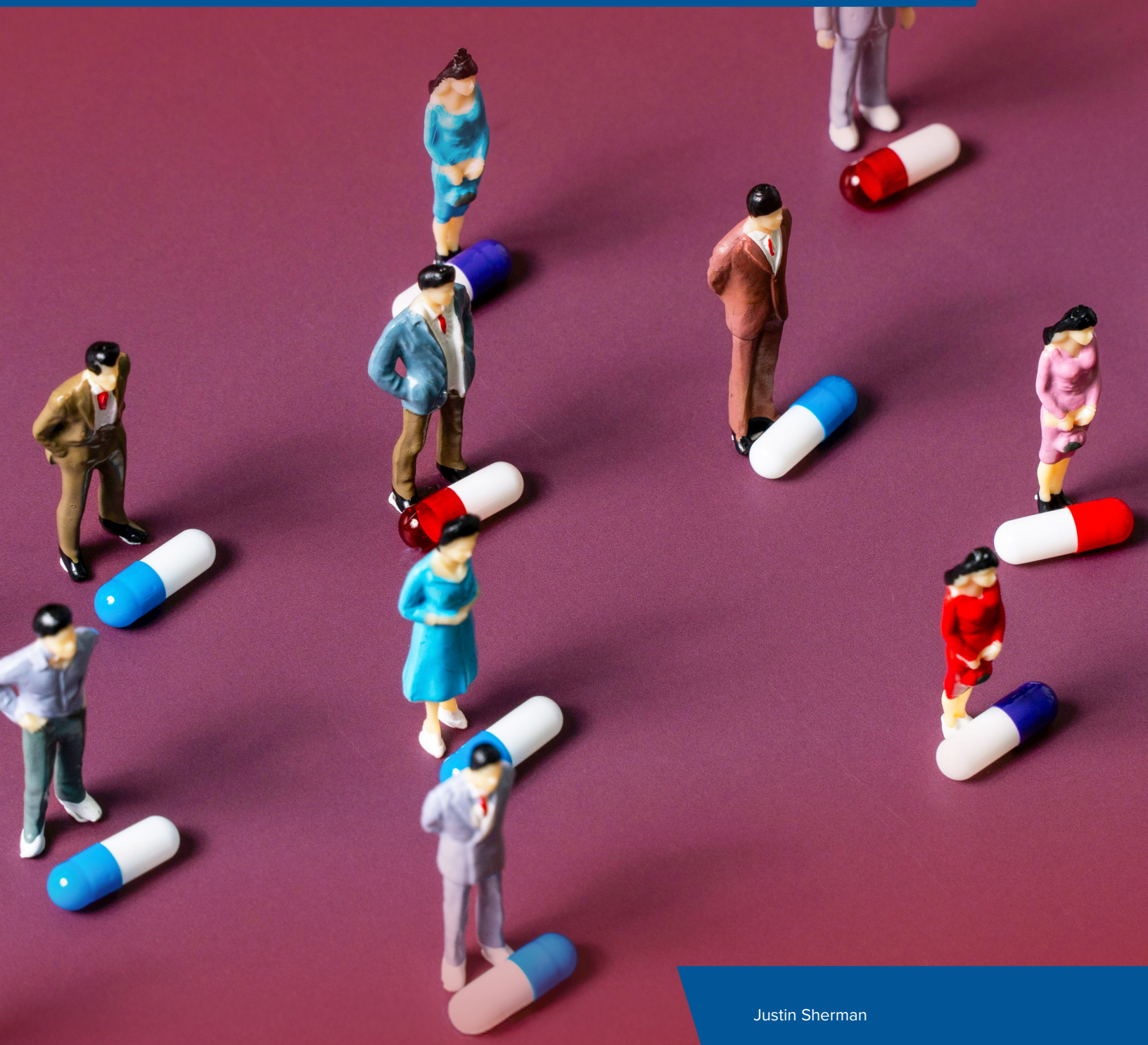
Atlantic Council
CYBER STATECRAFT INITIATIVE

ai x ac

AI Data Under the Microscope:

Accelerating and securing the AI data supply chain for the health and biopharmaceutical sectors



Justin Sherman

Atlantic Council

CYBER STATECRAFT INITIATIVE

The **Cyber Statecraft Initiative** works at the nexus of geopolitics, technology, and security to craft strategies to help shape the conduct of statecraft and to better inform and secure users. This work extends through the competition of state and non-state actors, the security of the internet and computing systems, the safety of operational technology and physical systems, and the communities of cyberspace. The Initiative convenes a diverse network of passionate and knowledgeable contributors, bridging the gap among technical, policy, and user communities.

Author

Justin Sherman

Please direct inquiries to:

Atlantic Council
1400 L Street NW, 11th Floor
Washington, DC 20005

September 2026

Acknowledgments

The author would like to thank Diana Pankevich, Lee Licata, Devin Lynch, Ranjit Kumble, Kenton Thibaut, and Trey Herr for their comments on earlier drafts of this report. Thanks especially to Kenton Thibaut for extensive assistance on China's sprawling regulatory apparatus and to Diana Panvkeich and Ranjit Kumble for discussion of health and biopharmaceutical industry subjects. Finally, thanks to Nitansha Bansal, Safa Shahwan Edwards, Nancy Messieh, and the rest of the team for support on the project and this report in particular.

Table of contents

Executive Summary

Health and biopharmaceutical companies occupy an important frontier in artificial intelligence (AI) research and development and the data that goes into it. Organizations in these sectors can, if behaving irresponsibly or harmfully, collect, analyze, use, and share data in ways that undermine patient privacy, create cybersecurity risks to individuals and society, threaten US national security, and inflict harm on people receiving care, particularly individuals in already marginalized or disadvantaged populations. At the same time, health and biopharmaceutical companies have tremendous opportunities to leverage data in line with established patient consent frameworks, apply data to developing novel AI technologies that advance research and development, and drive innovations that improve patient care, treat disease, and increase health equity—sometimes, without the tremendous costs that come with general-purpose large language model development. Given the importance of data and data-related components across the health and biopharmaceutical AI value chain, this report presents a framework for the data in the AI supply chain, maps it against the sectors in question, and offers five real-world use cases that illustrate varied opportunities, risks, and relationship dynamics. It then surveys fourteen major regulatory regimes around the world (AI-specific and non-AI-specific) that impact the data components of the AI supply chain for the health and biopharmaceutical sectors, describes current tensions and future issues, and ends with three high-level recommendations for how governments can navigate this frontier safely, securely, and responsibly for societal benefit.

Introduction

Health and biopharmaceutical (biopharma) industries continue to generate and collect significant amounts of data around the world and, increasingly, to build and deploy artificial intelligence (AI) technologies. Indeed, startups and research organizations in the health and biopharma sectors are using emerging AI models to predict the highly complex structures of proteins, improve early-stage diagnosis of diseases, bolster genetic sequencing and analysis capabilities, and potentially accelerate scientific discovery. These activities can further important functions—with much clearer societal benefit than many other use cases, such as general-purpose large language model chatbots in workplaces—including health condition diagnosis, disease research, and vaccine development. At the same time, the use of data as well as the development and deployment of AI raise critical questions about transparency, auditability, privacy, cybersecurity, inequity, national security, and more. Further, the globalized nature of the AI supply chain writ large as well as the globalized nature of the health and biopharma industries raise complicated geopolitical questions in an increasingly tense world, around cross-border access to citizen data (including personal health data), ideas of sovereignty, and national security.

As the companies and organizations in the health and biopharma sectors leverage AI technologies, they interact with many data components across the AI supply chain: training data, testing data, AI models, model architectures, model weights, application programming interfaces (APIs), and software development kits (SDKs). Health and biopharma companies do so subject to a varied, shifting patchwork of global laws and regulations, from the United States to the European Union (EU) to China to India, some specific to AI and others more broadly.

Altogether, the breadth and complexity of these components and the sometimes compatible, sometimes clashing nature of regulatory regimes raise many urgent policy questions for the future of health and biopharma AI-related activities. Poorly formed and implemented policies could expose patient data with unforeseen consequences, lead to the compromise of AI models, yield inadequately tested diagnostic or predictive AI systems, disproportionately and unnecessarily stall important health research, and much more. Thoughtful and carefully implemented policies, on the other hand, could accelerate the responsible, adequately protected use of health and biopharma data components in ways that genuinely advance patient outcomes—while often using far less energy, far less compute, and far more consent-permissioned data than many generic large language models currently depend upon. These patient outcomes could include earlier and more accurate prediction of disease, improved disease treatments, and more equitable treatment access and quality within and between populations.

This report unpacks the health and biopharma policy challenges across the data components of the AI supply chain, with the aim of identifying major tensions, flagging open questions, and offering potential next steps for future-proofed policymaking. The author begins by introducing a framework from a prior Atlantic Council publication[1] to review the core data components in the AI supply chain, including how they overlap and interact. Then, this framework is used to lay out the spectrum of health and biopharma examples of those data components in the AI supply chain, from a number of health and biopharma organizations around the world working on different or overlapping problems. After those examples of health and biopharma AI-related data components, mapped against the supply chain framework, the report details five exemplative case studies of health and biopharma use cases. These range from cancer prediction to the enhancement of genetic sequencing using large-scale, global datasets to cross-border activities from a company that raises US national security questions.

1. Justin Sherman, *Securing the Data in the AI Supply Chain*, Atlantic Council, September 2025, https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/securing-data-in-the-ai-supply-chain/.

Next, the report describes the fractured global regulatory landscape for the health and biopharma data components of the AI supply chain through the lens of fourteen regulatory regimes, including in the United States, the European Union, China, and India. The subsequent section details the current tensions in those regulatory regimes and lists a number of pressing, open questions for policymakers around the world vis-à-vis data privacy, data security, AI innovation, and the like as applied to the health and biopharma data spheres. Finally, the report concludes with three overarching, high-level recommendations for policymakers to identify current tensions, close current legal and regulatory gaps, and chart better paths forward.

- Governments should use existing best-practice privacy and cybersecurity principles when pursuing regulations on data collection, data use, cross-border data transfers, and AI model deployment that impact health and biopharma, including the healthcare profession's patient consent ethical framework and standards bodies' specifications for encryption, data minimization, and protection against reidentification.
- The United States, European Union, India, and China should issue clarifications about how the health and biopharma data components of the AI supply chain sit within their non-AI-specific and AI-specific laws, drawing on the tensions and open questions identified in this report.
- Governments should clarify criteria for any situation where they will treat health or biopharma AI-related data activities differently than AI-related data activities in other sectors.

Enabling the future of responsible innovation in the health and biopharma sectors, with the goal of delivering better and more equitable health outcomes for all, depends in part on policymakers and other stakeholders navigating these complexities. The time to do so is now.

Health and Biopharma Data in the AI Supply Chain

A complex supply chain that includes organizations, people, activities, information, and resources underpins AI technologies that enable AI research, development, deployment, and more.[2] The AI supply chain includes human talent, compute, institutional and individual stakeholders, and, data, the last of which is the focus of this report.

Data components fit into the AI supply chain because researchers, developers, deployers, users, maintainers, governors, securers, and attackers of AI systems depend upon and access different kinds of data that are transmitted, stored, and analyzed in different ways in order to make it all happen. They also fit into the AI supply chain because a wide range of entities around the world—from individuals who publish self-labeled datasets to corporations that analyze AI model outputs—provide, access, and use the underlying data as well. Conceptualizing this data as part of a supply chain echoes the concept of AI as a value chain (referring to the business activities that deliver value to customers), though focused specifically on data components.[3]

As detailed in a previous Atlantic Council report, the data in the AI supply chain covers many data types, sources, and formats—all of which need safeguards to enable competition, boost public trust, and protect against the leaks, exploitation, and other risks delineated above. The data in the AI supply chain includes the data describing an AI model's properties and behavior, as well as the data associated with building and using a model. It also includes the AI models themselves and the different systems that facilitate the movement of data into and out of models.[4] This report conceptualizes seven parts of the data and core data systems in the AI supply chain, which are laid out in the visualized framework and tables below:

- Training data
- Testing data
- Models (themselves)
- Model architectures
- Model weights
- Application Programming Interfaces (APIs)
- Software Development Kits (SDKs)

This framework draws, at a high level, inspiration from a November 2024 paper by Qiang Hu et al on the large language model (LLM) supply chain. The paper envisioned a framework for understanding the components and processes that go into LLMs.[5] Nonetheless, this paper differs, focusing on the data components themselves rather than the activities to produce them (like dataset processing); delineating data components by their properties and functional differences (such as distinguishing between training data and testing data); and looking at the data supply chain for AI technologies broadly (instead of just LLMs). This framework also differs in that it focuses specifically on the need for security.

Notably, the first five of these seven components are data per se, or the models themselves. The last two, however—APIs and SDKs—are neither data nor models themselves; instead,

2. This section is pulled and adapted from the earlier Atlantic Council report establishing this framework: Justin Sherman, *Securing the Data in the AI Supply Chain*, Atlantic Council, September 2025, https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/securing-data-in-the-ai-supply-chain/.
3. See: Beatriz Botero Arcila, *AI Liability Along the Value Chain* (San Francisco: Mozilla Foundation, April 2025), https://blog.mozilla.org/netpolicy/files/2025/03/AI-Liability-Along-the-Value-Chain_Beatriz-Arcila.pdf; Max von Thun and Daniel A. Hanley, *Stopping Big Tech from Becoming Big AI* (San Francisco: Mozilla Foundation, October 2024), https://blog.mozilla.org/wp-content/blogs.dir/278/files/2024/10/Stopping-Big-Tech-from-Becoming-Big-AI.pdf; SPEAR Invest, "Diving Deep into the AI Value Chain," NASDAQ, December 18, 2023, https://www.nasdaq.com/articles/diving-deep-into-the-ai-value-chain. See also: "The Value Chain," Harvard Business School: Institute for Strategy and Competitiveness, accessed June 17, 2026, https://www.isc.hbs.edu/strategy/business-strategy/Pages/the-value-chain.aspx.
4. While recognizing the necessity of evaluating AI in relation to the social, political, and economic systems that researchers, companies, and others operate within and use to build AI technologies—such as exploitative labor systems and the environmental system—this report focuses, for scope- and length-limitation purposes, on a typology of the digital and data elements themselves of relevance for AI R&D. For essential reading on other systems that generate data, move data into AI systems, and much more, see: Tamara Kneese, *Climate Justice and Labor Rights: Part I: AI Supply Chains and Workflows* (New York: AI Now Institute, August 2023), https://ainowinstitute.org/general/climate-justice-and-labor-rights-part-i-ai-supply-chains-and-workflows; Kashmir Hill, *Your Face Belongs to Us: A Tale of AI, a Secretive Startup, and the End of Privacy* (New York: Penguin Random House, 2023); Billy Perrigo, "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," TIME, January 18, 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/; Adrienne Williams, Milagros Miceli, and Timnit Gebru, "The Exploited Labor Behind Artificial Intelligence," *Noema Magazine*, Berggruen Institute, October 13, 2022, https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/.
5. Qiang Hu et al., "Large Language Model Supply Chain: Open Problems from the Security Perspective," arXiv, November 3, 2024, https://arxiv.org/abs/2411.01604. See: LLM supply chain map on page 2.

they are code and software systems that enable data to pass into, extract from, and otherwise collect around AI models. For example, business users of an LLM may use an API to submit questions to the chatbot in batches; mobile consumers using an AI image recognition application may, whether they know it or not, depend on an SDK to take their snapshot of a bird and submit to a cloud-hosted AI model, which then returns back through the SDK's code the species of the bird in question. The concept of data components of the AI supply chain does not list out every possible software system that could interact with AI data components, but it includes APIs and SDKs because of their prevalence and their security relevance in delivering AI data to and from cloud systems and mobile devices. After all, many of the major AI commercial companies in the United States, India, China, the European Union, and elsewhere offer access to APIs to use their models (including submitting queries to chatbots and uploading images to recognition models).

This supply chain is highly relevant to the health and biopharma industries. Companies in the healthcare and biopharma sectors are developing, deploying, procuring, using, testing, and maintaining AI systems. They must contend with enabling the value of the data in their AI supply chains while doing so securely. The costs of failing to innovate securely and responsibly are serious. Disruptions to data utility in this context can undermine activities that serve a societal benefit, such as disease research and vaccine development. In this context, compromises of data security can expose trade secrets (such as proprietary training datasets or testing methodologies), violate consumer or patient privacy (such as through the theft or leaking of genetic data), bring regulatory scrutiny (such as through violating data protection requirements), and detract from health research and other R&D goals (such as disrupting projects). More broadly, these compromises can undermine trust in the use of AI-related datasets and models specifically tailored for disease research and public health benefit, potentially forestalling important future health innovations.

Enabling the value of health and biopharma data in the AI supply chain, while securing it, first requires mapping such data in the AI supply chain itself. Table 1 lists each of the seven data components of the AI supply chain, defines them, and provides examples of health or biopharma companies and other entities involved in that component or its sourcing. Again, this focuses primarily on traditional AI models (e.g., excludes agents), and does not include other elements of the AI supply chain (e.g., human talent, compute). In addition, it does not consider the outputs and exhaust from AI model usage, including metadata. In doing so, this table lays the foundation for the discussion of use cases for health and biopharma AI-related data in the next section.

Table 1: Health and Biopharma Data Components of the AI Supply Chain

Component	Definition	Examples of involved companies and entities
Training data	Datasets used to shape a model's initial configuration in order to have the model recognize patterns in data, make statistical predictions based on data, generate content by iteratively updating the model to increase its predictive accuracy, and more.[6]	Biobank from the organization UK Biobank, which consists of data on half a million participants' biology, health, and lifestyle to drive health discoveries.[7] RCSB Protein Data Bank, funded by the US National Science Foundation, Department of Energy, National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences, consists of 3D structure data for large biological molecules (protein, DNA, and RNA) essential for biology, health, energy, and biotech R&D.[8]
Testing data	Datasets used to evaluate how a model performs on recognizing patterns in data, making statistical predictions based on data, and more, after the model has been initially trained. Testing data is often used by many different developers, such that results can be compared and "benchmarked" across models. The data could be a subset of training data, though not always—its functional distinction singles it out (e.g., perhaps a startup or government agency saves its most proprietary, sensitive, and realistic data examples for its testing phase, so it can observe how the model would perform in real-world environments).	CheXpert, from the Stanford Machine Learning Group, which consists of chest X-ray data along with uncertainty labels and radiologist-labeled reference standard evaluation sets.[9] Datasets from the US National Institute of Standards and Technology's Genome in a Bottle Consortium, which include reference data of whole human genome sequencing for benchmarking.[10]
Models (themselves)	The overall program with the data processing structure/architecture to take data as inputs; recognizes patterns in, makes predictions based on, or perform other mathematical analysis of the data; and generate a resulting output.[11] Models are comprised of both a model architecture (below) and model weights (below that).	AlphaFold from Google's DeepMind, which is designed to evaluate long chains of amino acids to predict protein structures in minutes.[12] The Nucleotide Transformer from the company InstaDeep's DeepChain,[13] which is a foundation model trained on DNA sequences from over 3,200 diverse human genomes and 850 genomes from a wide range of species to predict molecular phenotypes.[14]

6 See: Shigeyuki Takano, *Thinking Machines: Machine Learning and its Hardware Implementation* (Cambridge: Academic Press, 2021), https://www.sciencedirect.com/science/article/abs/pii/B9780128182796000116; Nayna Jaen, "How AI is Trained: The Critical Role of AI Training Data," *RWS*, March 26, 2024, https://www.rws.com/artificial-intelligence/train-ai-data-services/blog/how-ai-is-trained-the-critical-role-of-ai-training-data/; Michael Chen, "What Is AI Model Training & Why Is It Important?," Oracle, December 6, 2023, https://www.oracle.com/uk/artificial-intelligence/ai-model-training/.

7 "About Us," UK Biobank, accessed June 17, 2026, https://www.ukbiobank.ac.uk/about-us/.

8 "About RSCB PDB: A Living Digital Data Resource That Enables Scientific Breakthroughs Across The Biological Sciences," *RSCB.org*, accessed June 17, 2026, https://www.rcsb.org/pages/about-us/index.

9 "What is CheXpert?," Stanford ML Group, accessed June 17, 2026, https://stanfordmlgroup.github.io/competitions/chexpert/.

10 "Genome in a Bottle," National Institute of Standards and Technology, accessed June 17, 2026, https://www.nist.gov/programs-projects/genome-bottle.

11 See: "What Are AI models?," HPE, accessed April 29, 2025, https://www.hpe.com/us/en/what-is/ai-models.html; Chen, "What Is AI Model Training & Why Is It Important?"

12 "AlphaFold," Google DeepMind, accessed June 17, 2026, https://deepmind.google/science/alphafold/.

13 "InstaDeep—Decision-Making AI for the Enterprise," InstaDeep, accessed June 17, 2026, https://instadeep.com/; "Models," DeepChain, accessed June 17, 2026, https://deepchain.bio/models/.

14 "AI Foundation Models for Genomics," GitHub: instadeepai, June 17, 2026, https://github.com/instadeepai/nucleotide-transformer.

Table 1: Health and Biopharma Data Components of the AI Supply Chain (cont.)

Component	Definition	Examples of involved companies and entities
Model architectures	The layout of how data flows and processes within an AI model, such as a neural network, including vis-à-vis the inputs, hidden layer or layers (which analyze the previous layer's output to create the next layer's input), and outputs.[15]	Evoformer, a neural network block built and used in one of AlphaFold's iterations—specifically to view the prediction of protein structures as a graph inference problem in 3D space.[16] U-Net, an architecture that uses convolutional networks for biological image segmentation.[17]
Model weights	Numerical parameters that specify the strength between variables as a model trains and fine-tunes—that is, how the model numerically represents connections between pieces of data to achieve the desired output or result.[18]	The weights for Microsoft's BioGPT, made available via Hugging Face.[19] The weights for Google DeepMind's AlphaFold, made available via GitHub.[20]
Application programming interfaces (APIs)	Software systems that allow individuals to write code to input data to and pull data from a system, which could include a database or an AI model—for example, to download training data from an online repository or send queries or other data inputs to an externally hosted and deployed AI model for processing.	FHIR R4 APIs in Oracle's Health Millennium Platform, which allow authorized users to access electronic health record data via Oracle systems.[21] HealthLake APIs from Amazon Web Services, which enables organizations covered by the Health Insurance Portability and Accountability Act (HIPAA) to access machine learning and generative AI models.[22]
Software development kits (SDKs)	Prepackaged code that enables developers to easily add functionalities to their applications and programs—such as communicating with servers hosting an AI model.	Biopython, a set of openly available biological computation libraries and tools.[23] NVIDIA Holoscan SDK, a real-time runtime for building GPU-accelerated systems that ingest high-bandwidth, multimodal, multi-rate sensor data, such as for processing medical device data with AI models.[24]

15 See: "What Is a Neural Network?," AWS, accessed August 27, 2025, https://aws.amazon.com/what-is/neural-network/; "What Is Neural Network Architecture?," H2O AI, accessed August 27, 2025, https://h2o.ai/wiki/neural-network-architectures/; "The Essential Guide to Neural Network Architectures," V7 Labs, July 8, 2021, https://www.v7labs.com/blog/neural-network-architectures-guide.

16 John Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* (July 15, 2021), https://www.nature.com/articles/s41586-021-03819-2.

17 Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv, May 18, 2015, https://arxiv.org/abs/1505.04597.

18 See: Alisdair Broshar, "What Are LLMs? An Intro into AI, Models, Tokens, Parameters, Weights, Quantization and More," Koyeb, April 25, 2024, https://www.koyeb.com/blog/what-are-large-language-models; "AI Model Weights," US National Telecommunications and Information Administration, accessed April 29, 2025, https://www.ntia.gov/programs-and-initiatives/artificial-intelligence/open-model-weights-report/background.

19 "BioGPT," Hugging Face, accessed June 17, 2026, https://huggingface.co/microsoft/biogpt.

20 "AlphaFold," GitHub: Google DeepMind, accessed June 17, 2026, https://github.com/google-deepmind/alphafold.

21 "FHIR R4 APIs for Oracle Health Millennium Platform," Oracle, accessed June 17, 2026, https://docs.oracle.com/en/industries/health/millennium-platform-apis/mfrap/r4_overview.html.

22 "FHIR Storage and Interoperable Health Data Standards—AWS HealthLake," AWS, accessed June 17, 2026, https://aws.amazon.com/healthlake/.

23 "Biopython," Biopython, accessed June 17, 2026, https://biopython.org/.

24 "NVIDIA AI Platforms for Healthcare and Life Sciences," NVIDIA, accessed June 17, 2026, https://www.nvidia.com/en-us/industries/healthcare-life-sciences/.

Many health and biopharma data components of the AI supply chain span multiple categories of components at once. Google DeepMind's AlphaFold model, which helps to predict the structure of proteins, is entangled with multiple other component areas (including the model itself), the training data used to teach the model, and the testing data used to test the model. The U-Net architecture is a model architecture, but its use in many health- and biopharma-related AI systems—such as in recent work on medical image segmentation and multicellular biological systems[25]—makes it much more present throughout several aspects of the data in the AI supply chain. Health tech company Tempus advertises real-world, multimodal clinical, molecular, and imaging data for therapy development and trials,[26] at the same time as it uses AI models to build structured patient cohorts from unstructured patient records and data.[27]

Plenty of private-sector companies also maintain their own internal AI-related data components. Based on conversations with industry, while companies will ultimately release many of the products and services that stem from R&D using the AI-related data—such as a future vaccine—it varies how much of the underlying AI-related data components are publicly shared or disclosed to a third party, based on regulatory requirements, clinical trial obligations, and business decisions, among other factors. For example, the fact that the table above primarily contains examples of datasets themselves (e.g., training datasets, testing datasets) from public research consortiums, universities, and governments rather than private-sector biopharma companies may reflect the fact that many in the latter category build and maintain their own internal AI-related training and testing datasets that they cannot (e.g., for regulatory reasons) or will not (i.e., by choice) publish. This underscores, among other points, the limits of conceptually applying ideas about open-source large language models to the varied field of more focused health or biopharma AI models and related data components.

"Open" access to health and biopharma AI-related data is one thing, while the ability for any organization or individual to effectively make use of that data for the purposes of training, testing, or improving an AI model is another. It requires access to computing resources and data storage, medical knowledge, business or clinical research mechanisms through which to deploy and ethically test the model in practice, and so on. Similarly, having access on the public internet to health or biopharma AI-related data components, from model weights to APIs to training data, does not mean everyone can equally compromise the data component's security or the privacy of the people in the underlying data either—whether with malintent or with ethical intentions to help bolster subsequent protections. As other researchers and practitioners have explored,[28] there are many elements of the AI supply chain and the computing stack for AI that impact how an organization or individual could leverage any of these data components.

Other ways to conceptually parse the health and biopharma data components of the AI supply chain include:

- **Use restrictions:** Across this supply chain, AI-related health and biopharma data includes data that is open, free, and publicly available for download and unrestricted use. It includes data that is private, corporate, and internal-only—in other words, proprietary. It also includes data that is somewhere in the middle: licensed by a government agency to a particular set of companies for a defined set of purposes; sold by a data broker to a limited pool of customers for their own internal use; or, perhaps, shared under a research agreement between a university, a start-up, and a public interest research group. Similarly, health and biopharma organizations can make AI model weights or AI models themselves free for public use, free for noncommercial use, restricted to certain users, closed-source (for internal use only), and so forth. They can do the same with APIs and SDKs.
- **Consent type:** Health and biopharma organizations may functionally get or may be legally required to obtain different types of individual consent (or not) for AI-related data collection, use, and transfer. As discussed further in the regulatory section below, these constraints or lack thereof will depend on the jurisdiction, type of data, use case for the data in question, among other factors. Some AI-related datasets may require written consent

25. See, e.g., Nima Hassanpour and Abouzar Ghavami, "Deep Learning-based Bio-Medical Image Segmentation Using UNet Architecture and Transfer Learning," arXiv, May 24, 2023, https://arxiv.org/abs/2305.14841; Tien Comlekoglu et al., "Surrogate Modeling of Cellular-Potts Agent-Based Models as a Segmentation Task Using the U-Net Neural Network Architecture," *PLOS Computational Biology* (November 3, 2025), https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1013626.
26. "Real-World Multimodal Data for Therapy Development," Tempus, accessed June 17, 2026, https://www.tempus.com/solutions/real-world-data/.
27. "AI in Healthcare in 2026 and Beyond," Tempus, accessed June 17, 2026, https://www.tempus.com/content/article/ai-in-healthcare/.
28. See, e.g., David Gray Widder, Meredith Whittaker, and Sarah Myers West, "Why 'Open' AI Systems Are Actually Closed, and Why This Matters," Nature 635 (2024): 827-33, https://www.nature.com/articles/s41586-024-08141-1; Matt Davies and Jai Vipra, "Mapping Global Approaches to Public Compute," Ada Lovelace Institute, November 4, 2024, https://www.adalovelaceinstitute.org/policy-briefing/global-public-compute/; Sara Ann Brackett, Securing Cloud Infrastructure for AI, Atlantic Council, March 31, 2026, https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/securing-cloud-infrastructure-ai/; Noah Martin and Fahad Dogar, "Divided at the Edge—Measuring Performance and the Digital Divide of Cloud Edge Data Centers," Proceedings of the ACM on Networking 1, no. 16 (November 2023), https://dl.acm.org/doi/10.1145/3629138.

from a patient or other clear, explicit consent-acquisition steps. Other AI-related datasets may require none, with legal systems permitting health and biopharma companies to generate synthetic data, not tied to any real person, using statistical and machine learning-driven methods—or to buy health and genetic data on the commercial market. Secondary uses of biospecimens raise their own questions about data consent, because patients may agree to have their biospecimen collected for a particular purpose at a particular point in time, but where the organization may keep that biospecimen for future use is presently unknown (e.g., it might be determined through later scientific developments).

- **Functional use**: Companies and research organizations may employ AI data components for various purposes, ranging from early-stage R&D to regulatory approval.
- **Population scope:** Health and biopharma data components in the AI supply chain may relate to different populations, including by geography and by demographic. Geographically, distinct AI-related data components in the health and biopharma sectors might differently encompass the world, a few regions, one specific region, one country, and so forth. Demographically, different AI-related data components for those same sectors might differently encompass various age groups, races, ethnicities, nationalities, sexes, genders, and so on. The potential for population-scope variation mostly refers to dataset components of the AI supply chain, such as one medical image training dataset covering adults in the Middle East and another covering only teenagers in Brazil, because the data itself can be geographically or demographically bounded. But it is possible for the likes of an AI model or an SDK to be geographically bounded, too, if its use or export is restricted under laws (discussed more later) or if it has been developed specific to a region or population (and thus not trained on other data or built for use elsewhere, even if it remains technically available to other regions or groups).
- **Degree of reidentifiability:** Components differ in their degree of reidentifiability—that is, how quickly data that is supposedly disconnected from an individual's identity can be tied back to that identity. Decades of computer science and statistics literature have shown that many data types that are allegedly "deidentified" or "anonymized" can be quickly linked back to specific people, with variations in ease of reidentifiability based on the specifics of the data type, the dataset, other available data, the specific kinds of masking applied to the dataset, and so forth.29 Nonetheless, there are some important differences among health and biopharma data types. Genetic and genomic data sit at the high end of risk—uniquely identifying, unchanging, and implicating biological relatives who may have never consented to data collection and analysis. Aggregated datasets, or those to which differential privacy or other masking techniques have been applied, may sit lower on the risk scale. However, these dataset states are not fixed, and computer science and statistical advances, the growth of available third-party data, and other factors can increase the reidentification risk to these datasets over time.
- **Provenance and auditability:** Components differ in whether their origin and handling can be traced and independently verified. A training set built from documented, consented clinical sources with a recorded chain of custody is a different governance object than one aggregated from data brokers, scraped repositories, or undisclosed partners—even when the two are interchangeable for training. The same holds for models and weights, some released with datasheets, evaluations, and license terms and others arriving as opaque artifacts. Because downstream developers inherit whatever provenance gaps exist upstream, this axis determines whether an organization can actually attest to how a component was built—a prerequisite for both compliance and security assurance.

29. See, e.g., Latanya Sweeney, "Risks to Patient Privacy: A Re-Identification of Patients in Maine and Vermont Statewide Hospital Data," Harvard Kennedy School, October 2018, https://www.hks.harvard.edu/publications/risks-patient-privacy-re-identification-patients-maine-and-vermont-statewide-hospital.

- **Revocability:** Components differ in whether they can be corrected, withdrawn, or deleted once in the supply chain. A database row can be expunged, while its influence on an already trained model's weights generally cannot (at least not cleanly). The inability to "revoke" certain data components in the AI context creates tension with rights such as the EU's General Data Protection Regulation (GDPR) erasure and with later-revoked consent, and it compounds as data propagates into testing sets, fine-tuning sets, and synthetic generators. Such a tension surfaces where a downstream obligation (e.g., delete this patient's data) cannot be met by an upstream component that has already learned from it.
- **Retention and lifecycle:** At any large health or biopharma organization, data necessary for patient health outcomes, internal recordkeeping, or ongoing research could be retained indefinitely (barring legal or regulatory requirements that indicate otherwise). Conversely, internal retention policies from a data security and cybersecurity perspective could require a company to delete data it is not using and has held for some period of years to minimize the risk of it getting breached (and the associated liability and other costs of such an incident). Companies may add to publicly accessible data components as there is more data available, such as by uploading more public training or testing data to an existing dataset, while they may choose to overwrite other components, rendering the prior versions inaccessible—as AI model vendors typically do when they update their models to new versions. These issues of retention and data lifecycle management can vary by data type, organization, and use case, with significant implications for individual privacy, cybersecurity, data provenance, and the ability to responsibly innovate.

Use Cases for Health and Biopharma AI-Related Data

While the above section breaks down the health and biopharma data components in the AI supply chain, this section details a sample of five use cases for AI-related data in those same sectors. It showcases some of the many reasons why AI-related health and biopharma data can create opportunities for societal benefit when handled appropriately, but can create risks to security and privacy when designed or handled poorly. The use cases also illustrate how health and biopharma sector entities can use AI-related data with varied use restrictions, consent types, and geographic and demographic population scopes. In that vein, they underscore the complications of attempting to neatly bucket health and biopharma AI-related data components—such as publicly accessible versus not publicly accessible, or private-sector versus not-for-profit—when there can be much overlap and interconnection among the various buckets of data component accessibility and organization structure.

These use cases span:

- **Isomorphic Labs:** A British company leveraging publicly available, US government-funded protein data to enhance predictive AI functions for drug design;
- **AI Medical Service (AIM):** A Japanese company aggregating image training data from more than one hundred globally dispersed research institutes to develop a cancer-diagnosing AI model, with a focus on Japan initially;
- **Basecamp Research:** A network of companies and research organizations that feeds into the UK-based company's AI-powering, proprietary database that it has used, among others, for frontier bio models that could enable AI-programmable therapeutics;
- **Tempus:** An American AI medicine and genomic testing data company building proprietary datasets across multiple types and clinical areas, and then offering the data as a service; and
- **BGI Genomics:** A Chinese genomics company that operates globally, conducting AI-related data activity that might otherwise be innocuous or standard risk, but due to its headquarters and reported work with the Chinese government there are potential risks to US national security.

Use Case Example #1: Predictive Drug Design

The UK company Isomorphic Labs has built what it calls the Isomorphic Labs Drug Design Engine, a computational drug-design model that builds on AlphaFold 3, Google DeepMind's protein structure-predicting model that was released in 2024.[30] Proteins have complex 3D structures, and the original AlphaFold showcased in 2020 the ability to predict these structures in minutes, rather than several years.[31] Doing so with speed and accuracy supports advancement in fields such as drug discovery and disease understanding.[32] The new model more than doubles the accuracy of AlphaFold 3 on a "challenging protein-ligand generalization benchmark."[33] An application programming interface (API) makes AlphaFold protein structure predictions available for free to the public.[34]

AlphaFold was trained on structures from the Protein Data Bank.[35] The Protein Data Bank is funded by the US National Science Foundation, Department of Energy, National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences and consists of 3D structure data for large biological molecules (protein, DNA, and RNA) essential for biology, health, energy, and biotech R&D.[36] Visitors to the Protein Data Bank website can use its built-in tools, such as an advanced search query builder, to filter the data by determination methodology,

30. "The Isomorphic Labs Drug Design Engine Unlocks a New Frontier Beyond AlphaFold," Isomorphic Labs, February 10, 2026, https://www.isomorphiclabs.com/articles/the-isomorphic-labs-drug-design-engine-unlocks-a-new-frontier; "AlphaFold 3 Predicts the Structure and Interactions of All of Life's Molecules," Isomorphic Labs, May 8, 2024, https://www.isomorphiclabs.com/articles/alphafold-3-predicts-the-structure-and-interactions-of-all-of-lifes-molecules.
31. "AlphaFold," Google DeepMind, accessed August 11, 2026, https://deepmind.google/science/alphafold/.
32. Marios G. Krokidis et al., "AlphaFold 3: An Overview of Applications and Performance Insights," International Journal of Molecular Sciences 26, no. 8 (April 2025): 3671, https://pmc.ncbi.nlm.nih.gov/articles/PMC12027460/.
33. "Accurate Predictions of Novel Biomolecular Interactions with IsoDDE," Isomorphic Labs, February 10, 2026, https://storage.googleapis.com/isomorphiclabs-website-public-artifacts/isodde_technical_report.pdf, 1.
34. "AlphaFold Protein Structure Database," EBI, accessed August 13, 2026, https://alphafold.ebi.ac.uk.
35. Ibid., 16.
36. "About RSCB PDB: A Living Digital Data Resource That Enables Scientific Breakthroughs Across The Biological Sciences," RSCB, accessed June 17, 2026, https://www.rcsb.org/pages/about-us/index.

scientific name of the source organism, taxonomy, input data, polymer entity type, and many more variables.[37]

Isomorphic Labs' goals for the model include predicting protein-binding affinity to rank and optimize potential molecules across diverse chemical series during drug design programs.[38] "Our dedicated drug design teams," the company says, "are using these capabilities every day across our programs—to understand unseen structures, identify uncharacterized pockets, and create novel chemical matter in the pursuit of new medicines for patients."[39] This use case captures how private-sector health or biopharma companies can leverage publicly accessible protein data and already released AI models to create newly robust AI models tailored for predictive functions and future drug designs, fusing tailored data with public and private capabilities.

Use Case Example #2: Cancer Diagnostic Support

Japanese company AI Medical Service, or AIM, built an AI model called "gastroAI-model G" to aid in gastric cancer diagnostics.[40] The model processes images gathered from an endoscopy system's video processor, scans the candidate lesions to detect whether a gastric lesion is a candidate for biopsy, and, as needed, alerts the physician of its findings and provides diagnostic assistance by overlaying a rectangle on the image.[41] Since April 2021, AIM has collaborated with the National University Hospital of Singapore on this effort.[42] AIM trained the model on lesion image data capturing early gastric cancer from over more than one hundred collaborative research institutes, which it describes as "world-class medical institutions."[43]

This use case illustrates how health and biopharma organizations can today leverage image-based health data for diagnostic purposes. It also represents an example of a private-sector company requiring approval from a government before debuting an AI model that makes diagnostic assessments. In this case, AIM required approval from the Japanese government because Japan regulates software functioning for a diagnostic or therapeutic purpose differently than software that does not make diagnoses or inform therapy treatments.[44] This case also shows how cross-border collaborations and cross-border sharing of AI training data can enable a company to offer a relevant health or biopharma AI model in one country, and only one country to start, but with the subsequent goal of global expansion.

Use Case Example #3: Cross-Border, AI-Powering, Proprietary Database

The UK-based AI company Basecamp Research runs an initiative called the Trillion Gene Atlas. Launched in March 2026 with Anthropic and US biotech companies Ultima Genomics and PacBio, the database currently includes more than ten billion novel genes from over one million new species compared to current public data.[45] Its partners directly help with data processing for the effort. For example, Ultima Genomics uses its next-generation sequencing systems to deliver "high-throughput, whole-genome, and multi-omics sequencing" at scale and lower cost for the Trillion Atlas initiative.[46] It also draws on a network of scientific collaborators across thirteen countries to establish "a scalable evolutionary genomics pipeline purpose-built for AI training."[47]

A January 2026 Basecamp Research paper described foundation models developed with this data, which the paper says can enable AI-programmable therapeutics through predictive and generative genomic and protein benchmarks.[48] It was coauthored with experts from NVIDIA, the University of Pennsylvania, Johns Hopkins University, the University of Oxford, the Centre for Genomic Regulation in Barcelona, and

37. "Advanced Search Query Builder," RCSB, accessed June 18, 2026, https://www.rcsb.org/search/advanced.
38. "The Isomorphic Labs Drug Design Engine."
39. "The Isomorphic Labs Drug Design Engine."
40. "AI Medical Service Inc. Announces Regulatory Approval of Gastric AI-based Endoscopic Diagnostic Imaging Support System," AI Medical Service, December 26, 2023, https://en.ai-ms.com/news/product/20231226.
41. "AI Medical Service Inc. Announces."
42. "AIM Applies for Approval to Manufacture and Market the World's First Gastric Cancer AI," AI Medical Service, August 31, 2021, https://en.ai-ms.com/news/product/20210831.
43. "AIM Applies for Approval."
44. Yurika Inoue, "Japan's Evolving AI and Digital Health Regulations: Legal Developments and Outlook," *International Bar Association*, December 3, 2025, https://www.ibanet.org/japan-ai-digital-health-regulations.
45. "Our Data," Basecamp Research, accessed June 18, 2026, https://basecamp-research.com/bcr-data/.
46. "Our Data."
47. Basecamp Research, "Basecamp Research Launches Trillion Gene Atlas to Scale AI-Designed Therapeutics," March 18, 2026, https://basecamp-research.com/wp-content/uploads/2026/03/BCR-TGA.pdf, 2.
48. Geraldene Munsamy et al., "Designing AI-Programmable Therapeutics with the EDEN Family of Foundation Models," *Basecamp Research*, January 2026, https://basecamp-research.com/wp-content/uploads/2026/01/BCR_Designing-programmable-therapeutics-with-the-EDEN-family-of-foundation-models.pdf, 1-2.

other companies and research centers.[49] The model, called EDEN, was built on a Llama3-style architecture and trained on up to 9.7 trillion nucleotide tokens from the company's database.[50] For the collection of samples that went into the database used for EDEN's training, the authors acknowledged contributions from numerous organizations around the world, including the British Antarctic Survey, Heritage Malta, CENIBiot in Costa Rica, Balaton Limnological Research Institute in Hungary, and AJESH International in Cameroon.[51]

This use case illustrates that a private-sector company in the health or biopharma sector can leverage global networks of both corporate and nonprofit organizations (e.g., university) to build a dataset that is ultimately proprietary. Organizations across sectors can collaborate on AI-related data components and initiatives in ways where neither datasets nor other data components of the AI supply chain, such as AI models themselves, necessarily cleanly fall into public- or private-sector buckets.

Use Case Example #4: Proprietary but Licensable, AI-Enabling Data

American AI medicine and genomic testing data company Tempus advertises several real-world datasets. It offers more than 8.5 million "deidentified research records" for scientific discovery, around four hundred thousand records with whole transcriptomic profiles, more than two million records with imaging data, more than 1.5 million records where matched clinical data is linked to genomic information, and a database the company says is one hundred times larger than the public Cancer Genome Atlas database.[52] The company advertises two mechanisms by which other organizations can purchase access to Tempus data:

- **Data collaborations:** "Work with experienced Tempus scientists and biostatisticians to perform a scoped analysis based on your research objectives and criteria. Receive end-to-end support designing and executing retrospective data analyses."
- **Advanced analytics:** "Leverage Tempus Lens to conduct preliminary analyses and identify cohorts of interest, then migrate your records to Workspaces to generate additional insights in our cloud-based coding environment. Layer on AI agents for faster data curation and abstraction."[53]

Tempus adds that it has data pipelines with over 4,500 US hospitals, "enabling us to ingest, structure, and harmonize multimodal data at an unprecedented scale." Its use cases for this kind of data-meets-AI span oncology, cardiology, neurology, radiology, pathology, and other areas.[54] The company says that 95 percent of the top twenty pharma oncology companies (based on their publicly available 2024 segment revenue) and over 250 biopharma companies use its services.[55]

This example illustrates that companies can build proprietary datasets that are themselves data components of the AI supply chain (e.g., potential training data) and can also be used to develop other components of the AI supply chain (e.g., testing data, AI models). It also illustrates that even though the databases themselves are proprietary—they are ostensibly unique to a company and privately held—the company can still make them available to others, for a price. Organizations building proprietary, AI-related health and biopharma datasets and making them available as a service could additionally represent the data as "deidentified," potentially intending to speak to consent requirements or privacy laws in various jurisdictions.

Use Case Example #5: Cross-Border National Security Risk

This fifth use case is quite different, intended to showcase the security risks of particular activities within the health or biopharma sectors.

BGI Genomics, a Chinese company, leverages health and genetic data to train AI models and produces other data components in the AI supply chain (such as AI models for health and biopharma functions). Its Genos model, for instance, is a ten-billion-parameter open-source genomic foundation model that is trained on 636 genomes from diverse populations.[56] BGI has also debuted an AI platform that ingests genomic, epigenetic, microbiome, imaging, and lifestyle data to "generate actionable insights for disease prevention and long-term health planning" shortly after it launched a clinical genomics

49. Munsamy et al., "Designing AI-programmable therapeutics," 1.
50. Munsamy et al., "Designing AI-programmable therapeutics," 9.
51. Munsamy et al., "Designing AI-programmable therapeutics," 36.
52. "Tempus Real-World Data," Tempus, accessed June 18, 2026, https://www.tempus.com/solutions/real-world-data/#overview.
53. "Tempus Real-World Data."
54. Ryan Fukushima, "Advancing the Frontier of AI in Healthcare," Tempus, October 9, 2025, https://www.tempus.com/content/article/advancing-the-frontier-of-ai-in-healthcare/.
55. "Tempus Real-World Data."
56. "BGI Group CEO Dr. Yin Ye: AI Is Redefining the Efficiency and Boundaries of Genetic Testing," BGI, March 17, 2026, https://en.genomics.cn/en-news-7477.html.

laboratory in Riyadh.[57] "As [genetic] sequencing becomes more accessible, laboratories and clinicians face an explosion of data, while the capacity to interpret results with consistency and clinical relevance has not scaled at the same pace," the company has written.[58] And, as the company paraphrased its CEO saying, "the next phase of progress depends on moving from generating data to understanding it—quickly, reproducibly, and in a way that can support real clinical and public health workflows."[59]

There is a US national security risk dimension to these activities. The US National Counterintelligence and Security Center (NCSC) has warned that while the bioeconomy can yield many benefits to the food supply, healthcare, environmental quality, and other important areas, actors such as foreign governments can use genomic data and technology to "identify genetic vulnerabilities in a population" and surveil people.[60] In particular, the agency has stated that Chinese firms are collecting genetic data globally in an effort to develop the largest bio-database in the world.[61] For example, the US Department of Defense updated its formal list of entities identified as Chinese military companies (the so-called 1260H list) in October 2022 to add BGI Genomics.[62] BGI Group is affiliated, per a June 2026 update to the list, with China's People's Liberation Army (PLA) and "is a military-civil fusion contributor to the Chinese defense industrial base."[63] A bipartisan group of members of Congress have worried that BGI's access to the US market could provide it with genetic data that it can use to then undermine US national security.[64] For example, there are concerns that such data could be used to fuel the development of bioweapons or in the service of next-generation capabilities whose details and subsequent risks have yet to be realized.

This case is relevant for several reasons. BGI appears to be conducting activity that many companies and other organizations around the world conduct: collecting genetic data, using genomic insights to predict health outcomes, building AI models using health and genomic data, and more. It does so globally, as do many other firms, as evinced by its recent AI and AI-related data work in Saudi Arabia (among other places). But BGI is also a company headquartered in China, which lacks strong rule of law, meaning that Chinese security agencies can easily compel BGI to hand over data of interest.[65] Further, BGI reportedly cooperates with China's PLA on Chinese defense and national security objectives. These phrases ostensibly refer to activities related to biosecurity or bioweaponry. Combined, this raises the risk that what might otherwise be innocuous or standard-risk activity—a company collecting AI-related health and genomic data, transferring it across borders, potentially exposing it to breaches, and the like—becomes a risk to US national security due to the Chinese government acquiring and using American or other genetic data for Chinese security ends.

57. "BGI Genomics Taps AI to Power Healthcare Transformation in Saudi Vision," BGI, September 21, 2025, https://www.bgi.com/global/news/BGI%20Genomics%20Taps%20AI%20to%20Power%20Healthcare%20Transformation%20in%20Saudi%20Vision.
58. "AI Is Redefining the Efficiency and Boundaries."
59. "AI Is Redefining the Efficiency and Boundaries."
60. *Protecting Critical and Emerging U.S. Technologies from Foreign Threats*, National Counterintelligence and Security Center, October 2021, https://permanent.fdlp.gov/gpo173109/FINALNCSCEmergingTechnologiesFactsheet10222021.pdf, 6.
61. Julian E. Barnes, "U.S. Warns of Efforts by China to Collect Genetic Data," *New York Times*, October 22, 2021, https://www.nytimes.com/2021/10/22/us/politics/china-genetic-data-collection.html.
62. US Department of Defense. *Entities Identified as Chinese Military Companies Operating in the United States in Accordance with Section 1260H of the William M. ("Mac") Thornberry National Defense Authorization Act for Fiscal Year 2021 (Public Law 116-283).* Arlington: Department of Defense, October 2022. https://media.defense.gov/2022/Oct/05/2003091659/-1/-1/0/1260h companies.pdf, 1.
63. US Department of Defense. *Entities Identified as Chinese Military Companies Operating in the United States in Accordance with Section 1260H of the William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021 (Public Law 116-283, Section 1260H, as amended)* (Arlington: Department of Defense, June 2026. https://media.defense.gov/2026/Jun/08/2003945537/-1/-1/1/entities-identified-as-chinese-military-companies-operatiung-in-the-united-states-in-accordance-with-section-1260h.pdf, 3.
64. Ken Dilanian, "Congress Wants to Ban China's Largest Genomics Firm from Doing Business in the US. Here's Why," *NBC News*, January 25, 2024, https://www.nbcnews.com/politics/national-security/congress-wants-ban-china-genomics-firm-bgi-from-us-rcna135698.
65. Beyond many press stories to this effect, the US government has publicized this risk with respect to health and genetic data in particular; US National Counterintelligence and Security Center. *China's Collection of Genomic and Other Healthcare Data from America: Risks to Privacy and U.S. Economic and National Security*, National Counterintelligence and Security Center, February 2021, https://www.dni.gov/files/NCSC/documents/SafeguardingOurFuture/NCSC_China_Genomics_Fact_Sheet_2021revision20210203.pdf.

The Fractured Global Regulatory Landscape

Countries around the world are debating how to apply existing regulations to AI-related data and AI technologies themselves or pass new laws to handle their transparency, innovation, fairness, security, privacy, and other implications. Other countries have already made decisions about how to apply existing regulations or have already passed new laws for similar purposes.

Some of these laws and regulations are AI-specific, including the EU AI Act and China's rules for public-facing algorithmic and generative AI services. China's AI-specific rules operate alongside broader personal information protection, data security, cybersecurity, health-sector, and human genomics resources requirements. A single health-related AI activity may therefore trigger several of China's regimes at once.[66] In other cases, the laws and regulations in question are more general but can apply to AI data components and AI use cases. For example, the US Health Insurance Portability and Accountability Act (HIPAA), the US Federal Trade Commission's (FTC) unfair and deceptive business acts and practices (Section 5 of the FTC Act) authority,[67] and the Food and Drug Administration's (FDA) regulation of medical devices writ large can all regulate AI-related data components in the health and biopharma sectors, even though they are not focused solely on AI or are not focused on it at all.[68] And sometimes, countries and blocs are using laws in both buckets. Most prominently, the EU's GDPR applies to AI-related data as one part of a broader tapestry of data protection requirements, while the European Union has also passed and continues to amend its AI Act to specifically regulate AI systems themselves.

Researchers, companies, and even government agencies continue to develop privacy-enhancing technologies to operate amid these regulations. Federated learning is a decentralized approach to machine learning where individual devices keep personal data on the device and collectively work to train a centralized model.[69] Differential privacy is a mathematical definition of privacy, whereby adding a person's data into a dataset or taking it out of that dataset does not change the overall dataset's behavior—ensuring that individual-level information about participants is not leaked.[70] Synthetic data generation aims to produce datasets that mirror real-world data so they are useful for tasks like healthcare model training, but without containing real personal data per se. And homomorphic encryption, to give one more example, allows organizations to perform computation on encrypted data.[71] Not all of these measures would necessarily bring a company into compliance with all the regulations in question, but they are technological developments for data privacy, obfuscation, and security that continue to evolve in tandem with complex regulations.

For their part, health and biopharma sector organizations collecting AI-related data and developing, deploying, procuring, using, testing, and maintaining AI systems must contend with this fractured global regulatory landscape. Some requirements are similar across regimes. For example, India's Digital Personal Data Protection Act, like some other national-level privacy regimes, is loosely modeled after the EU GDPR. Health and biopharma companies operating under the EU GDPR may already be providing rights to consumers, such as the right to access their data, that also ensure compliance with the Indian law. In other cases, countries' requirements for health and biopharma AI-related data components are significantly different or even contradictory. The United States and China, for instance, have both implemented restrictions on the outbound transfer on genetic and genomic data for national security reasons, which are intended to limit or outright prohibit the transfer of such data from one to the other.

Table 2 below shows fourteen exemplative regulatory regimes around the world that govern or substantially implicate health and biopharma data components of the AI supply chain, from genetic training data to medical image testing data to biotech AI models. The table is not meant to be comprehensive and certainly does not cover every law and regulation in question, let alone all the nuances, relevant case law, potential legal

66. Kenton Thibaut, *Balancing Openness and Control: Cross-Border Health Data and AI Governance in China*, *Atlantic Council,* June 23, 2026, https://www.atlanticcouncil.org/in-depth-research-reports/report/balancing-openness-and-control-cross-border-health-data-and-ai-governance-in-china/.
67. See, e.g., US Federal Trade Commission, "FTC Announces Crackdown on Deceptive AI Claims and Schemes," September 25, 2024, https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes.
68. See, e.g., "Artificial Intelligence-Enabled Medical Devices," US Food and Drug Administration, updated June 16, 2026, https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices.
69. See, e.g., "What is Federated Learning?," IBM, accessed August 24, 2022, https://research.ibm.com/blog/what-is-federated-learning.
70. See, e.g., "Differential Privacy," *Harvard University Privacy Tools Project,* accessed August 13, 2022, https://privacytools.seas.harvard.edu/differential-privacy.
71. See, e.g., "What is Homomorphic Encryption?," IBM, accessed August 13, 2026, https://www.ibm.com/think/topics/homomorphic-encryption.

gaps, and so forth. However, it does capture the variety of laws and regulations that impact health and biopharma data components of the AI supply chain. In doing so, it shows how different countries approach these laws and regulations in different ways (such as regulating AI as its own specific issue area versus applying non-AI-specific laws and regulations to AI technologies) and how those decisions create a global patchwork.

The table specifies, for each of the fourteen exemplative regimes, the regime name; whether the regime is AI-specific or broad (but encompassing AI-related data components); whether the regime applies to a region, to a country, or to multiple countries in a cross-border fashion; a summary of the regime; the data components of the AI supply chain that are clearly covered by the regime; and a short description of the regime's implications for the health and biotech sectors.

Table 2: Exemplative Data- and AI-Related Laws and Regulations Impacting the Health and Biopharma Data Components of the AI Supply Chain

Exemplative regime	AI-specific or not?	Summary (High-Level)	Covered AI data components[72]	Health and biotech implications
US Health Insurance Portability and Accountability Act (HIPAA)	Non-AI-specific	Regulates hospitals, health insurers, healthcare clearinghouses, and their business associates. Some privacy and security rules, such as rules concerning disclosure of covered health data that is not (a) subject to certain "deidentification" standards or (b) authorized for disclosure in writing by the patient.[73]	Training data Testing data Models (themselves) Model architectures APIs SDKs	Sets privacy and security rules for how HIPAA-covered entities should handle electronic medical records and other data, whether for AI purposes or not. Prohibits many data uses, transfers, training cases, and such without sufficient consent. Does not cover numerous commercial health and biotech use cases, such as by plenty of biotech startups, pharma companies, and AI model vendors.
US FTC Health Breach Notification Rule	Non-AI-specific	Requires personal health record vendors and related entities to notify consumers, or service providers as applicable, of a breach of unsecured data.[74]	Training data Testing data	Mandates that any organization handling covered health or biopharma data elements of the AI supply chain disclose covered breaches. Covers many more entities than HIPAA. Does not regulate transfer of data.
US FTC Section 5	Non-AI-specific	Makes unlawful any US company's unfair or deceptive business acts or practices.	Training data Testing data Models (themselves) Model architectures APIs SDKs	Prohibits deceptive statements about data collection, use, etc., API and SDK privacy and security, and AI model behavior. Prohibits unfair collection, use, etc. of data and use of AI models.

72. For each regulatory regime, this table column lists the data components of the AI supply chain that are clearly covered by the regime. The column entry in question for HIPAA, for example, does not list model weights because they do not as clearly fall under HIPAA, even though someone could argue for it. Some regulatory regimes also govern data categories that fall outside the report's seven-part framework. China's AI and data rules, for example, regulate prompts and users inputs, usage logs, data annotations, generated outputs, and provenance metadata. The table discusses these categories in the "Summary" and "Health and biotech implications" columns but does not list them as separate AI data components.

73. See, e.g., "Summary of the HIPAA Privacy Rule," Health and Human Services, accessed June 21, 2026, https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.

74. "Health Breach Notification Rule," US Federal Trade Commission, accessed June 21, 2026, https://www.ftc.gov/legal-library/browse/rules/health-breach-notification-rule.

Exemplative regime	AI-specific or not?	Summary (High-Level)	Covered AI data components[72]	Health and biotech implications
US FDA medical device regulations	Non-AI-specific	Regulate firms that manufacture, repackage, relabel, and/or import medical devices sold in the United States, as well as radiation-emitting electronic products like lasers and X-ray systems.[75] Regulates certain software functions intended for use in the diagnosis of disease or other conditions, or the cure, mitigation, treatment, or prevention of disease.[76] The FDA also maintains a list of AI-enabled medical devices authorized for marketing in the country.[77]	Models (themselves)* *to the extent a model might be, for example, used in a covered medical device or constitute a covered software function	Governs certain AI models themselves when used in covered ways, such as in a covered medical device or as a covered software function.
US Protecting Americans' Data from Foreign Adversaries Act (PADFAA)	Non-AI-specific	Makes it unlawful for a US third-party data broker to sell, license, rent, trade, transfer, release, disclose, provide access to, or otherwise make available a US person's covered data to any "foreign adversary" country—North Korea, China, Russia, or Iran.	Training data Testing data	Prevents entities with covered health data—describing or revealing "past, present, or future physical health, mental health, disability, diagnosis, or healthcare" conditions or treatments[78]—that they did not collect directly from the consumers in question, from providing it to entities in the adversary countries. Likely impacts a subset of health and biopharma sectors and may limit some access to US health training data from subsidiaries in, say, China.

75. "Overview of Device Regulation," US Food and Drug Administration, February 2, 2026, https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/overview-device-regulation.
76. "Mobile Health App Interactive Tool," US Federal Trade Commission, accessed June 21, 2026, https://www.ftc.gov/business-guidance/resources/mobile-health-apps-interactive-tool.
77. "Artificial Intelligence-Enabled Medical Devices," US Food and Drug Administration, June 16, 2026, https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices.
78. Making Emergency Supplemental Appropriations for the Fiscal Year Ending September 30, 2024, and for Other Purposes, Pub.L.118-50, 138 Stat. 895 (April 24, 2024), https://www.govinfo.gov/content/pkg/PLAW-118publ50/html/PLAW-118publ50.htm.

Exemplative regime	AI-specific or not?	Summary (High-Level)	Covered AI data components[72]	Health and biotech implications
US Department of Justice (DOJ) Data Security Program (DSP)	Non-AI-specific	Restricts first- and third-party brokerage of bulk "sensitive personal data" of US persons as well as data tied explicitly to US government personnel, when dealing with China, Russia, Iran, North Korea, Cuba, or Venezuela. Prohibits transfer to those countries of bulk 'omic[79] data or of human biospecimens from which bulk 'omic data could be derived. Imposes security requirements on transfers of non-human 'omic, such as personal health, data under vendor, employment, or investment agreements.	Training data Testing data	Bulk data restrictions cover human genomic data on >100 persons, human 'omic data (genomic, epigenomic, proteomic, or transcriptomic data) on >1,000 US persons, and personal health data on >10,000 persons.[80] Restrictions on transfers as part of vendor, employment, and investment agreements apply to genomic and personal health data.
US-EU Data Privacy Framework	Non-AI-specific	Defines how US organizations can transfer personal data from the European Union, United Kingdom (and Gibraltar), and Switzerland to the United States while meeting the "adequacy" standard for equivalent protections under the privacy regimes of the bloc's and those countries'.[81]	Testing data Training data	Governs how US health and biopharma companies must transfer EU, UK, or Swiss citizen data to the United States. It is unclear how it impacts AI models trained on EU, UK, or Swiss citizen data, including their transfer out of the bloc.[82]
EU GDPR	Non-AI-specific	Governs the processing and protection of EU personal data. Establishes the right to be forgotten, the right to access one's data held by a data collector, and other personal rights.	Testing data Training data	Sets collection, processing, use, and other requirements for health and biopharma companies—and establishes personal rights that health and biopharma companies must respect—for data related to a person's past, current, or future physical or mental health status.[83] It is unclear how it impacts AI models trained on EU, UK, or Swiss citizen data, including their transfer out of the bloc. Other privacy measures, such as the right to be forgotten and data minimization requirements, inform how covered health and biopharma companies must handle patient and personal data.

79. 'omic is a term of art under the regulations.
80. Preventing Access to US Sensitive Personal Data and Government-Related Data by Countries of Concern or Covered Persons, 28 CFR Part 202 (Jan. 8, 2025), https://www.ecfr.gov/current/title-28/chapter-I/part-202.
81. "Data Privacy Framework (DPF) Program Overview," US International Trade Administration, accessed June 21, 2026, https://www.dataprivacyframework.gov/Program-Overview.
82. On this last point, see the discussion below of the December 2024 European Data Protection Board decision.
83. "Recital 35: Health Data," General Data Protection Regulation, https://gdpr-info.eu/recitals/no-35/.

Exemplative regime	AI-specific or not?	Summary (High-Level)	Covered AI data components[72]	Health and biotech implications
EU AI Act	AI-specific	Regulates developers and users of AI systems in the European Union as well as any AI developers operating in the EU market. Regulates AI uses in four risk tiers: unacceptable risk (banned), high risk (compliance requirements, such as assessments), transparency risk (disclosure obligations), and minimal or no risk. Phased enforcement. Ongoing amendments.	AI models (themselves)	Governs as "high risk" AI systems those for biometrics, to determine people's eligibility for public healthcare services, and for emergency healthcare patient triage[84] as well as AI systems tied to medical device and in vitro diagnostics already covered by EU law.[85] Exempts from the bans in the "unacceptable" risk tier any emotion recognition systems that are specifically intended for medical reasons.[86] References AI model weights, such as to determine an AI model's "size" based on parameter values, but does not regulate them per se. Indirectly covers health and biopharma firms' AI-related APIs and SDKs to the extent the AI models themselves are covered.
China Personal Information Protection Law (PIPL) and Cyberspace Administration of China (CAC) cross-border personal information rules[87]	Non-AI-specific	Governs the processing and overseas provision of personal information. Medical and health information and biometric information qualify as sensitive personal information. Processing sensitive personal information and providing personal information abroad can require a lawful basis, heightened safeguards, a personal information protection impact assessment, and separate consent.[88] Unless an exemption applies, an exporter must use a CAC security assessment, personal information protection certification, or standard contract.[89]	Training data Testing data	For a non-critical information infrastructure operator (CIIO) handler, exporting sensitive personal information concerning fewer than ten thousand individuals generally requires a standard contract or certification; exports concerning at least ten thousand individuals require a CAC security assessment. Counts accrue from January 1 of the current year. CIIO exports of personal information and any export of important data require an assessment. Human genetic resources (HGR) and sector-specific health rules may apply in parallel.

84. "High-Level Summary of the AI Act," EU Artificial Intelligence Act, February 27, 2024, https://artificialintelligenceact.eu/high-level-summary/.

85. "Annex I: List of Union Harmonisation Legislation," EU Artificial Intelligence Act, accessed June 21, 2026, https://artificialintelligenceact.eu/annex/1/.

86. "Article 5: Prohibited AI practices," European Commission, https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-5. Accessed June 21, 2026.

87. See important prior work on this subject, from which these table entries draw, at: Kenton Thibaut, "Balancing openness and control: Cross-border health data and AI governance in China," Atlantic Council, June 23, 2026, https://www.atlanticcouncil.org/in-depth-research-reports/report/balancing-openness-and-control-cross-border-health-data-and-ai-governance-in-china/.

88. Zhōnghuá Rénmín Gònghéguó Gèrén Xìnxī Bǎohù Fǎ (中华人民共和国个人信息保护法) [Personal Information Protection Law of the People's Republic of China] (adopted by 13th National People's Congress, August 20, 2021), Articles 38–39, Cyberspace Administration of China, https://www.cac.gov.cn/2021-08/20/c_1631050028355286.htm.

89. Cùjìn Hé Guīfàn Shùjù Kuà Jìng Liúdòng Guīdìng (促进和规范数据跨境流动规定) Guójiā hùliánwǎng xìnxī bàngōngshì lìng dì 16 hào (国家互联网信息办公室令 第16号) [Regulations on Promoting and Regulating Cross-Border Data Flow, Order from the State Internet Information Office No. 16] (promulgated and effective March 22, 2024), Cyberspace Administration of China, https://www.cac.gov.cn/2024-03/22/c_1712776611775634.htm; https://www.cac.gov.cn/2025-10/17/c_1762449728720008.htm.

Exemplative regime	AI-specific or not?	Summary (High-Level)	Covered AI data components[72]	Health and biotech implications
China Data Security Law (DSL), Cybersecurity Law (CSL), and important-data and CIIO rules	Non-AI-specific	Establish a security-based regulatory layer for data with implications for national security, public health, economic security, or major public interests. The DSL establishes a classification and grading framework;[90] national standard GB/T 43697-2024 formalizes the categories of general, important, and core data.[91] Important data and personal information or important data held by CIIOs face heightened protection and outbound security-assessment requirements.[92]	Training data Testing data	Large-scale, aggregated, genomic, public health, multimodal, or strategically significant health and AI-training datasets may attract important data scrutiny. Health data is not automatically important data. For export-assessment purposes, a handler generally need not treat data as important unless a regulator has notified the handler or publicly identified the data as important. Some major hospital systems, national health platforms, or public health systems may qualify as CIIOs.
China's public-facing generative AI and algorithm-governance rules	AI-specific	China regulates public-facing AI through a set of service-specific rules, including the Algorithm Recommendation Provisions, Deep Synthesis Provisions, Generative AI Interim Measures,[93] and AI-Generated Content Labeling Measures.[94] For covered services, these rules impose requirements concerning lawful training-data sourcing, personal information protection, data and annotation quality, content safety, security assessment or filing (where applicable), and synthetic-content labeling. The Generative AI Interim Measures generally do not cover internal R&D or non-public deployment.	Training data Models (themselves) APIs* SDKs* *when used to provide a covered, public-facing service	Public-facing health AI providers must protect user inputs and usage records, avoid collecting unnecessary personal information, and comply with PIPL when training data, prompts, logs, or outputs contain identifiable health information. AI-generated health guidance or diagnostic content may require visible and embedded labels. Internal drug discovery and non-public clinical R&D remain subject to the general data, HGR, health sector, and ethics regimes.

90. "Data Security Law of the PRC," China Law Translate, June 10, 2021, https://www.chinalawtranslate.com/en/datasecuritylaw/.
91. "Data Security Technology Data Classification and Grading Rules", National Public Service Platform for Standard Information, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F0C385EDC38CBF277AEC021F23126ADE&.
92. "Regulations on the Administration of Network Data Security", Office of the Central Cyberspace Affairs Commission, September 30, 2024, https://www.cac.gov.cn/2024-09/30/c_1729384452307680.htm.
93. See: Shēngchéng Shì Réngōng Zhìnéng Fúwù Guǎnlǐ Zhànxíng Bànfǎ (生成式人工智能服务管理暂行办法) lìng dì 15 hào (令第15号) [Interim Measures for the Administration of Generative Artificial Intelligence Services, Order No. 15] (promulgated July 10, 2023, effective August 15, 2023), Cyberspace Administration of China, https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.
94. "Notice on Issuing the "Measures for Identifying Artificial Intelligence-Generated and Synthetic Content"", Office of the Central Cyberspace Affairs Commission, March 14, 2025, https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm.

Exemplative regime	AI-specific or not?	Summary (High-Level)	Covered AI data components[72]	Health and biotech implications
China Human Genetic Resources (HGR) Regulation	Non-AI-specific	Regulates Chinese HGR materials and human gene or genome information generated from those materials. Foreign organizations, individuals, and institutions they establish or actually control cannot independently collect or preserve Chinese HGR and generally must work with a Chinese institution for covered international research. Approval, filing, external-provision reporting, information-backup, and other requirements may apply according to the activity. Clinical, imaging, protein, and metabolic data fall outside HGR information unless they contain covered human gene or genome information.[95]	Training data Testing data	Genomic datasets used for AI training or testing may trigger Chinese-partner, consent, ethics-review, approval, filing, or external-provision requirements. HGR compliance does not satisfy separate PIPL, CAC transfer, DSL/CSL, or biomedical ethics obligations. A May 2026 proposal would narrow some HGR-specific triggers and make institutional ethics review more central for lower-risk activity, but it currently remains a draft.[96]
India Digital Personal Data Protection Act	Non-AI-specific	Governs private (nongovernment) entities collecting, handling, and transferring, etc. Indians' personal data. Modeled roughly on EU GDPR.	Training data Testing data	Requires health and biopharma companies handling any covered data on Indians—"about an individual who is identifiable by or in relation to such data"—to provide access and other rights to Indians, implement certain data protections, and so forth.

95. See: Kēxué Jìshù bù líng dì 21 Hào Rénlèi Yíchuán Zīyuán Guǎnlǐ Tiáolì Shíshī Xìzé" (科学技术部令第21号 人类遗传资源管理条例实施细则) [Order No. 21 of the Ministry of Science and Technology: Detailed Rules for the Implementation of the Regulations on the Management of Human Genetic Resources] (promulgated by the Ministry of Science and Technology of the People's Republic of China, May 26, 2023, effective date July 1, 2023), Ministry of Science and Technology of the People's Republic of China, 306-10-2023-221, https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/bmgz/202306/t20230601_186416.html (China).; see also Todd Liao, Mudan He, and Sylvia Hu, "Strategic Compliance: Navigating HGR and Data Risks in China Life Sciences Transactions," Morgan Lewis, May 15, 2026, https://www.morganlewis.com/pubs/2026/05/strategic-compliance-navigating-hgr-and-data-risks-in-china-life-sciences-transactions.

96. "Guānyú rénlèi yíchuán zīyuánguǎnlǐ tiáolì shíshīxìzé (zhēngqiúyìjiàn gǎo) gōngkāi zhēngqiúyìjiàn de gōnggào" 关于人类遗传资源管理条例实施细则 (征求意见稿) 公开征求意见的公告 [Public Notice on Soliciting Opinions on the Draft Implementing Rules of the Regulations on the Management of Human Genetic Resources], National Health Commission of the People's Republic of China, May 8, 2026, https://www.nhc.gov.cn/wjw/yjzj/202605/9b38dffa8d8047ea94813f40aca8d282.shtml.

China, the European Union, and the United States have regulatory regimes that stack on top of one another. The Chinese case provides an illustrative example. A health dataset may qualify as sensitive personal information under China's Personal Information Protection Law (PIPL) and as important data under its Data Security Law (DSL) and Cybersecurity Law (CSL) framework, while genomic content may also trigger the HGR and hospital-held information may face sector-specific storage or cybersecurity rules. Compliance with one pathway does not displace the others. A clinical dataset excluded from HGR information, for example, may still require PIPL consent, a personal information protection impact assessment, and a Cyberspace Administration of China (CAC) transfer mechanism.

China's health-sector rules add a more targeted and granular regulatory layer. The 2014 Population Health Information Measures prohibit covered population health information from being stored on overseas servers.[97] The 2018 Health and Medical Big Data Measures and 2022 healthcare cybersecurity rules (and the 2026 update) impose additional storage, lifecycle, access-control, and security duties.[98] The coverage of these measures depends on the institution, dataset, and processing context.

China has also eased some lower-risk transfer procedures. The 2024 CAC provisions raised numerical thresholds, expanded exemptions, extended the validity of successful security assessments, and authorized free-trade-zone negative lists.[99] The resulting system reflects managed openness: lower-risk transfers may proceed through defined channels, while important data, data held by any organization legally classified as a critical information infrastructure operator (CIIO), large volumes of sensitive personal information, and covered HGR, remain subject to closer state review.

The terms in the laws and regulations themselves are also complex. Any organization complying with them will contend with both straightforward definitions and those that are opaque or vague, leaving much more up for interpretation. This, in turn, contributes to variations in regulatory enforcement and the need for health and biopharma companies to understand their regulators' current and future enforcement postures. In the United States, for example, the FTC regulates unfair or deceptive business acts or practices under its Section 5 authority. The Commission defines "deceptive" practices as practices involving a material representation, omission, or practice likely to mislead a reasonable consumer.[100] This definition is quite straightforward: companies wishing to avoid a deceptive practices claim should tell the truth to consumers.

By contrast, the Commission has long defined "unfair" practices as those that cause a substantial injury, which is not outweighed by any offsetting consumer or competitive benefits that the practice also yields, and which consumers could not reasonably have avoided.[101] On its face, the definition is easy to follow. However, the FTC has fluctuated widely in recent years in its approach to enforcing "unfair" data or AI practices under Section 5 of the FTC Act. It has recently argued, breaking with some previous Commission postures, that certain uses of particularly sensitive personal data are in and of themselves harmful enough to make them unfair—such as companies selling identifiable data about individuals' 24-7 movements, whether or not it is clear that the recipients of the data subsequently took any identifiable action with that data.[102] There are plenty of reasons why such disclosures in and of themselves cause harm.[103] However, debate persists as to whether a category of behavior (such as selling certain kinds of data) should be considered "unfair" underscores the variation in how future FTC enforcers could interpret this definition, and thus scrutinize health or biopharma company handling of AI-related data components, compared to the more obvious definition of a "deceptive" practice.

Most of these laws put controls on how health or biopharma companies and organizations could collect, analyze, transfer, and use any AI-related data components, such as training data composed of people's genetic information. Most of the above-listed regimes do not specify what happens when a company transfers an AI model out of the country that was trained on

97. https://en.nhc.gov.cn/2014-06/15/c_46801.htm

98. https://www.nhc.gov.cn/mohwsbwstjxxzx/s8553/201809/742308fad8ed47fb85b9675ee4b6c6eb.html; see "China's Hospitals Get Their Own Data Rulebook: Reading the 2026 Healthcare Data Security & PI Measures," Data Compliance China, June 4, 2026, https://datacompliancechina.com/posts/china-healthcare-data-rulebook-2026/

99. Cùjìn Hé Guīfàn Shùjù Kuà Jìng Liúdòng Guīdìng (促进和规范数据跨境流动规定) Guójiā hùliánwǎng xìnxī bàngōngshì lìng dì 16 hào (国家互联网信息办公室令 第16号) [Regulations on Promoting and Regulating Cross-Border Data Flow, Order from the State Internet Information Office No. 16] (promulgated and effective March 22, 2024), Cyberspace Administration of China, https://www.cac.gov.cn/2024-03/22/c_1712776611775634.htm.

100. "FTC Policy Statement on Deception," US Federal Trade Commission, October 1983, https://www.ftc.gov/system/files/documents/public_statements/410531/831014deceptionstmt.pdf, 1.

101. "FTC Policy Statement on Unfairness," US Federal Trade Commission, December 17, 1980, https://www.ftc.gov/legal-library/browse/ftc-policy-statement-unfairness.

102. "FTC to Ban Kochava and Subsidiary from Selling Sensitive Location Data to Settle Charges They Sold Location Data Linked to Millions of Mobile Devices," US Federal Trade Commission, May 4, 2026, https://www.ftc.gov/news-events/news/press-releases/2026/05/ftc-ban-kochava-subsidiary-selling-sensitive-location-data-settle-charges-they-sold-location-data.

103. For whatever it is worth, I agree with the FTC's argument in the cited Kochava case.

data that the country's laws and regulations *do* cover. For example, India's Digital Personal Data Protection Act does not clearly describe this issue. The US Data Security Program does not clearly describe this issue, either—perhaps an area for a future advisory opinion from the Justice Department. The European Union presents a thornier situation, given the overlap of the GDPR) with the AI Act, and the related complexities of the EU-US Data Privacy Framework.

EU data privacy and security regulations do not clearly specify whether biopharma companies that train AI models on EU data can or cannot transfer the models themselves out of the European Union, without implicating the GDPR or other requirements. For example, in December 2024 the European Data Protection Board (EDPB) issued Opinion 28/2024 on the processing of personal data in the context of AI models. The EDPB said that "claims of an AI model's anonymity should be assessed by competent [Supervisory Authorities] on a case-by-case basis, since the EDPB considers that AI models trained with personal data cannot, in all cases, be considered anonymous."[104] Essentially, it depends on the specific case whether or not an American, Japanese, or other organization could train an AI model on EU citizen data and transfer the resulting model out of the European Union without triggering GDPR obligations on the AI model itself. Doing so creates several complications.

On the one hand, there are plenty of reasons to take a case-by-case regulatory approach to organizations transferring AI models trained on EU citizen data out of the European Union. A health organization could use two similar health datasets to train two different AI models, making one a significant source of privacy concern (thus subject to some outbound-transfer restriction under GDPR) and the other not (thus unrestricted) due to variations in the model architectures. To give another example, two biopharma companies could train similar-functionality AI models on datasets with different levels of underlying data identifiability, likewise making the risks of subsequent training data leakage or model reverse engineering outside of the European Union much different.

However, making the rules vague and so case-specific creates uncertainty for health and biopharma companies seeking to plan for how to continue conducting clinical trials, vaccine development, and other such functions in an era of AI usage. This can lead to variability among each EU member state's enforcement, meaning that a health or biopharma organization that collects EU citizen data from multiple member states under a single GDPR framework could subsequently be subject to varied post-collection enforcement by country on the resulting model's transfer. Such enforcement variation can create a risk to privacy insofar as member states whose privacy enforcement bodies are particularly Big Tech-friendly could be incentivized to lower the floor of outbound transfer protections in cases where protections should be robust. It also leaves open-ended the question of what level of protection on the underlying data and the model itself is sufficient to mitigate risks of data leakage or reverse-engineering—because the risk cannot be eliminated entirely.

China provides one partial clarification relevant to this issue. CAC guidance treats data stored in China as exported when a person or organization accesses, queries, retrieves, downloads, or calls that data from outside China.[105] The location of the access or calling activity controls this determination. However, whether that rule reaches overseas use of an AI model trained on protected Chinese data remains unresolved—the answer would likely depend on whether the model, its outputs, or the access process provides or reveals covered personal information or important data.

Most major data privacy and security regimes, as well as emerging AI regulatory regimes, have some explicit coverage of data or AI models relevant to health and biopharma. Within these regulatory frameworks, countries commonly define personal health data as any data describing someone's past, present, and future conditions, treatments, and so on. Governments, even in countries with otherwise conflicting perspectives on tech regulation, frequently treat genetic data and genomic data as among the most sensitive, if not the most sensitive, data types. And many AI-specific and non-AI-specific measures cover as a sensitive category the use of AI models for health- and biopharma-related purposes, whether for public agency monitoring of disease outbreaks or private determination of diagnoses or treatments. The above-cited and other laws and regulations have varied terms, frameworks, restrictions, exemptions, and incentives for health and biopharma AI-related data components and use cases. But these treatments of health and biopharma data and AI technologies as important and sensitive are relatively consistent.

Lastly, as mentioned, some of these regimes represent clashing perspectives. EU politicians are increasingly discussing how to create regulations that hive off the European technology and data sphere from the United States and create a relatively autonomous European technology stack. Meanwhile, the current US federal government perspective, overall, is to remove and curtail enforcement of technology- and data-related regulations, with the goal of accelerating AI research and development—even if rolling back regulations creates substantial privacy, cybersecurity, and other risks to Americans, companies, and the country. The United States has implemented

104. "Opinion of the Board (Art. 64): Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models," European Data Protection Board, December 2024, https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf, 2.

105. "Q&A on Data Cross-Border Security Management Policies (October 2025)", Office of the Central Cyberspace Affairs Commission, October 31, 2025, https://www.cac.gov.cn/2025-10/31/c_1763633376984070.htm.

cross-border data transfer regulations for national security perspectives, especially focused on curtailing transfers of genetic and genomic data to Chinese organizations and significantly limiting transfers of personal health data to them.

China applies its own controls to outbound health and genetic data, while recent CAC rules and free-trade-zone pilots create more predictable channels for some lower-risk transfers. The Chinese government's approach combines selective facilitation with retained state discretion over important data, CIIO-held data, large-scale sensitive personal information, and human genetic resources. These regimes are in conflict with one another. Opposing views of what data should flow where feed into the overall patchwork of laws and regulations for health- and biopharma-related AI data and data supply chain components.

Current Tensions and Future Issues

Policymakers, companies, nonprofit research organizations, and other stakeholders charting the future of health and biopharma, data components of the AI supply chain, and responsible innovation should use the areas covered and gaps left by the legal and regulatory patchwork to identify paths forward. Both the benefits and the frictions that these laws and regulations create are instructive for identifying areas for improvement and harmonization.

This section starts by describing some of the reasons why the health and biopharma sectors' interactions with training data, testing data, models (themselves), model architectures, model weights, APIs, and SDKs are differentiated compared to other sectors. These are:

- Healthcare is not advertising. Health and biopharma functions related to the data components of the AI supply chain—compiling training datasets; transferring protein folding or genetic-analysis AI models across borders; using AI models to generate synthetic health or genetic data—can carry significant risks but, on the whole, still serve fundamentally important societal functions around healthcare, disease research, vaccine development, and the like.
- Some health and biopharma activities, either by science or regulation (or both), require or benefit from access to data from multiple countries to function effectively. Data from different populations across countries can ensure that treatments indeed work across populations, improve health equity, significantly drive down the subsequent costs to societies of treatments that only work for a subset of the population, and help better tackle global diseases and pandemics. Some regulatory structures can also facilitate or encourage this multi-country, multi-population data collection for maximally effective, ethical clinical trials.
- Health and biopharma activity related to AI data components is often guided by quite robust ethical frameworks that cover patients, health decisions, and patient data. Informed patient consent, for example, is a bedrock principle of healthcare that is quite relevant for laws, regulations, and practices related to the data components of the AI supply chain, such as how data could be used in AI model construction or subsequent AI-driven diagnostics or analysis.

After detailing what makes the health and biopharma sectors unique in this area—compared to, say, activities carried out by social media companies or general-purpose LLM chatbot vendors—this section highlights some of the open legal and regulatory questions for policymakers and other stakeholders. These span whether or how laws and regulations can or should reference existing data best practices for health and biopharma entities, such as from requirements for clinical trial data; if health and biopharma activities should be treated differently under the AI-related and AI data-related regimes; how organizations should approach the export of AI models trained on covered health or genetic data, as well as the provision of those models via API to entities outside the jurisdiction; how synthetic data-generating AI models fit into the regulatory picture; and, among others, whether national security restrictions on cross-border data transfers should or should not have exemptions for healthcare and biopharma activity.

Listing these questions is designed to serve two purposes. These questions connect the prior section's discussion of the global regulatory patchwork to this section's discussion of what is unique about the health and biopharma sectors. And these questions also tee up issues for the subsequent, concluding discussion and recommendations.

What is Unique to Health and Biopharma Versus Other Sectors?

Many of the tensions that the legal and regulatory patchwork for the AI supply chain's data components pose are not unique to the health and biopharma sectors. Countless global organizations seek to transfer data across borders. Many companies face regulatory uncertainty when building new technologies or deploying them with impacts on society or specific populations. But there are several reasons why these health and biopharma activities are different from other use cases, whether in transportation, finance, or advertising. Spelling them out facilitates a more nuanced conversation about future regulations.

The first, overarching point is that healthcare is not advertising. Disease research, clinical trial development for vaccines, and other activities are fundamentally different from the running of advertisements for retail goods, entertainment events, technology services, and the like. Certainly, these health- and biopharma-related activities (and sectors) have their fair share of problems, including critical issues around equity and privacy. This statement is not intended to downplay those issues. Nonetheless, a private-sector company collecting individuals' health conditions so it can target them with an ad on a social media platform is different than a private-sector company collecting individuals' health conditions so it can map patterns in surgical needs, improve early-stage diagnostics with AI models, or monitor for signs of potential disease outbreaks. The latter activities serve different purposes with far greater potential for societal benefit.

Again, the nature of health and biopharma work can make the AI-related and AI data-related risks of errors, biases, data leaks, and the like much more serious than in other contexts; a major error in a system's diagnostic finding has the potential to

hurt someone more than a poorly targeted social media ad.[106] But the fact remains that regulations recklessly accelerating,[107] responsibly guiding, or curtailing[108] health research, clinical trials, vaccine development, and disease surveillance will have more immediate impacts on human life, human wellness, and societal safety than those regulations impacting other sectors that intersect with the data in the AI supply chain.

Second, some health and biopharma activities, either by science or regulation (or both), require or benefit from access to data collected from multiple countries to function effectively. On the scientific front, having more data from multiple countries can improve health equity and ensure that proposed treatments are effective for different groups of people. Groups historically underrepresented in or excluded from clinical research can have "distinct disease presentations or health circumstances that affect how they will respond to an investigational drug or therapy," meaning that comprehensive, including globalized, datasets can help improve the generalizability of findings.[109] Beyond the moral implications of not covering these populations in health and biopharma work, failing to do so can also cost US society alone hundreds of billions of dollars, as medical problems that could have been avoided early on through inclusive clinical trials show up later and at much greater cost.[110] In this context, companies, nonprofit research organizations, and others gaining access to data from different countries is therefore not (or is not merely) a "surveillance" function, such as an advertising company or chatbot LLM vendor wishing to profile even more people through training data. Representative data can help to ensure that therapies work across populations,[111] which especially important in situations such as global pandemics. For example, the pharmaceutical companies that developed a COVID-19 vaccine leveraged clinical trial participants and data from around the world to understand the virus and develop a mitigation.[112]

Regulations can simultaneously lead health and biopharma companies to pursue multi-country data collection and analysis. The FDA issued draft guidance in September 2024, for instance, encouraging sponsors to use, where appropriate, clinical data from outside the United States for oncology clinical trials—to ensure that the data applies to the diverse US population and accounts for the prevalence, presentation, causes, or severity of a disease across countries or regions.[113] European regulators have developed a unified system through which clinical trial sponsors can submit applications to run a clinical trial in several European countries, rather than the prior process of submitting separate applications to national authorities and ethics committees in each country.[114]

Third, health and biopharma activity related to AI data components is often guided by quite robust ethical frameworks that cover patients, health decisions, and patient data. This is the case even if regulatory requirements and coverage can vary country to country. Informed consent, for one, is a cornerstone of medicine in many countries around the world, centering the right of patients to make informed and voluntary treatment decisions.[115] Under this ethical principle, providers should fully inform patients about the nature of procedures or interventions, the potential risks and benefits, and the available, alternative

106. Again, there are nuances, but this is meant as a general statement.
107. This includes policymakers significantly weakening regulations, eliminating them entirely, or choosing not to implement new ones to address novel risks.
108. This includes regulations that create enough friction or uncertainty to significantly and harmfully curtail important research and other activities.
109. Kirsten Bibbins-Domingo, Alex Helman, eds., "Why Diverse Representation in Clinical Research Matters and the Current State of Representation within the Clinical Research Ecosystem," in I*mproving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups* (Washington, DC: National Academies Press, 2022), https://doi.org/10.17226/26479.
110. Bibbins-Domingo, Helman, eds., "Why Diverse Representation in Clinical Research Matters,".
111. See, e.g., Annette S. Gross et al., "Clinical Trial Diversity: An Opportunity for Improved Insight into the Determinants of Variability in Drug Response," *British Journal of Clinical Pharmacology* 88, no. 6 (February 2022): 2700-2717, https://pmc.ncbi.nlm.nih.gov/articles/PMC9306578/.
112. See, e.g., "Pfizer and BioNTech Conclude Phase 3 Study of COVID-19 Vaccine Candidate, Meeting All Primary Efficacy Endpoints," Pfizer, November 18, 2020, https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-conclude-phase-3-study-covid-19-vaccine.
113. "FDA Issues Draft Guidance on Conducting Multiregional Clinical Trials in Oncology," US Food and Drug Administration, September 16, 2024, https://www.fda.gov/news-events/press-announcements/fda-issues-draft-guidance-conducting-multiregional-clinical-trials-oncology. For discussion of how the FDA's posture on this may effectively change or have changed since then, see, e.g., Jacqueline R. Berman, "Key Considerations for Foreign Clinical Trials When Looking Abroad for Product Development," Morgan Lewis, April 1, 2025, https://www.morganlewis.com/pubs/2025/04/key-considerations-for-foreign-clinical-trials-when-looking-abroad-for-product-development.
114. "Clinical Trials Regulation," European Medicines Agency, accessed June 21, 2026, https://www.ema.europa.eu/en/human-regulatory-overview/research-development/clinical-trials-human-medicines/clinical-trials-regulation.
115. See, e.g., Parth Shah et al., "Informed Consent," National Institutes of Health, November 24, 2024, https://www.ncbi.nlm.nih.gov/books/NBK430827/.

treatments.[116] These are notions of consent that far surpass that imposed in practice in many privacy laws and regulations, such as the United States' broken notice-and-choice approach to consumer disclosure through lengthy, hard-to-read, and take-it-or-leave-it privacy policies.

For clinical research in particular, the 1947 Nuremberg Code, the 1979 Belmont Report, the 1991 US Common Rule, the 2000 Declaration of Helsinki (updated last year), and the 2002 Council for International Organizations of Medical Sciences (CIOMS) guidelines (updated in 2016)[117] are used around the world to guide its ethical conduct.[118] Together, and along with other sources, professionals reference seven principles that should guide clinical research activities: social and clinical value, scientific validity, fair subject selection, favorable risk-benefit ratio, independent review, informed consent, and respect for potential and enrolled subjects.[119] The World Health Organization, to give another example, emphasizes centering public health needs and equitable health advancements; involving patients, the public, and communities to align research with public needs and sustain trust; ensuring randomized controlled trials are ethical, efficient, informative, risk-based, and proportionate; and addressing local health research needs and expanding cross-border collaborations, among others.[120] Although not surveyed comprehensively here, extensive literature exists on these subjects.

This is not to say that health and biopharma activity related to AI data components should not be regulated—far from it. But to underscore that unlike some other sectors, where AI-applicable ethical frameworks are still nascent, many relevant concepts already exist for the health and biopharma spaces. Some guidelines may fall short, and some organizations may not follow them. Nonetheless, having many ingrained codes of ethics adds a layer of both thinking about responsibility and AI-useful frameworks to the discussion. Policymakers building AI-related or AI data-related regulations, companies determining what uses of AI-related data components are ethical, and stakeholders interacting with the health and biopharma AI ecosystems can all reference these concepts to understand tradeoffs, identify opportunities and problems, and build solutions.

Additionally, existing heath and biopharma regulations could apply to AI use cases, serving as another layer of governance. There are instances in which companies or other organizations use AI to develop a medicine or vaccine, but where the medicine or vaccine still goes through its own regulatory approval. Use of AI data components or AI models can still introduce new risks—older health or biopharma regulations may not cover every scenario or harm—but an underlying layer of governance can be developed over it.

Open Legal and Regulatory Questions

The current legal and regulatory landscape leaves questions unanswered for health and biopharma organizations engaged in activities related to the data in the AI supply chain. These open questions, which policymakers and other stakeholders should seek to address in the coming months, include:

- How do data privacy, cross-border data transfer, AI model, and other regulations point to or draw lessons from health best practices, such as clinical trial data requirements?
- Should laws and regulations treat health and genetic data used in an AI context for health research, drug development, and similar public-interest functions differently than the same data used by organizations for purposes such as online advertising?
- Can organizations train an AI model on health or genetic data from one country (or bloc) and then transfer the model beyond that jurisdiction? Are there requirements they must satisfy before doing so? If the organizations need to obtain a case-by-case determination from a regulator, is there any standard that the regulator uses to make such a determination?
- What if the health or biopharma organization trains an AI model on health or genetic data from that jurisdiction and does not transfer the model from the jurisdiction per se—for instance, hosting it on a server in that jurisdiction—but makes it available to internal or external users outside of that jurisdiction—such as via API? Is that prohibited? What factors go into the determination, and what is the legal basis for the regulator's position?
- If an organization builds a synthetic data generator based on underlying health or genetic data that is covered under a regulatory regime, can the organization move the model outside of the jurisdiction? Can it access the model from outside the jurisdiction or make it accessible to others beyond the jurisdiction? Can the organization generate synthetic health or genetic data from the model, keep the model in the jurisdiction, but

116. Shah et al., "Informed Constent..
117. https://cioms.ch/news/new-cioms-international-ethical-guidelines-now-available/
118. "Ethics in Clinical Research," US National Institutes of Health, accessed June 21, 2026, https://www.cc.nih.gov/recruit/ethics.
119. "Ethics in Clinical Research."
120. "Guidance for Best Practices for Clinical Trials," World Health Organization, accessed June 21, 2026, https://www.who.int/our-work/science-division/research-for-health/implementation-of-the-resolution-on-clinical-trials/guidance-for-best-practices-for-clinical-trials.

move the generated, synthetic data beyond the jurisdiction?

- Is the current legal and regulatory predominant focus on the training data, testing data, and AI models of the health and biopharma AI supply chain, as captured in the table in the previous section, going to persist? Or will future laws and regulations increasingly, explicitly govern other health and biopharma data components of the AI supply chain, such as model weights or APIs, as well?
- Do regulators perceive that national security-driven restrictions on cross-border data transfers of health or genetic data apply to AI models trained on that data? If so, how, and with what factors in play? What if the AI models themselves present a marginally low risk of regurgitating the underlying training data itself but could much more easily generate inferences about individuals that are generated based on the underlying training data?
- On the flip side, should national security restrictions on cross-border data transfers for the most high-risk areas be subject to exemptions, just because the functions may relate to health or biopharma activities? Where do and should policymakers draw the risk threshold? And how do they handle data components of the AI supply chain, such as health training datasets or AI models for sophisticated analysis of genetic sequences, that are strictly for health purposes versus those that are dual-use—where the civilian (including commercial) application could be equally used in a military context?

Conclusion and Recommendations

Healthcare and biopharma AI is a promising area of AI research and development, at the same time as it raises critical questions about transparency, auditability, privacy, cybersecurity, inequity, national security, and more. These questions are relevant for the United States, the European Union, China, India, and many other countries and regions around the world. They are also relevant across every data component of the AI supply chain, as shown in this report: training data, testing data, models (themselves), model architectures, model weights, APIs, and SDKs. For policymakers to enable responsible innovation—safely and securely promoting vital healthcare services, disease research, vaccine development, and so forth while mitigating risks—they must tackle some of the current challenges and tensions between health and biopharma AI and AI-related data use cases, a patchwork of global laws and regulations, outdated regulatory concepts (such as definitions of "anonymization"), and a governance discussion driven by LLM chatbots rather than a wide range of other AI technologies and use cases.

This report makes three core recommendations for governments, companies, and nonprofit research organizations navigating this frontier of health and biopharma, the data components of the AI supply chain, and future considerations around healthcare progress, scientific innovation, data privacy, cybersecurity, economic competitiveness, and national security:

1. **Governments should use existing best-practice privacy and cybersecurity principles when pursuing regulations on data collection, data use, cross-border data transfers, and AI model deployment that impact health and biopharma, including the healthcare profession's patient consent ethical framework and standards bodies' specifications for encryption, data minimization, and protection against reidentification.** The countries' legislatures and their respective enforcers of broader data regulations—including the US Federal Trade Commission and DOJ, the EU member states' information commissioners, the Cyberspace Administration of China, and the Data Protection Board of India—should carry this out. They should ensure that the healthcare concept of informed consent is reflected in laws and regulations that impact AI-related data components for health and biopharma activities, drawing on global-consensus frameworks and documents to articulate that concept in writing for any organization handling the likes of health or genetic AI-related data components. Doing so should strengthen many laws' and regulations' notion of consent over the presently weak, advertising technology-driven "notice and choice" consent baseline—and could have the indirect effect of helping to harmonize notions of consent across jurisdictions, drawing on standard concepts in the medical field. Governments should amend any regulations that exempt "anonymized" or "deidentified" data from coverage to ensure that their definitions of what makes data "anonymized" or "deidentified" comport with the current and over-the-horizon computer science and statistics literature on reidentification attacks—as well as recent machine learning advances in reidentification. They should, generally speaking, ensure that any pre-transfer requirements for cross-border transfers of health or genetic data and AI models refer to globally recognized standards for system transparency, encryption, data privacy, and the like, drawing on standards from the International Organization for Standardization (ISO) and the US National Institute of Standards and Technology. For example, if a regulatory regime is going to require that a covered organization implement encryption before transferring patient data abroad, it should reference widely known and used encryption standards, such as from the ISO, rather than specify their own. Doing so would lower the burden for health and biopharma organizations complying with restrictions (including in multiple countries), make reference to already negotiated best practice standards, and potentially help to harmonize regulatory requirements across jurisdictions. And as part of these governance best practice principles, governments should, among others, make clear for any regulatory review processes the strict or average anticipated timelines for review, a process through which regulated organizations and individuals can seek advisory opinions or supplemental guidance on compliance, and whether there is a process through which organizations and individuals can appeal decisions and if so, what it is.
2. **The surveyed regions should issue clarifications about how the health and biopharma data components of the AI supply chain sit within their non-AI-specific and AI-specific laws, drawing on the tensions and open questions identified in this report.** This recommendation is made not just because this report is health- and biopharma-focused but because of the particular importance of these sectors to society. For example, the US Department of Justice should clarify whether AI models trained on bulk data and then exported to China, Russia, Iran, North Korea, Cuba, or Venezuela would, from its perspective, implicate the federal Data Security Program—or whether that would sit outside the program's scope. FDA regulators should also issue updated guidance on how they view health or biopharma AI models vis-à-vis their authority to regulate certain software functions, including what factors health or biopharma organizations might need to weigh to reasonably determine if the FDA would consider a particular AI model to be covered. To give another example, the European Data

Protection Board should further clarify its December 2024 opinion about AI models trained on EU citizen data to provide clearer, articulated standards for member countries as to whether an organization could train an AI model on EU health or genetic data and then export it beyond the European Union without triggering additional GDPR obligations. All jurisdictions should clarify whether they would handle the export of an AI model trained on their citizens' data differently from the provision of that same AI model to entities beyond its jurisdiction, via API: would the latter be permitted or subject to similar restrictions?

3. **Governments should clarify criteria for any situations where they will treat health or biopharma AI-related data activities differently than AI-related data activities in other sectors.** As covered in this report, there are several reasons—including due to the societal importance of health research and innovation—for policymakers to consider health and biopharma AI-related data activities differently than they would other kinds, such as transportation data generation or the creation of facial recognition AI models. At the same time, it is also difficult to disentangle health and biopharma AI-related data components by data type or by component use case. For data types, a hospital research center could theoretically use a database of patient records for AI training just as an advertising company could theoretically use that same database for AI training—meaning rules focused only on the data per se would rope in one activity that could be ethical with another activity that could be invasive and harmful. For data use cases, a company made up of clinical experts could theoretically develop a cancer diagnostic AI model grounded in sound methodologies and medical ethics (something a policymaker might want to happen, subject to responsible constraints), while another company made up of untrained hobbyists could seek to develop a "cancer diagnostic AI model" based on junk understandings of science (something policymakers might wish to block); hence, rules focused only on the use case and not the outcome could also rope in a range of activities, both desirable and undesirable. All to say, policymakers face complicated tasks in ensuring that scientifically sound, methodologically rigorous, and ethical health and biopharma innovation can advance while simultaneously preventing and mitigating the harms of other actors using health and biopharma AI-related data components. This report has offered several criteria that policymakers should consider in this effort: protections on the data itself, the ethical guidelines underpinning the AI research and development process, and the principle of informed consent cutting across all data component-related activities, among others. For example, policymakers could mandate that before any AI data component or AI use case is even considered for exemptions to accelerate health and biopharma research, the data components must satisfy strict data protection and informed consent requirements. Policymakers should at minimum strive to make clear to the public—in legislative development processes, regulatory oversight, and much more—what guides their thinking on how to carve out, if at all, health and biopharma activities from others. Doing so will be increasingly critical as data and AI regulations that impact health must also consider myriad other factors, from sovereignty to privacy to national security. Regulatory guidance could come from the enforcers in the listed countries of the broader privacy and data security rules—including the US Federal Trade Commission, the EU member states' information commissioners, the Cyberspace Administration of China, and the Data Protection Board of India. A broader strategy on this front should come from the respective executive branch or agency bodies.

About the authors

Justin Sherman is the founder and CEO of Global Cyber Strategies, a Washington, DC-based research and advisory firm; a contributing editor at Lawfare; and a columnist at Barron's. He is the author of the book Navigating Technology and National Security.

Atlantic Council Board of Directors

The Atlantic Council is a nonpartisan organization that promotes constructive US leadership and engagement in international affairs based on the central role of the Atlantic community in meeting today's global challenges.

Atlantic Council
1400 L Street NW, 11th Floor
Washington, DC 20005

(202) 463-7226

www.AtlanticCouncil.org