Viewpoint

Assessing the Evolution and Influence of Medical Open Databases on Biomedical Research and Health Care Innovation: A 25-Year Perspective With a Focus on Privacy and Privacy-Enhancing Technologies

Albert Yang[1,2,3], MD, PhD; Mei-Lien Pan[4], PhD; Henry Horng-Shing Lu[5,6,7,8,9], PhD; Chung-Yueh Lien[10], PhD; Da-Wei Wang[11†], PhD; Chih-Hsiung Chen[12], PhD; Der-Cherng Tarng[2,13,14,15], MD, PhD; Dau-Ming Niu[2,14,16], MD, PhD; Shih-Hwa Chiou[3,17], MD, PhD; Chun-Ying Wu[18,19], MD, PhD; Ying - Chou Sun[20], MD; Shih-Ann Chen[2,21], MD; Shuu-Jiun Wang[2,22,23], MD; Wayne Huey-Herng Sheu[24], MD, PhD; Chi-Hung Lin[25], MD, PhD

[1]Digital Medicine and Smart Healthcare Research Center, National Yang Ming Chiao Tung University, Taipei City, Taiwan
[2]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[3]Department of Medical Research, Taipei Veterans General Hospital, Taipei, Taiwan
[4]Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei, Taiwan
[5]Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[6]Biomedical Artificial Intelligence Academy, Kaohsiung Medical University, Kaohsiung, Taiwan
[7]Department of Artificial Intelligence in Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan
[8]Department of Medical Research, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan
[9]Department of Statistics and Data Science, Cornell University, Ithaca, NY, United States
[10]Department of Information Management, National Taipei University of Nursing and Health Sciences, Taipei, Taiwan
[11]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[12]Institute of Technology Law, National Yang Ming Chiao Tung University, Taipei, Taiwan
[13]Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan
[14]Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[15]Department and Institute of Physiology, National Yang Ming Chiao Tung University, Taipei, Taiwan
[16]Department of Pediatrics, Taipei Veterans General Hospital, Taipei, Taiwan
[17]Institute of Pharmacology, College of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[18]Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan
[19]Health Innovation Center, National Yang Ming Chiao Tung University, Taipei, Taiwan
[20]Department of Radiology, Taipei Veterans General Hospital, Taipei, Taiwan
[21]Department of Cardiovascular Center and Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan
[22]Department of Neurology, Neurological Institute, Taipei Veterans General Hospital, Taipei, Taiwan
[23]Brain Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan
[24]Institute of Molecular and Genomic Medicine, National Health Research Institutes, Taipei, Taiwan
[25]Department of Biological Science & Technology, National Chiao Tung University, Taipei, Taiwan
[†]deceased

Corresponding Author:

Albert Yang, MD, PhD
Digital Medicine and Smart Healthcare Research Center
National Yang Ming Chiao Tung University
No. 155 Sec. 2 Linong St., Beitou Dist
Taipei City 112
Taiwan
Phone: 886228267995
Email: accyang@nycu.edu.tw

Abstract

The integration of medical open databases with artificial intelligence (AI) technologies marks a transformative era in biomedical research and health care innovation. Over the past 25 years, initiatives like PhysioNet have revolutionized data access,

fostering unprecedented levels of collaboration and accelerating medical discoveries. This rise of medical open databases presents challenges, particularly in harmonizing research enablement with patient confidentiality. In response, privacy laws such as the Health Insurance Portability and Accountability Act have been established, and privacy-enhancing technologies have been adopted to maintain this delicate balance. Privacy-enhancing technologies, including differential privacy, secure multiparty computation, and notably, federated learning (FL), have become instrumental in safeguarding personal health information. FL, in particular, represents a significant advancement by enabling the development and training of AI models on decentralized data. In Taiwan, significant strides have been made in aligning with these global data-sharing and privacy standards. We have actively promoted the sharing of medical data through the development of dynamic consent systems. These systems enable individuals to control and adjust their data-sharing preferences, ensuring transparency and continuity of consent in the ever-evolving landscape of digital health. Despite the challenges associated with privacy protections, the benefits, including improved diagnostics and treatment, are substantial. The availability of open databases has notably accelerated AI research, leading to significant advancements in medical diagnostics and treatments. As the landscape of health care research continues to evolve with open science and FL, the role of medical open databases remains crucial in shaping the future of medicine, promising enhanced patient outcomes and fostering a global research community committed to ethical integrity and privacy.

J Med Internet Res 2026;28:e58954; doi: 10.2196/58954

Keywords: medical open databases; artificial intelligence; privacy-enhancing technologies; federated learning; dynamic consent systems

Introduction

The emergence of medical open databases, coupled with advances in artificial intelligence (AI), heralds a significant change in biomedical research and health care innovation, facilitating an era of enhanced accessibility and data sharing [1-3]. This movement toward open data science, augmented by AI technologies, enables researchers worldwide to access a wealth of data, including physiological signals [4,5], genomic [6], and health care information [7], and, most prominently, large-scale medical imaging archives. While this review covers the broad spectrum of medical data, the impact of open imaging databases has been particularly transformative for the application of AI. This movement toward open data science fosters collaboration and speeds up the pace of medical discoveries.

AI's role in analyzing vast datasets has been instrumental in uncovering patterns and insights that would be impossible for humans to detect unaided, leading to breakthroughs in understanding diseases and patient care. Initiatives like annual challenges and shared toolboxes have spurred the development of novel algorithms and techniques, leveraging AI to address complex biomedical challenges and advance medical diagnostics and treatments. This synergy between open medical databases and AI is transforming the landscape of health care, promising a future of more accurate, efficient, and personalized medicine.

Simultaneously, this rise in open data repositories brings to the forefront crucial privacy concerns [6,8]. The necessity to balance the imperative of research enablement with the protection of patient confidentiality has never been more pronounced. In this context, laws such as the Health Insurance Portability and Accountability Act (HIPAA) play a pivotal role in shaping the landscape of data deidentification and anonymization processes, ensuring that shared data comply with strict privacy standards [9]. Moreover, the introduction of privacy-enhancing technologies (PETs), such as differential privacy [10], synthetic data [11], homomorphic encryption [12], secure multiparty computation [13], and federated learning [14], represents a proactive approach to safeguarding personal health information. These technologies provide the means to conduct meaningful research while upholding the principles of data privacy and security.

In a country like Taiwan, strides in medical data sharing suggest the global shift toward interconnected health systems, highlighting both advancements and ongoing challenges in securing patient data. The implementation of dynamic consent frameworks reflects a growing recognition of the need for more flexible approaches to data privacy, particularly in an era of personalized medicine and digital health records [15]. As the landscape of medical research evolves with these developments, the interplay of data sharing, privacy, and technology continues to reshape the boundaries of what is possible in health care innovation, marking a critical junction in the journey toward more open, collaborative, and ethically responsible research environments.

This review aims to provide a 25-year perspective on the evolution of medical open databases, tracing their impact on biomedical research and health care innovation, and to examine how emerging PETs and data-governance frameworks, including dynamic consent systems, shape the ethical, technical, and collaborative landscape of digital medicine worldwide.

The Rise of Medical Open Database

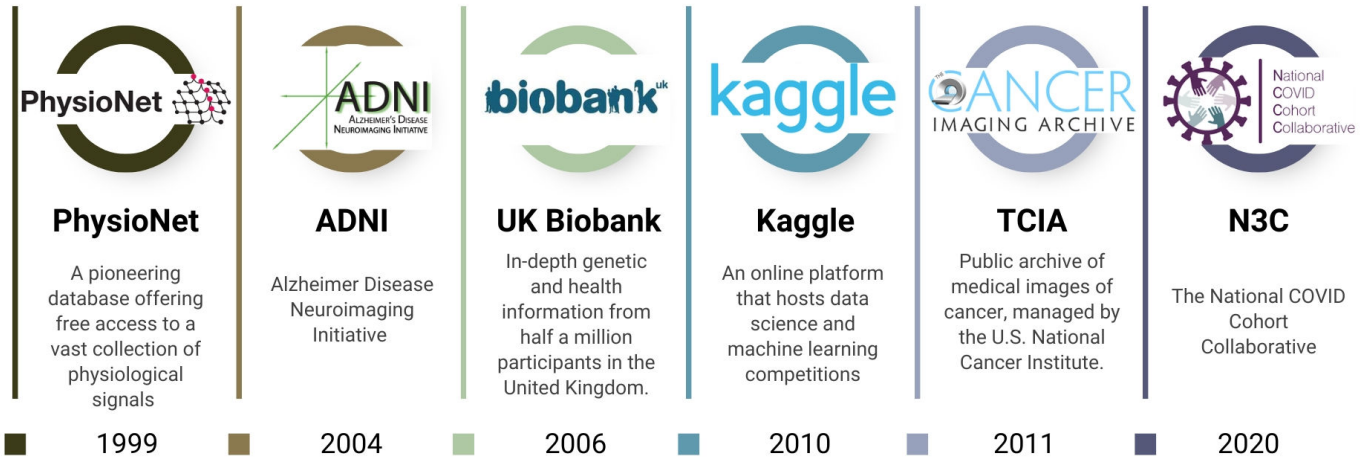
The rise of medical open databases represents a transformative shift in the landscape of biomedical research and health care innovation. Among the pioneers in this movement is PhysioNet. Established in 1999, PhysioNet is a pioneering open database that provides free access to a wide range of physiologic signals and related open-source software for research in medicine, physiology, and biomedical engineering [4,16]. It was initiated by a collaborative project involving

researchers from Boston's Beth Israel Deaconess Medical Center/Harvard Medical School, Boston University, McGill University, and Massachusetts Institute of Technology [4,17]. The database contains a diverse collection of physiological datasets, including those related to cardiovascular and other complex biomedical signals [4,18]. PhysioNet has had a significant impact on the development of medical open databases, serving as a model for the establishment of similar resources. It has also played a key role in promoting the dissemination and exchange of medical resources.

A significant contribution of PhysioNet to the scientific community is its annual PhysioNet Challenge, which has markedly influenced the field by promoting innovation and collaboration among researchers and clinicians. These challenges stimulate the creation of novel algorithms and methods aimed at solving complex biomedical problems, thus expanding the limits of what can be achieved in medical data analysis and application. For instance, the challenges have catalyzed the development of innovative algorithms capable of detecting obstructive sleep apnea from electrocardiograms [19], illustrating the practical impact of these competitions on advancing medical diagnostics and treatment strategies [20].

Since the success of PhysioNet, numerous other medical open databases have emerged globally, fostering a more cooperative and transparent research atmosphere (Figure 1). The impact of these large-scale databases on biomedical discovery is profound. One notable example is the UK Biobank [21], launched in 2006, which provides a vast repository of genetic and health information from half a million UK participants. This database has become an essential tool for unraveling the complex interplay between genetics, lifestyle, and disease, thereby enhancing our understanding of the factors influencing human health [22-25]. By leveraging large neuroimaging cohorts such as Alzheimer Disease Neuroimaging Initiative and the UK Biobank, researchers have developed AI-based models that generate an Alzheimer disease risk score from structural magnetic resonance imaging (MRI), enabling the identification of prediagnostic populations suitable for early intervention and preventive trials [26]. In cardiovascular research, analysis of the UK Biobank's genetic and imaging data has enabled the development of NeuralCVD, a neural network–based risk model that integrates polygenic and clinical predictors to estimate the 10-year risk of major adverse cardiac events, improving risk discrimination and reclassification beyond established clinical scores and Cox models, and highlighting the added predictive value of genetic predisposition in early prevention [27]. Similarly, the Cancer Imaging Archive, inaugurated in 2011 in the United States, offers a dedicated platform for the cancer research community, enabling access to a comprehensive array of imaging datasets [28].

Figure 1. Historical development of major medical open databases and data-sharing platforms over the past 25 years.



Beyond these examples, the ecosystem of medical open databases has diversified into numerous specialized domains. For instance, OpenNeuro provides a vast repository for neuroimaging data, particularly functional magnetic resonance imaging, electroencephalogram, and magnetoencephalography, supporting reproducible brain research [29]. The Neuroimaging Informatics Tools and Resources Clearinghouse offers a rich collection of imaging and data-processing tools [30]. Similarly, the National Database for Autism Research [31] and the Federal Interagency Traumatic Brain Injury Research informatics system [32] provide deeply phenotyped datasets crucial for research in their respective fields. These platforms underscore the field's shift toward creating specialized, high-quality resources to tackle specific biomedical questions. They display how shared resources can drive forward innovation and improve patient care worldwide, illustrating the critical role of collaborative environments in the advancement of health care research and application.

Additionally, Kaggle, an online platform for data science and machine learning (ML) competitions, has emerged as a pivotal player in the field of ML analysis and data sharing [33]. Launched in 2010, Kaggle facilitates collaboration and competition among data scientists and researchers by hosting challenges in various domains, including health care. These competitions often involve complex medical datasets, encouraging participants to develop innovative solutions and algorithms for disease prediction, medical imaging analysis [34], and other health-related issues. Kaggle has not

only democratized access to large medical datasets but has also fostered a global community where knowledge and techniques are openly shared. This environment has led to significant breakthroughs and advancements in medical research and analytics, further suggesting the importance of open data and collaborative problem-solving in improving health care outcomes and accelerating medical innovation.

Balancing Privacy and the Need for Medical Open Databases

The success of open medical databases such as PhysioNet poses the challenge of balancing patient confidentiality with research enablement on open platforms [35,36]. Research datasets in health care often contain protected health information (PHI), and the process of removing this information, a process known as deidentification or anonymization, can be challenging and prone to errors [37]. Despite the use of these datasets, the need for deidentification introduces a significant barrier to data sharing due to the effort and cost involved.

The HIPAA, established in the United States in 1996, plays a vital role in safeguarding patients' medical information. In response to the HIPAA mandate, U.S. Department of Health and Human Services published a final regulation in the form of the privacy rule in December 2000, which became effective on April 14, 2001. Central to this rule is the designation of 18 specific categories of PHI that, if disclosed, could be used to identify an individual (Textbox 1). These categories encompass a broad spectrum of personal data, including, but not limited to, names, geographic details smaller than a state, various identifiers like social security numbers, medical record numbers, and contact information, as well as certain biometric and photographic images [38].

Textbox 1. Eighteen categories of protected health information.

- Names
- All geographic subdivisions smaller than a State
- All elements of dates (except year) for dates directly related to an individual
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers
- Device identifiers and serial numbers
- Web URLs
- IP address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code

Additionally, HIPAA mandates that covered entities must ensure they do not possess knowledge that the remaining information could be used, whether alone or in conjunction with other data, to identify the subject. By strictly adhering to these guidelines, entities can share deidentified health information for broader uses, such as public health and research, without infringing on individual privacy rights, thus striking a balance between privacy protection and the beneficial use of health data [39]. Despite these efforts, the tension between the promise of big data and patient privacy in health care research remains a challenge [40].

PhysioNet is a pioneer in medical public databases, ensuring that the datasets it provides do not compromise individual privacy. This involves ensuring that any data shared does not contain PHI or has been sufficiently anonymized to prevent the identification of individuals. The challenges posed by the HIPAA privacy rule are not insignificant; they include the need for informed consent from data subjects and potential limitations on access to health information that can hinder clinical research [41,42]. Furthermore, the rule's interaction with other regulations, like the common rule, adds complexity to privacy concerns in research, leading to inconsistencies and additional burdens for researchers.

Despite these challenges, the privacy rule does allow for certain disclosures without patient authorization, particularly for public health purposes. This is intended to facilitate the use of medical data in important public health endeavors without undermining individual privacy protections [43]. The balance sought by the HIPAA privacy rule between protecting privacy and facilitating research is a critical aspect of its implementation, particularly in the context of medical open databases. By navigating these regulations successfully, repositories can contribute to the advancement of medical research while ensuring compliance with privacy standards [44].

The emergence of medical databases, such as the PhysioNet, UK Biobank, and the Cancer Imaging Archive,

has significantly advanced collaborative research in health care [45,46]. These databases have the potential to transform cancer research and improve patient outcomes [45]. However, the collection, linking, and use of data in biomedical research raise ethical concerns, particularly regarding privacy and security [36,47,48]. Despite these concerns, the benefits of open data in health care, including improved diagnostics and treatment, are substantial [48]. The push for data sharing in cancer trials by pharmaceutical companies further underscores the importance of open medical databases in driving innovation and improving patient care [49].

Privacy Enhancing Technology

Overview

A range of studies have been conducted to explore the increasing frequency and impact of health care data breaches, highlighting the rising number of incidents and their detrimental effects on patient privacy and health care providers [50-53]. These breaches are often caused by a combination of technical, organizational, and human factors [50-52]. Human vulnerabilities, such as lack of awareness and training, play a significant role in these breaches [51]. The use of the Swiss Cheese Model can help assess vulnerabilities and risks [50]. Cloud computing breaches are a particular concern, highlighting the need for digital forensic readiness [54]. Hacking and unauthorized internal disclosures are the most prevalent forms of attack [53]. Further studies may examine specific cases and the implications for digital forensic readiness, emphasizing the importance of adhering to regulations.

Below, we reviewed several PETs and their applications in enhancing data privacy and security in health care settings (Table 1). PETs, such as encryption, anonymization techniques, and secure multiparty computation, offer powerful mechanisms to protect sensitive health data. Implementing these technologies, alongside robust privacy policies and employee training, can significantly reduce the likelihood of data breaches and bolster the trust between patients and health care providers.

Table 1. Summary of privacy-enhancing technologies.

Technologies	Core principle	Advantages	Challenges and trade-offs
Differential privacy	Adds calibrated statistical noise to query results to make it impossible to determine if an individual's data were included.	Provides strong and mathematically provable privacy guarantees.	Inherent trade-off between privacy and data use; high privacy can reduce analytical accuracy.
Synthetic data	Creates an artificial dataset that mimics the statistical properties of the original data without containing real patient information.	High use for model training; no real patient data are shared, eliminating reidentification risk.	Can be difficult to generate high-fidelity data that captures all complex correlations; potential for model bias.
Homomorphic encryption	Allows computations to be performed directly on encrypted data without decrypting it first.	Offers extremely strong security, as the raw data are never exposed.	High computational overhead; currently too slow for many complex ML[a] tasks.
Secure multiparty computation	Enables multiple parties to jointly compute a function over their inputs while keeping those inputs private.	Allows for collaborative analysis without a central data repository; no single party sees another's data.	High communication overhead between parties; can be complex to set up and scale.
Federated learning	Trains a central AI[b] model across decentralized devices or servers holding local data samples, without exchanging the data itself.	Keeps raw data local, enhancing privacy and data sovereignty.	Vulnerable to model poisoning/inversion attacks; performance can degrade with heterogeneous data.

[a]ML: machine learning.
[b]AI: artificial intelligence.

Differential Privacy

Differential privacy, a method for protecting individual privacy in data analysis, has been increasingly applied in the health care sector. It involves adding noise to the data to prevent reidentification of individuals. This approach has been used in various areas of health research, including genomics, neuroimaging, and health surveillance [55]. However, there are challenges in its practical application, such as the theoretical nature of the privacy parameter epsilon [56]. To address these challenges, researchers have proposed differentially private data release strategies and noise mechanisms, such as the Laplace and exponential mechanisms [57]. However, a key challenge is the inherent trade-off between privacy and data use; increasing the amount of statistical noise to protect privacy can reduce the accuracy of analytical outcomes.

The application of differential privacy in medical questionnaires has also been explored, with the randomized response mechanism showing promise in improving privacy while retaining data use [58]. Furthermore, the use of differential privacy in geospatial analyses of standardized health care data has been demonstrated, with the development of geodatabase functions for privacy-aware analysis [59]. Finally, the combination of differential privacy and decision tree approach has been proposed for data publishing, and the differentially private mini-batch gradient descent algorithm for model publishing of medical data [60].

Synthetic Data

Synthetic data, generated through simulators, is increasingly used in health care to address the challenges of data availability and privacy [61]. PETs, such as differential privacy, are combined with synthetic data generators to create private synthetic data, preserving statistical properties while ensuring

privacy [62]. These technologies have been applied in various use cases, including clinical risk prediction [63] and medical research [64]. However, the evaluation of synthetic data's privacy and use metrics remains a challenge, with a lack of consensus on standard approaches [65]. Despite these challenges, the potential of synthetic data in preserving data use and patient privacy in electronic health care data is being explored [66].

Homomorphic Encryption

Homomorphic encryption, a powerful tool for preserving privacy in medical data, allows for computations to be performed on encrypted data without the need for decryption. It has been successfully applied in various medical data scenarios, including ML models for classification and training, secure genomic algorithms, and predictive analysis tasks [67-69]. For example, it has been used to securely manage personal health metrics data, process medical images [70,71], and enable secure medical computation [72]. The use of homomorphic encryption in these applications ensures that sensitive medical data remains private and secure. Despite its power, the primary limitation of homomorphic encryption is its significant computational overhead, which can make it slow and resource-intensive for complex computations on large datasets.

Secure Multiparty Computation

Secure multiparty computation is a cryptographic technique that enables data analytics without sharing the underlying data, making it a valuable tool for preserving privacy in medical data analysis [73]. It has been applied in various health care scenarios, including collaborative systems [74], statistical analysis of health data [75], and electronic medical record (EMR) data [75]. Secure multiparty computation has also been used in health care internet of things systems to handle privacy issues [76], prevent data disclosure in sensor networks [77], and enable the reuse of distributed electronic health data [75]. Furthermore, it has been applied to enable privacy-preserving query processing on EMRs [78]. Notably, secure multiparty computation has enabled research on highly sensitive data (such as HIV, rare diseases, and population genomics) that would otherwise be inaccessible due to privacy concerns.

Federated Learning

Federated learning (FL), a decentralized ML approach, is increasingly being applied in the medical field due to the sensitive and fragmented nature of health care data [14,79]. It allows for the collaborative development of ML models without sharing raw data, thus preserving privacy [80,81]. This approach has been used in various medical domains, including oncology and radiology, for tasks such as image analysis and disease prediction [81,82]. However, there are challenges to be addressed, such as data homogeneity and transparency [81]. Furthermore, FL can be vulnerable to security risks, such as model inversion attacks that attempt to reconstruct training data from the shared model updates, and require careful design to ensure robustness. Despite these challenges, FL shows promise in improving the efficiency and privacy of medical data processing [83-85].

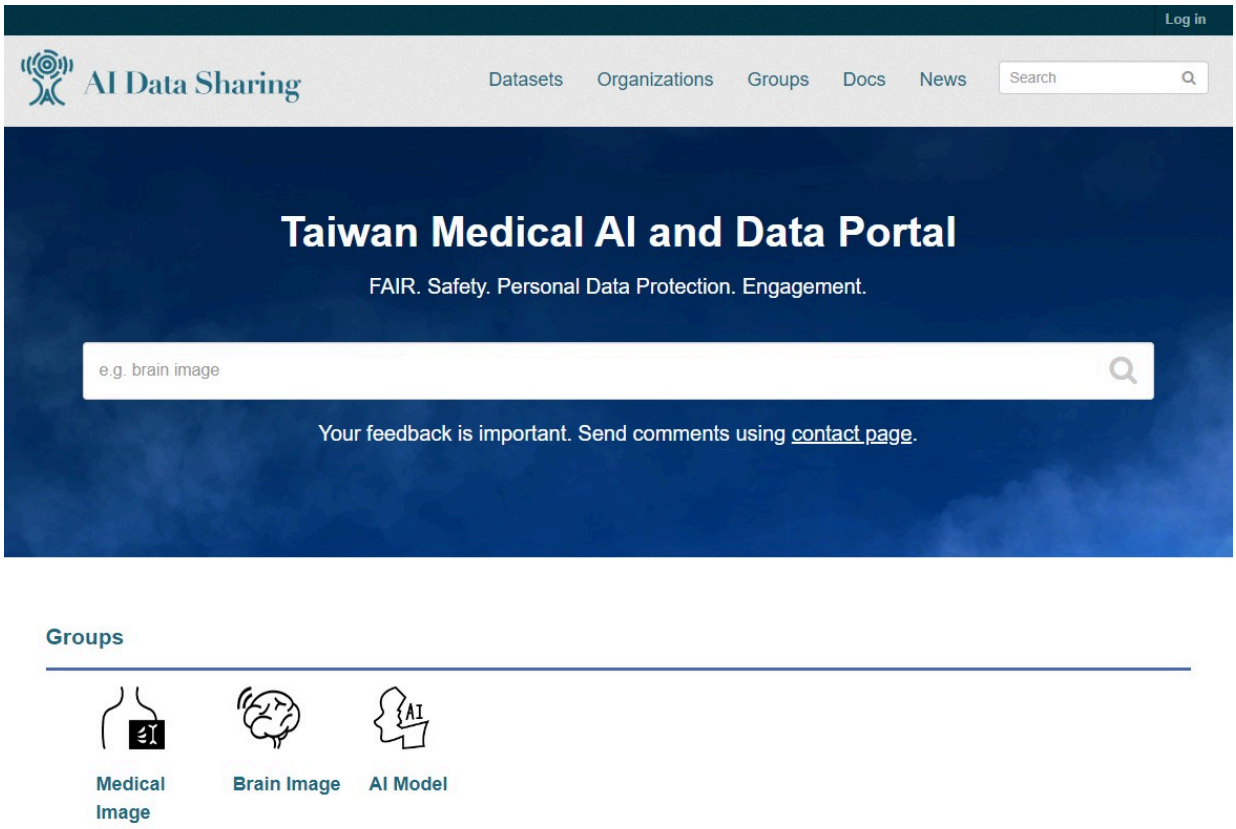
To address the challenge of data heterogeneity in FL, we have proposed the Dynamically Synthetic Images for Federated Learning method, significantly improving the conventional FL framework by integrating local information from local multiple institutions with heterogeneous data types [86]. The core principle of its implementation involves a dynamic process where, at the start of each training round, a client's local data are evaluated by the current global model to identify misclassified images. Using a synthetic minority oversampling technique, the system generates new, synthetic images based on these misclassified cases, which are then added to the local training set to compel the model to focus on features it previously failed to learn. In terms of effectiveness, experimental results demonstrated that Dynamically Synthetic Images for Federated Learning-based models achieve higher accuracy than conventional FL approaches and that their performance can be comparable to that of traditional centralized learning, proving especially beneficial for institutions with smaller or more heterogeneous datasets [86].

Taiwan Medical AI and Data Portal and Dynamic Consents System

Taiwan has made significant strides in medical data sharing, particularly in the areas of privacy protection and electronic health records exchange. The country's comprehensive embedded integrated circuit-based health insurance card system, implemented by the Bureau of National Health Insurance, Taiwan, allows for the secure sharing of health information [87]. The use of blockchain technology has been proposed as a means to further enhance the security and privacy of medical data sharing [88,89]. The Taiwan Electronic Medical Record Template and the National Electronic Medical Record Exchange System have been developed to facilitate the exchange of EMRs [90,91]. However, concerns about unauthorized access and secondary use of EMRs persist, particularly among highly educated individuals [92]. The country has also established guidelines for the security and privacy protection of health information, drawing on international best practices [93].

In the past 4 years, funded by the National Science and Technology Council of Taiwan, we have assembled teams from National Yang Ming Chiao Tung University, Taipei Veterans General Hospital, Academia Sinica, and National Taipei University of Nursing and Health Sciences to form a data repository task force known as the Smart Medical AI and Repository Taskforce Center. We launched a medical AI and data-sharing platform aimed at advancing the field of medical AI research in October 2023 [94]. This platform not only provides the public and researchers with access to a multitude of shared datasets but also ensures a meticulous evaluation process (Figure 2). Researchers can apply for access to the data by providing an abstract of their research proposal. Dataset managers assess applications based on their intended use, specific needs, and detailed research plans.

Figure 2. Taiwan medical artificial intelligence and data portal. AI: artificial intelligence.



Currently, seventeen datasets have been released on the platform, covering neuropsychiatric disorders, brain tumors, ophthalmic diseases, musculoskeletal disorders, and cardiopulmonary diseases. All datasets have undergone deidentification and delinking processes and include annotated information to facilitate AI training and validation. Specifically, our data-sharing platform includes: MRI images of vestibular schwannoma; computed tomography (CT) images of intracerebral hemorrhage; brain Fluorodeoxyglucose-Positron Emission Tomography/Magnetic Resonance Imaging images for dementia diagnosis; primary brain tumor MRI datasets, including meningioma, glioma, and pituitary adenoma; MRI data of brain metastases, which represent the largest collection nationwide; hand and foot X-rays of rheumatoid arthritis; X-rays of compression fractures; spinal X-rays of ankylosing spondylitis; chest CT images and clinical data of atrial fibrillation patients; chest X-rays for lung cancer screening; annotated preoperative liver CT images; neck lymph node CT images with postoperative pathology results; the Taiwan Aging and Mental Illness Cohort brain imaging database; the dementia molecular imaging database; fundus image datasets for glaucoma; and fundus image datasets of polypoidal choroidal vasculopathy.

The data sharing platform is built on a comprehensive architecture designed to support AI research by integrating 3 core systems: a CKAN-based sharing platform for dataset management, a data application system, and a dynamic authorization consent platform for patient privacy. Specific features include a robust user authentication and authorization mechanism, allowing dataset managers to grant access to specific users or collaborators. The platform ensures data integrity and ethical compliance through a multistep deidentification process for all medical images and by linking to the dynamic consent system (for sensitive clinical data), which allows patients to manage their data sharing preferences in real-time. To use the database, researchers first search for datasets on the platform, then apply for access through a formal registration and review process. Once approved and authorized by the dataset manager, users can obtain a login key to programmatically access the data through standardized protocols, such as DICOMweb, ensuring a secure, convenient, and interoperable environment for third-party AI applications.

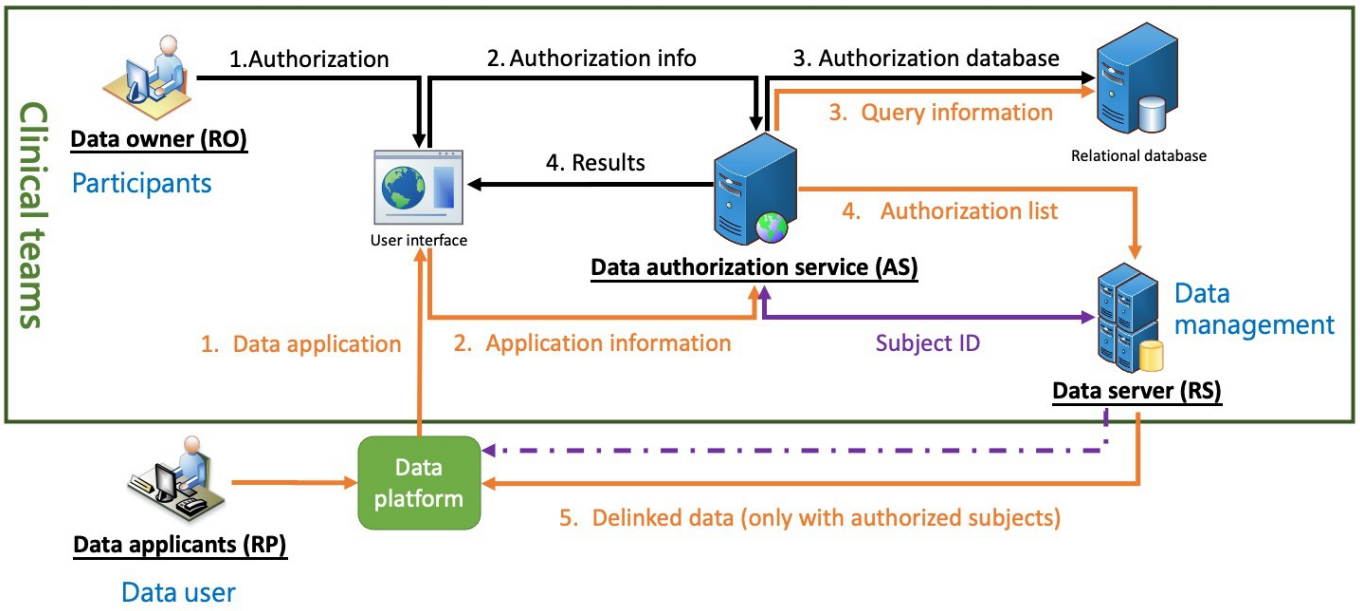
This effort aims to advance research across 7 crucial clinical areas that greatly benefit from AI technology, including heart disease, neurological disorders, mental illness, diabetes, cancer, genetic predispositions to complex diseases, and medical imaging. Moreover, the platform underpins collaboration between distinct teams specializing in AI methodology, science and law, and data governance, jointly fostering a robust data governance framework that emphasizes FL, cloud-based AI solutions, and trusted AI practices. Importantly, the system is designed to streamline the research process while maintaining a focus on ethical standards and participant privacy. In line with this, the platform incorporates a dynamic informed consent mechanism, especially for datasets that are anonymized but cannot be completely separated from their sources. This approach ensures that

participants' privacy is safeguarded while also enabling their informed and ongoing consent, reflecting our commitment to ethical research practices and the dynamic nature of consent in medical studies.

Dynamic informed consent, a concept that has been explored in various contexts, is a personalized, digital communication interface that allows participants to manage their consent preferences [95]. It has been proposed as a solution to improve patient confidence and trust in the use of electronic patient records in medical research [96]. In the context of privacy-aware pervasive health and well-being, dynamic consent enables granular data consent and management [97]. It has also been suggested as a potential solution to challenges in modern biomedical research, including participant recruitment, informed consent, and consent management [98]. The use of blockchain technology has been proposed to enhance the privacy-preserving aspects of dynamic consent in genomic data sharing [99]. The concept of dynamic informed consent has been further explored in the context of personalized medicine, emphasizing the need for a more dynamic and enriched consent model [100].

To enhance the privacy of participants contributing to the data in our platform and increase participants' engagement, we have developed a dynamic informed consent system, named the Health Data Authorization Service Platform (Figure 3). This collaborative effort involves our data governance, humanities, science and law, and clinical teams, aiming to facilitate scalable, dynamic consent operations suitable for complex data environments. The platform supports flexible data governance, allowing data owners to dynamically express their consent preferences, thereby making dynamic consent practical and sustainable.

Figure 3. Health data authorization service platform. This diagram depicts the architecture of a medical data-sharing platform that integrates dynamic consent for secure and transparent data sharing. The system centers on 4 key roles: the resource owner (RO), who owns the data; the resource server (RS), where medical data are stored and managed; the requesting party (RP), typically researchers seeking data access; and the authorization server (AS), the core component facilitating connections among the roles. This framework ensures data access is based on explicit consent from the data owner, allowing real-time adjustments to consent settings for different data types and uses, such as academic research. The platform's primary tasks are to establish individual preferences, maintain consent history, and ensure trust and transparency between data owners and users, thereby advancing responsible data science development.



The platform's architecture encompasses 4 key roles: the resource owner, usually the participants, who owns the data; the resource server, which stores and manages medical data; the requesting party, typically researchers seeking data access; and the authorization server, the backbone of our dynamic consent system, connecting the 3 roles. This system ensures that data are accessed only with the owner's express consent, respecting their preferences and enhancing data use transparency. Resource owners can modify their consent settings at any time, reflecting changes in their willingness to share different data types, like EMRs or medical imaging, for specific purposes such as academic research. This flexibility ensures that data use aligns with the owners' current preferences. The platform seamlessly integrates with our shared data framework, maintaining each citizen's consent history and enabling swift updates to consent forms as needed. By streamlining the consent process and ensuring data are shared according to owner permissions, our platform respects individual preferences while promoting responsible data science development. It exemplifies a forward-thinking approach to data governance, enabling real-time adjustments in consent and fostering a culture of trust and transparency between data owners and users.

These initiatives in Taiwan can be understood within the global context of evolving data privacy regulations. Unlike the "one-time, broad consent" model often used in US-based research under the Common Rule, Taiwan's move toward

dynamic consent aligns more closely with the principles of the European Union's General Data Protection Regulation [101]. The General Data Protection Regulation mandates that consent must be specific, informed, and easily revocable. The dynamic consent system builds on this by providing a technological interface for participants to manage their preferences granularly and continuously, representing a best-practice approach to balancing research needs with individual autonomy and privacy rights.

Acceleration of Medical AI Research Through Open Databases

The advent of medical open databases has significantly accelerated the field of medical AI research, fostering an environment of innovation and rapid development [102]. By providing researchers with access to vast amounts of health-related data, these databases have become a cornerstone for advancements in predictive analytics, diagnostic algorithms, and personalized medicine.

One of the most notable contributions of open medical databases to AI research is the democratization of data [103]. Historically, the scarcity and inaccessibility of medical data posed substantial hurdles to AI development. However, platforms like PhysioNet, established in 1999, have bridged this gap by offering a plethora of datasets ranging from physiological signals to clinical outcomes [104]. However, challenges remain, including the need for large datasets and the lack of external validation in perioperative medicine [105]. The use of open science approaches, including data liberation and crowdsourcing, can help address these challenges [106]. The integration of networked medical devices and clinical repositories based on open standards can further enhance AI research in high-acuity medical environments [107]. This enhanced availability allows researchers from diverse backgrounds and institutions to engage in health care innovation, leveling the playing field and stimulating a surge in AI-based solutions.

The availability of open databases has catalyzed the application of diverse AI methodologies to complex medical problems. Deep learning, particularly convolutional neural networks, has achieved state-of-the-art performance by leveraging large-scale imaging datasets; for example, researchers have trained convolutional neural networks on millions of images from The Cancer Imaging Archive to develop algorithms capable of detecting and classifying tumors in radiological scans with accuracy comparable to human experts [108]. For structured data such as the genetic and clinical information in the UK Biobank, traditional ML models like random forests and gradient boosting have been widely used, excelling at identifying complex patterns to predict disease risk, including the calculation of polygenic risk scores for coronary artery disease based on thousands of genetic variants [109]. In addition, natural language processing techniques have been applied to large repositories of unstructured clinical notes, such as those in the Medical Information Mart for Intensive Care version IV (MIMIC-IV) database (part of PhysioNet), to extract critical information on symptoms, treatments, and outcomes, thereby enabling large-scale retrospective studies that were previously infeasible [110].

Open medical databases encompass a wide variety of data types, including EMRs, imaging, genomic sequences, and more. This diversity enables AI researchers to explore multifaceted health care questions, from predicting disease trajectories to optimizing treatment plans. Moreover, the rich, varied datasets facilitate the training of more robust and generalizable AI models, capable of addressing complex medical scenarios across different populations and settings. The shared nature of open databases fosters collaboration across the global research community [111]. Through platforms that offer shared data, researchers can combine their expertise to tackle larger and more complex problems than they could individually. This collaborative approach has led to significant breakthroughs in AI, such as algorithms that can detect diseases from images with accuracy rivaling that of trained professionals [112,113].

Open databases also streamline the validation and implementation phases of AI development [114]. Access to diverse datasets enables researchers to rigorously test their algorithms under various conditions and patient demographics, ensuring their reliability and effectiveness. The expansion of these databases has significantly propelled medical AI research forward, marking a new phase of health care innovation with faster discoveries, collaborative efforts, and a commitment to ethical data use. As the field evolves, the role of open databases in shaping the future of medicine remains pivotal.

Conclusions

In conclusion, the evolution toward medical open databases, exemplified by the inception of platforms like PhysioNet in 1999 and their progression over the past 25 years, alongside the integration of PETs, marks a significant milestone in the domain of biomedical research and health care innovation. This journey not only fosters an unprecedented level of collaboration and accessibility but also emphasizes the crucial need to address privacy concerns and ethical considerations diligently. The ongoing efforts to balance data sharing with individual privacy protection are underscored by the adaptation of legal frameworks and the implementation of cryptographic and data management solutions. The introduction and growth of medical open databases have been pivotal, providing a wealth of data that has propelled research and innovation while highlighting the challenges and responsibilities of managing sensitive information. Specifically, the availability of open medical databases has significantly accelerated AI research, leading to breakthroughs in disease prediction, diagnostics, and personalized medicine. As we continue to explore the vast potential of open science and FL, the landscape of health care research is on the brink of remarkable transformations. These advancements promise enhanced patient outcomes, faster medical discoveries, and a more inclusive global research community, all achieved

by adhering to the highest standards of privacy and ethical integrity.

Acknowledgments

The authors thank Dr Watson Lin for supporting the infrastructure of the data-sharing platform and for providing valuable perspectives on the guidelines for medical data sharing. The authors also thank the National Science and Technology Council for offering legal insights on privacy policies related to medical data sharing.

Funding

This work was supported by grants from the National Science and Technology Council, Taiwan (grant number NSCT 114-2634-F-A49-006 and 113-2634-F-A49-003). ACY was also supported by the Mt. Jade Young Scholarship Award from the Ministry of Education, Taiwan, as well as Brain Research Center, National Yang Ming Chiao Tung University, and the Ministry of Education (Aim for the Top University Plan), Taipei, Taiwan.

Authors' Contributions

Conceptualization; writing – original draft: AY;
Writing – review and editing: All Authors;
Funding acquisition: CHL

Conflicts of Interest

None declared.

References

1. Merelli I, Pérez-Sánchez H, Gesing S, D'Agostino D. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. Biomed Res Int. 2014;2014:134023. [doi: 10.1155/2014/134023] [Medline: 25254202]
2. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Aff (Millwood). Jul 2014;33(7):1115-1122. [doi: 10.1377/hlthaff.2014.0147] [Medline: 25006136]
3. Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Change. Jan 2018;126:3-13. [doi: 10.1016/j.techfore.2015.12.019]
4. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. Jun 13, 2000;101(23):E215-20. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
5. Dean DA 2nd, Goldberger AL, Mueller R, et al. Scaling up scientific discovery in sleep medicine: the national sleep research resource. Sleep. May 1, 2016;39(5):1151-1164. [doi: 10.5665/sleep.5774] [Medline: 27070134]
6. Wylie JE, Mineau GP. Biomedical databases: protecting privacy and promoting research. Trends Biotechnol. Mar 2003;21(3):113-116. [doi: 10.1016/S0167-7799(02)00039-2] [Medline: 12628367]
7. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124-130. [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
8. Shahin MH, Bhattacharya S, Silva D, et al. Open data revolution in clinical research: opportunities and challenges. Clin Transl Sci. Jul 2020;13(4):665-674. [doi: 10.1111/cts.12756] [Medline: 32004409]
9. Ness RB, Joint Policy Committee, Societies of Epidemiology. Influence of the HIPAA privacy rule on health research. JAMA. Nov 14, 2007;298(18):2164-2170. [doi: 10.1001/jama.298.18.2164] [Medline: 18000200]
10. Dwork C, Roth A. The algorithmic foundations of differential privacy. FNT in Theoretical Computer Science. 2014;9(3-4):211-407. [doi: 10.1561/0400000042]
11. Rubin DB. Discussion: statistical disclosure limitation. J Off Stat. 1993;9:461-468. URL: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf [Accessed 2025-12-19]
12. Yi X, Paulet R, Bertino E. Homomorphic encryption. In: Homomorphic Encryption and Applications. Springer; 2014:27-46. [doi: 10.1007/978-3-319-12229-8_2] ISBN: 978-3-319-12228-1
13. Cramer RJF, Damgård IB, Nielsen JB. Secure Multiparty Computation and Secret Sharing. Cambridge University Press; 2015. [doi: 10.1017/CBO9781107337756] ISBN: 978-1-107-04305-3
14. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res. Mar 2021;5(1):1-19. [doi: 10.1007/s41666-020-00082-4] [Medline: 33204939]
15. Packer M. Data sharing in medical research. BMJ. Feb 14, 2018;360:k510. [doi: 10.1136/bmj.k510] [Medline: 29444885]

16. Moody GB, Mark RG, Goldberger AL. PhysioNet: physiologic signals, time series and related open source software for basic, clinical, and applied research. Annu Int Conf IEEE Eng Med Biol Soc. 2011;2011:8327-8330. [doi: 10.1109/IEMBS.2011.6092053] [Medline: 22256277]
17. Moody GB, Mark RG, Goldberger AL. PhysioNet: a research resource for studies of complex physiologic and biomedical signals. Presented at: Computers in Cardiology 2000; Sep 24-27, 2000; Cambridge, MA. [doi: 10.1109/CIC.2000.898485]
18. Henry IC, Goldberger AL, Moody GB, Mark RG. PhysioNet: an NIH research resource for physiologic datasets and open-source software. Presented at: 14th IEEE Symposium on Computer-Based Medical Systems CBMS 2001. Jul 26-27, 2001:IEEE Comput Soc. 245-250; Bethesda, MD. [doi: 10.1109/CBMS.2001.941728]
19. Thomas RJ, Mietus JE, Peng CK, Goldberger AL. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. Sleep. Sep 2005;28(9):1151-1161. [doi: 10.1093/sleep/28.9.1151] [Medline: 16268385]
20. Thomas RJ, Mietus JE, Peng CK, et al. Differentiating obstructive from central and complex sleep apnea using an automated electrocardiogram-based method. Sleep. Dec 2007;30(12):1756-1769. [doi: 10.1093/sleep/30.12.1756] [Medline: 18246985]
21. Collins R. What makes UK Biobank special? Lancet. Mar 31, 2012;379(9822):1173-1174. [doi: 10.1016/S0140-6736(12)60404-8] [Medline: 22463865]
22. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature New Biol. Oct 2018;562(7726):203-209. [doi: 10.1038/s41586-018-0579-z] [Medline: 30305743]
23. Allen N, Sudlow C, Downey P, et al. UK Biobank: current status and what it means for epidemiology. Health Policy Technol. Sep 2012;1(3):123-126. [doi: 10.1016/j.hlpt.2012.07.003]
24. Littlejohns TJ, Holliday J, Gibson LM, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun. May 26, 2020;11(1):2624. [doi: 10.1038/s41467-020-15948-9] [Medline: 32457287]
25. Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: opportunities for cardiovascular research. Eur Heart J. Apr 7, 2019;40(14):1158-1166. [doi: 10.1093/eurheartj/ehx254] [Medline: 28531320]
26. Azevedo T, Bethlehem RAI, Whiteside DJ, et al. Identifying healthy individuals with Alzheimer's disease neuroimaging phenotypes in the UK Biobank. Commun Med (Lond). Jul 20, 2023;3(1):100. [doi: 10.1038/s43856-023-00313-w] [Medline: 37474615]
27. Steinfeldt J, Buergel T, Loock L, et al. Neural network-based integration of polygenic and clinical information: development and validation of a prediction model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. Lancet Digit Health. Feb 2022;4(2):e84-e94. [doi: 10.1016/S2589-7500(21)00249-1] [Medline: 35090679]
28. Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. Dec 2013;26(6):1045-1057. [doi: 10.1007/s10278-013-9622-7] [Medline: 23884657]
29. Markiewicz CJ, Gorgolewski KJ, Feingold F, et al. The OpenNeuro resource for sharing of neuroscience data. Elife. Oct 18, 2021;10:e71774. [doi: 10.7554/eLife.71774] [Medline: 34658334]
30. Buccigrossi R, Ellisman M, Grethe J, et al. The neuroimaging informatics tools and resources clearinghouse (NITRC). AMIA Annu Symp Proc. Nov 6, 2008;1000:1000. [Medline: 18999128]
31. Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing heterogeneous data: the national database for autism research. Neuroinformatics. Oct 2012;10(4):331-339. [doi: 10.1007/s12021-012-9151-4] [Medline: 22622767]
32. Thompson HJ, Vavilala MS, Rivara FP. Chapter 1 common data elements and federal interagency traumatic brain injury research informatics system for TBI research. Annu Rev Nurs Res. 2015;33(1):1-11. [doi: 10.1891/0739-6686.33.1] [Medline: 25946381]
33. Kaggle. URL: https://www.kaggle.com/ [Accessed 2025-12-22]
34. Yang X, Zeng Z, Teo SG, Wang L, Chandrasekhar V, Hoi S. Deep learning for practical image recognition: case study on kaggle competitions. Presented at: KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Aug 19-23, 2018:923-931; London, United Kingdom. [doi: 10.1145/3219819.3219907]
35. Krishna R, Kelleher K, Stahlberg E. Patient confidentiality in the research use of clinical medical databases. Am J Public Health. Apr 2007;97(4):654-658. [doi: 10.2105/AJPH.2006.090902] [Medline: 17329644]
36. Kobayashi S, Kane TB, Paton C. The privacy and security implications of open data in healthcare. Yearb Med Inform. Aug 2018;27(1):41-47. [doi: 10.1055/s-0038-1641201] [Medline: 29681042]
37. Eze B, Peyton L. Systematic literature review on the anonymization of high dimensional streaming datasets for health data sharing. Procedia Comput Sci. 2015;63:348-355. [doi: 10.1016/j.procs.2015.08.353]
38. Centers for Disease Control and Prevention (CDC). HIPAA privacy rule and public health. Guidance from CDC and the U.S. Department of Health and Human Services. MMWR Suppl. May 2, 2003;52:1-17. [Medline: 12741579]

39. Prasser F, Kohlmayer F, Spengler H, Kuhn KA. A scalable and pragmatic method for the safe sharing of high-quality health data. IEEE J Biomed Health Inform. Mar 2018;22(2):611-622. [doi: 10.1109/JBHI.2017.2676880] [Medline: 28358693]
40. Sarpatwari A, Gagne JJ. Balancing benefits and harms: privacy protection policies. Pharmacoepidemiol Drug Saf. Aug 2016;25(8):969-971. [doi: 10.1002/pds.4048] [Medline: 27278106]
41. Kulynych J, Korn D. The new HIPAA (Health Insurance Portability and Accountability Act of 1996) medical privacy rule: help or hindrance for clinical research? Circulation. Aug 26, 2003;108(8):912-914. [doi: 10.1161/01.CIR.0000080642.35380.50] [Medline: 12939240]
42. Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. Annu Rev Med. 2006;57(1):575-590. [doi: 10.1146/annurev.med.57.121304.131257] [Medline: 16409167]
43. Gostin LO, Nass S. Reforming the HIPAA privacy rule: safeguarding privacy and promoting research. JAMA. Apr 1, 2009;301(13):1373-1375. [doi: 10.1001/jama.2009.424] [Medline: 19336713]
44. Wartenberg D, Thompson WD. Privacy versus public health: the impact of current confidentiality rules. Am J Public Health. Mar 2010;100(3):407-412. [doi: 10.2105/AJPH.2009.166249] [Medline: 20075316]
45. Green AK, Reeder-Hayes KE, Corty RW, et al. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist. May 2015;20(5):464-e20. [doi: 10.1634/theoncologist.2014-0431] [Medline: 25876994]
46. Conroy M, Sellors J, Effingham M, et al. The advantages of UK Biobank's open-access strategy for health research. J Intern Med. Oct 2019;286(4):389-397. [doi: 10.1111/joim.12955] [Medline: 31283063]
47. Anderson R. The collection, linking and use of data in biomedical research and health care: ethical issues. The Nuffield Council on Bioethics; 2015. [doi: 10.17863/CAM.31760]
48. Kostkova P, Brewer H, de Lusignan S, et al. Who owns the data? Open data for healthcare. Front Public Health. 2016;4:7. [doi: 10.3389/fpubh.2016.00007] [Medline: 26925395]
49. Bhattacharjee Y. Biomedicine. Pharma firms push for sharing of cancer trial data. Science. Oct 5, 2012;338(6103):29. [doi: 10.1126/science.338.6103.29] [Medline: 23042862]
50. Kamoun F, Nicho M. Human and organizational factors of healthcare data breaches: the swiss cheese model of data breach causation and prevention. Int J Healthc Inf Syst Inform. Jan 1, 2014;9:42-60. [doi: 10.4018/ijhisi.2014010103]
51. Nifakos S, Chandramouli K, Nikolaou CK, et al. Influence of human factors on cyber security within healthcare organisations: a systematic review. Sensors (Basel). Jul 28, 2021;21(15):5119. [doi: 10.3390/s21155119] [Medline: 34372354]
52. McLeod A, Dolezel D. Cyber-analytics: modeling factors associated with healthcare data breaches. Decis Support Syst. Apr 2018;108:57-68. [doi: 10.1016/j.dss.2018.02.007]
53. Seh AH, Zarour M, Alenezi M, et al. Healthcare data breaches: insights and implications. Healthcare (Basel). May 13, 2020;8(2):133. [doi: 10.3390/healthcare8020133] [Medline: 32414183]
54. Chernyshev M, Zeadally S, Baig Z. Healthcare data breaches: implications for digital forensic readiness. J Med Syst. Nov 28, 2018;43(1):7. [doi: 10.1007/s10916-018-1123-2] [Medline: 30488291]
55. Ficek J, Wang W, Chen H, Dagne G, Daley E. Differential privacy in health research: a scoping review. J Am Med Inform Assoc. Sep 18, 2021;28(10):2269-2276. [doi: 10.1093/jamia/ocab135] [Medline: 34333623]
56. Dankar F, Emam K. Practicing differential privacy in health care: a review. Trans Data Priv. 2013;6(1):35-67. [doi: 10.5555/2612156.2612159]
57. Tamane S, Solanki VK. Dey N, editor. Privacy and Security Policies in Big Data. IGI Global; 2017. [doi: 10.4018/978-1-5225-2486-1] ISBN: 978-1-5225-2486-1
58. Appenzeller A, Terzer N, Philipp P, Beyerer J. Applying differential privacy to medical questionnaires. Presented at: 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops); Mar 13-17, 2023:608-613; Atlanta, GA. [doi: 10.1109/PerComWorkshops56833.2023.10150373]
59. Harris DR. Leveraging differential privacy in geospatial analyses of standardized healthcare data. Proc IEEE Int Conf Big Data. Dec 2020;2020:3119-3122. [doi: 10.1109/bigdata50022.2020.9378390] [Medline: 35253022]
60. Sun Z, Wang Y, Shu M, Liu R, Zhao H. Differential privacy for data and model publishing of medical data. IEEE Access. 2019;7:152103-152114. [doi: 10.1109/ACCESS.2019.2947295] [Medline: 31328077]
61. McDuff D, Curran T, Kadambi A. Synthetic data in healthcare. arXiv. Preprint posted online on Apr 6, 2023. [doi: 10.48550/ARXIV.2304.03243]
62. Appenzeller A, Leitner M, Philipp P, Krempel E, Beyerer J. Privacy and utility of private synthetic data for medical data analyses. Appl Sci (Basel). Dec 1, 2022;12(23):12320. [doi: 10.3390/app122312320]

63. Qian Z, Callender T, Cebere B, Janes SM, Navani N, van der Schaar M. Synthetic data for privacy-preserving clinical risk prediction. medRxiv. Preprint posted online on May 24, 2023. [doi: 10.1101/2023.05.18.23290114]
64. Kokosi T, Harron K. Synthetic data in medical research. BMJ Med. 2022;1(1):e000167. [doi: 10.1136/bmjmed-2022-000167] [Medline: 36936569]
65. Kaabachi B, Despraz J, Meurers T, et al. A scoping review of privacy and utility metrics in medical synthetic data. medRxiv. Preprint posted online on Oct 21, 2024. [doi: 10.1101/2023.11.28.23299124]
66. Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: preserving data utility and patient privacy. Comput Intell. May 2021;37(2):819-851. [doi: 10.1111/coin.12427]
67. Wood A, Najarian K, Kahrobaei D. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Comput Surv. Jul 31, 2021;53(4):1-35. [doi: 10.1145/3394658]
68. Bos JW, Lauter K, Naehrig M. Private predictive analysis on encrypted medical data. J Biomed Inform. Aug 2014;50:234-243. [doi: 10.1016/j.jbi.2014.04.003] [Medline: 24835616]
69. Bocu R, Costache C. A homomorphic encryption-based system for securely managing personal health metrics data. IBM J Res & Dev. Jan 1, 2018;62(1):1. [doi: 10.1147/JRD.2017.2755524]
70. Krishnegowda P, M. Boregowda A. Efficient matrix key homomorphic encryption of medical images. IJEECS. Jul 1, 2023;31(1):406. [doi: 10.11591/ijeecs.v31.i1.pp406-416]
71. Kartit A. New approach based on homomorphic encryption to secure medical images in cloud computing. Trends Sci. 2022;19(9):3970. [doi: 10.48048/tis.2022.3970]
72. Khedr A, Gulak G. SecureMed: secure medical computation using GPU-accelerated homomorphic encryption scheme. IEEE J Biomed Health Inform. Mar 2018;22(2):597-606. [doi: 10.1109/JBHI.2017.2657458] [Medline: 28129194]
73. Veeningen M, Chatterjea S, Horváth AZ, et al. Enabling analytics on sensitive medical data with secure multi-party computation. Stud Health Technol Inform. 2018;247:76-80. [Medline: 29677926]
74. Marwan M, Kartit A, Ouahmane H. Applying secure multi-party computation to improve collaboration in healthcare cloud. Presented at: 2016 Third International Conference on Systems of Collaboration (SysCo); Nov 28-29, 2016:1-6; Casablanca, Morocco. [doi: 10.1109/SYSCO.2016.7831325]
75. Yigzaw KY, Bellika JG. Evaluation of secure multi-party computation for reuse of distributed electronic health data. Presented at: 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI); Jun 1-4, 2014:219-222; Valencia, Spain. [doi: 10.1109/BHI.2014.6864343]
76. Şahinbaş K, Catak FO. Secure multi-party computation-based privacy-preserving data analysis in healthcare IoT systems. In: Kose U, Gupta D, Khanna A, Rodrigues J, editors. Interpretable Cognitive Internet of Things for Healthcare. Springer; 2023:57-72. [doi: 10.1007/978-3-031-08637-3_3] ISBN: 978-3-031-08636-6
77. Tso R, Alelaiwi A, Mizanur Rahman SM, Wu ME, Hossain MS. Privacy-preserving data communication through secure multi-party computation in healthcare sensor cloud. J Sign Process Syst. Oct 2017;89(1):51-59. [doi: 10.1007/s11265-016-1198-2]
78. Tawfik AM, Sabbeh SF, EL-Shishtawy T. Privacy-preserving secure multiparty computation on electronic medical records for star exchange topology. Arab J Sci Eng. Dec 2018;43(12):7747-7756. [doi: 10.1007/s13369-018-3122-5]
79. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag. May 2020;37(3):50-60. [doi: 10.1109/MSP.2020.2975749]
80. Nguyen DC, Pham QV, Pathirana PN, et al. Federated learning for smart healthcare: a survey. ACM Comput Surv. Mar 31, 2023;55(3):1-37. [doi: 10.1145/3501296]
81. Crowson MG, Moukheiber D, Arévalo AR, et al. A systematic review of federated learning applications for biomedical data. PLOS Digit Health. May 2022;1(5):e0000033. [doi: 10.1371/journal.pdig.0000033] [Medline: 36812504]
82. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform. Apr 2018;112:59-67. [doi: 10.1016/j.ijmedinf.2018.01.007] [Medline: 29500022]
83. Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: systematic review and architecture proposal. ACM Trans Intell Syst Technol. Aug 31, 2022;13(4):1-23. [doi: 10.1145/3501813]
84. Guo K, Chen T, Ren S, Li N, Hu M, Kang J. Federated learning empowered real-time medical data processing method for smart healthcare. IEEE/ACM Trans Comput Biol Bioinform. 2024;21(4):869-879. [doi: 10.1109/TCBB.2022.3185395] [Medline: 35737631]
85. Pfitzner B, Steckhan N, Arnrich B. Federated learning in a medical context: a systematic literature review. ACM Trans Internet Technol. Jun 23, 2021;21(2):1-31. [doi: 10.1145/3412357]
86. Wu JCH, Yu HW, Tsai TH, Lu HHS. Dynamically synthetic images for federated learning of medical images. Comput Methods Programs Biomed. Dec 2023;242:107845. [doi: 10.1016/j.cmpb.2023.107845] [Medline: 37852147]

87. Liu HH. Use and disclosure of health information and protection of patient privacy in Taiwan. Med Law. Mar 2010;29(1):87-101. [Medline: 22458000]
88. Jin H, Luo Y, Li P, Mathew J. A review of secure and privacy-preserving medical data sharing. IEEE Access. 2019;7:61656-61669. [doi: 10.1109/ACCESS.2019.2916503]
89. Liang X, Zhao J, Shetty S, Liu J, Li D. Integrating blockchain for data sharing and collaboration in mobile healthcare applications. Presented at: 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC); Oct 8-13, 2017:1-5; Montreal, QC. [doi: 10.1109/PIMRC.2017.8292361]
90. Rau HH, Hsu CY, Lee YL, Chen W, Jian WS. Developing electronic health records in Taiwan. IT Prof. Mar 2010;12(2):17-25. [doi: 10.1109/MITP.2010.53]
91. Li YCJ, Yen JC, Chiu WT, Jian WS, Syed-Abdul S, Hsu MH. Building a national electronic medical record exchange system - experiences in Taiwan. Comput Methods Programs Biomed. Aug 2015;121(1):14-20. [doi: 10.1016/j.cmpb.2015.04.013] [Medline: 26001420]
92. Hwang HG, Han HE, Kuo KM, Liu CF. The differing privacy concerns regarding exchanging electronic medical records of internet users in Taiwan. J Med Syst. Dec 2012;36(6):3783-3793. [doi: 10.1007/s10916-012-9851-1] [Medline: 22527781]
93. Yang CM, Lin HC, Chang P, Jian WS. Taiwan's perspective on electronic medical records' security and privacy protection: lessons learned from HIPAA. Comput Methods Programs Biomed. Jun 2006;82(3):277-282. [doi: 10.1016/j.cmpb.2006.04.002] [Medline: 16730852]
94. Taiwan medical AI and data portal. NYCU Data Management Center. 2023. URL: https://data.dmc.nycu.edu.tw/ [Accessed 2025-12-22]
95. Kaye J, Whitley EA, Lund D, Morrison M, Teare H, Melham K. Dynamic consent: a patient interface for twenty-first century research networks. Eur J Hum Genet. Feb 2015;23(2):141-146. [doi: 10.1038/ejhg.2014.71] [Medline: 24801761]
96. Williams H, Spencer K, Sanders C, et al. Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. JMIR Med Inform. Jan 13, 2015;3(1):e3. [doi: 10.2196/medinform.3525] [Medline: 25586934]
97. Lee H, Lee U. Toward dynamic consent for privacy-aware pervasive health and well-being: a scoping review and research directions. IEEE Pervasive Comput. Oct 1, 2022;21(4):25-32. [doi: 10.1109/MPRV.2022.3210747]
98. Budin-Ljøsne I, Teare HJA, Kaye J, et al. Dynamic consent: a potential solution to some of the challenges of modern biomedical research. BMC Med Ethics. Jan 25, 2017;18(1):4. [doi: 10.1186/s12910-016-0162-9] [Medline: 28122615]
99. Albalwy F, Brass A, Davies A. A blockchain-based dynamic consent architecture to support clinical genomic data sharing (consentchain): proof-of-concept study. JMIR Med Inform. Nov 3, 2021;9(11):e27816. [doi: 10.2196/27816] [Medline: 34730538]
100. Goncharov L, Suominen H, Cook M. Dynamic consent and personalised medicine. Med J Aust. Jun 20, 2022;216(11):547-549. [doi: 10.5694/mja2.51555] [Medline: 35611469]
101. Vlahou A, Hallinan D, Apweiler R, et al. Data sharing under the general data protection regulation: time to harmonize law and research ethics? Hypertension. Apr 2021;77(4):1029-1035. [doi: 10.1161/HYPERTENSIONAHA.120.16340] [Medline: 33583200]
102. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Curr Cardiol Rep. Jan 2014;16(1):441. [doi: 10.1007/s11886-013-0441-8] [Medline: 24338557]
103. Pereira T, Morgado J, Silva F, et al. Sharing biomedical data: strengthening AI development in healthcare. Healthcare (Basel). Jun 30, 2021;9(7):827. [doi: 10.3390/healthcare9070827] [Medline: 34208830]
104. Paton C, Kobayashi S. An open science approach to artificial intelligence in healthcare. Yearb Med Inform. Aug 2019;28(1):47-51. [doi: 10.1055/s-0039-1677898] [Medline: 31022753]
105. Lim L, Lee HC. Open datasets in perioperative medicine: a narrative review. Anesth Pain Med (Seoul). Jul 2023;18(3):213-219. [doi: 10.17085/apm.23076] [Medline: 37691592]
106. Wilson JR, Prevedello LM, Witiw CD, Flanders AE, Colak E. Data liberation and crowdsourcing in medical research: the intersection of collective and artificial intelligence. Radiol Artif Intell. Jan 2024;6(1):e230006. [doi: 10.1148/ryai.230006] [Medline: 38231037]
107. Kasparick M, Andersen B, Franke S, et al. Enabling artificial intelligence in high acuity medical environments. Minim Invasive Ther Allied Technol. Apr 2019;28(2):120-126. [doi: 10.1080/13645706.2019.1599957] [Medline: 30950665]
108. Koh DM, Papanikolaou N, Bick U, et al. Artificial intelligence and machine learning in cancer imaging. Commun Med (Lond). 2022;2(1):133. [doi: 10.1038/s43856-022-00199-0] [Medline: 36310650]
109. Levin MG, Rader DJ. Polygenic risk scores and coronary artery disease: ready for prime time? Circulation. Feb 25, 2020;141(8):637-640. [doi: 10.1161/CIRCULATIONAHA.119.044770] [Medline: 32091922]

110. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. Jan 3, 2023;10(1):1. [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
111. Milham MP, Craddock RC, Son JJ, et al. Assessment of the impact of shared brain imaging data on the scientific literature. Nat Commun. Jul 19, 2018;9(1):2818. [doi: 10.1038/s41467-018-04976-1] [Medline: 30026557]
112. Patterson E, McBurney R, Schmidt H, Baldini I, Mojsilovic A, Varshney KR. Dataflow representation of data analyses: toward a platform for collaborative data science. IBM J Res & Dev. Nov 1, 2017;61(6):9. [doi: 10.1147/JRD.2017.2736278]
113. El Naqa I, Haider MA, Giger ML, Ten Haken RK. Artificial intelligence: reshaping the practice of radiological sciences in the 21st century. Br J Radiol. Feb 1, 2020;93(1106):20190855. [doi: 10.1259/bjr.20190855] [Medline: 31965813]
114. Kras A, Celi LA, Miller JB. Accelerating ophthalmic artificial intelligence research: the role of an open access data repository. Curr Opin Ophthalmol. Sep 2020;31(5):337-350. [doi: 10.1097/ICU.0000000000000678] [Medline: 32740059]

Abbreviations

AI: artificial intelligence
CT: computed tomography
EMR: electronic medical record
FL: federated learning
HIPAA: Health Insurance Portability and Accountability Act
ML: machine learning
MRI: magnetic resonance imaging
PET: privacy-enhancing technology
PHI: protected health information

Edited by Gunther Eysenbach, Tiffany Leung; peer-reviewed by Imran, Yanling Sun; submitted 31.Mar.2024; final revised version received 28.Oct.2025; accepted 29.Nov.2025; published 30.Jan.2026

Please cite as:
Yang A, Pan ML, Lu HHS, Lien CY, Wang DW, Chen CH, Tarng DC, Niu DM, Chiou SH, Wu CY, Sun YC, Chen SA, Wang SJ, Sheu WHH, Lin CH
Assessing the Evolution and Influence of Medical Open Databases on Biomedical Research and Health Care Innovation: A 25-Year Perspective With a Focus on Privacy and Privacy-Enhancing Technologies
J Med Internet Res 2026;28:e58954
URL: https://www.jmir.org/2026/1/e58954
doi: 10.2196/58954