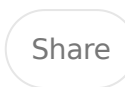
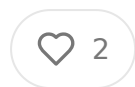


# NVIDIA's Healthcare Stack Is the Pick and Shovels Play You've Been Waiting For

MAR 31, 2026



## Table of Contents

Section 1: The Inflection Point Is Already Here

Section 2: BioNeMo and the Drug Discovery Revolution

Section 3: MONAI and the Medical Imaging Flywheel

Section 4: Isaac for Healthcare and the Robotics Buildout

Section 5: Holoscan and the Edge Intelligence Layer

Section 6: Parabricks and the Genomics Data Deluge

Section 7: Clara, NIM, and the Open Source Bet

Section 8: What This Means for Investors and Founders

## Abstract

This essay examines NVIDIA's healthcare and life sciences platform stack as an investment and competitive landscape thesis for health tech entrepreneurs and investors. Drawing on NVIDIA's 2026 State of AI in Healthcare and Life Science survey (600+ respondents, fielded Aug-Sept 2025) and the company's product ecosystem documentation, the piece argues that NVIDIA has quietly built the most comprehensive AI infrastructure layer in healthcare, and that understanding each

component, BioNeMo, MONAI, Isaac for Healthcare, Holoscan, Parabricks, Clara NIM, is now table stakes for anyone deploying capital or building companies in space.

## **Key data points from the survey:**

- 70% of healthcare/life sciences orgs actively using AI, up from 63% in 2025
- 69% using generative AI/LLMs, up from 54%
- 85% of orgs increasing AI budgets in 2026
- 85% of management-level respondents report AI increased annual revenue
- 44% of management say AI boosted annual revenue by more than 10%
- 57% of medtech orgs report ROI from medical imaging AI
- 46% of pharma/biotech report ROI from drug discovery AI
- 47% either actively using or assessing agentic AI
- 82% say open-source models are moderately to extremely important to their AI strategy
- Hybrid computing for AI workloads rose from 35% to 43% year over year

## **Section 1: The Inflection Point Is Already Here**

The 2026 NVIDIA State of AI in Healthcare and Life Sciences survey is the kind of data that should make any health tech investor put down whatever deck they're reading and pay attention. Seven out of ten healthcare and life sciences organizations are now actively using AI. That's not piloting, not exploring, not forming a committee to assess strategic readiness, actually using it. That number was 63% a year prior and that jump isn't noise. It's an industry crossing a threshold.

What's more interesting than the headline number is where the growth came from. The payers and providers segment, which historically moves about as fast as a fax machine, jumped 13 percentage points year over year, from 43% to 56% active AI usage. Hospitals and insurance companies are now majority AI users by this measure. That's not a small thing. Payers and providers represent the largest slice of U.S. healthcare spending by a wide margin, and they've been the laggard cohort in every digital health wave since the EMR rollouts of the early 2010s. When that segment starts moving, the infrastructure underneath it matters a lot.

The survey also surfaced something that more or less confirms what anyone building in health tech has been observing in the field: generative AI and LLMs blew past predictive analytics as the top AI workload category. Sixty-nine percent of respondents cited gen AI as their primary focus, up from 54% the prior year. Data analytics and data science came in at 65%, predictive analytics at 51%, and agent AI, newly tracked this year, debuted at 47%. That agentic number is going to be important. More on that later.

The revenue story is the part that should accelerate capital deployment. Eighty-four percent of management-level respondents said AI increased their annual revenue. Eighty percent said it reduced annual costs. Forty-four percent said the revenue increase exceeded 10% annually. For small companies specifically, 56% reported more than 10% revenue lift from AI. These are not marginal improvements on the margin of a spreadsheet. At the portfolio company level, these are the kinds of numbers that change valuations, extend runways, and make the next fundraising significantly easier.

Budget intentions for 2026 are equally unambiguous. Eighty-five percent of respondents said their AI budgets will increase. Nearly half said budgets will grow more than 10% year over year. The shift in how that money gets allocated is also telling: in 2025, 47% of respondents said identifying new AI use cases was a top spending priority. In 2026, that number dropped to 37%. Meanwhile, optimizing existing AI workflows and production cycles jumped from 34% to 47% as the top spending category. The industry has found its use cases. Now it's scaling them. That's a very different market than it was 18 months ago, and it means the infrastructure layer underneath those production deployments is about to get a lot more important.

That infrastructure layer is predominantly built on NVIDIA.

## Section 2: BioNeMo and the Drug Discovery Revolution

If you're an investor in biotech-adjacent AI or a founder building anything in the discovery or precision medicine space, BioNeMo is the framework you need to understand cold. It's NVIDIA's platform specifically built for AI-driven drug discovery, and it represents a genuine architectural shift in how preclinical R&D is done.

The traditional drug discovery pipeline is one of the most inefficient processes in the industry. Average time from target identification to approved therapy runs somewhere between 12 and 15 years. Average cost is north of 2 billion dollars depending on the study, with failure rates exceeding 90% in clinical trials. Most of that failure happens because the preclinical computational work simply wasn't good enough to predict what would happen in a human body. BioNeMo attacks that problem at the source.

The platform runs generative AI models on NVIDIA GPUs and is designed to let AI models navigate what the documentation calls vast biochemical universes, meaning the combinatorial space of possible molecular structures, protein interactions, and binding predictions that would take human researchers lifetimes to explore manually. The system can design candidate drug molecules and predict molecular interactions at atomic precision, compressing discovery timelines from years to months in favorable cases.

The architecture is worth understanding in some detail because it tells you what kinds of startups can actually build on top of it. BioNeMo has three distinct layers. The NIM microservices layer delivers pre-trained state-of-the-art models through standardized APIs, meaning a team of five engineers can access world-class molecular simulation capabilities without standing up massive infrastructure. The BioNeMo Framework layer is the adaptation layer, where scientists and developers can fine-tune models on proprietary molecular or genomic data. This is where the moat gets built.

for commercial companies. The BioNeMo Blueprints layer operationalizes entire workflows into what NVIDIA describes as self-learning loops of design, make, test, and learn, essentially autonomous research cycles that iterate without constant human input.

For investors, the strategic implication is that the old drug discovery model, where value was concentrated in massive R&D operations with thousands of bench scientists, is getting disaggregated. A small team with access to BioNeMo, a compelling proprietary dataset, and a focused therapeutic hypothesis can now do work that have required a mid-size pharma company's computational biology department a decade ago. That changes the addressable competitive landscape for biotech startups considerably. It also means the venture returns math on early-stage biotech AI companies looks different than it did five years ago, both in terms of capital efficiency going in and in terms of partnership and acquisition interest from large pharma coming the way out.

The NVIDIA survey data from pharma and biotech organizations reinforces this. Literature review and analysis was the top agentic AI use case at 55% for that segment. Drug discovery and biomarker identification came in at 48%. These aren't exploratory pilots anymore. Nearly half of pharma and biotech respondents have AI agents running in their discovery workflows. Forty-six percent of that segment reported ROI from their drug discovery AI investments.

## **Section 3: MONAI and the Medical Imaging Flywheel**

Medical imaging is the use case where healthcare AI ROI is most clearly established, and MONAI, the Medical Open Network for AI, is the open-source framework underpinning a significant share of that activity. Fifty-seven percent of medtech respondents in the NVIDIA survey reported ROI from AI in medical imaging. That's the highest confirmed ROI rate of any specific use case across any segment in the entire report.

MONAI has 6.5 million downloads as of the latest data, has been cited in over 4,000 peer-reviewed papers, and has won more than 20 international medical AI competitions, frequently outperforming proprietary tools. Those are credibility numbers that matter when you're trying to get a hospital system's IT governance committee to approve a new AI vendor. The fact that the underlying framework is open-source, well-documented, and academically validated is a genuine adoption accelerant in a segment that treats vendor risk with extreme caution.

The technical architecture of MONAI is what makes it interesting from a build-out perspective. It provides domain-optimized tooling across the full imaging pipeline from interactive 3D segmentation to multimodal vision-language models that can integrate imaging data with clinical text and other modalities. That last piece, multimodal integration, is where the real clinical value gets generated. A model that can look at a CT scan and simultaneously contextualize findings against a patient's clinical notes, lab values, and medication history is a fundamentally different tool than a model that just classifies images. MONAI's architecture is designed to support that kind of integration.

For the medtech segment specifically, the survey data shows medical imaging at the top as the top use case, followed by clinical decision support at 42% and diagnostic tools including disease diagnosis and risk prediction at 34%. The medtech segment was the one where computer vision ranked as the top AI workload area at 59%, ahead of generative AI. That's counterintuitive relative to the rest of the industry but makes perfect sense when you think about what medtech companies actually build. CT scanners, MRI machines, pathology slide analyzers, and ultrasound systems are fundamentally computer vision applications running on specialized hardware.

The imaging AI market is also one of the few areas in health tech where the reimbursement pathway is reasonably established. The FDA has cleared or authorized over 950 AI-enabled medical devices as of early 2026, and a growing number of them have CPT codes for reimbursement. Founders building imaging AI companies on top of MONAI-based infrastructure are entering a market with a defined regulatory playbook and actual payment mechanisms, which is not something you can say in most digital health categories. That combination of technical maturity, regulatory

precedent, and demonstrated ROI is why imaging continues to attract disproportionate capital relative to its share of total healthcare AI activity.

## **Section 4: Isaac for Healthcare and the Robotics Buildout**

The robotics angle on healthcare AI is probably the least appreciated opportunity in the current investment landscape, partly because the timelines are longer and the capital requirements are higher, and partly because most health tech investors do not come from a robotics background. NVIDIA's Isaac for Healthcare platform is well worth understanding regardless, because it's defining the development environment for what will likely be a very large market segment over the next decade.

Isaac for Healthcare is a simulation and deployment platform that gives medical robotics developers a complete end-to-end pipeline from virtual environment construction through AI model training to real-world hardware deployment. The workflows currently supported are a useful guide to where commercial activity is concentrating. Robotic surgery and surgical assistant robotics using the SO-ARM manipulator are the furthest along, with full pipelines for data collection, policy training, and deployment. Robotic ultrasound, telesurgery with haptic feedback, low-latency video streaming, and hospital automation workflows are also supported.

The concept of a digital twin of a hospital, where care teams can simulate procedures and train AI models before any patient is involved, is no longer a speculative idea. The Isaac platform makes it operational. Developers can build sim-ready assets that mirror real hospital environments, run AI model training against those synthetic environments, and then deploy trained models to actual hardware. The implications for clinical trial design, staff training, and surgical outcomes research are significant and mostly unappreciated outside of a relatively small circle of surgical robotics investors.

For investors, the key strategic question around surgical robotics AI is whether value accrues at the hardware layer or the software layer. The Intuitive Surgical platform has historically concentrated value at the hardware layer through platform lock-

but that model is under meaningful competitive pressure from newer entrants building software-defined surgical systems on open platforms. Isaac for Healthcare is explicitly designed to support the software-defined model, where the intelligence of the surgical system is continuously updated through software rather than through hardware replacement cycles. That's a fundamentally better business model for recurring revenue, and it's the direction the market is moving.

The hospital automation workflow, called Rheo in the Isaac documentation, is also worth flagging. Autonomous hospital logistics, medication delivery, specimen transport, and environmental services robots are a category that has been commercially challenging historically because the environments are complex and unpredictable. Simulation-trained robotics on Isaac infrastructure addresses the training data problem directly, generating synthetic environments that cover the cases a real-world training program would take years to encounter organically. The operational cost savings potential in hospital logistics is substantial given nursing labor costs and the administrative burden that non-clinical tasks impose on clinical staff.

## **Section 5: Holoscan and the Edge Intelligence Layer**

Holoscan is the NVIDIA product that gets the least attention from health tech investors, which is a mistake given its technical position. It's a multimodal AI sensor processing platform designed specifically for real-time inference on streaming data at the edge. In healthcare terms, that means running AI directly on medical devices during procedures, not sending data to a cloud, not adding latency, not depending on network connectivity, doing inference in the operating room or at the point of care in real time.

The surgical video workflow is the clearest current application. Holoscan enables low-latency AI processing of surgical video feeds with real-time tool detection and segmentation. The modular pipeline architecture means that medtech companies can integrate specific AI models for their procedure type without rebuilding the



underlying infrastructure. The Holoscan Sensor Bridge extends this to arbitrary sensor types, handling high-bandwidth data from diverse sensors over Ethernet standard API and open software built on an FPGA interface.

The HoloHub repository is where the reference applications live, and browsing it gives a good sense of where commercial development is concentrating. End-to-end surgical video, body pose estimation, integration with 3D Slicer for surgical planning and augmented reality volume rendering via Magic Leap are among the available workflows. The 3D Slicer integration in particular is interesting because 3D Slicer is already deeply embedded in surgical planning workflows at academic medical centers, meaning Holoscan can slot into existing institutional infrastructure rather than requiring greenfield adoption.

The broader strategic significance of Holoscan is that it enables a class of medical applications that cloud-dependent architectures fundamentally cannot support. Real-time intraoperative guidance, where a surgeon needs AI feedback within milliseconds of a camera movement, cannot tolerate cloud round-trip latency. Real-time monitoring of critically ill patients where response time is measured in seconds, not minutes, requires edge inference. These are the high-value, high-stakes applications where AI actually changes clinical outcomes rather than just administrative efficiency, and Holoscan is the platform specifically built for them.

The NVIDIA survey data on hybrid computing is relevant here. The shift from 34% to 43% using hybrid computing for AI workloads year over year, concurrent with cloud-only dropping from 41% to 35%, reflects exactly the trend Holoscan is positioned to capture. Healthcare organizations are figuring out that some workloads need to be at the edge, and they're building infrastructure to support that. For founders building device-adjacent AI companies, the question of which inference platform to build on has a fairly clear answer at this point.

## **Section 6: Parabricks and the Genomic Data Deluge**

Parabricks is NVIDIA's GPU-accelerated genomics software suite, and the market context for it is almost comically large. The genomics field is heading toward terabytes of sequencing data in the coming decade as sequencing costs continue to decline exponentially. The cost to sequence a human genome has dropped from over 100 million dollars in 2001 to under 200 dollars in 2026. The problem has completely flipped: getting the sequence is now the easy part. Making sense of it is where the bottleneck is.

Parabricks handles the secondary analysis layer, taking raw sequencing output and doing the alignment, variant calling, and related processing that turns raw reads into interpretable genomic data. The GPU acceleration cuts processing runtimes from hours to minutes for standard whole-genome sequencing pipelines. On a practical level, that's the difference between genomic results being available for clinical decision-making during a hospitalization versus arriving days after discharge. In neonatal intensive care, oncology, and rare disease settings, that timeline difference is clinically material.

The CUDA-X Data Science suite, formerly RAPIDS, handles single-cell and tertiary analysis, which is where population genomics and research applications live. Combined with Parabricks for secondary analysis and GPU-accelerated primary analysis during sequencing itself, NVIDIA now has coverage across the entire genomics computational pipeline.

The investment angle here is mostly about what Parabricks makes possible downstream rather than Parabricks itself as an investment target. When genomic data can be processed in near-real-time at scale, the market for clinical genomics interpretation, population health genomics, and genomically-informed drug target identification all expand substantially. Any startup operating in those application layers benefits directly from the infrastructure improvement. The pharma and biotech segment ranked genomic applications at 44% as the second most common AI use case in the NVIDIA survey, just behind drug discovery at 57%. That's a meaningful indicator of where R&D capital is flowing.

## Section 7: Clara, NIM, and the Open Source Bet

The open-source question in healthcare AI is worth addressing directly because NVIDIA survey data on it is unusually strong. Eighty-two percent of respondent open-source models and software were moderately to extremely important to the strategy. Fifty-seven percent said they were very or extremely important. This is fringe preference. It's the dominant strategic orientation of the people actually building and deploying healthcare AI.

The logic is not hard to follow. Healthcare AI applications tend to be highly specific. An imaging AI for detecting early-stage pancreatic cancer in CT scans has a fundamentally different training distribution than one for detecting pneumonia chest X-rays. A clinical documentation model tuned on oncology notes performs differently than one tuned on emergency department encounters. General-purpose foundation models are a starting point, not a finish line. Fine-tuning on proprietary clinical data, using open-source frameworks that allow full customization without licensing constraints, is how organizations build AI that actually works in their specific context.

Clara is NVIDIA's family of open models and tools purpose-built for scientific discovery, medical imaging, and biology and chemistry research. It includes model development recipes, and evaluation frameworks across imaging, biology, and drug discovery domains. The strategic positioning of Clara as an openly accessible platform is a deliberate move to accelerate ecosystem adoption, and it's working given the MONAI download numbers and the peer-reviewed citation count.

NIM microservices are the delivery mechanism for accessing NVIDIA's most capable models through standardized APIs without requiring teams to manage underlying infrastructure. For health tech startups, NIM is particularly valuable because it removes the infrastructure overhead from model deployment, letting small engineering teams focus on the application layer rather than on GPU cluster management. The combination of NIM for access to foundation capabilities and

or BioNeMo for domain-specific fine-tuning gives a startup team a genuinely competitive technical starting point without enterprise-scale infrastructure.

The implication for the build-vs-buy question that every health tech founder faces is increasingly clear. Buying access to general-purpose AI from large model providers and building on top of it without domain customization produces mediocre results in clinical applications. Building everything from scratch is prohibitively expensive for most startups. The open-source fine-tuning path, using NVIDIA's domain-specific frameworks as the foundation and layering in proprietary clinical data, is where the best risk-adjusted technical outcomes are happening. The survey data on where the market is concentrating supports this: the organizations reporting the highest AI ROI are the ones applying specific AI to distinct use cases, not the ones deploying general-purpose tools broadly.

## **Section 8: What This Means for Investors and Founders**

The picture that emerges from the NVIDIA platform ecosystem combined with the 2026 survey data is of an infrastructure stack that is simultaneously mature enough to build on and early enough that the application layer is not yet crowded. That's a genuinely rare combination in health tech, which tends to alternate between underdeveloped infrastructure that makes building too hard and overcrowded application markets where differentiation is nearly impossible.

For angel investors and syndicate leads specifically, the portfolio construction implications break down across a few dimensions. First, any company building clinical AI applications that is not building on top of GPU-accelerated infrastructure—whether NVIDIA-based or otherwise, should be asked hard questions about how it plans to compete as inference demands scale. The survey data showing hybrid computing adoption at 43% and climbing reflects a market that is normalizing GPU infrastructure costs. That normalization reduces the infrastructure moat of companies that built early on proprietary compute and increases the importance of application layer differentiation.

Second, the agentic AI debut at 47% usage or assessment is the number to watch. Agentic AI systems that can autonomously reason, plan, and execute multi-step healthcare tasks represent a qualitative jump in what AI can do in clinical and research settings. The current top use cases, knowledge management and retrieval at 46%, literature review at 38%, and internal process optimization at 37%, are mostly back-office. The interesting commercial territory is what happens when agentic moves into clinical workflows in earnest. The regulatory environment, HIPAA, FDA, and similar frameworks, remains the primary constraint on that transition, with 60% of respondents citing regulatory compliance as the top factor influencing their AI implementation approach. Founders who can navigate that regulatory surface and build defensible governance frameworks are building a real moat.

Third, the small company revenue data deserves more attention than it usually gets. Fifty-six percent of small healthcare AI companies reported more than 10% annual revenue growth attributable to AI, versus 44% for large companies. That's counterintuitive given that larger organizations have more resources, more data, more infrastructure. The explanation is probably that small companies are better at applying AI to a single well-defined problem rather than trying to boil the ocean, which maps to the survey's broader finding that specific AI applied to distinct use cases outperforms general-purpose deployment on ROI metrics. For early-stage investors, this is an argument for funding companies with narrow, well-defined applications over companies pitching broad platform plays.

The infrastructure inequality finding between large and small organizations is the genuine caution flag in the data. Forty percent of small healthcare AI companies cited budget as their top challenge. Thirty-three percent cited data size constraints for model training. These are structural disadvantages that don't go away on their own, and they're why the open-source ecosystem matters so much for small company competitiveness. NVIDIA microservices, MONAI, BioNeMo, Parabricks, and the CLIP model families collectively give small teams access to capabilities that would have required dedicated ML infrastructure teams just a few years ago. The playing field is leveling in terms of model access, but capital constraints on compute and data acquisition remain real.

The overall thesis is not complicated. NVIDIA has built the most comprehensive infrastructure stack specifically targeting healthcare and life sciences. It spans drug discovery through BioNeMo, medical imaging through MONAI, surgical and hospital robotics through Isaac for Healthcare, edge inference through Holoscan, genomics through Parabricks, and open model access through Clara and NIM. The survey shows an industry that has crossed the adoption inflection point, is generating measurable ROI, is increasing budgets, and is shifting from experimentation to production scaling. The companies building serious clinical AI applications in 2023 and beyond are building on this infrastructure whether they realize it or not, and investors who understand the stack have a meaningful edge in evaluating who is building on solid ground versus who is building on sand.

The picks-and-shovels metaphor gets overused in tech investing but it's actually very relevant here. In the 1849 California Gold Rush, the people who got rich selling picks and shovels did so because they got paid regardless of which miner struck gold. NVIDIA's position in healthcare AI infrastructure is structurally similar. Every drug discovery company that finds a novel molecule using BioNeMo, every radiology AI company that deploys on MONAI, every surgical robotics company that trains on Isaac, every genomics company running Parabricks in production, they all run on NVIDIA. The question for founders is how to build a durable application-layer business on top of that infrastructure. The question for investors is which of those application-layer businesses are most likely to generate asymmetric returns given a market that is clearly in the early innings of a sustained scaling cycle.

The NVIDIA survey's closing observation is probably the right one to end on. The researchers predict that by 2027, healthcare AI will shift from predominantly predictive analytics toward more consistent deployment of agentic systems capable of reasoning across patient populations, clinical trials, and care workflows simultaneously. That transition, if it happens anywhere close to that timeline, will be the most significant shift in how clinical decisions get made since evidence-based medicine became the standard of care in the 1990s. The infrastructure to support this already exists. The regulatory frameworks are catching up. The capital is moving. This is the moment to be paying very close attention.



2 Likes • 1 Restack

[← Previous](#)

## Discussion about this post

**Comments**

Restacks



Write a comment...

---

