CODE

BRIEFING PAPER

Balancing Privacy with Health Data Access

July 2019

Table of Contents

Introduction 3

Background: The Data Privacy Landscape in Healthcare 4

What kinds of risks are associated with health data disclosure? 4

How is health data privacy regulated in the United States? 5

What kinds of challenges impact health data privacy frameworks? 8

Risk and Rewards of Using Health Data 9

Approaches to Balancing Privacy with Health Data Access 14

Privacy by design and data minimization 14

Privacy balancing tests 14

De-identification 14

Differential access 16

Data governance structures 16

Permission-based approaches 17

Conclusion and Discussion Questions 19

Appendix: Health Data Terminology 20

INTRODUCTION

The increasing availability of health data is transforming the health sector. Researchers are using clinical and surveillance data to better prevent, diagnose, and treat disease. Technology companies are using patient-generated data from mobile phones and wearable devices to help individuals track their medical conditions and customize their treatment plans. And healthcare providers are using administrative and claims data in combination with data on the social determinants of health to better understand risk factors for health conditions and improve healthcare delivery.

At the same time, complex questions are emerging around data privacy and the use of individual-level health data. In the United States, the Health Insurance and Portability Accountability Act (HIPAA) and a patchwork of federal, state, and local laws and regulations govern the security and privacy of personal health information. But while these laws protect data collected by healthcare providers and health plans, they are not well-designed to handle the many other kinds of health data produced and collected today.

Patients and patient advocates must play a critical role in shaping the conversation around balancing privacy with appropriate health data access and use, particularly in the context of individual-level health data. For example, patients can benefit from research that uses individual-level health data to better diagnose disease and find new treatments. But if protected health information (PHI) is misused, patients may be at risk of discrimination, financial exploitation, or other harms.

The nonprofit Center for Open Data Enterprise (CODE) and the HHS Office of the Chief Technology Officer (CTO) will convene a Roundtable on Balancing Privacy with Health Data Access on July 15, 2019. This Roundtable, which is the second in a series of three, will explore a portfolio of approaches for balancing the privacy of sensitive health information with the need to analyze the data for public good. It will focus on approaches to data privacy, the individual's right to exercise control over their personal information, rather than data security, which pertains more to technical security and data encryption. The event will include HHS leaders, patients, and health data experts in federal and state government agencies, industry, law, and patient-advocacy organizations.

This Briefing Paper, which has been developed in preparation for the Roundtable, is divided into four sections. It outlines the broad risks of disclosing individual-level health data, the privacy frameworks that govern health data, the privacy tradeoffs associated with using different high-value health data, and current approaches to balancing health data privacy with appropriate access.

BACKGROUND: THE DATA PRIVACY LANDSCAPE IN HEALTHCARE

Appropriate access to individual-level health data, facilitated by technological advances such as cloud computing and artificial intelligence, can greatly benefit patients and other stakeholders across the healthcare system. These benefits range from improving diagnostic accuracy to increasing the understanding of complex genetic conditions.[1] If personal health data is misused, however, then patients may be at risk of financial discrimination, reputational damage, or other harms from their loss of privacy.

Policymakers have responded to these risks by adopting state and national regulatory frameworks that attempt to balance data access and risk. At the federal level, HIPAA governs the security and privacy of protected health information (PHI) collected by entities such as healthcare providers and health plans. PHI is health data that contains personally identifiable information (PII) and may include demographic histories, medical records, lab tests, and other personal attributes. This data does not necessarily include patient-generated health data from wearables and mobile applications. In addition to HIPAA, a patchwork of state laws and regulations govern ownership of electronic health records (EHRs), Medicaid reporting, and data sharing agreements. This section examines the major risks of breaching the privacy of individual-level health data, the laws that address these vulnerabilities, and ongoing gaps and inconsistencies within the healthcare data privacy landscape.

What kinds of risks are associated with health data disclosure?

The increasing diversity and availability of health data has made protecting data privacy more complicated. Although data scientists and researchers have made advances in de-identifying data, removing key identifiers from data may not be enough to safeguard its privacy. Recent research has demonstrated how companies and researchers can take anonymized datasets and re-identify individuals with a high degree of accuracy when the de-identified data is combined with other third party data. This process has come to be called the "mosaic effect." A 2018 study outlined how researchers were able to use machine learning techniques to re-identify anonymized physical activity data, such as running patterns and heart rate, collected from wearable devices.[2]

The risks of re-identifying protected health information (PHI) include:

- **Adverse financial effects and discrimination.** The release of sensitive PHI could make patients the targets of discriminatory pricing or financial exclusion. Insurance companies are not barred from using different kinds of information to adopt discriminatory pricing schemes, charging higher premiums to people who they believe are more likely to get sick. This information can include PHI, which can be combined with social and demographic data based on an individual's residence or other factors.
- **Negative psychological or reputational impact.** Releases of sensitive PHI can lead to negative stigmas in professional spaces or the loss of social status and reputation.[3] Researchers and news stories have shown that personal infractions of a patient's privacy can have large effects on that individual and his

or her family. For example, in 2018 a New Jersey woman sued a hospital after it shared information about her son's attempted suicide with his high school.[4]

- **Loss of public trust.** The collective effect of breaching privacy in healthcare can damage public trust. Experts have noted that participant confidence in a research trial or health program depends upon guaranteeing the privacy of its participants.[5] Moreover, many patients value not only the privacy of their EHRs but also transparency about how that information will be transmitted and shared and how its security will be guaranteed.[6] Large-scale privacy breaches have revealed the PII of thousands of patients, such as the October 2018 breach of healthcare.gov that caused 75,000 patient EHRs to be compromised.[7]

How is health data privacy regulated in the United States?

Given this range of risks, government agencies in the United States have sought to define the rules and regulations that safeguard the privacy of PHI. Congress enacted HIPAA and its subsequent amendments as the cornerstone of the federal government's framework to govern the exchange of PHI. Other federal legislation, such as the Genetic Information Nondiscrimination Act (GINA), governs how sensitive genomic and biomarker data can be shared. Lastly, a system of rules at the state level provide additional oversight for how patients and providers manage PHI in different jurisdictions.

Health Insurance and Portability Accountability Act (HIPAA)

HIPAA was designed primarily to create a federal floor for the privacy and security of personal health information. HIPAA defines "personal health information" as data that "includes the individual's past, present, or future mental or physical condition, the provision of healthcare to an individual, and any past, present, or future payment for the provision of healthcare to the individual."[8] HIPAA sets the standards for how covered entities must transmit this personal health information, which includes claims, enrollment, eligibility, payment, and coordination of benefits. The law also defines "covered entities" as qualified healthcare providers, healthcare clearinghouses, and health plans.

Congress amended HIPAA's Title II in the early 2000s to include the HIPAA Privacy Rule and the Security Rule. These provisions are enforced by the HHS Office of Civil Rights (OCR), which can administer financial penalties for rule violations by qualified health providers. Below are the major dimensions of HIPAA that impact the privacy of individual-level health data.

- **HIPAA Privacy Rule.** The Privacy Rule sets the standards for individually identifiable health information. It assures "that individuals' health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and wellbeing." The Privacy Rule also outlines six instances of permitted use and disclosure of PHI. Those include release on behalf of the individual, for healthcare treatment or payment, to give the individual an opportunity to agree to or object to the data, for the public interest, and for limited purposes of research, public health or health care operations.[9]
- **HIPAA Security Rule.** The HIPAA Security rule focuses on safeguarding electronic personal health information (EPHI). It dictates that healthcare providers that create, receive, maintain,

and transmit EPHI institute measures to protect this EPHI from anticipated threats, hazards, and impermissible uses of this data.[10] The rule aims to ensure the confidentiality, integrity, and availability of EPHI.

- **Additional Protections under Title 42 CFR Part II and FERPA.** In addition to HIPAA, personal health information about substance abuse disorders is protected under Title 42 CFR Part II and cannot be released without a patient's written consent. Similarly, the Family Educational Rights and Privacy Act (FERPA) establishes specific guidelines for protecting the privacy of personal health information in students' educational records, such as vaccinations and nurse visits.

The E-Government Act (2002) and Privacy Impact Assessments

The E-Government Act of 2002 mandates that any agency that collects personally identifiable data must evaluate the security of its systems to ensure adequate data protection. Most federal agencies achieve this by conducting a privacy impact assessment (PIA) of their operational and developmental systems. HHS publishes all of the PIAs from its various operating divisions and also shares the PIAs from its third party websites. The PIA follows a standard template and describes the systems of data collection, the technical security measures, and approaches to addressing any individual's concern that their data may have been inappropriately used.[11]

Genetic Information Nondiscrimination Act (GINA)

Some individuals avoid participating in genetic testing because they fear being subjected to discrimination based on their results. Passed in 2008, the Genetic Information Nondiscrimination Act (GINA) strives to protect Americans from such discrimination in both health insurance and employment.[12] Notably, insurance companies cannot use genetic information to make decisions related to eligibility, coverage, premium costs, or underwriting for members. GINA also governs individuals who participate in clinical trials by improving the informed consent procedure and mandating that researchers disclose any possible risks from the tests.

The Confidential Information Protection and Statistical Efficiency Act (CIPSEA)

CIPSEA establishes laws to govern confidentiality protections for data collected by U.S. statistical agencies and units. The National Center for Health Statistics and the Center for Behavioral Health Statistics and Quality are the two HHS entities covered under CIPSEA.[13] The law requires agencies to use its collected information solely for statistical purposes and violations are subject to five years imprisonment and up to $250,000 in fines. While the law is clear about maintaining the confidentiality of the data, it is less clear on the exact definition of statistical purposes.

State policies governing health data privacy

While HIPAA provides broad guidance for governing the protection and dissemination of PHI, states have additional frameworks in place to manage individual-level health data. States can either meet the baseline of HIPAA's requirements or institute policies that are more rigorous than HIPAA. States are also responsible for designating the owners for a patient's healthcare record, establishing the specific fields

for an EMR or EHR, and setting the procedures for sharing a patient's PHI with other providers. As a result, state healthcare laws vary widely. This patchwork of state laws governing health data privacy is described below:

- **State-level health information exchanges.** States can adopt health information exchange (HIE) programs to help improve flow of individual-level health information between hospitals, clinics, and payers. These programs may include provisions to protect privacy. Maryland, for example, has a state-designated HIE program that aims "to build the fundamental foundation for interoperability to communicate health data among Maryland physicians, hospitals, and other health care organizations and providers."[14] As a result, the Code of Maryland Regulations includes provisions on Health Information Exchanges: Privacy and Security of Protected Health Information, which is designed to "ensure the privacy and security of PHI accessed, used, or disclosed through an HIE; improve access to clinical records by providers; and support public health goals."[15]
- **Consent for "Health Information Exchange."** In addition to broad-based privacy measures, HIEs have specific implications for patient consent. Opt-in policies promote patient privacy by ensuring that patients explicitly grant permission to healthcare providers that access their EHRs. Florida, Nevada, California, New York, Vermont, Rhode Island, and Massachusetts currently maintain opt-in policies which require patient consent prior to sharing data with a qualified HIE.[16] Many states, however, have no opt-in or opt-out laws that govern health data exchange consent, which creates uncertainty for patients seeking to protect their PHI from inappropriate use.
- **Patient ownership of medical records.** Patient ownership of medical records builds trust by giving patients access to their full medical history, and increases trust between physicians and patients since both parties can access the same information. New Hampshire is the only state that grants patients direct ownership of their EHRs, although this only consists of digitally recorded PHI. Most states, however, instead grant EHR ownership to hospitals or other medical institutions.[17]
- **Consent for releasing genomic data.** The consent process to release genomic data and test results vary by jurisdiction. Definitions of genomic data also differ by state and many states do not specify acceptable reasons for sharing genomic data with patients. While most states ask patients to consent to the tests performed, New York and Massachusetts mandate that genomic results include a list of individual diseases tested. These policies impact patients who may not be aware that companies are using their genomic data for purposes outside of the requested tests.
- **Medicaid requirements.** Medicaid claims and records are a joint responsibility of the federal government and the states. States can implement their own rules and procedures to govern the exchange and use of individual-level Medicaid data by other agencies and partners. For example, the state of Georgia created a special data use agreement between its Department of Community Health and Department of Public Health that adhered to the HIPAA privacy rule and expanded the parameters of protecting PHI.[18]

In addition to health-related laws, states are implementing broader privacy frameworks to protect consumer data gathered by businesses. The 2018 California Consumer Privacy Act established a baseline of consumer protections allowing individuals to request the deletion of personal information, opt-out of

the sale of personal information, and access their personal information in a usable format.[19] New York is currently considering a similar law that would define personal data to include medical and biometric data.[20] These laws will have a profound impact on how patients access and protect their individual-level health data.

What kinds of challenges impact health data privacy frameworks?

Federal and state privacy frameworks have neither kept pace with each other nor the modern landscape of health-related data and technology. These frameworks have caused confusion among stakeholders, resulting in the following gaps and inconsistencies:

- **Inconsistent Rules for Patient Consent.** The lack of opt-in and opt-out clauses for data sharing leaves many patients uncertain about how they can protect or share their data at the state level.
- **Entities Not Covered by HIPAA.** HIPAA applies to "covered entities" such as health plans and healthcare providers, but does not apply to software and social media companies that may collect patient-generated data with sensitive health information. As a result, individual-level health data generated by wearable sensors and mobile applications often falls outside the purview of state and federal regulations. This data may be provided to third parties without the patient's consent.
- **Lack of Patient EHR Ownership.** Most states do not recognize patient ownership over their EHRs. Patient mobility and employment changes may complicate matters by leaving a trail of EPHI across a variety of different health insurers and providers.
- **Insufficient Technical Security Protections.** The HIPAA Security Rule governs the use of EPHI but data breaches have been growing in scale and scope. During the month of April 2019, U.S. healthcare providers reported a record 44 healthcare breaches that compromised thousands of patient records.[21] Healthcare data breaches can expose highly sensitive information like social security numbers and medical histories, and reduce patient trust in healthcare institutions.

As states and the federal government seek to amend this patchwork of policies and laws, it is often patients and healthcare providers who must navigate ambiguities in consent, protection, and access. The next section describes the types of data that these stakeholders have access to, and the risks and benefits of using it.

RISK AND REWARDS OF USING HEALTH DATA

Health data can take many forms, from granular individual-level data to aggregated population data. Stakeholders collect and use this information for different purposes, and should try to maximize benefits and minimize risks when sensitive health information is involved. A recent paper published by the Fragile Families Challenge, a collaborative project aiming to improve the wellbeing of disadvantaged families, outlines the challenges of balancing risk and utility of sensitive health data. As the graph below from their paper shows, these authors suggest that most people will only be willing to release data when the benefit to science is much greater than the risk. More research may shed light on how different groups and individuals see the tradeoffs of risk and benefit for different types of health-related data.

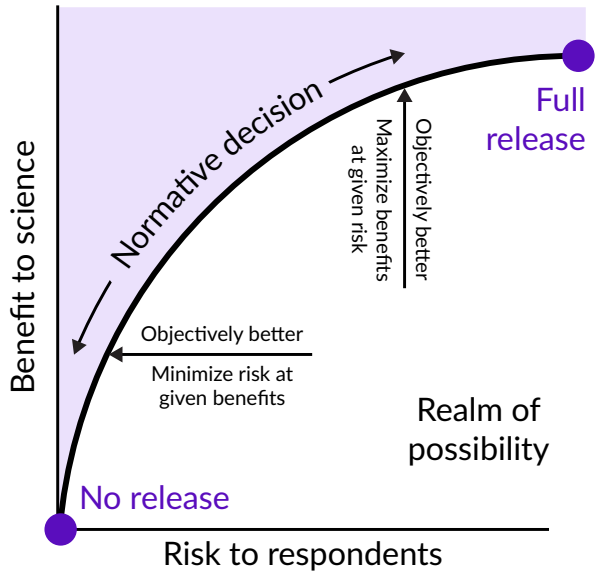

Figure 1. "Risk - Benefit Frontier Curve"[22]

Policymakers, researchers, and patients are responsible for balancing the possible benefits and risks of using health data. While healthcare providers often extol the benefits to science when using high-value health data, fewer stakeholders address the potential risks to patients.

The following section identifies potential risks and rewards of using different types of health data. (For more definitions of data types and relevant concepts, see Appendix, "Health Data Terminology" and CODE's report on **Utilizing and Sharing Health Data for AI Applications**)

Administrative and Claims Data generally comes from federal, state, and local government agencies as well as healthcare providers and insurers. This can range from hospital discharge summaries to records of payments to the healthcare system for insured patients.[23]

Potential Benefits of Use	Potential Risks of Use
■ Improving healthcare administration ■ Evaluating population health trends ■ Examining disparities in healthcare delivery ■ Monitoring drug adherence ■ Understanding patterns in hospital visitation and length of stay	■ Discriminating against patients through risk scores ■ Raising patient costs through higher insurance ■ Revealing sensitive PHI or patient medical history

Clinical Data is a broad term that encompasses different kinds of data generated "in a clinical setting and controlled by a clinician, as opposed to a patient or caregiver."[24]

- **Clinical Trials Data** includes registries and results from publicly and privately funded clinical studies. Large amounts of data, including sensitive information about participants, are generated over the course of a clinical trial. Researchers must obtain regulatory approval to collect and use this data.
- **EHR Data** is focused on individual patients, and can include information on routine checkups, prescriptions, and medical procedures. Physicians can draw upon EHR data to develop individual treatment plans and diagnose conditions. This data can also be combined with social determinants of health to develop rich longitudinal profiles of individual patients and populations.

Potential Benefits of Use	Potential Risks of Use
EHRs	
■ Quickly accessing a patient's medical history ■ Improving "Coordination of Care" through better communication ■ Building longitudinal patient profiles	■ Revealing sensitive PHI or patient medical history ■ Growing threat of data breaches or access by unauthorized users[25]
Clinical Trials Data	
■ Testing new therapeutics ■ Preventing the onset of disease ■ Expanding access to treatment for at-risk patients	■ Using data that falls outside of original consent or purpose ■ Revealing sensitive patient medical history ■ Re-identifying data when anonymized data is combined with other third party data[26]

Patient-Generated Data includes "health-related data created and recorded by or from patients outside of the clinical setting to help address a health concern."[27] This data type is becoming increasingly prevalent through the creation of mobile health applications and wearable health devices. Unlike clinical data, there are relatively few legal frameworks and guidelines that protect this data from misuse.

- **IoT Data** includes data from mobile software applications, voice assistants, and wearable devices such as smart watches. These technologies are part of the "internet of things," or IoT, which refers to the growing system of machines and devices connected to the internet. This data is generally collected under terms of service agreements and is not regulated by HIPAA. It provides important information on critical health indicators, such as heart rates, sleep cycles, and diet.
- **Social Media Data** includes interactions on social media platforms such as Facebook and Twitter. Researchers have noted that, "Social media may offer insight into the relationship between an individual's health and their everyday life, as well as attitudes towards health and the perceived quality of healthcare services," among other opportunities.[28] This data is also governed under terms of service agreements and company privacy policies.

Potential Benefits of Use	Potential Risks of Use
IoT Data	
■ Providing the best understanding of a patient's health through fitness trackers, wearables, and voice-operated assistants ■ Monitoring preventive health measures like exercise and healthy eating ■ Empowering patients to understand their own health profiles	■ Revealing sensitive PII or patient medical history ■ Sharing or selling patient PHI to third party providers[29] ■ Raising patient premiums based on an elevated risk profile[30]
Social Media Data	
■ Understanding fitness and health trends at the individual level ■ Developing community support options for patients with specific diseases ■ Recruiting patients for specialized clinical trials	■ Damaging patient-physician trust through improper access of data ■ Developing biased profiles of patients due to social media activity ■ Accessing incorrect or false information

Genomic Data can range from an individual's complete DNA sequences to data on individual DNA variants.[31] Genomic data is considered highly sensitive and must be shared and used under carefully controlled conditions.

Potential Benefits of Use	Potential Risks of Use
■ Using genetic sequencing to identify a patient's risk of developing a rare disease or defect ■ Understanding a patient's complete medical history ■ Empowering patients to take ownership of their health history	■ Revealing sensitive PHI by sharing results with a third party ■ Discriminating against a patient based on genetic risks ■ Receiving genetic information from private providers that is incomplete or lacks context[32]

Social Determinants of Health Data represent "conditions in the environments in which people are born, live, learn, [and] work...that affect a wide range of health, functioning, and quality-of-life outcomes and risks."[33] Examples of these social determinants include access to transportation, education, job opportunities, and availability of food and housing options. Social determinants of health data can come from many sources inside and outside of government, and its use is often unregulated given that it does not include individual-level information.

Potential Benefits of Use	Potential Risks of Use
■ Understanding population health trends ■ Designing programs tailored to patient or member needs ■ Identifying and treating at-risk patients before they are admitted to inpatient settings ■ Complementing clinical data to provide better clinical decision support for physicians[34]	■ Potential for discrimination against individuals or groups based on socioeconomic factors and expected risk of illness ■ Using socioeconomic data in the context of individual health rather than to address social problems[35]

Surveillance Data is a broad term that encompasses the "ongoing, systematic collection, analysis, and interpretation of health-related data essential to planning, implementation, and evaluation of public health practice."[36]

- **Registry Data** includes data shared voluntarily by individuals that is generally focused around a specific diagnosis or condition such as cancer or cystic fibrosis. This data can be used to track trends and better understand conditions over time. According to the NIH, this data "belongs

to the sponsor of the registry and may be shared with the participants and their families, and approved health care professionals and researchers. However, personal, identifying information is kept private."[37]

- **Survey Data** includes the results of surveys and studies conducted to assess population health. This data can help stakeholders monitor the spread of disease, track health insurance coverage across regions, and assess trends in nutrition and exercise, among other uses.[38]
- **Vitals Data** is generally collected and exchanged between local jurisdictions and the federal government. This data represents "vital events," such as births, deaths, marriages, divorces, and fetal deaths.[39]

Potential Benefits of Use	Potential Risks of Use
Registry Data	
■ Providing firsthand information to medical professionals ■ Tracking broader population health trends of specific diseases ■ Increasing information and knowledge about rare conditions ■ Creating a participatory path for patients opting in to share their information	■ Revealing sensitive PII or patient medical history ■ Using data for other reasons outside of its intended purpose
Survey Data	
■ Monitoring the spread of disease ■ Tracking health insurance coverage across regions ■ Assessing trends in nutrition and exercise[40] ■ Identifying barriers to healthcare access ■ Evaluating federal health programs[41]	■ Using data for other reasons outside of its intended purpose ■ Developing biased profiles of patients or their health issues ■ Discriminating against certain geographic areas based on risk profiles
Vitals Data	
■ Monitoring births, deaths, marriages, divorces, and fetal deaths[42] ■ Analyzing the causes of diseases and effectiveness of different interventions[43]	■ Revealing sensitive PHI or patient medical history ■ Disclosing data during exchange between federal agencies and local jurisdictions ■ Impacting the allocation of funds to poorer regions

APPROACHES TO BALANCING PRIVACY WITH HEALTH DATA ACCESS

Despite the adverse effects of privacy breaches, policymakers, researchers, and healthcare providers have adopted a variety of approaches to protect sensitive PHI while making these datasets available to interested parties. Different stakeholders should identify the right blend of approaches to suit their needs. This section outlines the primary approaches to ensure the privacy of PHI and also includes relevant examples. While this portfolio of approaches is not comprehensive, it outlines some of the most common approaches to balancing privacy with data access in the health sector.

Privacy by design and data minimization

"Privacy by design" and "data minimization" are two concepts embedded in the recently approved General Data Protection Regulation (GDPR), which governs data privacy in the European Union. Privacy by Design consists of different principles that businesses are expected to adhere to during the design phase of their products. Rather than evaluating privacy concerns at later stages of product development, Privacy by Design encourages organizations to think about the potentially adverse effects of using sensitive data from the beginning.

The data minimization principle emphasizes that data processing should only use as much data as necessary to accomplish a specific task.[44] The GDPR states that personal data shall be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed."[45] Companies and organizations must verify their collection of personal data is relevant and adequate for the stated purpose.[46] Data minimization is now gaining support in the United States as politicians and executives alike have called on companies to restrict their data collection efforts to only necessary personal data.

Privacy balancing tests

Privacy balancing tests can be used to map the tradeoffs between privacy protections and data utility for different types of data. Balancing frameworks may benefit from a "circumstance catalogue," which would list "circumstances, or factors, that should be considered when assessing whether, and under which conditions, a dataset should be released, as well as different options for how it should be released."[47] Balancing tests can enable healthcare companies and organizations to evaluate the possibilities for releasing their data. They can also create new justifications for researchers who hope to access data for a specific purpose.

De-identification

Although it is difficult to completely de-identify datasets, many organizations and healthcare companies have taken steps to scrub PII from datasets. Researchers may lose some detail and granularity from these datasets but can still use them for research and to analyze larger trends. Common approaches to technical de-identification include:

- **Providing Anonymized Identifiers:** These identifiers allow researchers to connect disparate datasets while preserving the privacy of individuals.[48]
- **Removing Non-Critical Information:** Researchers can remove a number of variables, including the last digits of a zip code, social security numbers, account information, and other identifying information to anonymize a dataset.
- **Leveraging Synthetic Data:** Synthetic data is produced by "a complex statistical model that generates a simulated population that has the same general features as the original data."[49] There is currently no gold standard for generating synthetic data, but many researchers are attempting to use these statistical models to ensure patient privacy.
- **Applying Differential Privacy:** Differential privacy places constraints on algorithms that rely on inputs from a database of information. This masks the personal information so an external user cannot determine if an individual's information was used in the computation process.

Example: Automating medical image de-identification

Physicians analyze medical images, such as CT Scans and MRIs, to identify patterns in disease formation and other medical issues. This visual information often falls under the purview of HIPAA since it may contain PHI. Amazon has released a machine learning program called "Amazon Rekognition" and a Natural Language Processing tool called "Amazon Comprehend Medical", which together remove strings of identifying text from any medical image.[50] Integrated through a Python script, these programs can remove sensitive information and be used for machine learning algorithms for quicker detection and diagnosis.

Example: Using blockchain for patient records

Most states recognize health providers and health insurers as the owners of medical records rather than patients themselves. As patients change locations and providers, their data becomes increasingly difficult to access. This problem is compounded by other data sharing issues such as health data interoperability. To address these issues, the Beth Deaconess Medical Center has partnered with the MIT Media Lab to pilot MedRec, a blockchain-based solution that enables patients to access their records through "special ownership and viewership permissions shared by members of a private, peer-to-peer network."[51]

Differential access

Differential access assumes that individual-level health data can be made accessible under controlled conditions even though release to the public is not appropriate. It grants access to datasets only under specific circumstances and for specific purposes. Approaches to differential access can include a federated data cloud model that grants trusted users specific credentials to access this data, and tiered access for multiple levels of access. Differential access can include the use of credentialing systems, which allows qualified and credentialed researchers to have access to certain tiers of data for their own use.

Example: Developing a genomic data commons

The National Cancer Institute maintains the Genomic Data Commons to "provide the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine."[52] The centralized data portal aggregates and standardizes genomic datasets from researchers around the United States and provides both controlled and open access to these datasets. Users who attempt to access the database must adhere to a standard set of policies, including a commitment to not identify individuals in the data and to provide attribution for the data use.

Example: Creating data containers

The research world currently uses "middle paths", or approaches between fully open and closed data access, which may provide a solution to growing concerns around compromised data privacy and security. Sage Bionetworks has proposed that HHS adopt a "Model to Data" (M2D) approach which allows researchers to query datasets without directly accessing that dataset. A data steward, responsible for the management of the native dataset, performs the query and then returns the results to the researcher. Rather than have direct access, the "data steward acts as a mediator and trusted partner for both data generators and data users, protecting data while maximizing its use."[53] This model relies largely on the use of standards, cloud computing, and container technologies.

Data governance structures

All organizations - including federal and state government agencies - can take steps to improve their data governance structures. This can include establishing common data models, creating disclosure review boards, and creating centralized data hubs.

Example: Advancing a trusted exchange framework

Health data interoperability remains a major issue for healthcare providers who want to share information across networks and for patients who hope to access their data. The HHS Office of the National Coordinator for Health Information Technology (ONC) drafted the Trusted Exchange Framework and Common Agreement as a potential solution to this problem. The framework creates "exchange modalities" to improve data sharing and technical interoperability of EHI between vetted Health Information Networks (HINs). The revised model should support patients by allowing patients to exercise "Meaningful Choice" to request that their PHI not be used or disclosed.[54] It also establishes other baseline security and privacy requirements, including rules for accessing data from mobile devices, and explaining how their data is being accessed, used, and shared. While still in its drafting phase, the Trusted Exchange Framework will allow healthcare providers to seamlessly access patient information and improve care coordination.

Example: Establishing privacy principles in precision medicine

The All of Us Precision Medicine Initiative established a set of privacy and trust principles during its launch in 2014. These principles range from "Participant Empowerment through Access to Information" to "Respecting Participant Preferences".[55] The research program follows strict protocols to protect patient data, with an emphasis on informed consent. All of Us allows participants to withdraw from future research and change their opt-in and opt-out preferences.

Permission-based approaches

Permission-based approaches enable individuals to grant access to their personal data for the benefits of public research. Patients may opt-in and provide consent to use their personal data for a specific purpose such as studying a rare disease or identifying genetic trends. As an emerging form of health data access, researchers are allowed to access this data based on the parameters of the patient's original consent. New programs like the One Million Veterans project by the Department of Veterans Affairs (VA)[56] and the Patients Like Me initiative[57] strive to create patient-driven databases where researchers and patients alike can access important health data about their conditions. The permission-based approach provides a new avenue for patients who wish to make their data both accessible and protected for appropriate access.

Example: Creating a national, volunteer research program

The Million Veteran Program (MVP) seeks to invite veterans receiving care at VA hospitals to share blood samples and health information to better understand how genes affect health.[58] Participating veterans can decide whether to participate and choose to enroll in the program through visiting a participating VA center. The program will store the information in a safe and secure manner using managerial, operational, and technical approaches to safeguarding veteran data. Researchers can use the aggregated data to study diseases like cancer and diabetes but also focus on military-specific diseases like post-traumatic stress disorder.

Example: Designing a patient-driven, personalized health network

PatientsLikeMe is the world's largest repository of patient-generated experiences and information on a wide range of diseases and conditions. Over 650,000 patients have signed up to the site and shared information about nearly 3000 conditions. New patients can search through a database of other patients' conditions and also access clinical research studies driven by real world experiences. The site, which empowers patients to learn how to better manage their own diseases, has adopted a strong privacy framework that encourages data sharing while granting patients specific rights, including the ability to access, delete, and correct personalized health data.[59]

CONCLUSION AND DISCUSSION QUESTIONS

This briefing paper sets the stage for deeper discussions about the future of privacy and access in the healthcare space. The Roundtable on Balancing Privacy with Health Data Access will provide an opportunity to discuss questions such as:

- What are the primary benefits and risks of using high-value data such as genomic data, claims data, and patient-generated data? What datasets are the highest priority for balancing risk and reward? And why?
- What are the best situations to use different approaches to health data access? Which kinds of strategies could best be applied to high priority data?
- How do we involve patients in shaping the recommended strategies for accessing high priority data? How do patients also help shape the laws and policies to balance privacy with access?

APPENDIX: HEALTH DATA TERMINOLOGY

Individual-Level Health Data. This is health data on individuals that includes a range of data types related to quality of life and wellbeing, and may come from survey data, EHRs, or other personal data. Individual-level health data encompasses Protected Health Information (PHI) but may also include data not formally gathered by healthcare professionals.

Protected Health Information (PHI). Also referred to as Personal Health Information, this data is collected by healthcare professionals and may include demographic histories, medical histories, lab tests and special procedures, insurance information, and other personal health attributes.[60]

Electronic Protected Health Information (EPHI). A large portion of PHI is not available digitally. EPHI refers to the electronic version of PHI in an accessible, electronic format which is often found in both EMRs and EHRs. Its confidentiality, integrity, and availability is governed by HIPAA.

Personal Identifiable Information (PII). PII is defined as information that can be used to distinguish or trace a person's identity, either alone or when combined with other personal or identifying information that is linked to an individual.[61]

Electronic Medical Record (EMR). EMRs are "digital versions of the paper charts found in clinician offices, clinics, and hospitals."[62] These notes are collected by and for clinicians in that office or hospital, and can be shared with healthcare providers for diagnosis and treatment.

Electronic Health Record (EHR). EHRs are broader records of a patient's full medical history, which are maintained by a provider over time, and may also include critical data such as clinical trials data, medications, vital signs, physician notes, and other key information.[63]

Health Information Exchange (HIE). Health Information Exchanges allow "health care professionals and patients to appropriately access and securely share a patient's medical information electronically."[64] There are three forms of HIEs including direct exchange and query-based exchange, which enable the sharing of PHI between insurers, and consumer mediated exchange, which allows consumers to aggregate PHI among providers.

Health Insurance Portability and Accountability Act (HIPAA). The primary legislation that governs privacy and security of EHRs in the United States. This Act was enacted by Congress in 1996 to modernize the flow of electronic health information and has been subsequently updated with privacy and security rules to govern how EPHI is protected and shared.

REFERENCES

[1] Miner, Luke. "Opinion | For a Longer, Healthier Life, Share Your Data." The New York Times, May 27, 2019, sec. Opinion. **https://www.nytimes.com/2019/05/22/opinion/health-care-privacy-hipaa.html**.

[2] Na, Liangyuan, Cong Yang, Chi-Cheng Lo, Fangyuan Zhao, Yoshimi Fukuoka, and Anil Aswani. "Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning Feasibility of Reidentifying Individuals by Their Protected Health Information; Feasibility of Reidentifying Individuals by Their Protected Health Information." JAMA Network Open 1, no. 8 (December 21, 2018).

[3] Lane, Julia, and Claudia Schur. "Balancing Access to Health Data and Privacy: A Review of the Issues and Approaches for the Future: Balancing Access to Health Data and Privacy." Health Services Research 45, no. 5, pt. 2(October 2010): 1456–67.

[4] Ornstein, Charles. "Small Violations Of Medical Privacy Can Hurt Patients And Erode Trust." NPR.Org. Accessed May 15, 2019. **https://www.npr.org/sections/health-shots/2015/12/10/459091273/small-violations-of-medical-privacy-can-hurt-patients-and-corrode-trust**.

[5] O'Hara, Kieron, Cabinet Office (2011) Transparent government, not transparent citizens: a report on privacy and transparency for the Cabinet Office London, GB. Cabinet Office, 84pp. **https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/61279/transparency-and-privacy-review-annex-a.pdf**.

[6] Snell, Elizabeth. "Why EHR Privacy, Transparency Are Crucial to Healthcare." HealthITSecurity, February 18, 2015. **https://healthitsecurity.com/news/why-ehr-privacy-transparency-are-crucial-to-healthcare**.

[7] Morse, Susan. "CMS Responds to Data Breach Affecting 75,000 in Federal ACA Portal." Healthcare Finance News. Accessed May 15, 2019. **https://www.healthcarefinancenews.com/news/cms-responds-data-breach-affecting-75000-federal-aca-portal**.

[8] Office for Civil Rights. "Summary of the HIPAA Privacy Rule." Text. HHS.gov, May 7, 2008. **https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html**.

[9] Office for Civil Rights. "Summary of the HIPAA Privacy Rule." Text. HHS.gov, May 7, 2008. **https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html**.

[10] Green, Katherine H. "HIPAA Security Rule." NIST, January 3, 2011. **https://www.nist.gov/healthcare/security/hipaa-security-rule**.

[11] Affairs (ASPA), Assistant Secretary for Public. "Privacy Impact Assessments." Text. HHS.gov, February 13, 2009. **https://www.hhs.gov/pia/index.html**.

[12] "Genetic Discrimination." Genome.gov. Accessed June 7, 2019. **https://www.genome.gov/about-genomics/policy-issues/Genetic-Discrimination**.

[13] U.S. Department of Health and Human Services Office of the Chief Technology Officer, "The State of Data Sharing at the U.S. Department of Health and Human Services," September 2018, **https://www.hhs.gov/sites/default/files/HHS_StateofDataSharing_0915.pdf**.

[14] Maryland Health Care Commission, "Health Information Exchange." Accessed May 15, 2019. **http://mhcc.maryland.gov/mhcc/pages/hit/hit_hie/hit_hie.aspx**.

[15] Maryland Health Care Commission, "Health Information Exchange," Accessed June 25, 2019. **http://mhcc.maryland.gov/mhcc/pages/hit/hit_hie/hit_hie.aspx**.

[16] "State Health IT Privacy and Consent Laws and Policies." Accessed May 14, 2019. **https://doi.org/apps/state-health-it-privacy-consent-law-policy.php**.

[17] "Who Owns Medical Records: 50 State Comparison | Health Information & the Law." Accessed May 15, 2019. **http://www.healthinfolaw.org/comparative-analysis/who-owns-medical-records-50-state-comparison**.

[18] "Improving Care for Medicaid Beneficiaries with Complex Care Needs and High Costs." Accessed May 20, 2019. **https://www.medicaid.gov/state-resource-center/innovation-accelerator-program/program-areas/beneficiaries-with-complex-needs/index.html**.

[19] Ghosh, Dipayan. "What You Need to Know About California's New Data Privacy Law." Harvard Business Review, July 11, 2018. **https://hbr.org/2018/07/what-you-need-to-know-about-californias-new-data-privacy-law**.

[20] Donelson, Baker. "The New York Privacy Act: A Consumer Privacy Bill to Monitor Closely." JD Supra. Accessed June 24, 2019. https://www.jdsupra.com/legalnews/the-new-york-privacy-act-a-consumer-20164/.

[21] HIPAA guide. "April 2019 Was the Worst Ever Month for Healthcare Data Breaches." HIPAA Guide (blog), May 22, 2019. **https://www.hipaaguide.net/april-2019-was-the-worst-ever-month-for-healthcare-data-breaches/**.

[22] Lundberg, Ian, Arvind Narayanan, Karen Levy, and Matthew Salganik. "Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge." Fragile Families Challenge, September 5, 2018. **http://www.fragilefamilieschallenge.org/privacy-ethics-and-data-access-a-case-study-of-the-fragile-families-challenge/**.

[23] University of Washington Health Sciences Library, "Data Resources in the Health Sciences," **http://guides.lib.uw.edu/hsl/data/findclin**.

[24] Office of the National Coordinator for Health Information Technology, Conceptualizing a Data Infrastructure for the Capture, Use, and Sharing of Patient-Generated Health Data in Care Delivery and Research through 2024, January 2018, Retrieved from **https://www.healthit.gov/sites/default/files/onc_pghd_final_white_paper.pdf**.

[25] Menachemi, Nir, and Taleah H. Collum. "Benefits and Drawbacks of Electronic Health Record Systems." Risk Management and Healthcare Policy 4 (2011): 47–55.

[26] Tucker, Katherine, Janice Branson, Maria Dilleen, Sally Hollis, Paul Loughlin, Mark J. Nixon, and Zoë Williams. "Protecting Patient Privacy When Sharing Patient-Level Data from Clinical Trials." BMC Medical Research Methodology 16 Suppl 1 (July 8, 2016): 77. **https://doi.org/10.1186/s12874-016-0169-4**.

[27] Office of the National Coordinator for Health Information Technology. "Conceptualizing a Data Infrastructure for the Capture, Use, and Sharing of Patient-Generated Health Data in Care Delivery and Research through 2024," Office of the National Coordinator for Health Information Technology, January 2018. Accessed June 6, 2019. https://www.healthit.gov/sites/default/files/onc_pghd_final_white_paper.pdf.

[28] Padrez, Kevin A. et al. "Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department," BMJ Quality & Safety 2016; 25: pp. 414-423. Accessed June 6, 2019.

[29] Huckvale, Kit, John Torous, and Mark E. Larsen. "Assessment of the Data Sharing and Privacy Practices of Smartphone Apps for Depression and Smoking Cessation." JAMA Network Open 2, no. 4 (April 19, 2019).

[30] Allen, Marshall. "Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates." ProPublica, July 17, 2018. https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates.

[31] PHG Foundation at the University of Cambridge, Identification and genomic data, December 2017, http://www.phgfoundation.org/documents/PHGF-Identification-and-genomic-data.pdf.

[32] Hu, Jane. "Genetic Tests Like 23andMe Promise the Moon and Stars—but What Can They Actually Tell Us?" Pacific Standard. Accessed June 25, 2019. https://psmag.com/social-justice/what-can-genetic-tests-really-tell-us.

[33] Office of Disease Prevention and Health Promotion. "Social Determinants of Health," Office of Disease Prevention and Health Promotion. Accessed June 6, 2019. https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health.

[34] LaPointe, Jacqueline. "How Addressing Social Determinants of Health Cuts Healthcare Costs." RevCycleIntelligence, June 25, 2018.

[35] Gottlieb, Laura M., and Hugh Alderwick. "Integrating Social and Medical Care: Could It Worsen Health and Increase Inequity?" The Annals of Family Medicine 17, no. 1 (January 2019): 77–81.

[36] World Health Organization, "Public health surveillance," Retrieved from https://www.who.int/topics/public_health_surveillance/en/.

[37] National Institutes of Health, "List of Registries," Retrieved from https://www.nih.gov/health-information/nih-clinical-research-trials-you/list-registries.

[38] National Center for Health Statistics, "Resources for Survey Participants," Retrieved from https://www.cdc.gov/nchs/nchs_for_you/survey_participants.htm.

[39] National Center for Health Statistics, "National Vital Statistics System," Retrieved from https://www.cdc.gov/nchs/nvss/index.htm.

[40] National Center for Health Statistics. "Resources for Survey Participants," National Center for Health Statistics. Accessed June 6, 2019 https://www.cdc.gov/nchs/nchs_for_you/survey_participants.htm.

[41] "NHIS - About the National Health Interview Survey," May 10, 2019. https://www.cdc.gov/nchs/nhis/about_nhis.htm.

[42] National Center for Health Statistics. "National Vital Statistics System," National Center for Health Statistics. Accessed June 6, 2019. https://www.cdc.gov/nchs/nvss/index.htm.

[43] Wartenberg, Daniel, and W. Douglas Thompson. "Privacy versus Public Health: The Impact of Current Confidentiality Rules." American Journal of Public Health 100, no. 3 (March 2010): 407–12.

[44] "Key Changes with the General Data Protection Regulation – EU GDPR." Accessed May 29, 2019. https://eugdpr.org/the-regulation/.

[45] Nelson, Hunter. "GDPR Principles: Data Minimization." Tortoise and Hare Software (blog), November 8, 2018. https://tortoiseandharesoftware.com/gdpr-principles-data-minimization/.

[46] Nelson, Hunter. "GDPR Principles: Data Minimization." Tortoise and Hare Software (blog), November 8, 2018. https://tortoiseandharesoftware.com/gdpr-principles-data-minimization/.

[47] Zuiderveen Borgesius, Frederik and van Eechoud, Mireille M. M. and Gray, Jonathan, Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework (November 24, 2015). Berkeley Technology Law Journal, Forthcoming; Institute for Information Law Research Paper No. 2015-04; Amsterdam Law School Research Paper No. 2015-46. Available at SSRN: https://ssrn.com/abstract=2695005.

[48] Van Schalkwyk, François, Stefaan Verhulst, Gustavo Magalhães, Juan Pane, Johanna Walker, Project Muse, and Project Muse. The Social Dynamics of Open Data, 2018. https://muse.jhu.edu/book/57819/.

[49] Callier, Viviane. "How Fake Data Protects Real People's Privacy." The Atlantic, July 30, 2015. https://www.theatlantic.com/technology/archive/2015/07/fake-data-privacy-census/399974/.

[50] "De-Identify Medical Images with the Help of Amazon Comprehend Medical and Amazon Rekognition." Amazon Web Services, March 19, 2019. https://aws.amazon.com/blogs/machine-learning/de-identify-medical-images-with-the-help-of-amazon-comprehend-medical-and-amazon-rekognition/.

[51] Ekblaw, Ariel, Asaph Azaria, John Halamka, and Andrew Lippman. "A Case Study for Blockchain in Healthcare: 'MedRec' Prototype for Electronic Health Records and Medical Research Data." White Paper. Boston, MA: MIT Media Lab, August 2016. https://www.healthit.gov/sites/default/files/5-56-onc_blockchainchallenge_mitwhitepaper.pdf.

[52] "About the GDC | NCI Genomic Data Commons." Accessed June 24, 2019. https://gdc.cancer.gov/about-gdc.

[53] "Regulations.Gov - Comment." Accessed May 16, 2019. https://www.regulations.gov/document?D=HHS-OCR-2018-0028-1156.

[54] "Trusted Exchange Framework and Common Agreement (TEFCA) Draft 2." The Office of the National Coordinator for Health Information Technology, April 2019.

[55] "Precision Medicine Initiative: Privacy and Trust Principles | National Institutes of Health (NIH) — All of Us." Accessed May 16, 2019. https://allofus.nih.gov/about/program-overview/precision-medicine-initiative-privacy-and-trust-principles#precision-medicine-initiative-privacy-and-trust-principles-4.

[56] "Million Veteran Program (MVP)." Accessed June 28, 2019. https://www.research.va.gov/mvp/.

[57] "PatientsLikeMe." PatientsLikeMe. Accessed June 28, 2019. https://www.patientslikeme.com/.

[58] "Million Veteran Program (MVP)." Accessed June 26, 2019. **https://www.research.va.gov/MVP/**.

[59] "PatientsLikeMe | About Us." PatientsLikeMe. Accessed June 28, 2019. **https://www.patientslikeme.com/about/privacy**.

[60] "What Is Protected Health Information (PHI) or Personal Health Information? - Definition from WhatIs.Com." SearchHealthIT. Accessed May 22, 2019. **https://searchhealthit.techtarget.com/definition/personal-health-information**.

[61] "Rules and Policies - Protecting PII - Privacy Act." Accessed May 22, 2019. **https://www.gsa.gov/reference/gsa-privacy-program/rules-and-policies-protecting-pii-privacy-act**.

[62] "What Are the Differences between Electronic Medical Records, Electronic Health Records, and Personal Health Records? | HealthIT.Gov." Accessed May 28, 2019. **https://www.healthit.gov/faq/what-are-differences-between-electronic-medical-records-electronic-health-records-and-personal**.

[63] Medicare, Centers for, Medicaid Services. "Electronic Health Records Overview," March 26, 2012. **https://www.cms.gov/Medicare/E-health/EHealthRecords/index.html**.

[64] "Health Information Exchange | HealthIT.Gov." Accessed May 22, 2019. **https://www.healthit.gov/topic/health-it-basics/health-information-exchange**.