5-1-2018

The Ethics in Synthetics: Statistics in the Service of Ethics and Law in Health-Related Research in Big Data from Multiple Sources

Sharon Bassan Ph.D.
Princeton University

Ofer Harel Ph.D.
University of Connecticut

THE ETHICS IN SYNTHETICS:

STATISTICS IN THE SERVICE OF ETHICS AND LAW IN HEALTH-RELATED RESEARCH IN BIG DATA FROM MULTIPLE SOURCES

SHARON BASSAN, PHD, JD AND OFER HAREL, PHD

I. INTRODUCTION 88
II. A DELICATE EQUILIBRIUM 89
III. DIFFERENT SCOPES OF PROTECTIVE REGULATIONS 93
IV. AN AUTHORIZATION TO USE INFORMATION FOR HEALTH-RELATED RESEARCH 97
V. POSSIBLE EXEMPTIONS FROM THE AUTHORIZATION REQUIREMENT 103
VI. SYNTHETIC DATA AS A MEANS TO FULFILL ETHICAL REQUIREMENTS 109
VII. A NEW RISK-BENEFIT BALANCE 112
VIII. CONCLUSION 116

I. INTRODUCTION

An ethical advancement of scientific knowledge demands a delicate equilibrium between benefits and harms, in particular in health-related research. When applying and advancing scientific knowledge or technologies, Article 4 of UNESCO's Universal Declaration on Bioethics and Human Rights, ethically justifiable research requires maximizing direct and indirect benefits, and minimizing possible harms.[1] The National Institution of Health [NIH] Data Sharing Policy and Implementation Guidance similarly states that data necessary for drawing valid conclusions and advancing medical research, should be made as widely and freely available as possible (in order to share the benefits), while safeguarding the privacy of participants from potentially harmful disclosure of sensitive information.[2] This paper discusses the challenges in the maximization of research benefit and the minimization of potential harms in the unique context of health-related research in Big Data from multiple sources, which are differently protected by the law.

Part I frames the ethical dilemma by discussing potential benefits and harms, showing the constant misalignment in health-related research in Big Data from multiple sources, between the benefits in the use of confidential information for scientific purposes, and the value in keeping confidentiality. In part II, the paper addresses existing regulations, their nature and legal coverage. It highlights the challenges prevailing when combining data from multiple sources that are differently protected by the law. Part III compares different requirements for consent or

[1] United Nations Educational, Scientific and Cultural Organization, *Universal Declaration on Bioethics and Human Rights* (2005).

[2] NIH, *NIH Data Sharing Policy and Implementation Guidance* (2003), http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

authorization to use persons' health information for research. It focuses on the difficulty of existing regulation to ensure those requirements when using multiple sources of data. Part IV investigates whether exemptions from the authorization requirement could prevail in the context of information that exceeds the protection of the HIPAA and the Protection of Human Subjects Regulations. In part V the paper proposes a solution is of a statistical nature, using the method of synthetic data to balance conflicting consideration. Part VI shows how the use of synthetic data can overcome some of the ethical challenges.

II. A DELICATE EQUILIBRIUM

The term "Big Data" is differently defined by users and policy makers. What it means is dramatically different to the media, business, health, or academic statistics communities, and to different regulatory bodies.[3] To our knowledge, there is no gold standard definition. Big Data is considered data on a massive scale in terms of volume, intensity, and complexity that exceed the ability of standard software tools to manage and analyze.[4] But also, "It is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets."[5] Laney coined the definition in the Big Data analytics world: volume (amount of

[3] Jordan JM & Lin DK, *Statistics for Big Data: Are Statisticians Ready for Big Data?* INT'L CHINESE STAT. ASS'N BULL (2014) 52: 133; Vlasses F et al., *Leveraging Technology for Research, in* NURSING INFORMATICS (2009), 692.

[4] *See, e.g.*, Snijders C et al., *"Big Data": Big Gaps of Knowledge in the Field of Internet Science*, INT'L J. INTERNET SCI. (2012), 7: 1.

[5] Boyd D & Crawford K, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, INFO. COMM. & SOC'Y (2012), 15: 662, 663.

data), velocity (speed of data in and out), and variety (range of data types and sources) – the 3V definition.[6]

As life is being recorded and quantified in ways hard to imagine a decade ago, there is great promise in Big Data research, in particular for health purposes. The literature often addresses medical records as the source for health-related research in Big Data,[7] for example, electronic records document multiple aspects of medical care: quantitative and qualitative data of patients, imaging records, providers' documentation of health care delivery (medication and other services), narratives and genetic information, all of which provide important information on a person's physical condition.[8] But as many details of our lives are documented and easily available for analysis, a variety of nontraditional or even unstructured data types contains different kinds of health-related information, which is combined with traditional medical databases. Medical or genetic data can be connected to data found on multiple sources: social media, surveillance videos, education, military service, exercise regimens, credit card payments for physician visit co-pays, visits to alternative practitioners, over the counter medications, home testing products, tobacco products, diet habits, or leisure time preferences.[9] Since a single database may not provide a complete picture of a patient's condition or health history, combining information from multiple sources is often necessary and allows ways of

---

[6] Laney D., *3D Data Management: Controlling Data Volume, Velocity, and Variety*, META GROUP RESEARCH NOTE 6 (2001), 70.

[7] *See, e.g.*, SHARONA HOFFMAN, ELECTRONIC HEALTH RECORDS AND MEDICAL BIG DATA: LAW AND POLICY Vol. 32 (Cambridge University Press, 2016).

[8] *See, e.g.*, Fan J. et al., *Challenges of Big Data Analysis*, NATIONAL SCI. REV. (2014), 1: 293, 295.

[9] Tasha Glenn & Scott Monteith, *Privacy in the Digital World: Medical and Health Data Outside of HIPAA Protections*, 16 CURRENT PSYCHIATRY REPORTS 494(2) (2014).

research previously not possible through traditional methods performed on a narrow spectrum of samples.[10]

On the one hand, health-related research combining multiple sources of Big Data offers the potential to explore hidden structures of the data, and extract important common features across data sets, in order to derive accurate results regarding complex questions in real-time.[11] Research can find correlation between multiple contextual variables found in public databases without having to interview a single patient. It contributes to a better understanding of people's life-style by creating an observational and even dynamic analysis, even when there are significant individual variations.[12] Such research can open the door for the promising world of personalized medicine and bring each individual customized treatments based on evidence drawn from their own lives. It can lead to the improvement of efficiency of health care delivery and public health decisions through standardized care and advance medicine, while saving costs nationally and globally.

On the other hand, there is a constant conflict between the benefit of using multiple sources of information and the value of preserving confidentiality of medical information. The need to maintain patients' privacy is an ethical obligation inherent in the physician-patient relationship, believed to be essential in order to generate better medicine – from diagnosis to treatment. The rationale underlying the doctrine of confidentiality of medical information is to enable patients to benefit

---

[10] *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research*, 177 (Lawrence O. Gostin, Laura A. Levit & Sharyl J. Nass, eds., 2009) (hereinafter *Beyond the HIPAA Privacy Rule*).

[11] Fan et al., *supra* note 8, at 294.

[12] Murdoch TB & Detsky AS, *The Inevitable Application of Big Data to Health Care*, 309 JAMA 1351, 1351 (2013).

from free and open communication regarding their status. Since information is essential to treatment, it is of no wonder that health service providers have access to all or portions of a patient's health records, however health care providers have a duty to avoid disclosure of medical information they obtain. In legal terms, the patient has a right of privacy, which aims to restrict the disclosure of confidential information.[13]

While the classical physician-patient model requires that information will be kept confidential, the more health information is present in electronic databases, the harder it is to maintain the privacy of individuals who are the subject of such information, and the greater is the potential misuse of information. Evidence of confidentiality breaches exist in State agencies in the US, healthcare organizations,[14] as well as in private organizations.[15]

Free communication is altered when patients fear that their sensitive health information might be electronically disclosed. Such fear may compromise the health

---

[13] For individual rights, see *Beyond the HIPAA Privacy Rule*, *supra* note 10, at 160.

[14] *E.g*., Glenn & Monteith, *supra* note 9 at 494(6) (reviewing a survey of 91 healthcare organizations, where 90 % reported at least one incident in the last two years while 38 % reported more than five incidents. When including only breaches involving at least 500 individuals, over 29 million patient health records have been compromised since 2009.); Goldman J, *91 Percent of Healthcare Organizations Suffered Data Breaches in the Past Two Years*, ESECURITY PLANET (May 12, 2015), http://www.esecurityplanet.com/network-security/91-percent-of-healthcare-organizations-suffered-data-breaches-in-the-past-two-years.html (The Ponemon Institute found that 91% of healthcare organizations have suffered at least one data breach since 2013, and 40% have suffered more than five); *Hacker Reportedly Selling Personal and Medical Data of 655,000 Patients*, INSURANCE FRAUD NEWS (June 27, 2016), http://www.insurancefraud.org/IFNS-detail.htm?key=22997 (In June 2016, a hacker was reported to sell copies of databases stolen from three unidentified U.S. healthcare organizations and one unnamed health insurer, containing data on nearly 10 million individuals on the dark web or prices ranging from about $96,000 to $490,000 in bitcoin for each database).

[15] Goldman J, *Data Breach at UCLA Health Exposes 4.5 Million People's Personal Information*, ESECURITY PLANET, (July 21, 2015), http://www.esecurityplanet.com/network-security/data-breach-at-ucla-health-exposes-4.5-million-peoples-personal-information.html (According to the Department of Health and Human Services, more than 120 million people have been compromised in more than 1,110 breaches since 2009 – a third of the U.S. population.)

care they seek. As studies show, for fear of disclosure to unauthorized persons, patients may withhold information, giving an incomplete or misleading description of their condition.[16] In recent surveys, a substantial number of people said they would withhold data from their physician due to privacy concerns related to technology.[17] Patients concerned with privacy violations are also less likely to seek care or return for follow-up treatment. They may seek care outside of their provider network, compromising the benefits of care coordination.[18]

Achieving ideal privacy or attempting to eliminate all possible breaches of confidentiality prevents society from the benefits inherent in research. However, researchers are often unaware of potential harms, especially given the large presence of health information in different sources of databases, some of which is voluntarily provided by users.[19] The next chapter reviews existing regulations, their nature and legal coverage. It highlights the challenges prevailing when combining data from multiple sources that are differently protected by the law.

III. DIFFERENT SCOPES OF PROTECTIVE REGULATIONS

In the case of health-related research in Big Data, policies and regulations should assure two dimensions of ethics: on the one hand, focused on harms, the

---

[16] Sankar P et al., *Patient Perspectives of Medical Confidentiality: A Review of the Literature*, 18 J GEN INTERN MED. 659 (2003).

[17] Agaku IT et al., *Concern about Security and Privacy, and Perceived Control Over Collection and Use of Health Information are Related to withholding of Health Information from Healthcare Providers*, 21 J. AM. MED. INFORMATICS ASSOC. 374 (2014) (over 10% of patients tend to withhold relevant medical information from their physicians).

[18] Glenn & Monteith, *supra* note 9 at 494(7) tb. 3; Scott C, *Is Too Much Privacy Bad for Your Health--An Introduction to the Law, Ethics, and HIPAA Rule on Medical Privacy*, 17 GA. ST. UL REV. 481, 529 (2000).

[19] Glenn & Monteith, *supra* note 9 at 494(2) (e.g.,one-third of U.S. consumers use YouTube, Facebook and Twitter for medical related discussions such as to check consumer reviews).

protection of information itself and of subjects whose information is used. On the other, focused on benefit, that the benefit from the research overweighs potential harms. The first dimension relates to confidentiality of personal information held in databases, whether or not they are health-related, Big Data, publicly available, or not. The second dimension addresses the equilibrium between privacy concerns and benefits from the research. Research in health-related data from multiple sources jeopardizes both aspects.

The legal frameworks addressing the release of data for health-related research purposes differ in levels of protection in terms such as the scope of the information covered and permitted disclosures: identified and de-identified information, held by entities covered or non-covered by laws, different uses of information, etc. This paper focuses on the HHS Protection of Human Subjects Regulations, and the Privacy Rule of Health Insurance Portability and Accountability Act (HIPAA), which represent the regulative framework of health-related research on the one hand, and of the operation of personal health information, on the other.[20]

The Protection of Human Subjects Regulations are the leading guidelines in ethical and regulatory issues in biomedical and behavioral research. According to definitions in the regulations, research on a human being is a systematic investigation, and research on a living individual about whom an investigator conducting research obtains identifiable private information constitutes research in a human being.[21] Typically, this regulation have not been applied to the core

[20] Protection of Human Subjects Regulations, 45 C.F.R. §46 (2007); The Health Insurance Portability and Accountability Act of 1996 (HIPAA), P.L. No. 104-191, 110 Stat. 1938 (1996).

[21] 45 C.F.R. § 46.102(d), (f).

disciplines of Big Data (computing, mathematics, and statistics) because in health-related Big Data research, researchers frequently do not interact with the individual subjects of their research. Big Data disciplines are assumed to be conducting research on systems, not people.[22] However, we chose to use them as reference due to their health research orientation and concern for research subjects whose personal health information is being used for research. These, we believe, should be guiding rationales in health related research in Big Data as well.

HIPAA's main goal is to assure that identified personal health information of people who seek care is properly protected under national standards, while allowing the flow of health information needed to provide and promote high quality health care and public health.[23] The HIPAA strikes a balance between the benefit in the use of information and safeguarding the privacy of individuals. Similarly to the Protection of Human Subjects Regulations, HIPAA broadly defines research as any "systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to generalizable knowledge".[24] While the HIPAA is inclusive of Big Data research in terms of the type of research, it is exclusive in terms of types of information covered, and nature of coverage.

HIPAA's scope is limited to covered entities (a health plan, a health care clearinghouse, or a health care provider who transmits any Protected Health Information (PHI) in electronic form[25]) in relation to treatment or healthcare

[22] Metcalf J, *Big Data Analytics and Revision of the Common Rule*, 59 COMMUNICATIONS OF THE ACM, 31 (2016).

[23] *Summary of the HIPAA Privacy Rule*, HHS.gov, https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html (last visited Dec. 7, 2017).

[24] 45CFR § 164.501 (2007).

[25] See 45 C.F.R. §160.103(4) (2007).

operation. Medical information held by third parties is not protected under the HIPAA.[26] Researchers are not themselves covered entities, unless they are also health care providers and engage in any of the covered electronic transactions. However, researchers who are not themselves covered entities, or who are not workforce members of covered entities, may be indirectly affected by the Privacy Rule if covered entities supply their data.[27] Also, any information or analytics derived from sources not subjected to HIPAA, although may be equally sensitive, is not legally protected under HIPPA safeguards.[28]

Health information used in research which combines multiple data bases may differ in the regulation it is subjected to, since different sources of personal information are differently regulated. Some may be subjected to more than one legislation, thus protected, while others subjected to none. However, both types of regulation specifically excludes two types of research, involving the collection of publicly available sources of data, e.g. information in social media, as well as de-identified information, i.e. information that was recorded by the investigator in such a manner that an individual subject could not be identified either directly or through identifiers linked to the subject.[29] Neither information people voluntarily share on

---

[26]For example, 23andMe, a direct-to-consumer genetic testing enterprise, holds more than a million people's genetic data, and there sample bank is constantly expanding. Social media and other commercial companies also have large Databases.

[27]*Health Services Research and the HIPAA Privacy Rule*, NIH, 10, (May 20, 2005), https://privacyruleandresearch.nih.gov/healthservicesprivacy.asp. *Cf.* HIPAA requires that any entity which is subjected to the law will obtain satisfactory contractual assurances from its business associates, stating that they will appropriately safeguard the PHI they receive or create on behalf of the covered entity according the requirements specified under the Privacy Rule. 45 C.F.R. § 160.300 (2007).

[28] Coy K & Hoffman NW, *Big Data Analytics Under HIPAA*, LEGAL INSIGHT (2016) http://www.agg.com/files/Publication/52c146ff-493e-41a9-84f9-a014e091ebfb/Presentation/PublicationAttachment/76dc5af9-fd45-4533-98ac-5beed1010278/Coy-Hoffman-Big-Data-Analytics-under-HIPAA.pdf.

[29] 45 C.F.R. § 46.101(b)(4) (2007); 45 C.F.R. § 164.514 (a) (2007).

social media, nor information that does not identify an individual is considered Protected Health Information. Traditionally, where the data is "publicly available," the research is not subject to IRB approval. Taking into consideration the amount of data voluntarily put in the public and commercial sphere, neither the purpose of use, nor the lack of explicit consent fall under the regulative data protections. The risk of identification that has been determined for one particular data set in the context of a HIPAA environment may not be appropriate for the same data set in a multiple source environment which exceeds the scope of the HIPAA.

IV. AN AUTHORIZATION TO USE INFORMATION FOR HEALTH-RELATED RESEARCH

According to the Protection of Human Subjects Regulations ("Common Rule"), in any government-funded research held in the U.S. individuals must express their voluntarily autonomous decision that their information will be used, by giving an informed consent based on adequate information and comprehension.[30] Adequate information should relate, among other things, to the purposes of the research, the expected duration of the subject's participation (in Big Data research it could address the expected duration of the use of the information), foreseeable risks, the extent, to which confidentiality of records identifying the subject will be maintained, the expected benefits to the subject or to others.[31] Obviously, the requirement for informed consent used in research done on human subjects is different from the authorization to use personal data in HIPAA.[32] The Common Rule is concerned primarily with the physical risks to humans associated with participation in a

[30] 45 C.F.R. § 46.116 (2007).

[31] *Id.*

[32] See *Beyond the HIPAA Privacy Rule*, *supra* note 10, at 164 (On informed consent vs. patient authorization); Scott, *supra* note 18, at 521.

research study, rather than with the protection of privacy of information, which is the focus of the HIPAA.[33] The HIPAA is concerned with the use of information. While no prior authorization is required from the individual before information about him is used or disclosed, any use of PHI requires an authorization (not informed consent), which is a signed permission to allow a covered entity to use or disclose the individual's information.[34] Individuals should be informed in advance how their information will be used or disclosed, therefore the authorization must contain at least one of the following elements: a description of the information to be used or disclosed, identification of persons authorized to make the requested use or disclosure; identification of the person(s) to whom the covered entity may make the disclosure; a description of each purpose of the disclosure.[35] They should have the opportunity to agree to, prohibit, or restrict the use or disclosure.[36]

According to both the Common Rule, a subject may discontinue to participate in a research at any time.[37] However, individuals have the right to revoke authorization only to the extent a covered entity has not taken action in reliance on that authorization.[38] While such revocation may affect a clinical trial and would compromise its future, in Big Data research, the option to revoke authorization for use and disclosure of PHI is therefore practical only before the data has been used. The covered entity can continue using and disclosing information obtained prior to

---

[33] *Beyond the HIPAA Privacy Rule, supra* note 10, at 186.

[34] 45 C.F.R. § 164.508 (a)(1) (2007).

[35] 45 C.F.R. § 164.508(c)(1) (2007).

[36] 45 C.F.R. § 164.510 (2007).

[37] 45 CFR 46.116(a)(8). 164.508(c)(2) (2007).

[38] 45 C.F.R. § 164.508(b)(5)(i), (ii) (2007). *Cf.* Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) art. 7(3).

the revocation for future research.[39] This is especially alarming given that the authorization does not have to state an expiration date or event, or limit the permission to use certain sensitive information.[40]

In particular when combined with information voluntarily provided by users, the authorization to use personal information for Big Data research compromises all three elements at the basis of a person's decision to allow the use of his or her information: information, comprehension and voluntariness. First, in most cases, authorization to use the data from public or commercial data bases for research is obtained when users are requested to accept a "Data Use Policy" or "Privacy Policy" of a website, an app, or a program in which they share their information, as a necessary condition to use the service. Information provided in most Data Use Policies does not comply with basic demands regarding the necessary information given to subjects before consenting that their health-related information will be used for research. With regards to research, Data Use Policy may simply read as follows: "We transfer information to vendors, service providers, and other partners who globally support our business, such as providing technical infrastructure services, analyzing how our Services are used, measuring the effectiveness of ads and services, providing customer service, facilitating payments, or conducting academic research and surveys."[41] Unlike regulation regarding research in human subjects or even the authorization to use PHI under HIPAA, regulation of other sources of Big Data research does not necessarily specify details

[39] *Withdrawal of Subjects from Research Guidance*, NIH (2010), https://www.hhs.gov/ohrp/regulations-and-policy/guidance/guidance-on-withdrawal-of-subject/index.html.

[40] Mark A. Rothstein, *HIPAA Privacy Rule 2.0: Currents in Contemporary Bioethics*, 41 J. L., MED. & ETHICS 525, 526 (2013).

[41] Data Policy, FACEBOOK, https://www.facebook.com/about/privacy/.

about what sort of information should be given to subjects of data used for research. They are not study specific, although information regarding the purpose of the research may play a role when authorizing use of information, and should be part of the information provided to subject authorizing the use of their information, and often do not succeed explaining the risks of data sharing to the public.[42] It has been argued that privacy policies should provide subjects providing them with detailed information about what, why and how personal data will be collected, processed, stored, used and in cases, disclosed.[43] Moreover, in these agreements, most social media companies retain the right to revise their Data Use Policies at any time without providing notice to the user regarding the specific change.[44] In such cases the user may miss the opportunity to opt out or refuse to be included.[45] These policies are difficult to understand and most people do not even read their terms.[46] Usually, they serves more as liability disclaimers than as assurances of consumer privacy, yet, by providing these notices, data owners comply, at least formally, with

[42] Certain required information, such as the sort of findings or expected implications, cannot always be foreseen in Big Data research, as statistical algorithms might find unexpected correlations and generate unexpected implications and risks. See, *e.g.*, Joshua Berlinger & Maegan Vazquez, *US Military Reviewing Security Practices After Fitness App Reveals Sensitive Info*, CNN DIGITAL EXPANSION SHOOT, (Jan. 29, 2018), https://www.cnn.com/2018/01/28/politics/strava-military-bases-location/index.html (recent data about running routes, collected from the users of Strava, which bills itself as "the social network for athletes", ended up revealing location patterns of security forces working out at military bases in remote locations.)

[43] Shara Monteleone, *Addressing the Failure of Informed Consent in Online Data Protection: Learning the Lessons from Behavior-Aware Regulation*, 43 SYRACUSE J. INT'L L. & COM. 69, 79 (2015).

[44] Michael L. Rustad & Maria Vittoria Onufrio, *Reconceptualizing Consumer Terms of Use for a Globalized Knowledge Economy*, 14 U. PA. J. BUS. L. 1085, 1086-87 (2012).

[45] Lauren B. Solberg, *Complying with Facebook's Terms of Use in Academic Research: A Contractual and Ethical Perspective on Data Mining and Informed Consent*, 82 UMKC L. REV. 787, 792 (2013).

[46] Monteleone, *supra* note 43, at 75.

their information obligations.[47] When the subject of the information used never learns about the research, presumed authorization can be considered a deceptive practice.[48]

With regards to comprehension, most users are not aware of the multiplicity algorithms that are gathered, stored analyzed from their data by multiple agents.[49] Many data brokers combine information about habits, behaviors, or attributes that is derived from Big Data sets based on individual's online and offline accounts and devices, and create predictive profiles and modeling for multiple purposes, including commercial.[50] Such profiles may be valuable to medical purposes, they also influence employment, insurance and may serve as grounds for discriminatory behaviors.[51] Although these practices compromise the right to privacy of individuals and bear irreversible consequences related to important aspects of their lives, current legal framework does not address predictive models using data outside of HIPAA.

---

[47] *Id.* at 80.

[48] U.S. DEP'T OF HOMELAND SECURITY, THE MENLO REPORT: ETHICAL PRINCIPLES GUIDING INFORMATION AND COMMUNICATION TECHNOLOGY RESEARCH, 11, (2012), http://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803.pdf.

[49] Glenn & Monteith, *supra* note 9 at 494(3); Monteleone, *supra* note 43, at 88.

[50] Glenn & Monteith, *supra* note 9 at 494(4-5); US Senate Committee on Commerce, Science, and Transportation, A review of the data broker industry: collection, use, and sale of consumer data for marketing purposes, (2013), http://www.commerce.senate.gov/public/?a=Files.Serve&File_id=0d2b3642-6221-4888-a631-08f2f255b577; Satish Garla et al., *What Do Your Consumer Habits Say About Your Health? Using Third-Party Data to Predict Individual Health Risk and Costs*. SAS INSTITUTE 6, (2013), http://support.sas.com/resources/papers/proceedings13/170-2013.pdf.

[51] Ann Reilly Dowd, *Protect Your Privacy: A Money Investigation Reveals the Five Biggest Threats to Your Privacy and How You Can Safeguard Yourself Against the Most Serious Types of Snooping*, MONEY 104, 107 (1997). (In studying the privacy practices of three hundred Fortune 500 companies, David Linowes, former chair of the President's Commission on Privatization and the U.S. Privacy Protection Commission finds that "35% of employers said they use personal medical information as a basis for hiring, promotion, and firing decisions".) Sharona Hoffman, *Employing E-Health: The Impact of Electronic Health Records on the Workplace*, (Case Legal Studies Research Paper No. 2010-1, 2010), https://ssrn.com/abstract=1531265.

Moreover, private confidential data is traded all over the world. Some corporation's sole purpose is to collect such data in order to capitalize on it.[52] A company can sell de-identified individual-level data if consumers sign the research consent document – which 80 percent of consumers do.[53] Turning profit from data for research purpose may play a role and influence the willingness of the subject of the data to agree to its use, thus such information should be provided to data subjects. It is not far-fetched to assume that some users would not give authorization to use their information for research commercial targeted advertising, purposes, even if they have agreed that it will be used for a health-related issue.

With regards to voluntariness, while the Privacy Rule generally prohibits covered entities from conditioning treatment, payment, enrollment, or eligibility on an individual's provision of an authorization,[54] in social media as well as commercial online activities the user gives away the right for the data when using the app. Fitbit and Facebook are only a few examples. Most consumers are willing to pay for online services with personal information rather than with money.[55] In these contexts, consumers do not have the right to control what personal information is collected, maintained, used, and shared by data brokers. It can be argued that the use of

---

[52] Hal Hodson, *Revealed: Google AI Has Access to Huge Haul Of NHS Patient Data*, NEW SCIENTIST. (29 Apr., 2016), https://www.newscientist.com/article/2086454-revealed-google-ai-has-access-to-huge-haul-of-nhs-patient-data/.

[53] Kayte Spector-Bagdady, *Why You Should Worry About the Privatization of Genetic Data*, (Sep. 8, 2016); 8.17pm AEST, https://theconversation.com/why-you-should-worry-about-the-privatization-of-genetic-data-62591 (last visitedMarch 2, 2017).

[54] 45 C.F.R. § 164.508(b)(4) (2007).

[55] Christine Bauer, et al., *On The Value Of Information–What Facebook Users Are Willing To Pay*, (20th European Conference on Information Systems proceedings (ECIS 2012)); Aleecia M. McDonald et al., *Americans' Attitudes About Internet Behavioral Advertising Practices* 63 (Proceedings of the 9th annual ACM workshop on Privacy in the electronic society, 2010).

conditioned authorization does not meet the requirement of voluntary authorization from subjects in health-related research.

V. POSSIBLE EXEMPTIONS FROM THE AUTHORIZATION REQUIREMENT

In some cases the HIPAA's Privacy Rule enables the covered entities to obtain a complete waiver of the authorization requirement from an Institutional Review Board (IRB), following three criteria[56]: 1) The use or disclosure of PHI involves no more than a minimal risk to the privacy of individuals, 2) the research could not practicably be conducted without the waiver, or 3) without access to and use of the PHI.[57] This is another expression of the ethical requirement to justify any research by demanding that the anticipated benefits and importance of the research will overweigh the potential harms, using the principle of proportionality. Meaning, exceptions to the demand for individual expression of authorization are given only when it is practically impossible to get individual authorization, and only when the risks to the research subject are minimized and balanced with the research benefits. Part of what the researcher will have to argue in front of the IRB is that the research has no harmful impact.

Often, researches use the method of de-identification to show that the research poses minimal harms on individuals whose information is being used. Data de-identification methods refer to the process of removing or obscuring any personally identifiable information from individuals' records in order to minimize the risk of unintended disclosure of identity and information. Under the Privacy Rule, creating de-identified data or a limited data set is a health care operation of the

[56] 45 C.F.R. § 164.512 (i2ii) (2007).

[57] 45 C.F.R. § 164.512 (i)(1)(i)-(ii) (2007). *Cf.* 45 C.F.R. § 46.116(d) (2007).

covered entity and, thus, does not require an individual's authorization, even if the limited data set or de-identified data will function as a database for research.[58] HIPAA provides two methods for de-identification.[59]

The Safe Harbor method specifies eighteen direct identifiers that when removed, HIPAA's Privacy Rule's restrictions do not apply.[60] The de-identifying code can be assigned to the de-identified information for purposes of re-identification, but cannot be derived from or related to identifiers of the individual (e.g., Social Security number).[61] Specific steps and methods used to de-identify information have been considered appropriate to protect the confidentiality of the individuals.[62] However, data de-identified using the Safe Harbor method may lack critical information needed and become unusable for research purposes, or at least of lesser value.[63] The technical means legally required to protect individual information lack proof or guarantees of privacy. There is an increasing area of work indicating that de-

[58] *Health Services Research and the HIPAA Privacy Rule*, NIH, 10, (May 20, 2005), https://privacyruleandresearch.nih.gov/healthservicesprivacy.asp.

[59] For possible mitigation methods such as suppression techniques (withholding information in selected records from release), generalization of the information, perturbation etc. See *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, HHS.gov, https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#_ednref5 (last visited Dec. 8, 2017).

[60] Art. 164.514(e). (the identifiers are: name, address (all geographic subdivisions smaller than state, including street address, city county, and zip code), all elements (except years) of dates related to an individual (including birthdate, admission date, discharge date, date of death, and exact age if over 89), telephone numbers, fax number, email address, social security number, medical record number, health plan beneficiary number, account number, certificate or license number, any vehicle or other device serial number, web URL, internet protocol address, finger or voice print, photographic image, any other characteristic that could uniquely identify the individual.)

[61] *Beyond the HIPAA Privacy Rule , supra* note 10, at 173.

[62] Scott Burris et al., *The Role of State Law in Protecting Human Subjects of Public Health Research and Practice*, 31 J. L. MED. & ETHICS, 654, 656 tbl 2 (2003) (discussing U.S. states (48), which allow release of de-identified data to outside researchers.)

[63] *Beyond the HIPAA Privacy Rule , supra* note 10, at 175.

identification is no-longer sufficient to protect privacy of subjects, and that data can be re-identified and name the data subject.[64] When using data from multiple sources, there is a need to create links between different databases, protected and un-protected. As the dimensions of data increase, and the number of different variables on each individual increases, it is more complicated to protect the identity of participants. Studies indicate that even after the removal of the identifiers, public data may contain personal identifying information that can lead to re-identification.[65] Privacy policies apply to data collected from users information through registration forms or cookies, and not to the content the users post.[66] When this information is combined with de-identified data subjects can be re-identified. "The more information available about a person, the easier it is to re-identify the person in the future."[67] Once the information can be connected directly to data owners, even when using encryption to protect security, the Safe Harbor Method can no longer serve as

---

[64] Gregory J. Matthews & Ofer Harel, *Data Confidentiality: A Review of Methods for Statistical Disclosure Limitation and Methods for Assessing Privacy*, 5 STAT. SURVEYS, 1 (2011); Khaled El Emam et al., *A Systematic Review of Re-Identification Attacks on Health Data*, 6 PLOS ONE, p.e28071 (2011); Latanya Sweeney et al., *Identifying Participants in the Personal Genome Project by Name* (2013), https://arxiv.org/ftp/arxiv/papers/1304/1304.7605.pdf; Health IT Policy Committee, *Health Big Data Recommendations* 1, 14 (HITPC Privacy and Security Workgroup, Aug. 16, 2015) *available at* https://www.healthit.gov/facas/sites/faca/files/HITPC_Draft_PSWG_Big_Data_Transmittal_2015-08-11.pdf; Federal Committee on Statistical Methodology. 2005. *Statistical Policy - Report on Statistical Disclosure Limitation Methodology*, 10 , *available at* https://www.hhs.gov/sites/default/files/spwp22.pdf.

[65] *Beyond the HIPAA Privacy Rule , supra* note 10, at 175. E.g., Dov Greenbaum et al., *Genomics and Privacy: Implications of the New Reality of Closed Data for the Field*, PLoS Comput Biol. (2011) p.e1002278, 7; Bradley Malin & Latanya Sweeney, *How (not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-Identification to Evaluate and Design Anonymity Protection Systems*, 37 J. BIOMED INFORMATICS 179 (2004); Sarah Zhang, *Scientists Are Just as Confused about the Ethics of Big-Data Research as You*, WIRED, (May 20, 2016), https://www.wired.com/2016/05/scientists-just-confused-ethics-big-data-research/.

[66] Glenn & Monteith, *supra* note 9 at 494(5).

[67] *Id.* at 494(4).

a magic word to waive the need for authorization to use private information, or to automatically consider the harms minimal.

Scholars approach three aspect of the harms: security of sensitive information,[68] the consequences and ramifications involved with the violation of privacy,[69] or the decision-making process of subject whose information could be used,[70] but so far the challenges is greater than the success. In a complex research environment, regulation does not extend far enough to resolve the ethical issues presented by advanced research technology using data retrieved from multiple sources. Eighty percent of security executives in North America believe conventional network security solutions are insufficient to protect their companies' cloud storage environments, where most of the information is located.[71] It seems that neither researchers nor users are happy with the current practice. On the one hand, formal trials to adapt to the regulation does not necessarily bring an improvement in terms of better awareness and compliance of the users.[72] It has been argued that researchers have reported that after the implementation of the Privacy Rule, the time

[68] Murdoch & Detsky, *supra* note 12, at 1352 (suggesting to extend security similarly to what is required to protect confidential financial data in other sectors).

[69] Sharona Hoffman, *Big Data and the Americans with Disabilities Act* (2017). 68 Hastings Law Journal 777 (2017); Case Legal Studies Research Paper No. 2016-33. SSRN: https://ssrn.com/abstract=2841431 (suggesting to use existing regulative mechanism, such as the one in the Americans with Disabilities Act of 1990, to broaden it in order to deal with new implications, subjecting state and local governments, employment agencies and labor unions to non-discriminatory practices against qualified individuals with disabilities).

[70] Monteleone, *supra* note 43, at 79 (suggesting complementary measures for users' better decision-making as regards data protection).

[71] James Bourne, *Four in Five Execs Think Conventional Security is Not Enough for Cloud Environments*, CLOUDTECH, (1 July 2015)., http://www.cloudcomputing-news.net/news/2015/jul/01/four-five-execs-think-conventional-security-not-enough-cloud-environments/ (last visited March 2, 2017).

[72] *See*, Annie Anton et al. *HIPAA's Effect on Web Site Privacy Policies*, 5 IEEE SEC. & PRIVACY 45 (2007) (A comparison of privacy policies for nine healthcare websites before and after HIPAA legislation found that after the legislation, the policies were more descriptive but longer and more difficult to comprehend).

and cost per recruited participant became about 30% higher than before,[73] or that is will cost the industry billions to come into compliance with the regulations.[74] Still, this regulation cannot prevent privacy violations. On the other hand, critiques claim that individuals are given limited control and expression of autonomy over the use of their information due to the extended scope of authorizations, which are often conditioned in commercial services, open ended, and based on information that is insufficiently detailed or clear.

The second way to de-identify PHI is to have a qualified statistician determine, using generally accepted statistical and scientific principles and methods, that the risk is very small that the information could be used to identify the subject of the information.[75] The notion of expert certification is not unique to the health care field and have been used to mitigate risk prior to sharing data. There are several experts (usually from computer sciences) who fill similar functions.

In our view, expert certification is an aspect of accountability. Accountability is the acknowledgment and assumption of responsibility for actions, decisions, and policies made. Recent regulations shift the focus from the individual whose rights are violated toward collective responsibility and solidarity, represented by the principle of accountability.[76] With great potential of benefit comes great responsibilities, therefore, the shift of focus lays even more duties on researchers to

---

[73] Roberta B. Ness & Joint Policy Committee, *Influence of the HIPAA Privacy Rule on Health Research*, 298 JAMA 2164, 2167 (2007).

[74] Scott, *supra* note 18, at 513.

[75] 45 C.F.R. § 164.514(b) (2013)..

[76] Solon Barocas & Helen Nissenbaum, *Big Data's End Run Around Anonymity and Consent* 44(2014); Boyd & Crawford, *supra* note 5, at 672; Joanna Stjernschantz Forsberg et al., *Changing Perspectives in Biobank Research: from Individual Rights to Concerns About Public Health Regarding the Return of Results*, 17 EU. J. HUM. GENETICS 1544 (2009); Andrej Zwitter. *Big Data Ethics*, 1 BIG DATA & SOC'Y 1,3 (2014).

justify the importance of the research they are responsible for. While the "classical rights model" recognizes individual rights but assumed state enforcement for their violation, this flexible model is a more optimistic model, which trusts the capacity and willingness of private organizations to effectively police themselves emphasizes accountability and self-regulation in order to comply with legal, but also social duties they have toward others.

The Privacy Rule does not prescribes specific safeguards, but rather, requires that the covered entities will design their own policies and procedures.[77] It requires that covered entities would implement safeguards to protect identified health information, and establishes the conditions under which they can use or disclose PHI.[78] Researchers' accountability contains private implications - on the individual whose information it is, and social implication on society in which such information is available. Accountable researchers must have responsibility over the use of information in a specific research, and over the ramification of their research. Such determination involves a person with appropriate knowledge and experience with generally accepted statistical, scientific principles and methods. A qualified expert may apply generally accepted statistical or scientific principles to compute the likelihood that a record in a dataset is expected to be unique, or linkable to only one person, within the population to which it is being compared. If an expert determines that the risk of identification is greater than "very small", the expert may modify the information to mitigate the identification risk to that level, as required by the de-identification standard. The expert must reduce the risk that the data sets could be combined with prior versions of the de-identified dataset or with other publically

[77] 45 C.F.R. 45 § 164.530 (a)(1)(i) (2013).

[78] 45 C.F.R. § 164.530 (c)(1) (2013).

available datasets to identify an individual. In general, the expert will adjust certain features or values in the data to ensure that unique, identifiable elements no longer, or are not expected to, exist. However, there is neither a specific professional degree or certification program for designating who is an expert at rendering health information de-identified particular, nor a method for assessing the level of risk and assure that it meets the "very small" level indicated by the method. This paper explores the statistical techniques of synthetic data to minimize the risks entailed in releasing individuals' data for research purposes.

VI. SYNTHETIC DATA AS A MEANS TO FULFILL ETHICAL REQUIREMENTS

The idea of synthetic (or simulated) data is introduced by Rubin.[79] It is a method of statistical disclosure limitation based on the missing data technique of Multiple Imputation (MI).[80] Synthetic data views sensitive data as missing values in the original data (hereinafter "the observed data"). In any synthetic data set, each data row represents a real individual. However, sensitive attributes would be fully or partially replaced by random draws from an appropriate posterior predictive distribution using MI techniques. In fully synthetic data, all the information is synthesized, and in partially synthetic data only some of the information is synthesized.

Ignoring the size of the data for a moment, let us consider micro-data as a dataset containing information about individuals. Consider a dataset in which each row represents a subject, and each column represents a variable. The method can

---

[79] D.B. Rubin, *Statistical Disclosure Limitation*, 9 J. OFFICIAL STAT. 461 (1993).

[80] D.B. RUBIN, MULTIPLE IMPUTATION FOR NONRESPONSE IN SURVEYS (1987); Ofer Harel & Xiao-Hua Zhou, *Multiple Imputation: Review of Theory, Implementation and Software*, 26 STAT. IN MED. 3057 (2007); D.B. Rubin, *Multiple Imputation After 18+ Years*, 91 J. AM. STAT. ASSOC. 473 (1996).

measure different variables (age, gender, disease status, blood pressure (BP)) or measure the same variable over time (BP pre-treatment, BP during treatment, BP after treatment, unprotected sex, needle sharing). Essentially, our complete data will look like a rectangle (we assume no missing data, however extensions to incomplete data are available). The observed micro-data (D) can be considered as a random sample from a specific population (P). One can use MI to create several populations ($m > 1$, e.g., $m = 10$) based on the observed data.[81] Next, from each imputed population we can draw a random sample of pre-specified size (n), which results in m sets of complete synthetic data.[82]

While Rubin introduced fully synthetic data sets where all data is replaced with imputed data and none of the released data is real, Little proposed partially synthetic data sets.[83] In partially synthetic data the observed micro-data is still considered a random sample from a population (P). Each partially synthetic dataset consists of the non-sensitive data, which is the same across all m synthetic datasets, and the imputed values of the sensitive data. The use of MI here is similar to the one mentioned above, except that unlike fully synthetic data where all variables are treated as potentially sensitive, in partially synthetic data the only variables synthesized are those pre-specified. For example, if a dataset contains information about gender, race, marital status and income, only the variable income may be considered as sensitive, and only this variable is synthesized ($m > 1$) times. The m

[81] *Id.*

[82] Trivellore E. Raghunathan et al., *Multiple Imputation for Statistical Disclosure Limitation*, 19 J. OFFICIAL STAT. 1 (2003); Rubin, *supra* note 79.

[83] Roderick J.A. Little, *Statistical Analysis of Masked Data*, 9 J. OFFICIAL STAT. 407 (1993).

sets are then combined using the combining rules specifically appropriate to partially synthetic datasets.[84]

In order to synthesize data, either fully or partially, researchers must rely on an observed dataset, using MI.[85] MI was developed originally to deal with incomplete data. When using MI for the purpose of synthesis, we assume that either the whole data (for fully synthetic data), or part of the data (for partially synthetic data) is considered missing for the user. Conventional MI includes three steps: imputation, analysis and combining results. After the imputation stage, the observed data ceased being used and we start working in the synthetic dataset. In the imputation stage, a statistical model is fitted to the data, and based on that model the imputed data is generated as random draws from the data's predictive distribution, resulting in m complete data sets. In the analysis stage, the researcher runs a pre-specified analysis on all "complete" data and saves the estimates and variances of interest. Finally, simple arithmetic rules permit combining the m sets of results into an aggregated final solution. The combining rules will be different depending on the type of problem at hand.[86]

The choices of imputation models are of extreme importance and have implications on the biases involved. We would like the relationships between the variables in the synthetic data to resemble the relationships between variables in the observed data so that analyses will results in similar conclusions to analyses run on the observed data. For example, consider continuous data that can be represented as

---

[84] Jerom P. Reiter, *Satisfying Disclosure Restrictions with Synthetic Data Sets*, 18 J. OFFICIAL STAT. 531 (2002).

[85] Rubin, *supra* note 80 (on Multiple Imputation).

[86] Jerom P. Reiter& Trivellore E. Raghunathan, *The Multiple Adaptations of Multiple Imputation*, 102 J. AM. STAT. ASS'N. 1462 (2007).

a joint normal distribution, where the correlation matrix contains the information about the relationships between the variables. If the correlation matrix of the observed data indicates high correlation between two variables, we expect the correlation matrix from the synthetic data to represent similar high correlation as well. But if the models are wrong, and the imputation model neglects to control some variables of interest, the correlations of these variables with others will diminish to zero, which means results are not accurate.

As we move from conventional size micro-data to Big Data, the thought process stays the same, but the complication due to the size of the data needs to be taken into consideration. When considering Big Data, it is not always clear if fully synthetic data is plausible to construct. The main advantage of partially synthetic data over fully synthetic data is the ability to deal with a larger amount of data and more simplified statistical models for the imputation models. Mathematically, it is simpler to build a model for the partially synthetic data compared with fully synthetic data due to the size of the models and the datasets in place. The larger the model is, the more parameters there are to estimate, and the complexity increases dramatically.

VII. A NEW RISK-BENEFIT BALANCE

Using synthetic data enables researchers to balance potential benefits from data analytics for health-related research with methods of protecting the privacy of data subjects. The procedure of synthetic data is based on the premise that releasing the data will maintain the privacy of the individuals in the sample. Synthetic data can rely on databases from one source or several sources, with different scopes of privacy protections, because eventually in this approach, although the analyses on

these datasets will represent the population it was sampled from, all variables in the dataset are synthesized. Meaning, on the one hand individuals' information is not being disclosed, and on the other, the scientific and social benefits are maximized. The confidentiality level is slightly different in each method (fully or partially synthesized data).[87] Utility is maximized when all synthetic data is identical to the original data (no noise was added), as there is no actual variability between observed and synthetic data. These datasets are highly useful for analysts, but pose great risks because they may contain individuals' sensitive information. As the variability between the synthetic dataset and the original dataset increases, the confidentiality increases and the utility decreases. When using fully synthetized data, the data released represents data of individual in the observed database, but effectively it is simulated data, not the actual data of that individual. The data analyzed is only a hypothesized representation of individuals, therefore the privacy of actual individuals is preserved. Since all the data released is synthetic, fully synthetic data can therefore be considered more secure. Partially synthetized data still contains some of the information of de-identified individuals, thus the risk in disclosure of information is higher in comparison to fully synthetic data that do not contain any individuals' real data. However, all the sensitive information for these individuals is synthesized, therefore even in partially synthetic data, the ability to connect these data to an individual is reduced dramatically in comparison with the risks in using the observed data. For example, age, gender and race are not synthesized but HIV

[87] For different procedures to deal with privacy when analyzing micro data, see e.g., Chris Clifton & Tamir Tassa, *On Syntactic Anonymity and Differential Privacy*, In DATA ENGINEERING WORKSHOPS (ICDEW), 88 (2013 IEEE 29th International Conference on IEEE); Cynthia Dwork & Jing Lei, *Differential Privacy and Robust Statistics*, IN PROCEEDINGS OF THE FORTY-FIRST ANNUAL ACM SYMPOSIUM ON THEORY OF COMPUTING, 371 (2009); Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*, 9 FOUNDATIONS AND TRENDS® IN THEORETICAL COMPUTER SCIENCE, 211 (2014).

status, and related biomarkers are synthesized. It does not seem like the use of personal information to create a synthetic data can differently satisfy the "impracticably to achieve individual consent" required for a waiver than any other use in Big Data research. The requirements of informed consent remain the same, as the process is done after the information has been shared. However, individuals will most likely be more willing to allow the use of their information for a synthesized database, relatively to non-synthesized one, because the risks are lower.

While this paper focuses on potential harms that are entangled in health-related research in Big Data research, the equilibrium between risks and benefits from the research begs wide accessibility to Big Data sets that may be useful for future research. There is extensive motivation of the government and industrial officials to gather and extract maximal value from existing data for public good, and facilitate access to available resources that can be utilized at no risk to individuals. Researchers receiving federal funding often have to submit their data to a public data bank, such as the National Institutes of Health's database of Genotypes and Phenotypes (dbGaP), which offers other researchers access to data for future work for little to no fee.[88] Other public and private funds go into the collection and maintenance of databases, which become less accessible, reducing potential benefit. Given the value large scale information holds, Big Data, broadly defined, is producing increased institutional powers for those who own them. At the moment, large data companies create monopolies. Commercial companies do not necessarily have an obligation to make their data available to others, as they do not follow the same policy regarding their information. They have complete control over what they share, whom they share it with, and for what purpose. Private companies may restrict

[88] Spector-Bagdady, *supra* note 53.

access to their data entirely or offer small data sets only to well-resourced bodies.[89] Limited access creates a digital barrier between those who have tools and access to data and those who do not.[90] Moreover, when those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access, the validity of the research is affected. It may also affect the areas and agenda of future research: those privileged enough to get access to proprietary data sets will not risk their access being cut and chose topics that are appealing to the database owners. Less accessibility therefore compromises potential benefit from research that is not necessarily favorable to the interests of database owners and minimizes, rather than maximizes, the benefit in research in Big Data. Synthetic datasets may improve this situation. Owner of the data can choose whether to give access to data whether it is synthesized or not. However, they are more likely to be willing to share synthesized data, because the risks involved with sharing such data are lower.

Moreover, when using synthetic data, researchers' access to data from which they can conduct a variety of secondary analyses, increases dramatically. In order to avoid privacy infringement, commonly the owners (data collection agencies) release some summary statistics of the data (such as tables of counts, means, variances, correlations, etc.). This is inefficient to other researchers since they cannot conduct most secondary analyses using only summary statistics. In particular, if only summary statistics are reported, results such as regression coefficients, odds ratios

[89] Boyd & Crawford, *supra* note 5. *See, e.g.*, Glenn & Monteith, *supra* note 9 at 498, Spector-Bagdady, *supra* note 53; James Rogers, *Hacker Looks To Sell 655,000 Alleged Patient Healthcare Records On The Dark Web*, FOXNEWS.COM, (June 27, 2016), http://www.foxnews.com/tech/2016/06/27/hacker-looks-to-sell-655000-alleged-patient-healthcare-records-on-dark-web.html (23andMe recently announced that the drug company Genetech offered to pay up to $60 million to use its database to conduct Parkinson's research. Prices range from about $96,000 to $490,000 for each database.)

[90] Boyd & Crawford, *supra* note 5, at 673.

and other frequently used parameters are not possible to obtain and may minimize potential benefits from existing databases, rather than maximizing them. The main advantage in synthetic data is allowing the use of extensive information, unrelated to actual individuals, across a wide range of research settings and substantive areas. Each synthetic dataset can be analyzed based on the common complete data statistical analysis procedure (i.e. regression) required by the secondary researchers. For example, secondary researchers who obtain the m sets of synthetic data might wish to analyze the role of some of the variables in a new regression. The results from each dataset are saved and combined for a final result taking into account the fact that the data were synthesized.[91] If the methodology used for imputation is general enough, (almost) any statistical procedure will be applicable for secondary users and there are no limitations to the specific secondary statistical analysis researchers can use.

VIII. CONCLUSION

Health-related research in Big Data holds not only a promise, but also a peril as a result of privacy violations, particularly when data is used from multiple resources. Neither regulatory nor security solutions adequately administrate risks involved in health-related research that relies on traditional health-related data as well as non-traditional sources, in particular when research combines health related and public data voluntarily provided by users.

This paper explores the statistical techniques of synthetic data to minimize the risks entailed in releasing individuals' data for research purposes. The method of synthetic data is not an exclusive method to conduct research, but demonstrates the

[91] For combining rules see, Raghunathan, Reiter &Rubin, *supra* note 80. For fully synthetic data see, Rubin, *supra* note 79.

importance in using multi-disciplinary scientific methods in the design and analysis of health-related research in Big Data. Statistical models and methodologies of analysis still leave many open questions and may require adaptation according to the social, medical, financial context of each source, similar to the diverse methodological and analytical approaches used in social science. If not adapted, we risk misunderstanding the results and as a result misallocate important public resources. This is true for any statistical analyses regardless to the database in question. While this model is imperfect, when dealing with multiple sources in health-related research, it addresses many of the challenges posed by the Safe Harbor method of de-identification in a better way, increasing privacy and accessibility to usable useful data.

The multi-disciplinary method may serve as a basis for coherent regulation, complying with the principle of researcher's accountability infused with technology. Where known harms can be minimized and benefits increased, it may be unjustifiable to conduct research without using such methods. We join the call to develop social media policies pursuant to the HIPAA Privacy Rule,[92] as well as an ethical Big Data policy, which will encompass information rules that manage the appropriate flows of information in ethical ways.[93]

---

[92] Kimbra Ratliff, *HIPAA Violations on Social Media: Will HHS Continue to Ignore*, 45 U. MEM. L. REV. 633, 638 (2014).

[93] Richards, Neil M and King, Jonathan H, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393, 395 (2014).