Journal of the Society for Clinical Data Management

Vulcano D. Compliance with US Privacy Regulations when Using Health Records Data for Real World Evidence Purposes. *Journal of the Society for Clinical Data Management*. 2023; 3(3): 3, pp. 1–9. DOI: https://doi.org/10.47912/jscdm.233

EDUCATION AND PROFESSIONAL DEVELOPMENT

Compliance with US Privacy Regulations when Using Health Records Data for Real World Evidence Purposes

David Vulcano

The research use of healthcare records in the United States is protected by a wide variety of regulations and ethical constructs. While healthcare providers are intimately familiar with these, life science companies and other researchers are often not. This lack of awareness often results in protocols and contracts that, although well intentioned, can cause increased delays and costs due to resistance from the curators of the electronic health records. The better that protocols, contracts and budgets are written with these considerations in mind, the better opportunity we have to generate real world evidence using electronic health records with less cost, greater speed and most importantly in a manner that does not compromise individual privacy to achieve societal benefits.

Keywords: Manage Clinical Research Data; Data Collection Structure; Design Form; Import

Introduction

Electronic health records can be used in research for many purposes, such as improving protocol design and feasibility; as a synthetic control arm to a clinical trial in which placebo controls would be unethical; and as raw material for artificial intelligence and machine learning engines to develop algorithms for various prediction and prevention reasons, replicating previous studies and other reasons.[1] While there are benefits to electronic health records, researchers are reminded that they are being entrusted with some of people's most private and sensitive information. The research use of health care records in the United States is protected by a wide variety of regulations. Concerning privacy and security of the regulations, at the federal level there primarily is the Health Insurance Portability and Accountability Act of 1996, (HIPAA) codified primarily in 45CFR164 (to which health care providers subject to the law are called "Covered Entities") as well as the lesser-known Confidentiality Of Substance Use Disorder Patient Records codified at 42CFR2 (also known as "SAMHSA Part 2"). There are also a wide variety of state regulations that provide additional privacy protections, especially surrounding behavioral health conditions and HIV status. While the privacy regulations govern all use and disclosure of protected health information, there are also additional regulations that surround the secondary use of data for research purposes (specifically 45CFR46, also known as the "Common Rule"). While the Common Rule technically applies only to federally funded research, many institutions apply the same principles and criteria consistently across all data research, regardless of the funding source. Unlike HIPAA, which only governs the use of protected health information, the Common Rule protects the private identifiable information about any human subject of the research, noting that oftentimes the subject(s) of the research may be employees, providers or other non-patient individuals affiliated with the health system. This article describes some high-level obligations to which health systems must adhere in order to support the secondary use of their data. Sponsors and other researchers should be aware of these obligations when writing protocols and establishing budgets that are sufficient to cover the activity(ies).

Common Misuse of the Word "De-identified"

All too often, a health system will be presented with protocols and/or contracts that state something to the effect that, "the data will be de-identified prior to sending it to the Sponsor". When it pertains to data governed by HIPAA, the word "de-identified" takes on a regulatory definition over and above what the scientific and/or vernacular definition entails. Specifically, for protected health information to be considered "de-identified", HIPAA's "safe harbor" for de-identification requires the removal of 18 specific data elements from the data.* Similarly, the description "limited data set" also takes on regulatory meaning over and above the scientific and/or vernacular use. In essence, unlike using the dictionary definition of the word "limited" to define a dataset as one that is restricted in size or amount, to be qualified as a Limited Data Set under HIPAA requires certain criteria to be met. See **Table 1** and **Figure 1** for more information on the HIPAA definitions of these terms.

HCA Healthcare, Society for Clinical Research Sites (SCRS), SCDM Advisory Board, US
david.vulcano@hcahealthcare.com

Table 1: De-identified Data Set and HIPAA Limited Data Set.

The following "HIPAA identifiers" of the individual or of relatives, employers, or household members of the individual, are removed:	**De-Identified Data**	**Limited Data Set**
1. Names	Not Allowed	Not Allowed
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000	Not Allowed (unless excepted for in the criteria)	No postal address information, other than town or city, State, and zip code;
3. All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older	Not Allowed (unless excepted for in the criteria)	All elements of date and all ages are allowed
4. Telephone numbers	Not Allowed	Not Allowed
5. Fax numbers	Not Allowed	Not Allowed
6. Email addresses	Not Allowed	Not Allowed
7. Social security numbers	Not Allowed	Not Allowed
8. Medical record numbers	Not Allowed	Not Allowed
9. Health plan beneficiary numbers	Not Allowed	Not Allowed
10. Account numbers	Not Allowed	Not Allowed
11. Certificate/license numbers	Not Allowed	Not Allowed
12. Vehicle identifiers and serial numbers, including license plate numbers	Not Allowed	Not Allowed
13. Device identifiers and serial numbers	Not Allowed	Not Allowed
14. Web Universal Resource Locators (URLs)	Not Allowed	Not Allowed
15. Internet Protocol (IP) addresses	Not Allowed	Not Allowed
16. Biometric identifiers, including finger and voice prints	Not Allowed	Not Allowed
17. Full-face photographs and any comparable images	Not Allowed	Not Allowed
18. Any other unique identifying number, characteristic, or code, except as permitted by [the regulations regarding "Re-identification"]; and	Not Allowed	Allowed
19. The Covered Entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.	Included in requirements	Not included in requirements

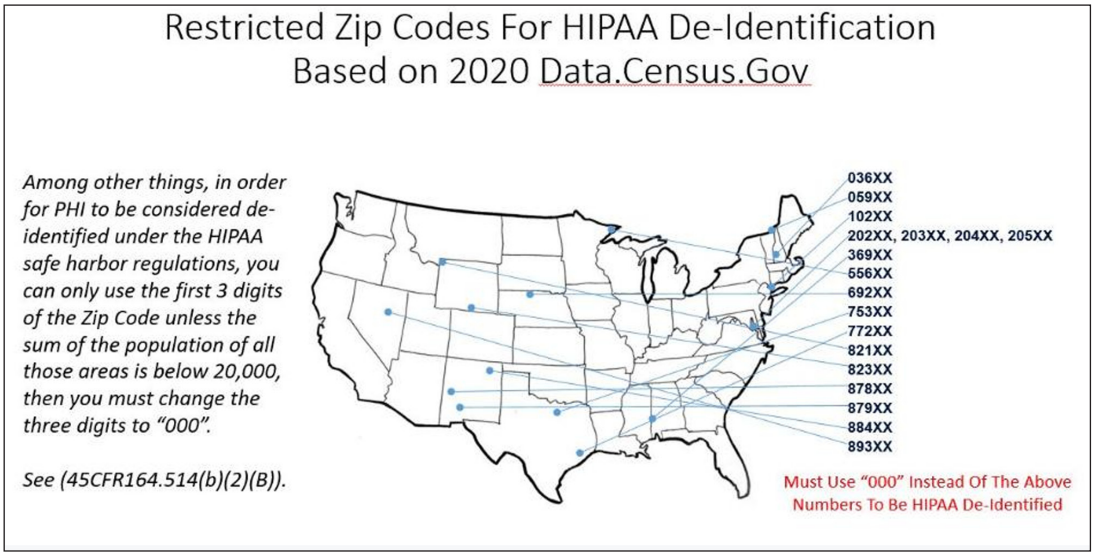

Figure 1: Restricted Zip Codes in A HIPAA De-Identified Data Set.

While researchers may be well versed in research regulations, such as 45CFR46 (also known as The Common Rule) that offer privacy protections, the Common Rule's concept of identifiability is described in general terms and not as descriptive as HIPAA.[2] Thus, while most researchers are not surprised to see items such as name, street address and Social Security Number on that list, they are often surprised to learn that elements of date other than year (ie, date of birth, date of consent, date of admission, age if over 89, patient's zip code and medical device serial number) are on the list. Strategies are often available to get the information needed without compromising the needs of the science. For example, if the protocol required length of hospital stay, just asking the provider to disclose that calculated number instead of asking for Date of Admission and Date of Disclosure solves the issue. Similarly, using relative dates (ie, Day −1, Day 0, Day 2, Day 5) instead of asking for the actual dates can also accomplish many tasks. Some solutions for Age>89 are also available in that the database can make accommodations to assure age over 89 is not abstracted. This is usually accomplished by the system replacing "ages over 89" with a grouping in the Age field (ie, ">89" or "90+", noting that this may require the alteration of the Age field to be alphanumeric which can affect the ability to do calculations) or leaving the Age field blank and creating a separate Boolean field of "Over 89". Regardless, the Electronic Health Record (EHR) query and/or Case Report Form (CRF) should hardwire in a mechanism to ensure that this identifier is not disclosed if the protocol or contract requires the data to be de-identified.

Problems occur when there are inconsistencies between the protocol, contract and CRF/database elements and other study-related documents. When the protocol or contract states something to the effect of "patient data will be de-identified prior to sending to Sponsor", a Covered Entity will generally interpret that as meaning meeting the HIPAA criteria of de-identification. A diligent Covered Entity will then turn to the data elements to be submitted and validate that it meets HIPAA's requirements. All too often there are required data fields, such as Date of Consent, Date of Visit/Procedure, or Age that do not have the ability to accommodate for age over 89. When this occurs, the diligent Covered Entity will require the Sponsor to either alter the database requirements or the protocol, contract and any other governing documents to eliminate this inconsistency. **Table 2** displays examples of such inconsistencies and methods to address them. Assuring internal consistency will prevent the need for costly protocol and contract amendments, especially in multicenter protocols.

An easy fix for data points, such as elements of date other than year, age if over 89, and zip code, is to not reference the data set in the protocol, contract and affiliated documents as "de-identified" but as a HIPAA Limited Data Set. Similar to the term "de-identified" a "limited data set" also has regulatory meaning in HIPAA quite different from the dictionary definition of "limited". A HIPAA Limited Data Set does allow for those three fields (elements of date other than year, age over 89, and elements of geography smaller than a state that include town or city, State, and full zip code); however, due to the large increase of the ability to re-identify individuals using those data points, HIPAA requires specific language to be in a data use agreement that, among other things, imposes security and privacy obligations upon the recipient. The necessity becomes apparent when realizing that using off-the-shelf software and publicly available information (such as voter registration databases or U.S. Census data), the combination of Year of Birth, Gender, and 3-Digit ZIP Code (ie, a HIPAA de-Identified Data Set) is unique for approximately 0.04% of residents in the United States;[3] however, the combination of a patient's Date of Birth, Gender, and 5-Digit ZIP Code (ie, a HIPAA Limited Data Set) has been found to be unique to anywhere between 63% and 87% of residents in the United States.[4,5]

Table 2: Examples of Consistencies and Inconsistencies of the word "De-identified".

Document	Data Fields Disclosed	Consistent?
Protocol states "Information will be de-identified prior to sending…"	[Date of Admission] and [Date of Discharge]	No. Elements of date other than year are not in the HIPAA safe-harbor de-identification regulation.
Contract states "Information will be de-identified prior to sending…"	[Year of Admission] and [Length of Stay]	Yes. Neither field is excluded from the HIPAA safe-harbor regulation of de-identification.
Protocol states "Information will be de-identified prior to sending…" and inclusion/exclusion criteria does not exclude ages over 89	[Age]	Possibly. To meet the HIPAA safe-harbor regulation of de-identification, age over 89 must be eliminated. Since the protocol allows records from those aged over 89, methods need to be used to mask that age.
Contract states "Information will be de-identified prior to sending…"	[Date of Admission] and [Date of Dosing]	No. Elements of date other than year are not in the HIPAA safe-harbor de-identification regulation.
Protocol states "Information will be de-identified prior to sending…"	[Year Of Admission} and [Number of Days From Admission To Dosing]	Yes. Neither field is excluded from the HIPAA safe-harbor regulation of de-identification.

Disclosing Identifiable Data

There are essentially two ways to disclose identifiable protected health information (ie, information that is not de-identified or in the form of a HIPAA Limited Data Set without the accompanying required elements in the data use agreement) and those are with or without the HIPAA Authorization of the patient to which the data pertains. In many cases of data-only research (ie, not clinical trials in which the patients sign informed consent documents with accompanying HIPAA Authorization language) the seeking of prior informed consent and HIPAA Authorization causes the impracticality to achieve the research goals; meaning that it would be impracticable to *perform the research* (eg when the obtaining of such consent leads to invalid study outcomes such as due to selection bias or when the consent process adds potential risk to the subjects), not simply impracticable to obtain consent and HIPAA Authorization due to financial or administrative burdens. T regulations therefore allow for the overseeing Institutional Review Board (IRB) or institutional privacy board to waive the HIPAA Authorization of the patient. Of note, this is a separate action to waiving the requirement of informed consent to participate in the research. While the criteria are similar, the criteria to waive informed consent speaks to the waiver of obtaining a prior understanding of the research as a whole and an affirmative voluntary agreement to participate whereas the waiver of HIPAA Authorization speaks to the waiver of obtaining the individual's permission to disclose their identifiable health information for research purposes.

The regulation to waive such HIPAA Authorization, among other things, requires the IRB to concur that (i) the research could not practicably be conducted without access to and use of the identifiers; (ii) the research could not practicably be conducted without the waiver or alteration of the HIPAA Authorization to release those identifiers; and (iii) there is an adequate plan to a) protect the identifiers from improper use and disclosure, b) destroy the identifiers at the earliest opportunity, and c) written assurances are in place that the identifiers, outside of defined exceptions, will not be reused or disclosed to (shared with) any other person or entity.[6] In addition to the IRB waiver, the regulations also require documentation of this release that must be disclosed to the patient upon their request. Thus, when identifiable Protected Health Information (PHI) is released for research purposes under an IRB waiver of HIPAA Authorization, the Covered Entity must log this release on the patient's Accounting of Disclosure log, and indicate things that include but are not limited to (i) the date of the disclosure; (ii) the name and other contact information about the receiving research entity; and (iii) descriptive information about the research protocol.[7] In essence, although the patient does not need to give prior permission or be actively notified of the release due to the privacy protections by design that are vetted by the IRB, they do have the right to request the list of all the research protocols their identifiable health information was disclosed for and who to contact about it.

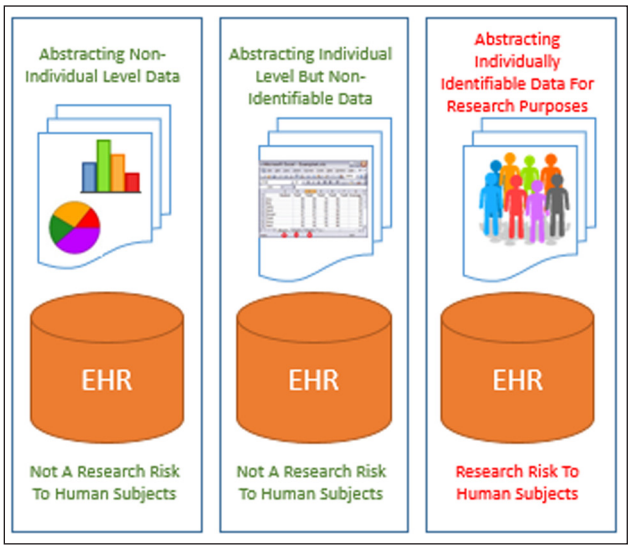

Figure 2: When Data Abstraction for Research Is A Risk To Human Subjects.

The research regulations, specifically those under 45CFR46 (also known as "The Common Rule") jointly regulate research privacy and while in theory are complimentary, they are not the same.[8] The abstraction of a research data set from a source is a research risk to human subjects when that information is identifiable. **Figure 2** provides a visual on different kinds of abstractions and whether or not each introduces a research risk.

While the Common Rule protects living individuals and their private identifiable information with obligations such as IRB oversight and informed consent, the 2018 revision to the rule does recognize that non-interventional data research (even identifiable data) done under HIPAA regulations has sufficient protections to be exempt from all Common Rule requirements.[9] While this does not apply to identifiable data not protected by HIPAA (such as identifiable data on employees and providers), this revision did alleviate a lot of duplicative protections that increased cost and delays in research. The US Department of Health and Human Services (HHS) also describes that even if the researchers are engaging in non-exempt research, that if the sole involvement of the health care provider is releasing the data (even identifiable data) that they are not engaged in the research and are exempted from most Common Rule requirements including not having to certify IRB oversight of the research.[10] Assuming the health care provider interprets the Common Rule term of "the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects" with HIPAA definition of "de-identified" then the Common Rule and HIPAA regulations are fairly complementary and do not complicate each other.

Identifiable Data in EHRs That Is Not Patient Data

From time to time the data gathered for the research is not only patient data but also that of the health system's providers and employees themselves (eg attending

physician name). When this occurs, although their identifiers are in the EHR, their data is not protected by HIPAA and so other regulations must be turned to. While the US does not have a law equivalent to the European Union's General Data Protection Regulation (GDPR), some US laws exist that may come into play for certain situations (such as The Family Educational Rights and Privacy Act (FERPA) for educational records and other human resources laws). Other laws are emerging that may offer additional protections for individually identifiable data but only for focused situations. Although the regulatory framework for the protection of provider and employee data is less structured than HIPAA regulations, many health care entities impose similar protections and definitions (eg such as defining it as a de-identified data set) in policies, procedures and contracts for the protection of their provider's and employee's private information.

Alternatives To Full Disclosure

There are alternatives to a health system in providing the data to an outside researcher. One option is for the health system to provide only the aggregated results to the external researchers. This can be achieved by either the health system doing the analysis itself or, for simpler analyses, building a front-end user interface allowing the researcher to input basic query parameters and see the results without ever seeing or having direct access to the raw data. Health systems may charge a fee to offset the cost for providing such an infrastructure. While this may help get the question answered without the need to disclose the raw data, this may not be optimal for some study designs, especially when data sets from multiple providers are desired to be combined.

Certain kinds of research require the matching of individuals across multiple health care providers and the temptation is to use patient identifiers to be that matching key. For example, to match a patient across multiple providers, a researcher may request the patient's Social Security Number accompany the EHR data with the promise to delete that information once they have collected all data and matched the patient across providers. Even with an IRB Waiver of HIPAA Authorization, health care providers may be hesitant to disclose highly sensitive data elements like Social Security Number. In these cases instead of using high risk matching keys such as Social Security Number or a combination of identifiers (such as "initials + last 4 digits of the Social Security Number" which would also make the data set not de-identified under HPAA standards), the varying covered entities can agree on a one-way cryptographic hash that tokenizes the identifiers into a reidentification key. This way, allowable under HIPAA[11] as a re-identifier, all HIPAA covered entities are releasing a de-identified data set that allows for high quality matching of individual records across those providers. Although the researcher can see the unrecognizable token, they would not be able to reverse it to generate the identifiers. Note that it is best that the external entities receiving the data set not know the hashing algorithm or the variables going into it as while they may not be able to reverse engineer the token, they may be able to create a matching token table for re-identification purposes. For example, if the nefarious individual knows the hash algorithm and that it was based only on Social Security Number, they can easily create a table tokenizing all possible Social Security Numbers with that hash to re-identify individuals. While there are many technical ways (eg increasing the number of variables for the hash, a process called "salted hashing" and others) as well non-technical ways of decreasing this risk (eg contractual obligations that the recipient will not attempt to re-identify individuals), cryptographic can further reduce but never completely eliminate the risk of re-identification attack.

"Who Cares If It's Legal. We Still Can't Do It"

Ethical Obligations: In general, health care providers take the fact that people have entrusted them with some of their most sensitive information very seriously. Covered entities are required by HIPAA to make available their Notice of Privacy Practices and included in that notice is often a statement on their use and disclosure of EHR data for research. Although a health care provider can, for example through de-identification, disclose EHR data without many regulatory obligations, they are often judicious in their disclosure, taking into consideration the protocols' alignment with the provider's mission and their written and unwritten promises to their patients.[12] For example, a hospital operating under a Catholic banner considers intentional selective abortion or in vitro fertilization to be against their adherence to the Catholic faith's doctrines on bioethics.[13] Thus, while the de-identification of that hospital's data could legally be used for research supporting those procedures, ethically it would arguably violate an implied covenant with their patients. Specifically, any patient may feel wronged by their health provider if their records were not protected from use or disclosure in research they find morally objectionable.[14] Additionally, one may ethically take issue with the purpose of the research; such as while one may see the disclosure of a data set for the purposes of research pursuant to the publication of evidence to improve patient care as consistent with their mission, they may not view the disclosure of the exact same data set for other kinds of research purposes such as research for marketing, sales or political purposes. This ethical construct applies not only to the business decision to not release data but also to impose prohibitions from any unauthorized secondary use of disclosed data to the extent practical. It is worth noting that with the emerging data transparency laws and data sharing obligations of medical journals,[15] this is getting harder for a health provider to prevent, thus increasing their risk and discomfort.

Business Considerations: As health care systems are continually challenged with the sustainability of their core business and the ever increasing demands of regulatory and business/cybersecurity requirements in data protection, there are often fewer resources available to support a third party's research interests. Even despite offers of adequate compensation for their efforts, the human and technical

resources may simply not be available. It is understood that there has been some mapping of standards between Health Level Seven International (HL7) and Clinical Data Acquisition Standards Harmonization (CDASH) as well as the use of interoperability resources such as Fast Healthcare Interoperability Resources (FHIR). However many requests from sponsors still impose additional burdens on the health care providers that prohibit the willingness to take on the effort. To the extent scientifically possible, restricting protocol information needs to the data that is most easily exchanged by health care providers (specifically, United States Core Data for Interoperability (USCDI)) may help to alleviate some of the burdens that prevent a provider's ability to contribute.

Prohibition from Subsequent Use

Using data for purposes other than it was originally gathered for (or permitted for use) is a growing global public concern in many areas (ie, social media, big tech, etc.), and such use of health care data (and biospecimens[16,17]) is no exception. In research, this problem is manifested in a researcher using data permitted for Purpose A for Purposes B, C and D. Such new purposes could include additional research projects, commercial purposes, and others. Even if a secondary use is permissible by law, a health care provider will often impose contractual obligations to prevent a data recipient from secondary use, specifically calling out the prohibition of re-identifying the data without permission as the ability to successfully match a HIPAA de-identified data set (or a HIPAA Limited Data Set) by combining it with other data (which has become much easier since the HIPAA law written in 1996).[18,19] The contractual obligations may even go so far as to specify the prevention of manipulations with the intent to prepare for secondary use, such as stripping the identifiers, making derivatives of the data set, or other strategies to render the data set unregulated. Such efforts are reaffirmed by the health care provider taking the stance that they are only licensing the data for the stated use and thus reinforcing that the recipient never has nor gains any ownership over the data. Regardless, the ownership of the data and its derivatives post-release is often a contentious issue in the intellectual property provisions of research and data use agreements, and can cause challenges and delays for all parties.

Living vs. Deceased Individuals

The Common Rule protections (ie, need for IRBs and research consents) are for living individuals.[20] HIPAA protected health information is governed up to 50 years post-decease.[21] HIPAA does allow for certain research uses and disclosures on identifiable PHI of decedents without a HIPAA Authorization (or other HIPAA compliant patient-directed request) from the legally authorized representative during the 50 year post-death protection rule if the "covered entity obtains from the researcher: (A) Representation that the use or disclosure sought is solely for research on the protected health information of decedents; (B) Documentation, at the request of the covered entity, of the death of such individuals; and (C) Representation that the protected health information for which use or disclosure is sought is necessary for the research purposes."[22] Despite the permissibility, there remain challenges in implementation, particularly germane to consensus between the researcher and health provider on acceptable documentation of death as well as ethical involvement of next of kin despite the allowance.[23]

Preparatory To Research

Preparing a research protocol (or preparing to conduct a final research protocol) is fundamentally different to conducting a research protocol, from both a HIPAA and a research regulatory perspective, but the line is somewhat difficult to define and/or observe. Although the HIPAA regulations use the term "preparatory to research", the regulation does not put forth a definition on what the phrase means and how its accompanying regulations differentiate from the regulations regarding the conduct of research with PHI. One can, however, infer two key concepts of the intended meaning from the examples given in HHS's guidance documents. The relevant guidance implies that an activity is "Preparatory To Research" essentially when 1) there is not yet a final written protocol and the researcher is engaging in activities needed to prepare one (or to prepare an amendment to an existing protocol); or 2) a written protocol does exist and the researcher is either conducting a feasibility assessment of that protocol (eg, to see if the data and/or subject population exists to support such a protocol) and/or engaging in activities to recruit subjects into that protocol (such as contacting potential individuals for purposes of seeking their HIPAA Authorization (or other HIPAA compliant patient-directed request) to use or disclose their PHI for the research study).

In most cases, protocol feasibility involves simply running a query that returns non-individual level aggregated results (eg, writing a query to return the total of patients in the database that meet inclusion/exclusion criteria). However in cases where the researcher is seeking the return of identifiable PHI for purposes Preparatory To Research, the researcher(s) must document three HIPAA required attestations prior to the access.[24] For reference, the three required statements are as follows: 1) use or disclosure is sought solely to review PHI as necessary to prepare a research protocol or for similar purposes preparatory to research; 2) no PHI is to be removed from the Covered Entity* in the course of review, and 3) the PHI for which use or access is sought is necessary for the research"

"To BAA or not to BAA. That is the Question"

Although it seems that the debate is over regarding if and when a research sponsor is a HIPAA defined "Business Associate" of the Covered Entity and thus requiring a Business Associate Agreement (BAA), from time to time this question comes up. The issue as to whether or not a receiving entity is a Business Associate imposes obligations and liability on both parties, so the decision must be made carefully. While each Covered Entity must make this determination on their own, there are certain aspects

of HIPAA that should be reviewed to prevent defining a relationship as a Business Associate relationship simply to move PHI to the external researching entity when there are other, more appropriate strategies of defendable data flow (eg, de-identification, Limited Data Set, IRB Waiver of HIPAA Authorization. Signed patient HIPAA Authorizations etc.). Relevant text to this discussion can be found in the following Federal Register quotes.

> "Disclosures from a Covered Entity to a researcher for research purposes as permitted by the Rule do not require a business associate contract. This remains true even in those instances where the Covered Entity has hired the researcher to perform research on the Covered Entity's own behalf because research is not a covered function or activity."[25]

> "A person or entity is a business associate only in cases where the person or entity is conducting a function or activity regulated by the HIPAA Rules on behalf of a Covered Entity, such as payment or health care operations, or providing one of the services listed in the definition of "business associate," and in the performance of such duties the person or entity has access to protected health information. Thus, an external researcher is not a business associate of a Covered Entity by virtue of its research activities, even if the Covered Entity has hired the researcher to perform the research. See *http://www.hhs.gov/ocr/privacy/hipaa/faq/business_associates/239.html*. Similarly, an external or independent Institutional Review Board is not a business associate of a Covered Entity by virtue of its performing research review, approval, and continuing oversight functions. However, a researcher may be a business associate if the researcher performs a function, activity, or service for a Covered Entity that does fall within the definition of business associate, such as the health care operations function of creating a de-identified or limited data set for the Covered Entity. Where the researcher is also the intended recipient of the de-identified data or limited data set, the researcher must return or destroy the identifiers at the time the business associate relationship to create the data set terminates and the researcher now wishes to use the de-identified data or limited data set (subject to a data use agreement) for a research purpose."[26]

With all that said, it remains a legal and risk-based decision of the Covered Entity as to if they believe the receiving party is being engaged by them as their Business Associate and if that is insisted upon, it is up to the recipient as to if they desire to take on the added legal risk that this arrangement imposes upon them.

Conclusion

Life science and health care delivery coexist in similar arenas, however the privacy and research regulations differ based on intent. While the internal *use* of information in electronic health records by a Covered Entity and other health care providers is relatively unrestricted, both fortunately and unfortunately the *disclosure* of the same data to outside researchers often crosses a regulatory and ethical line that invokes additional protections. The nuances of the disclosure of electronic health records for research purposes is much deeper than this article can explore; however, these high level considerations do provide the foundation for much of the discussion and planning. The better that protocols, contracts and budgets are written with these considerations in mind, the better the opportunity we have to generate real world evidence using electronic health records with less cost, greater speed, and, most importantly, in a manner that does not compromise individual privacy to achieve societal benefits.

Note

* There is an alternative to this "Safe Harbor" method to classify a data set as de-identified under HIPAA. Often referenced as the "Expert Determination" method, a statistician must apply complex statistical tests and analysis on the data set using common re-identification techniques to demonstrate that the re-identification risk is very small. However due to the subjectivity of this method, lack of definition on what is "very small", the documentation requirements and the regulatory risk of post facto challenges, this method is rarely used in the United States. Expert guidance should be sought if there is interest in classifying a data set as "de-identified" under the Expert Determination method as opposed to the Safe Harbor method.

Competing Interests

I am an employee of HCA Healthcare, the Honorary President of the Society for Clinical Research Sites (SCRS) and on the SCDM Advisory Board. Other affiliations not in the life science industry can be found on my LinkedIn profile.

References

1. **Vulcano D.** Understanding De-Identified Patient Data, Limited Data Sets, and Data Use Agreements. Policy & Medicine Compliance Update. Volume 7.1, January 2021.
2. **Rothstein MA.** Is deidentification sufficient to protect health privacy in research? *Am J Bioeth*. 2010 Sep; 10(9): 3–11. PMID: 20818545; PMCID: PMC3032399. DOI: https://doi.org/10.1080/15265161.2010.494215
3. **Sweeney L.** Testimony before the National Center for Vital and Health Statistics Ad Hoc Workgroup for Secondary Uses of Health Data. National Committee on Vital and Health Statistics. August 23, 2007. Accessed June 2022. https://ncvhs.hhs.gov/transcripts-minutes/transcript-of-the-august-23-2007-ncvhs-ad-hoc-workgroup-for-secondary-uses-of-health-data-hearing/
4. **Sweeney L.** Simple Demographics Often Identify People Uniquely. LIDAP-WP4, 2000 (2000). Accessed

August 17, 2023 https://dataprivacylab.org/projects/identifiability/paper1.pdf

5. **Golle P.** Revisiting the uniqueness of simple demographics in the US population. In Proceedings of the 5th ACM workshop on Privacy in electronic society (WPES '06). *Association for Computing Machinery, New York, NY, USA*. 2006; 77–80. DOI: https://doi.org/10.1145/1179601.1179615
6. **US Department of Health and Human Services.** Disclosures for Public Health Activities 45 CFR 164.512(i)(2)(ii). Revised April 2003. Accessed June 2022. https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/disclosures-public-health-activities/index.html
7. **US Department of Health and Human Services.** Accounting of Disclosures of Protected Health Information 45CFR164.528(b). Accessed June 2022. https://www.govinfo.gov/content/pkg/CFR-2002-title45-vol1/pdf/CFR-2002-title45-vol1-sec164-528.pdf
8. **Rothstein MA.** Currents in contemporary ethics. Research privacy under HIPAA and the common rule. *J Law Med Ethics*. 2005 Spring; 33(1): 154–9. PMID: 15934672. DOI: https://doi.org/10.1111/j.1748-720X.2005.tb00217.x
9. **US Department of Health and Human Services.** 45 CFR 46. 104(d)(4). Reviewed March 2021. Accessed June 2022. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html
10. **US Department of Health and Human Services.** HHS Guidance: Engagement of Institutions in Human Subjects Research (2008). Reviewed March 2016. Accessed June 2022. https://www.hhs.gov/ohrp/regulations-and-policy/guidance/guidance-on-engagement-of-institutions/index.html
11. **US Department of Health and Human Services.** Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule November 26, 2012. Accessed June 2022. https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf
12. **Kulynych J, Korn D.** The effect of the new federal medical-privacy rule on research. *N Engl J Med*. 2002 Jan 17; 346(3): 201–4. PMID: 11796857. DOI: https://doi.org/10.1056/NEJM200201173460312
13. **The Vatican.** INSTRUCTION DIGNITAS PERSONAE ON CERTAIN BIOETHICAL QUESTIONS (2008). September 2008. Accessed June 2022. https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_con_cfaith_doc_20081208_dignitas-personae_en.html
14. **Rothstein MA.** Is deidentification sufficient to protect health privacy in research? *Am J Bioeth*. 2010 Sep; 10(9): 3–11. PMID: 20818545; PMCID: PMC3032399. DOI: https://doi.org/10.1080/15265161.2010.494215
15. **International Committee of Medical Journal Editors (ICMJE) data sharing statement policy.** Accessed June 2022. https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html
16. **Mello MM, Wolf LE.** The Havasupai Indian tribe case--lessons for research involving stored biologic samples. *N Engl J Med*. 2010 Jul 15; 363(3): 204–7. Epub 2010 Jun 9. PMID: 20538622. DOI: https://doi.org/10.1056/NEJMp1005203
17. **Garrison NA, Cho MK.** Awareness and Acceptable Practices: IRB and Researcher Reflections on the Havasupai Lawsuit. *AJOB Prim Res*. 2013 Oct 1; 4(4): 55–63. PMID: 24089655; PMCID: PMC3786163. DOI: https://doi.org/10.1080/21507716.2013.770104
18. **Cohen IG, Mello MM.** Big Data, Big Tech, and Protecting Patient Privacy. *JAMA*. 2019; 322(12): 1141–1142. DOI: https://doi.org/10.1001/jama.2019.11365
19. **Rocher L, Hendrickx JM, de Montjoye YA.** Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun*. 2019 Jul 23; 10(1): 3069. PMID: 31337762; PMCID: PMC6650473. DOI: https://doi.org/10.1038/s41467-019-10933-3
20. **US Department of Health and Human Services.** 45 CFR 46.102(e)(1). Reviewed March 2021. Accessed June 2022. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html
21. **US Department of Health and Human Services.** 45 CFR 160.103. See paragraph (2)(iv) of the definition of "protected health information". Accessed June 2022. https://www.govinfo.gov/content/pkg/CFR-2013-title45-vol1/pdf/CFR-2013-title45-vol1-sec160-103.pdf
22. **US Department of Health and Human Services.** Disclosures for Public Health Activities 45 CFR 164.512(i)(1)(iii). Revised April 2003. Accessed June 2022. https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/disclosures-public-health-activities/index.html
23. **Huser V, Cimino J.** Don't take your EHR to heaven, donate it to science: legal and research policies for EHR post mortem. *J Am Med Inform Assoc*. 2014 Jan–Feb; 21(1): 8–12. Epub 2013 Aug 21. PMID: 23966483; PMCID: PMC3912713. DOI: https://doi.org/10.1136/amiajnl-2013-002061
24. **US Department of Health and Human Services.** Disclosures for Public Health Activities 45 CFR 164.512(i)(1)(ii). Revised April 2003. Accessed June 2022. https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/disclosures-public-health-activities/index.html
25. **Federal Register.** Vol. 67, No. 157 August 14, 2002. Page 53252.
26. **Federal Register.** Vol. 78, No. 17. January 5, 2013. Pages 5574-5575.

How to cite this article: Vulcano D. Compliance with US Privacy Regulations when Using Health Records Data for Real World Evidence Purposes. *Journal of the Society for Clinical Data Management*. 2023; 3(3): 3, pp. 1–9. DOI: https://doi.org/10.47912/jscdm.233

Submitted: 14 December 2022 **Accepted:** 16 May 2023 **Published:** 08 November 2023

Journal of the Society for Clinical Data Management is a peer-reviewed open access journal published by Society for Clinical Data Management.

OPEN ACCESS