The Pattern Always Repeats: Why Healthcare's Next Revolution Runs or Electricity, Not Software

MAR 02, 2026 • PAID

Abstract

This piece argues that the three horsemen of every major economic revolution h always been communication, energy, and transportation, and that understanding historical sequencing helps predict where healthcare goes from here.

Key claims:

- Every major economic upheaval from the printing press forward follows the sa three-part unlock pattern

- LLMs, specifically GPT-4 and successors, represent healthcare's communicatio unlock, equivalent in magnitude to Gutenberg or Morse

- Healthcare has had its communication revolution. The energy revolution is con next and will dwarf what software alone can do

- Nvidia has quietly become one of the most important energy infrastructure companies on the planet, and almost nobody is framing it that way

- Quantum computing, nuclear fusion, next-gen electrical infrastructure, and GI efficiency gains are the mechanisms

- For investors and founders, the implication is that the 2030s will look nothing the 2020s in terms of what healthcare can actually compute, model, and deliver i time

> The time to position is before the energy unlock, not after

Table of Contents

The Three-Part Pattern (and why most people miss it)

Historians love to argue about what causes economic revolutions. Was it a charis leader? A lucky war outcome? A policy shift? Occasionally it was all of those thi but underneath virtually every transformational economic leap in the last five centuries, you find the same boring trio doing the heavy lifting: something chan how people communicated, something changed in how they generated or moved energy, and something changed in how they moved physical things through spac Communication, energy, transportation. Every time. Like clockwork, except the

runs on about a hundred-year cycle, which is inconveniently longer than most investment horizons.

The reason most people miss the pattern is that they tend to fixate on the sexy individual invention, the printing press, the steam engine, the microchip, and tr like a standalone miracle. But none of those things worked in isolation. The prir press mattered because it happened alongside early capitalism and paper supply chains. The steam engine mattered because coal extraction had gotten good enou actually fuel it consistently. The internet mattered because fiber optics, server fa and the end of the Cold War all conspired to make it globally deployable. When zoom out far enough, the individual invention looks less like a cause and more li symptom of three underlying systems converging at once.

Healthcare is not exempt from this pattern. In fact healthcare is one of the best c studies available for understanding how the pattern plays out in a domain that is simultaneously information-intensive, energy-hungry, and physically distributed When you look at the history of medicine through the lens of communication, en and transportation, you get a completely different and arguably more useful stor the standard "great scientists and great discoveries" narrative.

The Printing Press Didn't Save Lives, B It Started the Chain

Johannes Gutenberg finished his press somewhere around 1440, give or take son academic debate, and it would be historically generous to say it immediately imp healthcare outcomes. It did not. What it did was democratize the written word, v meant that medical texts, previously copied by monks at enormous expense and distributed to maybe a few dozen institutions across Europe, could now be print the hundreds and eventually thousands. Andreas Vesalius published "De Humar Corporis Fabrica" in 1543 with detailed anatomical illustrations and it spread ac European universities in a way that would have been physically impossible a cen earlier.

This is the communication unlock doing its first pass on medicine. The knowled existed in fragments before Gutenberg. What printing did was aggregate, standa and distribute it. Physicians in Padua and physicians in Paris could now argue fi the same text, which sounds trivial until you realize that standardized knowledg sharing is the prerequisite for anything resembling evidence-based practice. You cannot have a scientific consensus if nobody can read each other's work.

But notice what printing could not do. It could not make surgery safer because i could not sterilize instruments. It could not speed up the transportation of sick patients. It could not power the equipment that did not yet exist. The communic unlock planted seeds that would take centuries to fully germinate because the en and transportation unlocks had not yet arrived. The lag between a communicatio revolution and its full downstream effects on something as complex as healthcar not a bug, it is a feature of how these systems build on each other.

Steam, Coal, and the First Time We Industrialized Medicine

The first industrial revolution, roughly 1760 to 1840, was predominantly an ener story. Coal became cheap and extractable at scale, the steam engine became relia enough to run factories, and for the first time in human history, productive outp decoupled from human or animal muscle. The economic consequences were staggering and well-documented. The healthcare consequences were, for a while genuinely terrible.

Urbanization happened faster than sanitation infrastructure could accommodate Workers flooded into cities that had no sewage systems, no clean water, and no understanding of germ theory. Life expectancy in English industrial cities durin early 1800s was lower than in rural areas, which is a remarkable historical fact g that cities had more access to doctors and hospitals. The energy revolution creat conditions for epidemic disease faster than medicine could respond to it.

But then the energy revolution started to pay its forward-looking dividends. Rai networks, powered by steam, meant that pharmaceutical compounds and medica

supplies could move across national territories in days rather than weeks. The sa coal-powered industrialization that was killing urban workers was also enabling manufacturing of medical instruments at a scale and quality level that artisan craftsmen could not match. Surgical tools got better and cheaper. Hospitals coul source supplies reliably. The first real pharmaceutical supply chains emerged.

The transportation unlock arrived right on schedule, hand in hand with the ene unlock, and together they started converting the communication unlock of the printing press into something operationally useful. The diffusion of medical knowledge that Gutenberg made possible started to translate into standardized clinical practice because hospitals could now get the same instruments, the sam compounds, and train physicians from the same texts. The three systems were fi talking to each other.

Electricity, Railroads, and the Birth of t Modern Hospital

The second industrial revolution, from roughly 1870 to 1914, is where healthcar really started to look like something modern. This one had a different energy pr electricity rather than steam, and the internal combustion engine rather than ra the primary transportation unlock. The communication unlock of this era was t telegraph and then the telephone, which meant that for the first time you could transmit information at something close to the speed of thought rather than the of horses.

The hospital as an institution we would recognize today is fundamentally a prod this era. Electric lighting made surgery viable after dark and in poorly lit spaces. Autoclaves powered by electric or steam heat made sterilization systematic rath than occasional. The telephone meant that physicians could consult with specia real time across distances that would have previously required a multi-day journ rays, discovered in 1895, required reliable electrical current to operate. The EK developed in the early 1900s, was entirely dependent on electrical infrastructure

It is worth pausing on how much of what we take for granted in clinical medicin essentially an artifact of the second electrical revolution rather than a product o scientific genius. The germ theory of disease was understood intellectually befor Pasteur and Koch formalized it, but it became clinically actionable only when hospitals had the energy infrastructure to actually sterilize things reliably. The knowledge arrived before the energy capability. The energy capability converted knowledge into practice. This is the pattern doing exactly what the pattern alwa does.

Transportation in this era deserves its own moment because the emergence of re rail and early automobile networks changed the geography of healthcare delivery ways that are still structurally embedded in the system today. The hospital as a centralized institution only makes sense if patients can travel to it reliably. Befo reliable transportation, healthcare was necessarily distributed because you coul move sick people safely at any scale. Rail and automobile networks made centralization economical, and the modern hospital system was built on that assumption. Spoiler: that assumption is about to get challenged again.

The Internet as a Communication Unlo and Why Healthcare Barely Felt It

The internet represents what should have been healthcare's third major communication unlock, following the printing press and the telegraph, and by m measures it underperformed its potential spectacularly. The commercial interne arrived in the early 1990s. By the mid-2000s, almost every other information-int industry had been structurally transformed. Music, retail, financial services, trav media were all fundamentally reorganized around digital information flow. Heal spent twenty years fighting over fax machines.

The reasons for healthcare's internet-era lag are well-rehearsed among the inves and operators reading this, so no need to dwell on them extensively. Regulatory structure, reimbursement misalignment, liability concerns, the sheer complexity existing workflows, all contributed. But there is a structural reason that goes be

the policy and economic friction. The internet as a communication technology w
built for moving structured data and documents efficiently. Healthcare's core pr
has never been moving structured data efficiently. Its core problem has been
interpreting unstructured clinical information, the physician note, the patholog
report, the radiology read, the patient narrative, and doing something clinically
with it. EHRs got good at storing documents. They got terrible at understanding
them.

What digital health as an industry did from roughly 2010 to 2022 was largely tak
existing clinical workflows, digitize them, and add a consumer-facing layer. That
not nothing, but it is a long way from the kind of structural transformation that
happened to music or financial services. The communication unlock the internet
provided was a translation of analog documents into digital ones, which improve
efficiency at the margins but did not change the fundamental epistemological pr
of clinical medicine, which is that most of the meaningful information in health
embedded in language and context that requires genuine interpretation rather th
just retrieval.

LLMs Are Healthcare's Gutenberg Moment

GPT-4 launched in March 2023. Within six months, every serious health tech inv
had at least ten portfolio companies claiming they were building the AI layer for
healthcare. Within a year, the more thoughtful founders had figured out that the
interesting question was not whether LLMs could pass the USMLE (they could,
but rather what specific clinical tasks they could automate reliably enough to tru
scale. By 2024 and into 2025, the actual use cases started separating into things t
clearly worked and things that clearly did not.

What clearly worked: prior authorization automation, clinical documentation
reduction, patient-facing triage and navigation, coding and revenue cycle
optimization, literature synthesis for clinical decision support. What remained h
anything requiring genuine reasoning under uncertainty with life-or-death stake

anything involving subtle physical examination findings that have no digital representation, anything where model hallucination was clinically intolerable. T pattern holds. LLMs are a communication unlock, not a universal problem solve

Here is the historical analogy that is most useful. The printing press did not immediately cure any diseases. What it did was create the infrastructure for scie consensus-building that eventually made evidence-based medicine possible abou years later. LLMs are doing something structurally similar on a compressed time They are not curing diseases in 2024. What they are doing is creating the infrastructure for a completely different relationship between clinical knowledg clinical action. For the first time, unstructured clinical language can be interpre synthesized, and acted upon at computational speed. That is the printing press moment for medicine, not the cure moment.

The investor implication of framing it this way is important. Companies that are building on LLMs right now are building during the Gutenberg era, which was genuinely transformative but which also preceded the steam engine by a couple centuries. The returns from communication-layer companies will be real and son will be generational. But the companies that capture the majority of the econom value from this broader revolution will likely be companies that combine the communication unlock with the next energy unlock, and we are just now startin see what that looks like.

Why Communication Alone Never Finishes the Job

There is a useful exercise here which is to ask what the specific computational a physical limitations are that prevent the full deployment of what LLMs can theoretically do in healthcare. Not the regulatory limitations, those will resolve time, and not the workflow limitations, those will also resolve. The genuine phys limitations.

The first one is compute cost. Running a frontier model at clinical scale, meanin continuous inference across a large health system's entire patient population in

time, is currently prohibitively expensive. The math on GPU inference costs ver reimbursement rates in most clinical workflows does not close unless you are in high-value specialty or a revenue cycle application. This is not a software proble is an energy and hardware problem. The cost of inference will fall as hardware improves, but the rate of improvement is constrained by the energy density of cu semiconductor architectures.

The second one is data latency. Clinical AI that actually improves outcomes in a settings needs to process information faster than current infrastructure allows. ICU early warning system that processes vitals every fifteen minutes is better tha nothing, but genuinely transformative clinical AI needs continuous real-time processing of physiological signals, imaging, lab values, and contextual clinical r simultaneously. The compute demands for this, at scale, across an entire health system, are currently not achievable with economically viable infrastructure.

The third one is the physical environment. A significant fraction of the highest-v clinical applications for AI involve scenarios where the AI needs to be present at point of care in a physically demanding environment, the operating room, the ambulance, the remote clinic, and where current power infrastructure is either unavailable or insufficient. Inference at the edge requires local compute. Local compute requires power. Power in clinical edge environments is a serious engine constraint that is not being solved by software.

All three of these limitations point at the same underlying bottleneck: energy de and energy cost. This is not a coincidence. It is the pattern telling you something

Nvidia and the Quiet Energy Revolutio Inside the Chip

Before getting to fusion and quantum, there is a company already mid-revolutio does not get nearly enough credit for being an energy story rather than just a ch story. Nvidia is, depending on how you want to think about it, either the most important infrastructure company of the current decade or the most underappreciated energy efficiency company in history. Probably both.

The H100, Nvidia's flagship data center GPU released in 2022, delivers roughly 2 teraflops of FP16 performance at around 700 watts. That sounds like a lot of pov until you compare it to what the same compute workload would have required fi years earlier. The A100, released in 2020, was already a massive leap over its predecessors. The H100 was another leap over that. The trajectory of compute-p watt across Nvidia's data center GPU generations over the last decade is roughly doubling every two to three years, which is actually faster than classical Moore's and represents one of the most significant energy efficiency improvements in the history of computing.

The Blackwell architecture, released in 2024, pushed this further with the GB20 NVLink system delivering roughly five times the inference performance of H100 roughly comparable power envelopes for many workloads. For large language mo inference specifically, Nvidia has published benchmarks showing the GB200 delivering around 30 times better performance per watt on certain inference tas compared to the H100 generation. Even discounting for benchmark optimism, th directional story is real and it is large.

Why does this matter specifically for healthcare AI? Because the economics of cl AI deployment are directly downstream of inference cost, and inference cost is directly downstream of compute-per-watt. Every time Nvidia's efficiency curve m use cases that were previously economically unviable become viable. The continu ICU monitoring application that did not pencil out in 2022 might pencil out in 2 on Blackwell infrastructure. The real-time surgical assistance system that requi cost structure unavailable outside of major academic medical centers might be deployable at community hospital scale within the next two GPU generations.

There is also a subtler point worth making about what Nvidia has done for the e consumption profile of AI at the national level. Data center electricity consumpt the US has been growing rapidly alongside AI workload growth, and projections groups like the Lawrence Berkeley National Laboratory have estimated that US d centers could consume somewhere between 6 and 12 percent of national electric the late 2020s depending on efficiency improvement trajectories. Nvidia's efficie gains are one of the primary mechanisms by which that growth is being partially

offset. The company is simultaneously making AI more powerful and making AI energy-intensive per unit of output, which is the exact combination of factors yo need to make clinical AI at scale economically rational.

Nvidia has also made less-publicized but important investments in the software for energy-efficient inference. The TensorRT inference optimization library, the microservices architecture announced in 2024, and the specific optimizations Nv has built for healthcare AI workloads through its Clara platform are all part of a coherent strategy to make the GPU do more clinical work per kilowatt-hour. Jen Huang has been explicit in public forums that Nvidia views itself as an energy efficiency company as much as a semiconductor company, and the product roadr supports that framing more than the typical chip company narrative does.

The healthcare-specific implication of Nvidia's trajectory is worth isolating. Nvi Clara Parabricks platform, which handles genomic sequencing computation, red the time to run a full genome variant analysis from roughly 30 hours on CPU to 30 minutes on GPU, at dramatically lower energy cost per analysis. At current sequencing volumes that is a meaningful efficiency gain. At the sequencing volu projected for the late 2020s as genomic medicine becomes more mainstream, the difference between CPU and GPU infrastructure for genomics is the difference between the economics working and not working. Nvidia is not incidentally part that story. It is central to it.

None of this means Nvidia is risk-free as a bet or that the current valuation is obviously rational. It means that when thinking about the energy unlock for healthcare AI, Nvidia belongs in the same conversation as fusion and quantum, because it is already delivering energy efficiency improvements that are changin what is computationally possible in clinical settings right now, not in fifteen yea

The Energy Unlock Is the Missing Piece

There are three additional energy technologies in active development that have genuine potential to change the constraints described above within the next dec fifteen years: next-generation nuclear fission via SMRs, nuclear fusion, and quan

computing as a distinct computational energy paradigm. Together with Nvidia's efficiency curve, these represent four different mechanisms converging on the sa outcome: dramatically cheaper and more abundant compute for clinical AI.

SMRs are the nearest-term story on the generation side. Companies like NuScale Kairos Power, and a handful of others are building reactors designed to deploy a small scale with dramatically reduced construction timelines compared to tradit nuclear. The DOE has been funding these programs aggressively and several are in late-stage regulatory review. For healthcare specifically, the implication of SM deployment is not directly about powering hospitals. It is about powering the da center infrastructure that clinical AI depends on. If the marginal cost of electric large-scale computing falls significantly, the economics of clinical AI at scale ch dramatically, and they compound with the efficiency gains Nvidia is already deli on the demand side.

Fusion is a longer bet. Commonwealth Fusion Systems, with its SPARC project, achieved a significant magnet milestone in 2021 and has credibly moved the exp timeline for net energy gain to sometime in the late 2020s or early 2030s. TAE Technologies, Helion Energy backed by Sam Altman (make of that what you will the LLM connection), and a handful of others are pursuing different confinemen approaches. The probability of commercially viable fusion electricity by 2035 is longer negligible. If it arrives, the downstream effects on compute-intensive applications like clinical AI are hard to overstate, and they land on top of an Nvi efficiency curve that will have continued improving in the interim.

Quantum computing is the most speculative of the three but deserves inclusion because its implications for healthcare go beyond cheaper computation. Quantu systems have theoretical advantages in molecular simulation that classical comp will never match efficiently. Drug discovery, protein folding beyond what Alpha can currently do, and the optimization problems involved in personalized dosing treatment sequencing are all domains where quantum advantage, when it arrives could compress development timelines by years. The energy story with quantum slightly different since current quantum systems require extreme cooling to nea

absolute zero which is itself an enormous energy cost, but that constraint is exac
what the next generation of room-temperature qubit approaches is designed to s

The healthcare energy unlock is not a single technology. It is a cluster of converg
developments in how we generate, transmit, and apply energy to computation an
physical environments, arriving more or less simultaneously within a fifteen-yea
window. Nvidia is the part of that cluster that is already here and already changi
math. SMRs, fusion, and quantum are the parts that are loading. This is exactly l
the second industrial revolution worked. Electricity, the internal combustion en
and the telephone were not planned to converge. They just did, because the
underlying physics and engineering were all mature enough at the same historic
moment.

What Quantum and Fusion Actually Me for Healthcare (Practically)

Getting concrete matters because the history lesson is useful but ultimately inve
and founders need to know what to actually build or fund.

The first practical implication is that the cost curve for clinical AI inference wil
faster than most current DCF models assume. Healthcare AI company valuation
on current GPU costs as a structural assumption are going to look wrong within
years, probably in a positive direction for pure AI application companies and in
complicated direction for companies whose moat is primarily compute efficiency
current energy prices. Nvidia's roadmap alone would drive this conclusion. Add
deployment to the grid and it accelerates.

The second practical implication is that real-time physiological AI becomes
economically viable in a low-energy-cost world. Right now the most interesting
continuous monitoring applications, the ones that can genuinely predict a sepsis
or a cardiac deterioration hours before current early warning systems, are not
deployed at scale because the compute cost of continuous inference on rich sens
data does not pencil out against DRG reimbursement. Drop the inference cost b
eighty percent, which is a plausible outcome of the combined Nvidia efficiency c

plus grid energy improvements over the next decade, and that math changes. Th companies building sensor infrastructure and clinical workflow integration for t applications right now are building something that will be worth dramatically m when the energy unlock arrives.

The third practical implication is drug discovery timelines. Quantum molecular simulation, even at relatively early fidelity levels, will let computational chemist experiments in silico that currently require months of wet lab work. The compa building platform chemistry capabilities with serious computational infrastruct assumptions are positioning for a world where the rate-limiting step in drug development shifts from synthesis and testing to regulatory and clinical validati That is a fundamentally different R and D economics problem with different cap requirements and different risk profiles.

The fourth implication, and this one is underappreciated, is what cheap abunda energy means for healthcare delivery infrastructure in the developing world. A significant fraction of the global disease burden sits in environments where heal delivery is constrained not just by communication and knowledge but by the ph impossibility of running diagnostic equipment reliably. Point-of-care diagnostic require consistent power. Cold chains for vaccines and biologics. Hospital-level surgical capability in rural settings. These are not software problems. They are e problems. A world with dramatically cheaper and more distributed energy prod is a world where a lot of global health infrastructure problems become tractable ways they simply are not today. Notably, Nvidia's edge inference platforms are pa this story too: the ability to run meaningful clinical AI models on low-power edg hardware in resource-constrained environments is already improving and will continue to.

How to Invest Ahead of an Energy Revolution You Can't Fully Predict

The honest version of this section starts with admitting that the timing of the fu energy unlock is genuinely uncertain. Fusion could be 2032 or it could be 2045.

regulatory timelines have a way of extending. Quantum decoherence problems m take another decade to crack at room temperature. What is not uncertain is Nvic efficiency trajectory, which is documented, roadmapped, and already changing t economics of clinical AI deployment in real time. That is the part of the energy s you can bet on now without requiring a fusion breakthrough.

Given that range of uncertainty across the different mechanisms, the sensible investment posture is to look for healthcare companies whose value proposition becomes dramatically stronger in a world of cheap abundant compute and cheap abundant energy, and to prefer those companies when they can generate returns on the slower timeline. Find the businesses where the energy unlock is option va rather than the base case.

Clinical AI application companies in revenue cycle, prior auth, and documentati already generating real returns on current energy economics. They get meaningf better with cheaper inference. That is a good risk profile and it improves with ev Nvidia generation regardless of what happens with fusion.

Continuous monitoring and early warning companies with serious data and worl integration built already are interesting asymmetric bets. The underlying clinica value is not in question. The economics are. An energy unlock, whether it comes primarily from Nvidia's efficiency curve or from grid-level changes, resolves the economics.

Computational drug discovery platforms that have built serious proprietary che data and molecular modeling infrastructure are positioning correctly for quantu advantage. The ones generating near-term revenue from classical compute appro on current Nvidia infrastructure have a survivable intermediate timeline.

Healthcare infrastructure companies focused on emerging markets, especially th building modular or low-power diagnostic and delivery capabilities, are playing game that the energy unlock potentially pulls forward by a decade.

What to be cautious about: healthcare AI companies whose primary defensibility claim is inference efficiency or proprietary compute optimization. If Nvidia's

efficiency curve plus eventual grid improvements commoditize inference, that m narrows or disappears. Also worth scrutinizing: any business model that require current energy constraints to persist in order to maintain competitive advantage

The meta-lesson from the historical pattern is that the people who made the mo from the second industrial revolution were not primarily the people who invente technologies. They were the people who built the applications, systems, and institutions designed for a world powered by electricity before most of the worl The hospital system, the pharmaceutical supply chain, the medical device indust of these were built by people who assumed electricity was real and was coming a designed forward from that assumption.

Nvidia is telling you, with published silicon roadmaps and real benchmark data, the next several years of compute economics look like. Fusion and quantum are you, with less certainty but with increasing credibility, what the decade after tha might look like. The same opportunity exists right now for the generation of hea tech founders and investors willing to design forward from the assumption that energy unlock is real and is compounding. The ones who assume a world of dramatically cheaper and more abundant compute, who assume that the physica constraints on clinical AI will be lifted within their company's scaling timeline, assume that the geographies currently excluded from modern healthcare by ener constraints will be included: those are the founders building the institutions tha define healthcare delivery in the second half of the 21st century.

The pattern always repeats. The communication unlock just landed. The energy unlock is already partly here and the rest of it is loading.

1 Like • 1 Restack

← Previous

Nex

Discussion about this post

Comments Restacks

Write a comment...