Public Datasets: A Foundation to Artificial Intelligence in Health Care

SPENCER FERGUSON, PATRICIA M. TILLE

ABSTRACT

The use of artificial intelligence (AI) in health care is predicated on its safety and efficacy. AI is a technical field of study and is fast evolving. It will affect everyone, so it is important that stakeholders, especially providers and legislators, understand the mechanisms of how AI works so they can make competent decisions to ensure patient safety. There are examples of successful AI systems in health care, but widespread application and adoption suffer due to several issues regarding the type of training data used. All AI systems must be trained using data, and the quality and quantity of this data are at the foundation of their success. Open data addresses issues of validation, reproducibility, and bias within AI systems. Initiatives from private and government agencies, including funding and legislation, support the propagation of open data for research use. The sharing of curated data and trained AI models will exponentially increase AI development in health care. Despite hurdles, open data is the key to implementing safe, reproducible AI models in health care.

ABBREVIATIONS: AI - artificial intelligence, CMS - Centers for Medicare and Medicaid Services, DL - deep learning, EHR - electronic health record, EU - European Union, HHS - Department of Health and Human Services, HIPAA - Health Insurance Portability and Accountability Act of 1996, HITECH - Health Information Technology for Economic and Clinical Health Act of 2009, IT - information technology, LLM - large language model, MACRA - Medicare Access and CHIP Reauthorization Act of 2015, ML - machine learning, NFDI - Nationale Forschungsdateninfrastruktur, NLP - natural language processing, PHI - protected health information, PSQIA - Patient Safety and Quality Improvement Act of 2005.

INDEX TERMS: artificial intelligence, public databases, health care, laboratory diagnostics.

Clin Lab Sci 2024;00(0):1–8

***Spencer Ferguson**, University of Cincinnati*

***Patricia M. Tille**, University of Cincinnati*

***Address for Correspondence: Patricia M. Tille**, University of Cincinnati, tillepm@ucmail.uc.edu*

INTRODUCTION

Artificial intelligence (AI) is the next technological leap forward for humanity. AI's application in health care is steadily growing, offering significant advancements in disease diagnoses, administrative tasks, and public health assessments.[1-5] However, the foundation of a particular AI model's success relies on the data with which it is programmed. Open data, also called public data, provides the foundation for implementing AI in health care.[1]

Broader datasets reduce bias when training AI models, and open datasets allow AI models to be rigorously validated and used across patient populations.[6,7] With more personal data available online, the public is poised to share the data needed to support open data initiatives. Electronic health records (EHRs) share medical records between healthcare providers, mobile health apps track individuals' data daily, and social media provides an outlet for the public sharing of personal information.[8,9]

With stakeholders at every level, it is critical to explain the function and importance of open data. Open data for AI training comes from patients, engineers use it to develop AI, and it ultimately impacts the results of an AI system used in health care. There are hurdles to open data, but heterogeneous data availability is foundational to implementing safe reproducible AI models in health care.

BACKGROUND

Health and Digital Literacy

AI in health care will affect everyone, requiring key stakeholders to understand how it works. Patients, providers, healthcare professionals, technology companies, and legislators all play a role in developing and accepting technology. There are high stakes in health care, resulting in a requirement to understand how AI works to ensure its safety. Mistakes in diagnoses or treatment can ultimately cost a patient's life. To implement any new technology in health care, it is paramount that it is safe and effective. Understanding a new technology like AI is the first step in implementation.

Scientific articles on AI are jargon heavy, reducing the application to only those with specific informatics training. Healthcare professionals, particularly primary care providers, will not trust a system they cannot understand. Patients will also be hesitant to accept technology their provider does not. Legislators, working to guide technological advancement and protection of patients, need to

understand AI's impact to legislate effectively. With legislation moving at a significantly slower pace than technological innovation, it is more likely that legislative efforts will always be one step behind. Despite the speed of technological innovation, these key stakeholders need a working understanding to effectively participate in AI's healthcare implementation.

Digital literacy is required in addition to a recent push for improving health literacy among patients. Health literacy is necessary to understand the impact of technology on health care, and digital literacy is required to understand the mechanisms of how AI affects patients. A general understanding of AI, combined with health and digital literacy, will ensure that providers and legislators make competent decisions to ensure patient safety.

AI in Health Care

While there are concerns in the public domain about AI in general and additional considerations for its application in health care, there is a strong case that it should be embraced. AI models have already shown significant improvements in health care. In emergency medicine, advancements include patient triage optimization, risk stratification, and patient outcome predictions.[1] AI in medical imaging provides interpretations that are at least equivalent to those of a radiologist.[2] While an all-purpose AI system is not available for widespread use, these examples show how specific implementations of AI in health care are incrementally improving patient care.

There are limitations to AI in health care that must be overcome. Transparency, or how an AI system is "thinking," is a concern for validating new systems and ensuring quality control.[1] Patient data must be protected, which adds liability to technology companies.[10] Ultimately, the quality of data used to train AI systems plays a foundational role in AI validation, reproducibility, and bias.[10] Open data provides the foundation for solving these limitations. Despite patient safety and data anonymity concerns, AI systems are being built in all industries. To ensure safe and effective implementation in health care, AI innovations should be openly embraced to identify and resolve limitations.

AI Modalities and Training

There are several different AI training modalities and definitions for AI. For this review, AI will be referred to in general terms rather than as a specific type. Some common modalities include machine learning (ML), deep learning (DL), and large language models (LLMs). The functionality of a specific AI model may supersede others in health care, but the importance of open data for training remains central to effective implementation and patient safety.

ML is a common model for developing AI systems for specific tasks. The system is trained on a sufficient dataset, which allows the system to recognize patterns or perform regression and classification.[11] The human training component of ML and the complexity of the problem it must solve limit this model.[11]

DL is a subset of ML that expands upon ML's capabilities but introduces the black-box issue of transparency, or how the model makes predictions. This is a key issue in health care when considering patient safety. DL solves significantly more complex problems than ML by using a mathematical framework to automatically derive representations from given data.[11] This method's limitations in health care are centered on the copious amounts of data required for training and the black-box nature of the decision-making.

LLM or natural language processing (NLP) are subsets of ML based on text generation.[12] These models also require a significant amount of training data but have seen promising use in the public with applications like ChatGPT. In health care, ChatGPT has also been effective at enhancing healthcare-associated infection surveillance.[13] LLMs may also provide efficiencies in administrative tasks like clinical note-taking or patient chatbots that may provide medical advice. Consider a healthcare patient chatbot trained on existing medical knowledge with access to a patient's entire medical record. The AI will surpass a doctor's ability to correlate all the information. Additionally, depending on the training data, LLMs can already produce near authentic scientific articles.[14] There are concerns about LLM AI being trained using copyrighted material like medical texts or scientific articles that are not public. Despite this concern, LLMs show great potential application in science and health care.

Regardless of the model or application, training plays a role in the success of AI. The training of AI systems can be supervised, unsupervised, or trained with reinforcement. Supervised and unsupervised learning is contingent on the data being labeled or unlabeled, respectively.[2] Reinforcement learning incorporates user input as feedback to the system.[2] Supervised and reinforcement learning requires additional human input, making these systems more trustworthy because of the human validation of the algorithm. Unsupervised training, as seen in DL, allows systems to identify patterns unrecognizable by humans. However, the black-box nature of this process is not acceptable in health care.

The type of data used for training depends on the AI model and function. The goal is to make the data open to the public for effective training of AI in health care. In health care, the data required can contain clinical notes, diagnoses, medical imaging, patient history, and patient demographics, such as age, sex, and race. Information about patients that may be used to identify them is termed protected health information (PHI). PHI used in training data is a concern for patient privacy. The training data will also include medical knowledge. This can come from various sources, such as textbooks, research, and expert opinions. Combined, the data form the knowledge base of the system being trained. To effectively train an AI system, the data are labeled by engineers.

Labeling data is important in supervised learning. The data label provides context to the system when training; this is primarily the training technique for ML.[2,15] For example, an AI system trained to detect pneumonia in chest x-rays will be trained using chest x-ray images labeled by a radiologist to indicate the presence of pneumonia. Additional data, such as patient demographics and medical history, aid the system in identifying patterns. Conversely, unlabeled data are used in DL models. DL models require copious amounts of data to build comprehensive AI systems in unsupervised learning. In DL models, the system is allowed to identify patterns without human intervention. Unlabeled data reduce the burden of human input required for ML but are also the impetus for the black-box decision-making of more complex AI systems.

Training data may be used in different ways, but regardless, the quality and quantity of data used for training will affect the outputs of the AI model. The adage, "garbage in, garbage out," applies to the training of AI models. These are issues that open data can solve. Patient population bias and lack of interoperability are frequently seen in training sets. Consider an AI model programmed to assist with disease diagnosis. A dataset built by hospital X that contains information on healthy individuals and those with common chronic illnesses may not produce reliable results for uncommon diagnoses. Training data from a single healthcare facility in a rural setting, hospital X, will also not be interoperable with a large healthcare facility in an urban setting. These examples of homogeneous datasets show how the training data can hamper interoperability or create a bias toward the patient population contained within the training data. Open data solves this issue by pooling training data across many patient populations.

Homogenous data from a single source prohibit reproducibility and validation and increase the risk of bias.[1] A frequent issue with new AI models in health care is that a research facility or hospital will use internal data to train the system. The results of a system trained with specific internal data to a facility are not reproducible with different datasets or patient populations. This prevents the independent validation of a new model and significantly reduces the interoperability of AI models in varying health systems. AI models trained with open data will reflect better reproducibility.

OPEN DATA

Need for Open Data

AI training is time consuming and often restrictive for software companies or healthcare research teams starting from scratch.[6] Copious amounts of data are required for an AI system to function properly. More data than what currently exists are required to train DL algorithms.[6] The labeling of data is time consuming and labor intensive. Using open data removes the data collection step. If labeled data are shared with open data sets, then this removes an additional step. The sharing of trained AI models and curated data increases the quantity and quality of open data while accelerating the growth rate of AI in medicine.[6]

AI models trained with homogenous data may contain bias. Data from a single research facility or hospital will always be limited compared to open data pooled from diverse patient populations and sources. The patient population used to train AI will ultimately create a bias toward those patients. Using homogenous data rather than open data results in training data with similar patient demographics, chronic diseases, environments, and nutrition. These factors will affect the system's predictions or pattern recognition.

New software or medical devices in health care must be validated and approved for use. Developers and research facilities may be reticent to share training data for additional facilities to use for validation due to their cost to curate the data or for proprietary concerns. If the data are shared, they may not represent a new healthcare facility's patient population. AI models trained on open data are interoperable because the data are available for validation and, as a more extensive data pool with better diversity, are more likely to be accurate across differing patient populations.

Open Data Initiatives

Despite a historical lack of a data-sharing culture, open data is often used in research settings. Open science includes initiatives supporting the emergence of open data and open software.[6] Funding also plays an important role in science and is significant in open data sharing.

Open science initiatives support open data in research. The F1000 Research model is an excellent example for future research. F1000 Research is an open research publishing platform that supports transparent and timely research publication in all study areas.[16] The 3 tenets of F1000 Research are open access, open data, and open peer review. All articles published by F1000 Research are freely available.[17] Researchers must include all data or, at minimum, list what data cannot be shared and why (eg, confidential patient identifiers).[17] Finally, an open peer review process allows for faster review of articles and citation of peer reviews.

Research funding, in addition to the structure of research peer review and publishing, is available to support open data initiatives. In 2018, Plan S, launched by cOAlition S, mandates that any research using public grants must be published in compliant Open Access journals starting in 2021.[18] Founders of Plan S include well-known organizations, such as the Bill and Melinda Gates Foundation, the Howard Hughes Medical Institute, and London-based funder Wellcome.[7] In addition, the European Commission, the US government, and the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles all leverage funding to drive the open science initiative.[17]

There are many initiatives to collect and promote open data. Both governments and private institutions support open datasets. Some examples of open data available include Germany's Nationale Forschungsdateninfrastruktur (NFDI); the United States' data.cdc.gov, data.gov, and healthdata.gov; and Massachusetts Institute of Technology's MIMIC program.[6,8,19]

These large databases support the collection and storage of data for research use. The NFDI contains branches for many areas of study, but the NFDI 4 Health branch works explicitly to "build a comprehensive inventory of German epidemiological, public health, and clinical trial data."[20] Data.cdc.gov, healthdata.gov, and data.gov provide US-specific data collected by government entities that may be used to inform the public, drive innovation, and support a transparent government.[19] These examples of government databases provide secure collections of copious data that can be regulated by the departments that are gathering them.

Data.gov is less helpful for healthcare AI datasets because it focuses on broad population datasets published by US agencies, such as the Centers for Medicare and Medicaid Services (CMS) and the National Oceanic and Atmospheric Administration. Population and environmental data collected by these agencies may be important for some models, especially those used for public health tracking and prediction. Still, the open data provided by data.cdc.gov is better for AI training in health care. Data.cdc.gov includes open data sorted by categories, such as disease, injury, and disability. It also includes instruction for developers and integrates the Socrata Open Data API software to provide access to open data resources from governments, nonprofits, and global nongovernmental organizations.[21]

Open data is also available for specific diseases. During the COVID-19 pandemic, several datasets were developed specifically for the training of AI models to detect COVID-19 in chest x-rays: COVID-19 Image Data Collection, COVID-19 Chest X-ray Dataset Initiative, ActualMed COVID-19 Chest X-ray Dataset Initiative, and COVID-19 Radiography Database (Italian Society of Medical and Interventional Radiology).[22] These types of highly focused datasets may be used or combined to train AI systems with the goal of highly accurate disease prediction. These 4 COVID-19 open datasets are frequently combined with pre–COVID-19 open data, such as the Radiological Society of North America Pneumonia Detection Challenge dataset. Combined, these make up an open data source termed COVIDx.[22]

Google has provided a great example of the power of open data when compiled like the COVIDx data. In 2020, the Google DeepMind team released a DL system called AlphaFold. With the help of over 50 years of open data, this chemistry-based system predicted 98.5% of the human proteome. Previous years of research only managed to produce 17% of the human proteome.[6] The success of AlphaFold supports the conclusion that open data infrastructures are the key to broader AI applications.[6]

Challenges

Open datasets are not without challenges. Patients, primary care providers, and legislators should know the challenges and the solutions for implementing AI in health care. Several issues to address include privacy, bias, confounding variables, and using datasets with errors.[17,22,23]

Privacy is arguably the first concern for patients' PHI. If not addressed appropriately, patients may refuse to share their data, which would negatively affect open data collection. Existing legislation, such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA), protects patient's privacy. However, despite the legal protection of PHI, data used to train AI models can require PHI. Cybercriminals can target the training data by tricking AI systems into divulging information about the training data.[23] These types of cyberattacks compromise patient data and undermine the trust of AI systems in health care.[9]

Differentially private ML is a possible technical solution to the issue of privacy.[23] Originally proposed by Dwork et al in 2006, this programming method adds "noise" to the data, obscuring targeted personal information.[24] This noise obscures PHI but also requires extra due diligence from the researchers to ensure the accuracy of the results. Another option is to ensure the use of anonymized patient data in training sets to prevent targeted attacks that leak PHI.[23]

After considering the privacy of patient data, we must look at the quality of the data. Open data can lead to bias and propagation of errors or confounding variables if not used carefully.[22] Confounding variables are presented when patient information is missing from datasets. This introduces variables that cannot be accounted for without that data. Open data with errors or confounding variables may be used repeatedly when not properly scrutinized, spreading bias into all models. Repeated publication of "successful" AI models trained with bad data may further reduce the trustworthiness of AI in health care.[22] The repeated use of COVID-19 data containing errors demonstrated this.

The use of open data during the COVID-19 pandemic demonstrates how poor datasets can affect AI systems output and trust in AI in health care. One study analyzed the use of publicly available COVID-19 datasets to train AI models that detected COVID-19 from chest x-rays. Bias was demonstrated in COVID-19 prediction models trained on publicly available data.[22] A significant issue with the training data was its labeling. Training data labeled for screening are not equivalent or appropriate for AI models intended for diagnosing.[22] Chest x-rays labeled to indicate the presence of pneumonia will not help an AI system understand if the pneumonia is specifically due to COVID-19. Therefore, these data will help screen chest x-rays for pneumonia but will not be able to diagnose COVID-19

infection specifically. As a result, the interoperability of this COVID-19 AI model on external data was only 45%, highlighting the importance of standardization and quality review of data in any open database.[22]

These hurdles to open data are not unsurmountable. Technology and legislation address privacy concerns. Process improvements related to data quality are essential to resolving issues with bias, confounding variables, and errors within datasets.

LEGAL PROTECTION OF PHI AND PROMOTING INTEROPERABILITY

Privacy of PHI and patient safety are the most salient issues with any technology in health care. Despite the slow speed of legislation, compared to technological innovation, there is significant legislation that protects PHI and improves data sharing. US legislation supports open data to meet the needs of AI regulation. International legislation also addresses these issues more specifically and serves as an example for future legislation. While not an exhaustive list, these legislative efforts lay the foundation for open data.

US Legislation

HIPAA protects PHI in several ways. The 3 parts of this legislation are commonly known as the Privacy Rule, the Security Rule, and the Breach Notification Rule. The Privacy Rule defines and protects PHI.[9] The Security Rule addresses electronic PHI by requiring covered entities (healthcare providers, healthcare clearing houses, or health plans) to take steps to prevent cyberattacks and set limits to who can access PHI within their system. Finally, the Breach Notification Rule requires covered entities to notify patients or the public in the case of a data breach. However, HIPAA does not directly regulate business associates of covered entities, nor does it cover newer technologies, such as mobile health applications (eg, Apple Health or Samsung Health) or AI. These limitations of HIPAA are addressed in later legislation.

Subsequent applicable US legislation incentivizes the use of EHRs and promotes interoperability. The Patient Safety and Quality Improvement Act of 2005 (PSQIA) and the Health Information Technology for Economic and Clinical Health Act of 2009 (HITECH) are examples. PSQIA and HITECH have different goals but work together to build on HIPAA and set the foundation for heterogeneous open data.

PSQIA focuses on quality and patient safety, while HITECH supplements the legislation provided by HIPAA. PSQIA initiates collecting and storing deidentified patient data for research and quality assurance of patient safety and errors.[9] This is not open data for research but is legislation that requires collecting and storing patient information. HITECH supports the framework of HIPAA by incentivizing the use of electronic health information technology (IT). Under HITECH, business associates of covered entities are responsible for HIPAA rules. It also enhances the risk assessment and breach notification portions of HIPAA, covering the gaps in the protection of PHI by business associates of covered entities.

CMS supports HIPAA and HITECH legislation with the Promoting Interoperability Programs. The Medicare Access and CHIP Reauthorization Act of 2015 (MACRA) and Merit-based Incentive Payment System provide financial incentives to providers for promoting interoperability of EHRs.[9] Financial incentives may include higher reimbursement rates for services by providers meeting the higher level of care defined by this legislation and lower rates for providers who are not meeting expectations.[25]

The CMS Information Blocking Rule took effect in 2020 under the 21st Century Cures Act. This rule removes barriers to information access by giving patients control over how, when, and with whom patient data are shared.[9] It does this by creating new standards and mechanisms for healthcare providers, insurers, and software developers to enable patients access to their data.[9] As of 2021, 88% of office-based physicians have adopted an EHR.[26] This represents a doubling of adoption since 2008.[26] The improved accessibility to information supports the quantity of data required for adequate AI training.

While this legislation supports EHR adoption and information sharing, AI in health care is not addressed. In December 2023, the Department of Health and Human Services (HHS) finalized a rule to advance health IT interoperability and algorithm transparency. This rule first establishes transparency requirements for AI and other predictive algorithms.[27,28] It also adopts the United States Core Data for Interoperability version 3. This data standard focuses on the accuracy and completeness of patient data with the goal of "promot[ing] equity, reduc[ing] disparities, and support[ing] public health data interoperability."[28] The rule also enhances information blocking requirements to "encourage secure, efficient, standards-based exchange of electronic health information under the Trusted Exchange Framework and Common Agreement^{SM} (TEFCA^{SM})."[28] Lastly, the rule implements the 21st century Cures Act's "Insights Condition," requiring focused reporting metrics by health IT program developers to further support interoperability. This timely legislation is the key to the structured collection of open data and regulation of AI transparency in health care to ensure patient safety.

International Legislation

Serving as a model for many other countries, the General Data Privacy Regulation (GDPR) passed by the European Union (EU) in 2016 is similar to the protections provided by HIPAA and HITECH in the United States. Organizations may use patient data for legitimate purposes, such as diagnoses or billing, and maintain it until that purpose is complete.[9] Patients must be informed of how the data

are being used and be able to amend them if there are errors or updates to their medical history.[9] The GDPR requires security provisions but allows data sharing with third parties.[9] As with HIPAA, GDPR does not cover the third party; the organization sharing the data is responsible for the third party's use. This is demonstrated with healthcare organizations that outsource tasks like billing, requiring the sharing of PHI for the third party, the billing company, to bill the patient.

Canada protects PHI like the EU, with the addition that individual provinces maintain specific privacy laws. The Personal Information Protection and Electronic Documents Act requires any organization holding personal data to abide by 10 principles. The principles are accountability, identifying purposes, consent, limiting collection, limiting use, disclosure and retention, accuracy, safeguards, openness, individual access, and challenging compliance.[9] These principles effectively provide the same safeguards seen in the United States and EU.

Health care is typically limited to a patient's home country, but as globalization continues, the value of having open data that includes other countries is apparent. From a medical perspective, more data provide a more robust dataset. From the point of view of AI developers, systems trained on global datasets open the possibility of worldwide software interoperability.

DISCUSSION

AI is a disruptive technology rapidly changing the complexity of how computer systems improve upon the computational limitations of the human brain. The observed and potential benefits of using AI in health care are apparent. AI systems trained to read radiology imaging can perform as well as radiologists.[2] The accuracy and efficiency of emergency medicine are improved with ML models focused on patient outcome predictions, risk stratification, and triage optimization.[1] The availability of better training data and AI models to perform more tasks will only expand the possibilities for AI to improve health care.

AI is a technology that will impact everyone, so everyone must have a general understanding of how it works and affects health care. Health and digital literacies are vital to competent regulation, health professionals' utilization of technology, and patient participation as a component of patient-centered care.

At the foundation of any AI system are the data used to train it. Homogenous, or single-source, data are insufficient but have played an essential role in early AI development in health care. Data collected by a single institution are a cost-effective way to develop a system for in-house use. However, the data are limited in quantity, likely to contain bias, and lack reproducibility.[1] Open data enhances the quality and quantity of data by including patient data across different geographic locations.

Many open databases exist to support AI research, including in health care. As with COVIDx, some databases may be combined to provide the data needed to train new AI systems adequately.[22] Open data solves the issue of bias demonstrated by AI systems trained on internal data. It also provides the ability to independently validate new systems, improving the development of new AI models across healthcare facilities. While the case study of COVIDx demonstrated how unidentified errors in open data can be propagated through continued research use, regulatory efforts continue to improve the quality of the data, reducing this risk.[22,28]

The scope and complexity of AI in health care dictate a level of regulatory oversight to ensure patient safety. Interoperability legislation supports data transmission to an open database. HIPAA and MACRA address privacy, security, and confidentiality concerns. Most recently, the HHS rule to advance health IT interoperability and algorithm transparency addresses the salient issues of AI in health care. Requiring transparency of AI models and improving the quality of training data support the interoperability of open data.

International governments and major research funders also support open data while addressing the security of patient information. Models, such as the EU's GDPR, may guide future legislation. International open data efforts like NFDI support global research and worldwide interoperability. Funding programs like Plan S redirect all science initiatives to open data and open science. The availability of specialized datasets shows a willingness to pool and share data. The COVIDx dataset shows how data are shared during critical public health emergencies. These initiatives are all essential to successfully creating open data for use in health care.

As open data is compiled, concerns for patient privacy and the security of stored data remain essential. Training data containing PHI can and must be protected. Anonymizing data is the first step to patient privacy.[23] AI programming methods, such as differentially programmed ML, can further protect data by obscuring patient data with noise, making it unidentifiable.[23,24]

Despite significant advancements in AI in health care, it is not ready to be universally applied. The security and quality of training data must be prioritized. Community standards must be developed for efficient data sharing.[6] High-quality data and data standards are essential for sharing code and data. Data standardization can allow independent datasets to be merged into large open datasets. The HITECH Act laid the foundation for interoperability in health care, and interoperability between hospitals and state databases already exists in infectious disease reporting. However, applying the HHS rule to advance health IT interoperability and algorithm transparency will ensure the universal application of transparency and data quality standards.

In addition to the HHS rule, a single reference for open data is needed. Studies are available via PubMed and other

sources that exist to identify and evaluate open data sources based on the healthcare application.[29-31] Researchers would benefit from a single regulatory body that indexes all open data initiatives and holds them accountable to the HHS rule.

Sharing curated data and trained AI models will exponentially increase AI development.[6] AI tools and sharing practices that aid data gathering should be embraced to support this growth. NLP and computer vision are AI tools that review human-readable text and transform it into a machine-readable format.[6] These technologies can improve data gathering from nontypical datasets like clinical notes or scientific articles.

Public grants should be directed at research on AI models in health care that utilize Plan S and the fundamentals of F1000 research. The requirement to provide all data will support using open data rather than homogenous training data, and the funding will direct the study specifically toward AI in health care.

CONCLUSION

AI is designed to advance health care in the same way the invention of the microscope opened possibilities previously unimagined. Several uses of AI in health care have already seen success and approval for use in clinical settings. The widespread success of AI in health care will ultimately depend on safety and efficacy. Researchers and governments highlight the importance of open data with significant strides to improve access to data. Ultimately, using heterogeneous data from public databases to implement safe, reproducible AI models in health care is essential.

REFERENCES

1. Piliuk K, Tomforde S. Artificial intelligence in emergency medicine. A systematic literature review. *Int J Med Inform*. 2023;180:105274. doi: 10.1016/j.ijmedinf.2023.105274
2. Mueller B, Kinoshita T, Peebles A, Graber MA, Lee S. Artificial intelligence and machine learning in emergency medicine: a narrative review. *Acute Med Surg*. 2022;9(1):e740. doi: 10.1002/ams2.740
3. Chan SL, Lee JW, Ong MEH, et al. Implementation of prediction models in the emergency department from an implementation science perspective-determinants, outcomes, and real-world impact: a scoping review. *Ann Emerg Med*. 2023;82(1):22–36. doi: 10.1016/j.annemergmed.2023.02.001
4. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. *Comput Biol Med*. 2023;166:107555. doi: 10.1016/j.compbiomed.2023.107555
5. Allgaier J, Mulansky L, Draelos RL, Pryss R. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artif Intell Med*. 2023;143:102616. doi: 10.1016/j.artmed.2023.102616
6. Brinkhaus HO, Rajan K, Schaub J, Zielesny A, Steinbeck C. Open data and algorithms for open science in AI-driven molecular informatics. *Curr Opin Struct Biol*. 2023;79:102542. doi: 10.1016/j.sbi.2023.102542
7. Else H. A guide to Plan S: the open-access initiative shaking up science publishing. *Nature*. Published online April 8, 2021. doi: 10.1038/d41586-021-00883-6
8. Electronic health records. U.S. Centers for Medicare & Medicaid Services. September 6, 2023. Accessed February 14, 2024. https://www.cms.gov/priorities/key-initiatives/e-health/records.
9. Cypko MA. *Development of Clinical Decision Support Systems Using Bayesian Networks: With an Example of a Multi-Disciplinary Treatment Decision for Laryngeal Cancer*. 1st ed. Springer Vieweg; 2020.
10. Jeyaraman M, Balaji S, Jeyaraman N, Yadav S. Unraveling the ethical enigma: artificial intelligence in healthcare. *Cureus*. 2023;15(8):e43262. doi: 10.7759/cureus.43262
11. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: an overview for clinical practitioners - saliency-based XAI approaches. *Eur J Radiol*. 2023;162:110787. doi: 10.1016/j.ejrad.2023.110787
12. Kumar V, Srivastava P, Dwivedi A, et al. Large-language-models (LLM)-based AI chatbots: architecture, in-depth analysis and their performance evaluation. In: *International Conference on Recent Trends in Image Processing and Pattern Recognition*. Springer Nature Switzerland; 2023:237–249.
13. Wiemken TL, Carrico RM. Assisting the infection preventionist: use of artificial intelligence for health care-associated infection surveillance. *Am J Infect Control*. 2024;52(6):625–629. doi: 10.1016/j.ajic.2024.02.007
14. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *J Med Internet Res*. 2023;25:e46924. doi: 10.2196/46924
15. Valverde JM, Imani V, Abdollahzadeh A, et al. Transfer learning in magnetic resonance brain imaging: a systematic review. *J Imaging*. 2021;7(4):66. doi: 10.3390/jimaging7040066
16. F1000Research. f1000research.com. 2024. Accessed March 17, 2024. https://f1000research.com/.
17. Rodgers CM, Ellingson SR, Chatterjee P. Open data and transparency in artificial intelligence and machine learning: a new era of research. *F1000Res*. 2023;12:387. doi: 10.12688/f1000research.133019.1
18. European Science Foundation. "Plan S" and "cOAlition S" – Accelerating the transition to full and immediate Open Access to scientific publications. Plan S. 2024. Accessed February 17, 2024. https://www.coalition-s.org/.
19. U.S. General Services Administration. Open government. Data.gov. Accessed February 18, 2024. https://data.gov/open-gov/.
20. Consortia NFDI4Health. National Research Data Infrastructure Germany. Accessed March 19, 2024. https://www.nfdi.de/consortia-nfdi4health/?lang=en.
21. SODA Developers. Tyler Technologies. Accessed March 19, 2024. https://dev.socrata.com/.
22. Harkness R, Hall G, Frangi AF, Ravikumar N, Zucker K. The pitfalls of using open data to develop deep learning solutions for COVID-19 detection in chest X-rays. *Stud Health Technol Inform*. 2022;290:679–683. doi: 10.3233/SHTI220164
23. Gong M, Xie Y, Pan K, Feng K, Qin AK. A survey on differentially private machine learning. *IEEE Comput Intell Mag*. 2020;15(2):49–64. doi: 10.1109/MCI.2020.2976185
24. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: *Theory of Cryptography Conference*. 3rd ed. Springer Berlin Heidelberg; 2006.
25. The U.S. Centers for Medicare & Medicaid Services. Quality Payment Program Overview. Quality Payment Program.

Accessed March 24, 2024. https://qpp.cms.gov/about/qpp-overview.
26. Office of the National Coordinator for Health Information Technology. Office-based physician electronic health record adoption. HealthIT.gov. Accessed March 3, 2024. www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption.
27. Federal Register. *Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing*. 2024. 89 FR 1192.
28. The Office of the National Coordinator for Health Information Technology. Health data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing (HTI-1) final rule. HealthIT.gov. December 13, 2023. Accessed March 3, 2024. www.healthit.gov/topic/laws-regulation-and-policy/health-data-technology-and-interoperability-certification-program.
29. Lim L, Lee HC. Open datasets in perioperative medicine: a narrative review. *Anesth Pain Med (Seoul)*. 2023;18(3):213–219. doi: 10.17085/apm.23076
30. Sauer CM, Dam TA, Celi LA, et al. Systematic review and comparison of publicly available ICU data sets-a decision guide for clinicians and data scientists. *Crit Care Med*. 2022;50(6):e581–e588. doi: 10.1097/CCM.0000000000005517
31. Fong N, Langnas E, Law T, Reddy M, Lipnick M, Pirracchio R. Availability of information needed to evaluate algorithmic fairness - a systematic review of publicly accessible critical care databases. *Anaesth Crit Care Pain Med*. 2023;42(5):101248. doi: 10.1016/j.accpm.2023.101248