Legislative Council Staff
Nonpartisan Services for Colorado's Legislature

Memorandum

January 30, 2026

TO: Interested Persons

FROM: Samantha Lattof, PhD MSc, Science and Technology Policy Program Fellow

SUBJECT: Artificial Intelligence and Health

Overview

Expanding on Legislative Council Staff's Overview of Artificial Intelligence (AI) memo, this memo focuses on the application of AI for health uses. It details model training, algorithmic equity, and data privacy and security from a health lens. Furthermore, the memo discusses challenges in data privacy and security, highlighting risks such as unauthorized monetization of patient records. Finally, it reviews AI applications in mental health, workforce support, and radiology that demonstrate the transformative role of AI in healthcare.

Background

AI has immense potential to revolutionize health. From discovering and designing new drugs (Xu et al. 2025) to detecting early signs of over 3,000 diseases using biomarkers (Garg et al. 2024), researchers and developers continuously explore new, creative uses for AI in healthcare. While newer developments may exist within the realm of clinical trials and proof of concept, as opposed to more widespread commercial use, AI has already changed how many aspects of public health and healthcare operate. AI applications for health have assisted in surveilling diseases, increasing diagnostic accuracy, reducing clinician burnout, and improving outbreak response, for instance.

While AI technologies for health have developed rapidly, the regulatory environment has taken more time. AI for health uses has attracted growing attention from legislators across the United States, with introduced legislation on AI and health surging from 15 bills in 2023 to 168 bills in 2025. At least 41 states have introduced health and AI bills since 2023 on topics ranging from the private sector to responsible use.[1]

[1] For an analysis of state legislation on AI and health, read the Legislative Council Staff memo, "State Legislative Trends in Artificial Intelligence and Health (2023-2025)."

While waiting on the regulatory environments to crystalize, it is largely up to developers, the private sector, and implementing bodies (e.g., healthcare facilities and providers) to ensure that they are developing and implementing AI in ways that uphold patient safety and privacy.

Model Training

AI models are computer systems that, when trained and designed to a high standard, process and organize data in such a way that allows the models to make meaningful connections and identify patterns. To help AI learn patterns that can predict health outcomes, people developing the AI train the model by giving the algorithm[2] large amounts of health data to analyze (e.g., heart rate data, text data from patient charts, images of skin lesions). This training or learning process can occur in different ways:

- "Supervised learning" is a type of learning used frequently for healthcare applications. By using labeled data (e.g., photos of skin lesions marked "melanoma" or "healthy"), a model learns to associate certain features with the correct diagnosis.
- "Unsupervised learning" trains a model with no human input, and "semi-supervised learning" trains a model with minimal human input. These training processes are the default method for large language models (LLMs), at least in the early stages when models are trained on massive amounts of primarily textual data.
- "Predictive AI" uses models to forecast future events (e.g., a patient's risk of infection following surgery).
- "Locked models" do not change once they are deployed, whereas "adaptive models" learn and change in response to the real-world data they process over time.
- "Deep learning models" use complex neural networks[3] to produce what may be highly accurate results. However, these models exhibit what is known as the "black box" challenge, meaning that people deploying the model are unable to understand why the model reached its conclusion.

[2] In AI, an "algorithm" refers to the programming instructions that tell the model how to learn, analyze information, and make decisions based on that information.

[3] "Neural networks" are methods for teaching computers to process data that are loosely modeled on the human brain and nervous system.

Risks of Bias

AI experts disagree about whether AI systems are "intelligent," and if they are intelligent, how intelligent these systems are. A model's "intelligence" depends how it was built and trained, which depends on the people involved in those processes. As a result, models can produce faulty results because of a variety of factors: poor training data, the general nature of AI models being statistical distributions, and biases inherent to the people creating the models.

Google's AI Overviews that appear at the top of searches, for example, use generative AI to answer over two billion people's health questions each month. Yet, instead of citing reputable medical sources, a study from Germany found that Google prioritized content from YouTube. YouTube, owned by Google, is the second most visited website in the world after Google. An investigation by The Guardian found that these Google AI Overviews have also provided health information that is false, misleading, and at times, harmful.

Another well-known example in health involves the use of a deep learning AI model to detect melanoma (skin cancer). Researchers trained an AI model to detect which images of skin lesions were malignant, but they missed a limitation of their training dataset. When dermatologists suspect a skin lesion is cancerous, they typically place a ruler next to it to track how it grows and changes over time. The people training the AI model did not take this fact into account.

Since the training data included more images of malignant cases with rulers than benign cases with rulers, the AI used "shortcut learning" to sort malignant images from benign images based on the presence of a ruler. The ruler, rather than the growth, became the diagnostic feature of cancer. Yet, while the presence of a ruler was correlated with cancer, the ruler did not cause the cancer. This bias in the AI model resulted from unseen patterns in the training data that were not initially apparent to researchers who developed the model. Once the researchers noticed this bias in their algorithm, they published a follow-up letter to share what they learned (Narla et al. 2018).

Algorithmic Equity

"Algorithmic equity" refers to designing AI so that it provides the same level of care and the same diagnostic accuracy to all patients, regardless of factors like race, sex, or socioeconomic status. An algorithm's performance depends heavily on its training data. If the data are biased, the algorithm will perpetuate those biases. Algorithmic bias has appeared across medical disciplines, such as cardiology, dermatology, and radiology.

Revisiting the example of using AI to detect melanoma, if a dataset featured only one group of people (e.g., an AI model to detect melanoma included only photos of light-skinned people),

then the model would be less accurate for other groups (e.g., people with darker skin, people with sunburns). As a result, patients with melanoma would be diagnosed correctly or incorrectly with varying degrees of accuracy, and they would receive different levels of care for the same condition.

While the AI model to detect melanoma depended on the presence of a ruler to diagnose melanoma, algorithms may also inadvertently discriminate against patients by other proxy variables specifically related to patients' socioeconomic and demographic characteristics. For example, an algorithm using healthcare spending as a proxy for medical need might conclude that poor patients are healthier than equally sick rich patients due to the fact that rich patients spent more on healthcare. Healthcare providers and insurance companies using algorithms that reflect these biases may deny care and insurance claims based on factors other than patients' health status.

Another challenge with health algorithms is historical inequity, where AI models replicate biases present in historical datasets. Old medical records might reflect unequal access to care or different standards of treatment for marginalized populations, for instance. Had the algorithm used newer data, it might reflect changes in treatment standards or improved access to care that would result in a more accurate model. However, it may be cheaper or easier to use historical datasets, especially when newer or better datasets are proprietary, making them unavailable for training purposes.

Mitigating Biases

To ensure that AI models work for all patients, researchers and developers have explored and adopted a variety of strategies to mitigate biases in their models. Starting with training data, selecting inclusive, diverse datasets improves the likelihood that a model can be used accurately across the entire population. Researchers are also working to help minimize bias by developing new methods to identify and correct hidden biases in datasets before the datasets are used to train models. An MIT SOLVE semi-finalist, for instance, developed a framework for mitigating dataset biases called AEquity that measures potential biases and allows developers to change their dataset so that it more closely reflects their population of interest.

Mitigating bias also requires strategies for after AI models are launched. To identify biases that emerge after an AI model is in use, healthcare providers and developers are implementing continuous monitoring strategies. These strategies range from implementing protocols for regular auditing of AI outputs to helping physicians understand the reasons behind AI decisions, so that the providers have an easier time identifying when a model is using biased shortcuts.

Data Privacy and Security

As AI models for health frequently involve sensitive patient data like medical histories and genetic data, data privacy and security involve protecting patient records and securing patients' identities. The onus for addressing data privacy and security within the private sector and clinical environments is on technology companies, researchers, and healthcare systems. No federal legislation regulating AI has passed yet, so the national regulatory landscape[4] is primarily defined in Executive Orders that may change from one administration to the next.

For health applications, AI developers are increasingly using models that can learn without using raw patient data. AI models previously had to be trained on a central server, which required moving sensitive medical data to that server. With federated learning, this direction is reversed. Hospitals or even patients with smartphones receive AI models that train locally (i.e., on the hospital servers, on the patient's phone), and the resulting insights are then sent back to the researchers. This technique maintains individual confidentiality while still enabling population-level learning.

Researchers can also strengthen privacy of AI models by using differential privacy, which uses mathematical formulas to help obscure data. As a result of differential privacy, an AI model that identifies a rare disease pattern would be unable to trace that pattern back to a specific patient.

Companies developing AI models work to strengthen their data privacy through hackathons to test models for vulnerabilities and by adopting frameworks to detect abnormalities in real time (e.g., changes in performance, malicious activity). That said, even with existing data privacy and security efforts, third parties have proven that data used for a company's model training are extractable without much effort.

Data breaches, model failures, and security lapses are particularly high-stakes for AI models used by patients, healthcare providers, and hospitals. In these applications, a breach or failure can be life-threatening. To address patient safety and security, some hospitals and health systems have hired Chief AI Officers to direct AI governance, implementation, and oversight.

In the private sector, companies are increasingly pursuing privacy-first infrastructure, meaning that security is embedded into the entire process from training through deployment. Some companies use "data clean rooms" when using external datasets so that they can learn from the data without the risk of exposing personal health information. Companies are also turning to

[4] For an overview of the national regulatory landscape, read the Legislative Council Staff issue brief, "Artificial Intelligence: The National Regulatory Landscape."

new certification programs, like AI Bill of Materials (AIBOM), to make AI systems more transparent, auditable, and secure. Even insurance companies providing cyber insurance are implementing measures to strengthen model security, such as requiring documented evidence of simulated attacks to test AI models' vulnerabilities before providing insurance coverage.

Impact of Data Breaches Involving Patient Data

AI companies have been involved in a series of data exposures involving patients. Most recently in January 2026, Epic, a major medical records company, and several of its health system customers filed a federal lawsuit against Health Gorilla, alleging that the clinical data platform and qualified health information network improperly accessed and monetized over 300,000 patient records. These records contained particularly sensitive patient data like genetic, mental health, and reproductive health information.

Other breaches involved health facility staff who used unvetted AI tools to handle clinical data (i.e., shadow AI). As a result of staff using shadow AI, sensitive patient data can be leaked to external AI companies. Even AI models themselves have been involved in data exposures after hackers introduced specific threats. Model or data poisoning, for instance, can result in tampered datasets that disrupt diagnoses. Since this type of attack can be quietly hidden, it could potentially deliver incorrect information about thousands of medical cases before someone identified the breach. Ethicists have started exploring the possibility of AI-induced medical manslaughter (Bartlett 2023), such as in cases where AI automation could lead to medical errors.

The impacts of these breaches may include financial loss to healthcare facilities or health systems. However, due to the ways AI is used for health, such breaches may also result in physical harm to patients or even death.

AI Applications for Health

The following applications for health reflect areas where AI has moved beyond experimentation to produce measurable outcomes. These tools are publicly available, whether in smartphone apps for personal use or in programs used in a hospital. These applications are increasingly used as core infrastructure to support clinical precision, workforce efficiency, and access to care. However, their use may come with challenges.

Mental Health

Within the mental health field, AI is being used to support clinicians, identify effective treatments, and deliver therapeutic interventions around the clock. Like in other health fields, mental health providers increasingly use clinician support tools to record sessions and generate notes, making it easier for them to focus on patient care. Other tools unique to the field, like Kintsugi, provide real-time feedback on patients' emotional distress.

AI applications for mental health are also focusing on digital phenotyping, where smartphone sensor data is used to predict health. These data from passive sensing can assist providers in identifying behavioral concerns, and changes in digital biomarkers (e.g., typing speed, keystroke dynamics) can predict early signs of mental health conditions like Parkinson's disease and mild cognitive impairment (Alfalahi et al. 2022). The application of these new data sources and AI tools to neuroscience has enabled more individualized treatment.

Chatbots are also an increasingly popular tool for delivering mental health care. Woebot Health, a company that spun out of Stanford, is one of many examples. The company offers a "relational agent" that "has the potential to rapidly develop a bond with users" in order to deliver Cognitive Behavioral Therapy (Darcy et al. 2021). Woebot Health's clinical research on this tool once appeared on the company's research webpage, but the company has since removed that page. The company also terminated study on its digital therapeutic for postpartum depression that it had registered on ClinicalTrials.gov. While these decisions were made internally for reasons that remain unknown, the Woebot Health example illustrates that the field is rapidly changing. Apart from pivoting and adapting in response to newer research and clinical guidance, chatbots in particular are under increased scrutiny from the United States House of Representatives Subcommittee on Oversight and Investigations.

In response to the influx of companies using LLMs to provide therapy services, researchers at Stanford investigated the use of LLMs to replace mental health providers (Moore 2025). Using therapy guides implemented by major medical institutions, the researchers assessed the ability of LLMs to deliver care in accordance with these best practices. They found that, in contrast to best practices, "LLMs 1) express *stigma* toward those with mental health conditions and 2) respond inappropriately to certain common (and critical) conditions in naturalistic therapy settings—e.g., LLMs encourage clients' delusional thinking, likely due to their sycophancy. This occurs even with larger and newer LLMs, indicating that current safety practices may not address these gaps." As a result, the researchers concluded that "LLMs should not replace therapists," and they explored alternative roles for LLMs in clinical therapy.

Model quality and patient safety are especially important when using AI for mental health or in populations with mental health issues. At least two people in Colorado have died by suicide following their use of chatbots. A 40-year-old man died after extensive emotional interactions with ChatGPT's chatbot about his mental health struggles. His family filed a lawsuit against OpenAI alleging that ChatGPT served as a "suicide coach." A teenage girl using Character.AI also died, resulting in her family filing a lawsuit against the company for sexual abuse and psychological manipulation.

Workforce Support

To address challenges like burnout and worker shortages, healthcare providers and institutions are turning to AI for assistance with administrative tasks. AI scribes are one of the most rapidly adopted tools, recording patient visits and automatically drafting clinical notes. The American Medical Association reports that departments with high documentation burdens (e.g., mental health, primary care, emergency medicine) were most likely to adopt AI scribes. This technology saved providers hours of paperwork and after-hours charting. It also improved physician wellbeing, work satisfaction, and patient communication.

AI tools are particularly effective for automating repetitive tasks like medical coding, billing, and scheduling. Patient communication systems and chatbots can schedule appointments while offering multilingual support, thus reducing some of the administrative staff's burden.

Health facilities are also turning to AI for operational support. Predictive staffing tools that can forecast seasonal illnesses and peak patient volumes help organizations optimize their shift planning, surgical scheduling, and bed utilization. In human resources, AI tools can assist in recruitment by scanning thousands of resumes quickly. Other tools can help predict which units are at risk for burnout by analyzing employee overtime and wellness data.

Radiology

Within the field of radiology, AI is skilled at handling high-volume, repetitive, and time-sensitive tasks. These tasks include assisting with triage, computer-aided detection, image enhancement, and automated measuring.

Historically, radiologists often processed scans in the order they were taken. AI can now "pre-scan" images once uploaded, flagging emergencies like brain bleeds and collapsed lungs. By identifying these life-threatening conditions and moving them to the top of the radiologist's reading list, AI can reduce these patients' time to treatment.

AI is also skilled at identifying subtle patterns that humans might miss, such as early-stage breast cancer lesions that may appear invisible to the human eye on scans. When specialist radiologists are unavailable, AI tools in emergency rooms can help providers spot subtle bone fractures in X-rays. AI can also analyze changes in brain tissue volume over time, making it easier to identify early indications of degenerative diseases like Alzheimer's or grading tumors.

Behind the scenes, AI can improve the performance of scanning machine hardware. From removing blurriness on images caused by a moving patient to producing high-quality images from low-dose radiation computed tomography (CT) scans in pediatric patients, these applications benefit both patients and providers. AI measurement and segmentation tools make it easier to more accurately measure the size of an organ or a tumor.

Takeaway Messages

- AI models can produce faulty clinical results if trained on biased datasets, such as an AI model for skin cancer that learned to assess for the presence of rulers rather than malignant growths.
- The integration of AI into the healthcare sector requires rigorous oversight to mitigate inherent biases and security vulnerabilities that could compromise patient safety and privacy.
- Healthcare institutions are increasingly adopting AI for repetitive tasks like clinical notetaking, medical coding, billing, scheduling, and image processing and enhancement. These uses help reduce clinician burnout, address worker shortages, and accelerate patient care.