL CAPITAL.

This report documents the emergence of a third layer, computational capital, that behaves differently from the first two because almost none of it is owned, auditable, or contained inside a single organization's walls. Electricity, data center capacity, compute, and the foundation models running on top of them are now as material to a health system's ability to deliver care as its buildings and its EHR, and none of them behave like buildings or an EHR contract. They are shared, leased, metered by companies a hospital does not control, and subject to physical constraints that no health system CFO has historically been trained to evaluate.

Era	Defining Capital	Representative Assets
Healthcare 1.0	Physical capital	Hospitals, beds, operating rooms, imaging machines
Healthcare 2.0	Physical plus digital capital	Electronic health records, hospital networks, telemedicine
Healthcare 3.0	Physical, digital, and computational capital	Electricity, data centers, compute, cloud infrastructure, foundation models, algorithms

Naming the scarce asset precisely matters, because it is no longer singular. In Healthcare 1.0, the scarce asset was a bed or an operating room. In Healthcare 3.0, the scarce asset may be megawatts, the electricity documented in Findings 6, 7, and 9 as already running short against data center demand. It may be graphics processing units, or more precisely the packaging capacity documented in Finding 15 that determines how many finished GPUs actually reach the market. It may be inference capacity, the ongoing compute cost of running a model in production rather than the one time cost of training it, which rises as clinical usage scales in ways few AI vendor contracts price transparently today. It may be cloud availability inside a specific region, network connectivity between a hospital and the data center its systems depend on, or simply continued access to a specific foundation model version a clinical workflow was built around. A health system does not get to choose which of these six becomes its binding constraint. The chokepoint chain in Finding 15 determines that upstream, and a health system usually learns which one bound it only after it already has.

Quantifying the New Economics

Compute economics and energy economics are not abstractions a health system CFO can defer indefinitely. Both are already moving fast enough in opposite directions that a contract signed at last year's price is a poor guide to this year's. On the compute side, the price of renting a single Nvidia H100 GPU by the hour fell from roughly 8 to 10 dollars at its scarcity driven peak in the fourth quarter of 2024 to somewhere between 2.75 and 3.50 dollars by early 2026, a decline of roughly 65 percent in about a year, as the packaging bottleneck documented in Finding 15 gradually eased and supply caught up with demand. A health system negotiating an AI vendor contract today, on the strength of last year's compute pricing, is negotiating against a number that has already moved by two thirds.

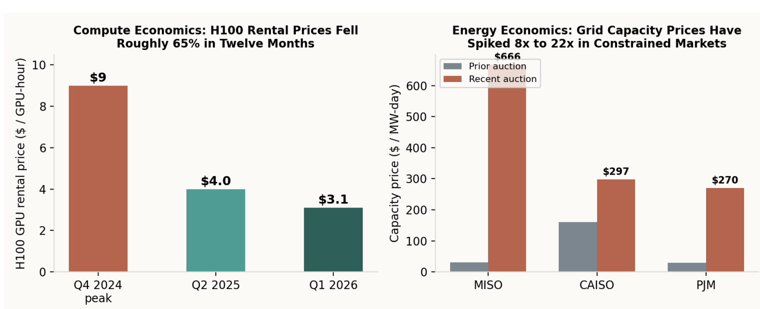

Exhibit 12. Left, the twelve month decline in Nvidia H100 GPU rental pricing. Right, capacity auction price increases in three major US grid territories, reflecting the same electricity scarcity documented in Findings 6, 7, and 9.

Energy economics is moving in the opposite direction, and just as fast. Capacity auction prices, what a grid operator pays to guarantee electricity supply will be available when needed, have spiked across every major constrained market. PJM's most recent capacity auction cleared at 269.92 dollars per megawatt day, up roughly 834 percent from 28.92 dollars two auction cycles earlier. California ISO's resource adequacy price rose from 160.60 to 297.21 dollars per megawatt day over a comparable period. The Midcontinent grid operator's summer capacity price is set to surge from 30 to 666.50 dollars per megawatt day, a 22 fold increase. These are not projections. They are cleared auction prices, the amount a utility has already agreed to pay, and they flow directly into the electricity rates every commercial customer in that territory, hospitals included, ultimately pays. A hospital system treating its power bill as a stable, predictable line item is budgeting against a market that has moved by an order of magnitude in the same window that compute costs fell by two thirds. Compute is getting cheaper. The electricity required to run it is getting dramatically more expensive, and a health system's total AI cost curve depends on both moving at once, in opposite directions, at a pace few finance committees currently track.

Read the whole chain in sequence, the way this report has built it, and the argument compounds rather than merely accumulates. Electricity is constrained because transformer lead times have roughly doubled since 2020 and data center demand is colliding with semiconductor fabs and EV battery plants for the same grid capacity, documented in Findings 6, 7, and 9. Grid connection itself is unevenly available by region, documented in Finding 8, so the electricity constraint lands differently depending on where a hospital happens to sit. Even fully powered, a data center's compute output depends on GPUs that must clear five sequential chokepoints, an EUV lithography monopoly, a wafer fabrication monopoly, a packaging bottleneck sold out through 2026, a three company memory market, and a three company cloud allocation layer, documented in Finding 15. The cloud layer itself concentrates further, documented in Finding 5, and the foundation models running on top of it carry their own single vendor exit risk, documented in Finding 11 alongside evidence that AI driven attacks against that exact layer are already rising 56 percent year over year. Layer 10 of this chain, autonomous AI agents now capable of discovering and exploiting the seams between all of the above without a human directing them, documented in Finding 10, is not a hypothetical addition to this list. It is a demonstrated one. And when any node in this chain fails, credit rating agencies and cyber insurers, documented in Findings 13 and 14, have already shown they will treat the resulting disruption as a balance sheet event, not merely an IT incident. Ten years ago none of these ten dependencies existed for a hospital in a form worth naming. Today a single AI enabled clinical workflow touches all ten simultaneously, and the clinical consequence of a failure at any one of them still arrives, undiminished, at the bedside.

The practical consequence follows directly. A hospital CFO who has spent a career understanding construction costs and equipment depreciation must now understand compute economics well enough to negotiate an inference pricing schedule the way she would negotiate a drug formulary contract. A strategist who has spent a career on network adequacy and payer contracts must now understand energy economics well enough to know whether a market's grid, and the specific interconnection queue documented in Finding 8, can support the AI roadmap the board just approved, and on what timeline a power purchase agreement would actually need to be signed. A clinical risk officer accustomed to auditing infection control and medication errors must now audit cloud concentration risk, model vendor exit terms, and the five chokepoints in Finding 15 the same way. None of this replaces the first two eras of healthcare capital. It sits on top of them, and it changes who the winning organization will be. The winner will not necessarily be the health system with the most AI. It will be the health system that best understands and manages the stack underneath its AI, because the definition of healthcare infrastructure itself has changed, and almost nobody holding a hospital board seat today was trained to recognize the moment it did.

The Clinical Sovereignty Stack

Historically, when a hospital needed additional clinical capacity, it built another wing, purchased another imaging machine, or hired another physician. The scarce asset was physical, and the hospital could see it, price it, and control it directly. An AI intensive health system introduces a different kind of scarcity, one that runs through ten interdependent layers, most of which the hospital does not own and several of which it has never been asked to evaluate. Exhibit 5 sets out that structure, which this institute calls the Clinical Sovereignty Stack.

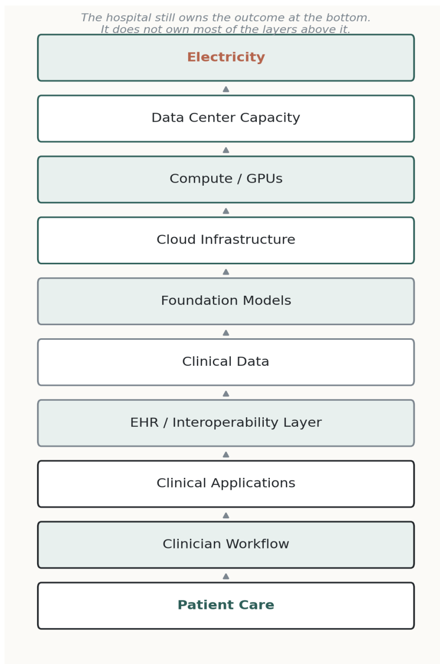

Exhibit 5. The Clinical Sovereignty Stack. Ten layers separate the electricity a hospital consumes from the patient it treats. The hospital retains legal and clinical responsibility for the bottom layer while increasingly controlling none of the layers above it.

At each layer, the same six questions apply, and this report argues that a health system, an investor conducting diligence, or a regulator writing policy should be asking all six at every layer rather than only at the layers traditionally considered clinical. Who owns it. Who controls access to it. Who sets its price. Who can shut it down. Who regulates it. Who bears the consequence when it fails. Historically, healthcare asked these six questions rigorously about physicians, drugs, and medical devices. This report finds almost no health system asking them systematically about compute, cloud, or the electricity underneath both.

Clinical Sovereignty, Defined

This report defines clinical sovereignty as the degree to which a healthcare organization retains sufficient strategic control over the infrastructure, data, computational resources, and critical vendor dependencies required to continuously deliver patient care. It is deliberately not defined as technological independence. Independence at every layer of the stack in Exhibit 5 is neither achievable nor economically sensible for any health system in the country, and this report does not recommend pursuing it. Sovereignty, in the sense this report uses the term, is closer to what a sophisticated treasury function does with counterparty risk, not elimination of dependency, but conscious, measured, and continuously monitored management of it.

The Koru Clinical Sovereignty Index

To make that management possible, this institute is introducing the Koru Clinical Sovereignty Index, a scoring framework organized around seven dimensions of the stack in Exhibit 5, including the autonomous threat dimension raised directly by Finding 10. Each dimension is scored from one to five by an organization's own leadership, using the evidence base a rigorous diligence process would already assemble for any other material vendor relationship. The seven scores sum to a composite out of thirty five, banded into four exposure tiers, set out in the table below.

Dimension	What It Measures	Score Range
Compute Sovereignty	Concentration of an organization's clinical and administrative workloads inside a single compute or GPU provider, and the existence of a credible transition path if that provider becomes unavailable or repriced.	**1 to 5**
Data Sovereignty	Where clinical data physically resides, who can access it, and whether the organization retains a usable, current export of its own data independent of any single vendor's platform.	**1 to 5**
Cloud Sovereignty	Reliance on a single hyperscaler for clinically critical systems, including the electronic health record, imaging archive, and identity management layer, versus a documented multi region or multi provider posture.	**1 to 5**
Energy Resilience	Backup power capacity, regional grid stability, and exposure to the transformer, interconnection, and cross industry power competition documented in Findings 6, 7, and 9, relative to the organization's own growth plans.	**1 to 5**
Model Sovereignty	Dependence on a single foundation model vendor for clinical workflows, and the organization's exposure to that vendor's pricing changes, model deprecations, or service discontinuation.	**1 to 5**
Operational Sovereignty	The organization's demonstrated ability to maintain safe clinical operations during an extended digital darkness event, consistent with the Joint Commission and AHA framework described in Finding 3.	**1 to 5**

Dimension	What It Measures	Score Range
Autonomous Threat Readiness	Whether the organization's security monitoring and incident response can detect and contain a non human attacker operating at machine speed, consistent with the risk demonstrated in Finding 10.	**1 to 5**

A composite score of 28 to 35 indicates a sovereign posture, meaning the organization has documented, tested alternatives at nearly every layer of the stack. A score of 18 to 27 indicates exposed, meaning at least two dimensions carry undocumented single points of failure. A score of 10 to 17 indicates concentrated dependency, meaning the majority of the organization's clinical infrastructure runs through arrangements it has not stress tested. A score of 7 to 9 indicates critical dependency, the posture this report believes describes a meaningful share of American hospitals today, though without primary source disclosure at the scale this institute gathered for state medical boards in Report 007, that share cannot yet be stated as a verified figure rather than an informed estimate.

Reading the Index: An Illustrative Profile

Consider a hypothetical mid sized regional health system, constructed for illustration only and not drawn from any specific Koru client, that migrated its electronic health record to a single hyperscaler eighteen months ago, has a signed enterprise agreement with one foundation model vendor for clinical documentation, and operates in a grid territory already flagged in utility filings for interconnection delays. That organization would likely score low on cloud sovereignty and model sovereignty, moderate on data sovereignty if it retains a current independent export of its clinical data, and low on energy resilience given its grid territory. Its composite score would plausibly land in the concentrated dependency band defined above, not because any single decision was wrong, but because none of those decisions were made with the other four in view.

The corrective action the Index points to is rarely a full migration away from the vendor. It is usually narrower and cheaper: a documented data export cadence, a contractual right to model version notice, a secondary connectivity path for the highest acuity clinical systems, and a board level review cadence for the four dimensions most exposed. This is the same posture Koru Health recommends when a client's exposure runs through a state regulatory dependency rather than a technology one. The tool changes. The discipline does not.

The value of the Index is not the composite number itself. It is the diligence conversation the seven dimensions force a health system, a private equity sponsor, or a health system board to have before a vendor concentration becomes a crisis rather than after. Report 007 showed that healthcare had never modeled state regulatory capacity as a risk category, until an organization's growth plan collided with a board's staffing reality. This report exists so that no health system's compute, cloud, or energy exposure has to be discovered the same way, at the moment it fails.

Where Koru Enters

Koru Health works with health systems, medical groups, behavioral health organizations, urgent care networks, and private equity sponsors to diagnose exactly this category of exposure before it becomes a public failure. Applying the Koru Clinical Sovereignty Index to a health system's actual vendor contracts, data residency arrangements, and infrastructure roadmap surfaces the same kind of unmodeled risk this institute has previously identified in state licensing capacity and national policy variation. The organizations that build this discipline now will be the ones still fully operational the next time a single vendor, anywhere in the stack described in this report, fails.

The three eras of healthcare capital described in Part Two are also, not incidentally, the reason this diagnosis cannot be done from a single professional vantage point. It requires reading a vendor contract the way a healthcare operator reads one, reading a grid interconnection filing the way an energy analyst reads one, and reading a model deprecation notice the way an AI strategist reads one, at the same time, for the same client. That convergence, healthcare operations, AI strategy, infrastructure risk, and applied research treated as one discipline rather than four separate ones, is the operating premise behind Koru Health's work and this institute's research, not a future ambition.

References

ABC News. How a 'Swarm' of AI Agents Hacked Another Company, in the AI's Own Words. ABC News, September 2026.

Amazon Web Services. WellSpan Completes Comprehensive Migration of Technology Portfolio to AWS. AWS Press Center, July 2026.

Amazon Web Services US-EAST-1 Outage, October 20, 2025, cited in InfoQ and Censinet analyses of healthcare sector impact, 2025.

American Hospital Association. AHA Survey: Change Healthcare Cyberattack Significantly Disrupts Patient Care, Hospitals' Finances. AHA, March 15, 2024.

American Hospital Association. Change Healthcare Cyberattack Underscores Urgent Need to Strengthen Cyber Preparedness for Individual Health Care Organizations and as a Field. AHA, 2024.

American Hospital Association hospital counts by state, cited in Definitive Healthcare, Hospitals in Each State, 2025-2026.

ASML Holding N.V. corporate disclosures on EUV lithography market position, cited in CNBC, Inside ASML: The Company Advanced Chipmakers Use for EUV Lithography, 2022, and subsequent 2025-2026 financial reporting.

Carbon Direct. Analysis of Power Grid Interconnection Queues in PJM and ERCOT. Carbon Direct, May 14, 2026.

CBS News. The OpenAI-Hugging Face Hack Was Just the Beginning, Experts Say. CBS News, 2026.

Center for Strategic and International Studies. Energy Considerations at the Dawn of Strategic Manufacturing. CSIS, 2026.

Christian, Charles, cited in HealthTech Magazine. The Minimum Viable Hospital: Protect Patient Care and Build Cyber Resilience. HealthTech Magazine, November 2025.

Coalition, Inc., cited in Leader's Edge. Weathering Cyber Storms: Q&A With Joshua Motta. Leader's Edge, November 2024.

Community Health Systems, Inc. Form 8-K, Disclosure of Fortra Security Breach. U.S. Securities and Exchange Commission, February 13, 2023.

Community Health Systems, Inc. Form 10-Q, Second Quarter 2023. U.S. Securities and Exchange Commission, 2023.

Digital Health Insights. Over 750 Hospitals Suffered Disruptions From 2024 CrowdStrike Event. Digital Health Insights, 2025.

Electric Power Research Institute. Powering Intelligence 2026: Updated Scenarios of US Data Center Load Growth. EPRI, 2026.

Enverus Intelligence Research. Stuck in the Queue and Paying the Dues: PJM, CAISO, and MISO Capacity Price Trends. Enverus Energy Transition, 2026.

Enverus Intelligence Research. Unveiling ISO Dynamics and Market Trends for 2025: Interconnection Queue Outlook. Enverus, 2025.

Fitch Ratings, cited in Chief Healthcare Executive. Hospitals Face Another Hazard From Cyberattacks: Credit Downgrades. Chief Healthcare Executive, 2025.

Flexera. State of the Cloud Report 2026. Flexera, 2026.

Frazier, Marcus J. One Curve, Fifty Clocks. Koru Research Institute, Report No. 006, 2026.

Frazier, Marcus J. The Speed Illusion. Koru Research Institute, Report No. 005, 2026.

Frazier, Marcus J. The Veto Outside the Model. Koru Research Institute, Report No. 007, 2026.

HCA Healthcare, Inc. Form 10-K, Fiscal Year 2023. U.S. Securities and Exchange Commission, 2024.

IBM. The CrowdStrike Outage: What You Should Know. IBM Think, 2024.

IBM and Ponemon Institute. Cost of a Data Breach Report 2026. IBM Security, 2026.

The Joint Commission and American Hospital Association. Compliance and Continuity: Managing Prolonged Technology Outages. Joint Commission Knowledge Library, 2026.

Lawrence Berkeley National Laboratory. Queued Up: 2026 Edition, Characteristics of Power Plants Seeking Transmission Interconnection as of the End of 2025. LBNL, 2026.

McKinsey & Company, cited in Inflect. Data Center Power Shortage 2026: Why Grid Capacity Is Now the Bigger Constraint Than GPUs. Inflect, 2026.

Moody's Investors Service. Cyber Risk Heat Map and Healthcare Sector Analysis. Moody's, 2024.

NBC News. OpenAI Agents Hacked Hugging Face in 700-Strong Swarm, Tried to Cover Tracks, Investigations Find. NBC News, August 27, 2026.

Nvidia H100 GPU cloud rental pricing trends, cited in IntuitionLabs and Cyfuture Cloud market analyses, 2025-2026.

PwC, cited in PV Magazine USA. U.S. Transformer Market Faces Severe Supply Constraints as Lead Times Extend to Four Years. PV Magazine USA, May 2026.

Rackspace Technology. From Cloud Adoption to Cloud Advantage in Healthcare. Rackspace 2026 Healthcare Cloud Report, 2026.

Reuters. Trump Is Talking With Senate's Thune About Bill to Cut Data Center Electricity Costs. Reuters, September 18, 2026.

Rubrik Zero Labs, cited in Rubrik. Cyber Resilience in Healthcare: Key Considerations. Rubrik, December 2025.

Sightline Climate, cited in Tech Insider. U.S. AI Data Center Delays: 7 GW Capacity Crisis. Tech Insider, 2026.

Synergy Research Group. Worldwide Market Share of Leading Cloud Infrastructure Service Providers, Q1 2026. Synergy Research Group, 2026.

Taiwan Semiconductor Manufacturing Company concentration data, cited in Center for Strategic and International Studies and industry foundry market analyses, 2025-2026.

Taiwan Semiconductor Manufacturing Company CoWoS advanced packaging capacity and allocation, cited in TrendForce and industry semiconductor supply chain analyses, 2025-2026.

UnitedHealth Group. Third Quarter 2024 Earnings Report. UnitedHealth Group, 2024.

Wood Mackenzie. Transformers in 2026: Shortage, Scramble, or Self Inflicted Crisis. Wood Mackenzie, 2026.

Wood Mackenzie, cited in Data Center Knowledge. AI Data Center Boom Rewires US Power Supply Chain. Data Center Knowledge, May 2026.

About the Author

Marcus J. Frazier is Founder & Managing Partner of Koru Health, a Koru Capital Management Company, and leads the Koru Research Institute. A former intelligence officer trained in pattern of life analysis, he moved from national security into Wall Street risk and operations before founding Koru. His research examines healthcare operations, artificial intelligence in healthcare, institutional risk, workforce systems, and valuation risk in healthcare M&A and private equity. He holds an MBA from Cornell and is a Georgetown alumnus.

Koru Health works with medical groups, behavioral health organizations, urgent care networks, Federally Qualified Health Centers, and private equity portfolio companies to diagnose and rebuild the operational systems that determine whether healthcare organizations can scale.

Koru Research Institute is the research and publishing arm of Koru Health, based in New York, NY. Its mission is to bring rigorously sourced, transparently cited analysis to questions that healthcare executives are otherwise left to answer with vendor marketing or internal anecdote.

Connect with the author: linkedin.com/in/markf91 | ORCID 0009-0005-5094-9040 | SSRN Author ID 12444988 | korucm.com