The NEW ENGLAND JOURNAL of MEDICINE

Please review the Supplemental Files folder to review documents not compiled in the PDF.

HIPAA and the Torrential Leak of 'De-Identified' Electronic Health Record Data

Journal:	*New England Journal of Medicine*
Manuscript ID	21-02616
Article Type:	Perspective
Date Submitted by the Author:	12-Feb-2021
Complete List of Authors:	Mandl, Kenneth; Boston Children's Hospital, Computational Health Informatics Program; Perakslis, Eric; Duke Clinical Research Institute
Abstract:	

HIPAA and the Torrential Leak of 'De-Identified' Electronic Health Record Data

Kenneth D. Mandl, MD, MPH

Computational Health Informatics Program, Boston Children's Hospital, Boston, MA

Department of Biomedical Informatics, Harvard Medical School, Boston, MA

Eric D Perakslis, PhD

Duke Clinical Research Institute, Duke University Medical Center, Durham, NC

Correspondence to:

Kenneth D. Mandl, MD, MPH

Director, Computational Health Informatics Program

Boston Children's Hospital

300 Longwood Avenue-LM5506, Mailstop BCH3187

Boston, Massachusetts 02115

617.355.4145

kenneth_mandl@harvard.edu

Word count: 1209

The permissible sharing of identified data to be used for treatment, payment and operations, by covered entities with their business associates, has led to a torrent of electronic health record (EHR) data flowing out of healthcare provider silos. The Health Insurance Portability and Accountability Act (HIPAA) also permits business associates to de-identify data on behalf of the covered entity, and once those data are de-identified, the business associate may use them freely, unless contractually prohibited from doing so. These circumstances enabled the rise of a multibillion-dollar industry comprising dozens of health data aggregation companies and hundreds of health data tool and technology companies aggregating, linking, and monetizing EHR data.

This phenomenon has been amplified by the explosion of data production after the HITECH Act promoted widespread adoption of EHRs, predicated on the aspiration for accurate and complete patient records, rapid real-time learning, better coordinated care, accelerated biomedical discovery, and digital and data exchange directly with patients. It is ironic that while physicians and their patients still find it difficult to obtain complete medical record information in a timely fashion, a provision in the HIPAA privacy rule permits massive troves of their digital healthcare data to traverse the medical-industrial complex unmonitored and unregulated.

While some may question the necessity or viability of privacy in the digital age, privacy is essential for limiting government and industry power, bolstering self-determination and individual preferences, and preserving reputations. Though the HIPAA Privacy Rule governs covered entities' uses of identifiable data, it does not apply when data are considered to be de-identified by expert determination or the safe harbor method, which requires that 18 specified identifiers are removed.

On the one hand, assembly of vast datasets advances the laudable objectives of a learning healthcare system, whereby the data routinely collected during the care delivery process continuously drive ever more intelligent treatment decisions. However, markets for secondary use of patient data, as configured today, do not always serve the best interests of patients or the public. Consider, for example, a data aggregation company that drives up the price of drugs by targeting physicians and patients for pharmaceutical detailing.

Though deidentification is often treated as infallible, it is not. Individuals can often be reidentified using only a handful of attributes.[1] Further, de-identification technologies relying on encryption may yet be highly vulnerable to advances in computing. When there are no contractual controls to govern data produced by the covered entity that is shared with a business associate, if something goes wrong, the harms accrue only to the patients; there is no comprehensive data privacy law in the U.S and none of the patchwork of laws or regulations protects patients from potential uses of de-identified data. Absent are duty to report when data has been re-identified or linked to external data sources (e.g., financial records). There is no opportunity for an individual seeking redress for re-identified data, and no external way to verify or ensure adequacy of de-identification.[4]

Research on de-identified data rarely is under institutional review board (IRB) oversight. Further, the free market for de-identified data can preclude desirable rigor on data quality, especially as an analyst or researcher cannot go back to the original data source for removal of duplicates, proper data validation, or correction of errors. These shortcomings contribute to potentially error-prone and non-reproducible research.[2]

As the practice of medicine becomes more digitally-driven, because very few healthcare provider organizations have data on large enough populations for even basic functions, such as diagnosis and decision support, it will be increasingly in their interest to procure access to multi-institutional data. Currently, when healthcare providers see short term gains through the sale of data, they often cede this benefit to third parties with no data-driven knowledge accruing back. At some point affordability of these multi-institutional datasets may become a challenging issue for health care providers attempting to practice in a digital medicine ecosystem.[3] If the goal of a healthcare institution is simply commercial gain, the extant health data market might offer short term advantages. If the goals are a learning health system, high quality research, and entry into durable and transparent compacts with patients, there are better ways.

Broadly speaking, two approaches can address the torrential leak of de-identified health record data. One is to establish best practice among the data providers—usually HIPAA covered entities. The other is to strengthen legal and regulatory protections.

Healthcare institutions can strengthen protection of their interests and their patients' privacy by treating de-identified data more like protected health information. First of all, patients should be informed by their healthcare providers, through consents to treat and privacy notices, that upon arrival to the hospital that their data may be used in a learning healthcare system, and when appropriate shared with commercial parties.

Secondly, when data must be externalized, proper contractual controls should be implemented to ensure that the data never passes outside the construct of the specific arrangement, cannot be linked with other datasets, and that re-identification is prohibited. This approach also enables disclosures to patients allowing full transparency of how the institution is

sharing and using their data. Ideally, the covered entity data provider is a genuine collaborator in the effort, not just a data supplier.

Thirdly, a preferred approach to permitting use of data is so-called "behind the glass" access where data do not leave the institution and instead the analytics are brought to the data. One approach is to create shared analytic workspaces, or enclaves, where health systems can interact directly with data and technology innovators but without the need to externalize patient data. Data can be combined on a project-by-project basis, under protective contracts or data use agreements. To aggregate multi-institutional datasets, federated analytic models permit data to stay at the originator sites.

To improve legal and regulatory oversight, first of all, states, and the federal government, should join California in making re-identification of de-identified health data illegal.[4] This would have an important effect on the market, but notably might not prevent a malicious actor from exposing a patient's medical information.

Secondly, it is worth examining closely the advantages and drawbacks of the European Union's Global Data Protection Regulations (GDPR) 'Right to Erasure,' which ensures a data subject can choose to be erased from a dataset upon request, without undue delay when the data are being used for purposes other than for which it was initially collected; the data subject withdraws consent or objects to the use; the data are being used unlawfully, required for issues of legal and/or regulatory compliance and other similar situations.[5] Right to erasure shifts the burden of risk from the patient back to institutions. However, if the patient does not opt out at the beginning of the project, they should understand that their records cannot be subsequently located in a truly de-identified dataset. In crafting laws or regulations inspired by GDPR

provisions, positive uses should be preserved. In a learning healthcare system for example, an opt-out model might severely bias the dataset and prevent accurate analytics.

HIPAA and its privacy rule were crafted in the pre-electronic health record era. There is now an opportunity for health systems, legislators, and regulators to protect health record data beyond what HIPAA currently mandates, while promoting and supporting beneficial uses toward improving health and optimizing health care delivery.

References

1. Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 2019;10(1):3069.
2. Mehra MR, Ruschitzka F, Patel AN. Retraction-Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. Lancet 2020;395(10240):1820.
3. Mandl KD, Bourgeois FT. The Evolution of Patient Diagnosis: From Art to Digital Data-Driven Science. JAMA 2017;318(19):1859–60.
4. California Legislature Adopts CCPA Exemption for Information Deidentified in Accordance with the HIPAA Privacy Rule [Internet]. 2020 [cited 2021 Jan 19];Available from: https://www.insideprivacy.com/ccpa/california-legislature-adopts-ccpa-exemption-for-information-deidentified-in-accordance-with-the-hipaa-privacy-rule/
5. Bovenberg J, Peloquin D, Bierer B, Barnes M, Knoppers BM. How to fix the GDPR's frustration of global biomedical research. Science 2020;370(6512):40–2.