Journal of Law and the Biosciences, 1–39
doi:10.1093/jlb/lsw061
Original Article

Clinical genomics, big data, and electronic medical records: reconciling patient rights with research when privacy and science collide

Jennifer Kulynych[1,*,†] and Henry T. Greely[2,‡]

1. Legal Department, The Johns Hopkins Hospital and Health System, 1812 Ashland Ave., Suite 300, Baltimore, MD 21205, USA
2. Law School, Stanford University, 559 Nathan Abbott Way, Stanford, CA 94305-8610, USA
*Corresponding author. E-mail: jkulyny1@jhmi.edu

ABSTRACT

Widespread use of medical records for research, *without* consent, attracts little scrutiny compared to biospecimen research, where concerns about genomic privacy prompted recent federal proposals to mandate consent. This paper explores an important consequence of the proliferation of electronic health records (EHRs) in this permissive atmosphere: with the advent of clinical gene sequencing, EHR-based secondary research poses genetic privacy risks akin to those of biospecimen research, yet regulators still permit researchers to call gene sequence data 'de-identified', removing such data from the protection of the federal Privacy Rule and federal human subjects regulations. Medical centers and other providers seeking to offer genomic 'personalized medicine' now confront the problem of governing the secondary use of clinical genomic data as privacy risks escalate. We argue that regulators should no longer permit HIPAA-covered entities to treat dense genomic data as de-identified health information. Even with this step, the Privacy Rule would still permit disclosure of clinical genomic data for research, without consent, under a data use agreement, so we also urge that providers give patients specific notice before disclosing clinical genomic data for research, permitting (where possible) some degree of choice and

† Senior Counsel, The Johns Hopkins Hospital and Health System Corporation.
‡ Professor of Law, Professor, by courtesy, of Genetics; Director, Center for Law and the Biosciences, Stanford University.

control. To aid providers who offer clinical gene sequencing, we suggest both general approaches and specific actions to reconcile patients' rights and interests with genomic research.

KEYWORDS: genomics, big data, research, privacy, HIPAA, de-identification, electronic medical records

INTRODUCTION

With the broad adoption of electronic medical record (EMR) systems, researchers can mine vast amounts of patient data, searching for the best predictors of health outcomes. Many of these predictors may lie in the genome, the encoded representation of each person's DNA. As gene sequencing continues to evolve from a complex, expensive research tool to a routine, affordable screening test, most of us are likely to have our DNA fully digitized, vastly expanding the already large store of electronic health data already preserved in or linked to our EMRs. In parallel, genomic researchers will, increasingly, seek out EMRs as an inexpensive source of population-wide genome, health, and phenotype data, thus turning patients into the subjects of genomic research. This will often occur without the patients' knowledge, let alone their consent, in a research climate where the privacy risks are routinely discounted and data security can be uncertain. The implications, both for research and for privacy, are profound, but the prospect has received little attention in the literature.[1]

The widespread re-use of health information in EMRs is already commonplace, but those records typically don't include detailed genomic information.[2] The landscape is changing, however, as technical advances make sequencing and storing patient genomes increasingly affordable, and as providers and academic medical institutions—along with government, science, and industry—envision using genomic data to enable 'precision medicine'.[3] As more patients have genomic data linked to their medical records, absent a change in policy or practice we will see the same non-consensual re-use of these data already allowed for other forms of health information.

Advocates of the status quo argue either that there is little real re-identification risk for genomic data (the 'privacy through obscurity' theory) or in the alternative, that if the risk is real, the consequences are minor, because relative to other forms of health data, information about genetic variation is less stigmatizing, less valuable, and, therefore, less attractive to hackers and criminals.[4] The net effect of these rationales is a privacy standard for DNA sequences much lower than what currently applies to data elements

1 See eg Ribhi Hazin et al., *Ethical, Legal, and Social Implications of Incorporating Genomic Information Into Electronic Health Records*, 15 GENET. MED. 810–816, 810–816 (2013).(Noting the need for a framework to govern secondary use of genomic information in the electronic health record.)

2 See eg Junji Lin et al., *Application of Electronic Medical Record Data for Health Outcomes Research: A Review of Recent Literature*, 13 EXPERT REV. PHARMACOECON. & OUTCOMES RES. 191–200 (2013) (describing new trends in research use of EMRs).

3 See Precision Medicine Initiative — National Institutes of Health (NIH), US NATIONAL LIBRARY OF MEDICINE, https://www.nih.gov/precision-medicine-initiative-cohort-program (accessed Apr. 19, 2016).

4 See eg Michelle Meyer, FORBES, http://www.forbes.com/sites/michellemeyer/2015/12/31/no-donating-your-leftover-tissue-to-research-is-not-like-letting-someone-rifle-through-your-phone/#5c1806da19df (accessed Apr. 19, 2016). See also Yaniv Erlich & Arvind Narayanan, *Routes for Breaching and Protecting Genetic Privacy*, 15 NAT. REV. GENET. 409–421, 409–421 (2014) (describing theory of privacy by obscurity as applied to genomic data).

such as URLs, fingerprints, and zip codes—each enumerated as an identifier under the Privacy Rule and protected when linked to health information.

Moreover, even assuming *arguendo* that genome sequence data don't constitute particularly sensitive health information, it is becoming difficult to maintain that a gene sequence (or substantial subset thereof) is not an 'identifier' that places any associated health or demographic information at risk, when databases of identifiable sequence data are proliferating and researchers are exploring ways to sequence DNA rapidly for use as a biometric identifier.[5]

And, finally, at the heart of this issue lies an important ethical, and practical, question: Should the scientific and provider communities continue to disregard the accumulating evidence from repeated studies that patients expect to be told about, and to control, research uses of their genomic and health information?[6]

The prospect of eventual, widespread EMR-based genomic research under current privacy practices drove us to write this paper. The paper proceeds in five parts: setting out the problem, reviewing the current status of records-based biomedical research, noting other secondary uses of medical records, describing the conflict between individual rights and societal interests implicated in genomics-based research, and providing our recommendations for a balanced approach.

We acknowledge the vigorous debate over almost every aspect of the problem of genomic privacy: whether genomic data are identifiable, whether it is likely that anyone would try to re-identify a subject of genomic research, whether patients have an obligation to participate in such research regardless of personal preference. Our paper builds on the 2008 recommendations of the Personalized Health Care Work Group of the US Department of Health and Human Services ('DHHS') American Health Information Community, which advocated special protections for the research use of genomic data in EMRs, arguing that such data are exceptional relative to other sensitive information due to their uniqueness and potential for re-identification.[7] Without engaging the debate over 'genetic exceptionalism', we maintain that it is still useful here to draw a line—even if it is in sand—and to insist that if patients have any genuine right to understand and influence the uses of any of their sensitive medical information, such a right must include their genomes. That all bright lines are imperfect does not mean no lines are useful.

Although we do not call for legal or regulatory changes, we question whether current federal health privacy law, properly interpreted, actually permits health care providers, whether clinicians or academics, to treat whole genome sequence data as 'de-identified' information subject to no ethical oversight or security precautions, especially when genomes are combined with health histories and demographic data. We recognize that pending amendments to the federal Common Rule might affect and even further strengthen our argument, especially if, as proposed, IRBs would no longer oversee

5 See DNA Biometrics, http://www.nist.gov/mml/bmd/genetics/dna_biometrics.cfm (accessed Apr. 19, 2016) (describing work of the NIST Human Identity Project team to evaluate DNA biometric identifiers).

6 Mildred K. Cho et al., *Attitudes Toward Risk and Informed Consent for Research on Medical Practices*, 162 Ann. Intern. Med. 690, 690 (2015). See also, Holly K. Tabor et al., *Genomics Really Gets Personal: How Exome and Whole Genome Sequencing Challenge the Ethical Framework of Human Genetics Research*, 155 Am. J. Med. Genet. 2916–2924, 2916–2924 (2011).

7 Amy L McGuire et al., *Confidentiality, Privacy, and Security of Genetic and Genomic Test Information in Electronic Health Records: Points to Consider*, 10 Genet. Med. 495–499, 495–499 (2008).

much secondary research involving medical records (as discussed below in Section II.A.2). We do not discuss those proposed changes in detail. The Common Rule amendments have been pending for half a decade, since the Advance Notice of Proposed Rulemaking (ANPR) was published in July 2011, so we do not assume that relevant regulatory changes are imminent or that their final form is predictable.

We conclude by offering standards (versus new regulations), for individual providers and provider institutions (eg academic medical centers, HMO, and large medical practices) to follow in dealing with both patients and researchers interested in genomic data of those patients. In these standards, we propose a model point-of-care notice and disclosure form for EMR-based genomic research. We call for rigorous data security standards and data use agreements (DUAs) in all EMR genomic research, but note that DUAs are relatively toothless without the means to audit compliance and penalize non-compliance.[8] We acknowledge the limitations of any model of permission or consent, recognizing that such models can't anticipate every legitimate use or disclosure occurring in connection with research. At the same time we do *not* agree that, at least in American culture, there is popular support for the view that all patients have a legal or ethical obligation to become subjects of *all* secondary records research, however, valuable the science. Finally, we consider how researchers might encourage patient participation by sharing more information about the research, more quickly, with the patients whose data they obtain.

The stakes are high and time is limited. There are compelling reasons why researchers want and need to combine EMRs with genomic data. Without new steps to promote disclosure and awareness, one day the public will discover that medical and genomic information it assumed was confidential is in fact used widely, and at some privacy risk, in research the subjects neither consented to nor even knew about. This discovery could become an ethical, practical, and political landmine—one that we can, and should, avoid.

I. THE PROBLEM

A health care provider must protect any health information associated with identifiers such dates of treatment, zip codes, and URLs, but that same provider may, under current federal law, give a patient's genome to anyone who asks for it. This is because the federal medical Privacy Rule, promulgated under HIPAA (the federal Health Information Privacy and Accountability Act of 1996), includes dates and URLs among a list of 18 enumerated identifiers whose use and disclosure is regulated, but doesn't specify that DNA sequence data constitute an identifier.[9] In a subsequent regulation implementing the federal Genetic Information Non-Discrimination Act (GINA), federal regulators amended the Privacy Rule, clarifying (in response to arguments to the contrary) that genetic information is considered health information under the Rule, but left

[8] See Yann Joly, Nik Zeps & Bartha M. Knoppers, *Genomic Databases Access Agreements: Legal Validity and Possible Sanctions*, 130 Hum. Genet. 441–449, 441–449 (2011). (noting that DUAs in genomics appear to be untested in legal fora, may be unenforceable against international parties, and are probably less effective at deterring lax security than a process involving 'community sanction').

[9] The federal medical Privacy Rule, 45 C.F.R. § 164.514(b)(2) (Dec. 5, 2016) (deidentification safe harbor), promulgated under The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Pub. L. No. 104–191, 110 Stat. 1938 (1996).

open the question of when such information becomes identifiable absent links to other enumerated HIPAA identifiers.[10]

Currently, actual genomic datasets, whether obtained through gene sequencing, exome sequencing, or whole genome sequencing (WGS), typically are not linked to clinical medical records, although genomic test reports and summary data already appear in an ever-increasing number of EMRs. Consider, for example, the hundreds of thousands of *BRCA1* and *BRCA2* tests performed annually for clinical purposes (the results of which will appear in the medical record), as well as the burgeoning practices of sequencing children with mysterious illnesses (and their parents) in an attempt to determine whether a given condition is linked to a genomic mutation. Most often stored separately on research servers, the genomic data obtained for these purposes is likely to remain linked to patient identities and medical data and preserved for future interrogation as researchers find new, disease-linked variations in the human genome. Notably, the National Human Genome Research Institute is funding a number of pilot projects to explore clinical sequencing in populations ranging from oncology and primary care patients to cardiac patients and those with intellectual disabilities.[11]

Aside from potential clinical uses, gene sequencing is common in research, where it often occurs without the specific consent of the persons whose DNA is sequenced. In fact, current medical research norms permit a scientist who has access to previously collected samples of a patient's blood or tissue to sequence that patient's genome *without* asking the patient to consent to sequencing. (At best, the patient whose clinical specimens are sequenced for research may have signed a clinical consent form containing an inconspicuous, somewhat vague disclosure that samples and data may be shared for unspecified future research.) The scientist then may, and in some cases (eg if a recipient of NIH funding for the sequencing) *must*, share the resulting genomic data with others, including sending the dataset for inclusion in federal government databases used by researchers and companies worldwide, usually without any additional notice to the patient.[12]

The main ethical and legal justification for this practice is the long-standing assertion that a genome constitutes 'de-identified' information, the disclosure of which poses no

[10] See DHHS, Office of the Secretary, Modifications to the HIPAA Privacy, Security, Enforcement and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules; Final Rule, 78 Fed. Reg. 17, 5565, at 5689 (Jan. 25, 2013).

[11] See Clinical Sequencing Exploratory Research, https://www.genome.gov/27546194 (accessed Apr. 26, 2016).

[12] See Final NIH Genomic Data Sharing Policy, 79 Fed. Reg. 51345 (Aug. 28, 2014) (referencing requirement to submit data from NIH funded GWAS to an NIH-designated data repository); see also Jane Kaye et al., *Data Sharing in Genomics—Re-shaping Scientific Practice*, 10 Nat. Rev. Genet. 5 (2009) 331(describing impact of funders' genomic data sharing requirements on scientific practice and participant privacy). NIH policy has been updated to require that for specimens collected after Jan. 25, 2015, specific consent will be required for genomic sequencing, and that consent must include a statement to the effect that, 'Because it may be possible to re-identify de-identified genomic data, even if access to data is controlled and data security standards are met, confidentiality cannot be guaranteed, and re-identified data could potentially be used to discriminate against or stigmatize participants, their families, or groups. In addition, there may be unknown risks'. See NIH Guidance on Consent for Future Research Use and Broad Sharing of Human Genomic and Phenotypic Data Subject to the NIH Genomic Data Sharing Policy, July 13, 2015, https://gds.nih.gov/index.html (accessed Dec. 5, 2016).

significant privacy risk.[13] Yet quietly, but with increasing urgency, medical researchers are debating whether subjects of genomic research can reasonably expect to remain anonymous, as some new studies suggest future re-identification is increasingly possible, if not probable.

Meanwhile, the focus of genomic research is shifting from individuals to populations, from small laboratory collections of DNA to vast databases of genomic and health information, with corresponding privacy implications for increasing numbers of people. The Precision Medicine Initiative, announced with fanfare by President Obama in January 2015, is accelerating this shift. Chief among the databases of interest to researchers will be the burgeoning EMR systems maintained by the nation's health care providers—the physicians, hospitals, laboratories, and insurers who create and maintain health care data.

Technology is changing not only how researchers study DNA, but also how providers manage clinical data. Due in part to federal financial incentives, EMRs have now become the standard for US medicine, replacing the familiar paper chart.[14] This is becoming true even for physician groups, which have lagged hospitals, laboratories, and insurers in adopting EMRs. In digital format, this immense, increasingly cross-institutional and networked collection of health information, from medical histories and patient demographics to treatment outcomes and laboratory test results, affords researchers new opportunities to amass and study large volumes of health outcomes data.

These trends in genomics and data storage are converging, as it becomes apparent that the data used by medical providers will eventually include rich genomic information.[15] No less an expert than Dr. Francis Collins, the director of the National Institutes of Health, has expressed his anticipation that once storage in the EMR becomes possible, patients' genomes can and should be sequenced and the data made available for clinical care and research.[16]

---

13 As further justification, researchers may in some cases be relying on vague language, found in the fine print of surgical and other procedural consent forms, 'informing' any patients who happened to read, and understand, the form in detail, that data or tissue might be used or shared with unspecified parties for unspecified future research.

14 See Federal Support for Health Information Technology in Medicaid: Key Provisions in the American Recovery and Reinvestment Act. *Issue Brief*. THE HENRY J. KAISER FAMILY FOUNDATION, Aug. 2009. Web. Apr 27. 2016. http://kff.org/medicaid/issue-brief/federal-support-for-health-information-technology-in-medicaid-key-provisions-in-the-american-recovery-and-reinvestment-act/.

15 See William Gregory Feero, *Clinical Application of Whole-Genome Sequencing:Proceed with Care*, 311 JAMA 1017, 1017 (2014). See also, *Q&A: Mt. Sinai's Erwin Bottinger on Linking Patient Sequence Data with Electronic Medical Records*, GENOMEWEB, https://www.genomeweb.com/sequencing/qa-mt-sinais-erwin-bottinger-linking-patient-sequence-data-electronic-medical-re (accessed Apr. 26, 2016); see also, *Hospitals Launch Genome Sequencing Programs to Get Ready for the Future of Medicine*, MODERN HEALTHCARE, http://www.modernhealthcare.com/article/20131214/magazine/312149990 (accessed Apr. 26, 2016).

16 See Francis Collins, *Francis Collins Says Medicine in the Future Will Be Tailored to Your Genes The Director of the National Institutes of Health Says Cheaper DNA Sequencing Will Make Personalized Care Routine*, THE WALL STREET JOURNAL, July 7, 2014. Web. Apr.27, 2016 (stating, '[O]ver the course of the next few decades, the availability of cheap, efficient DNA sequencing technology will lead to a medical landscape in which each baby's genome is sequenced, and that information is used to shape a lifetime of personalized strategies for disease prevention, detection and treatment'.

A. The coming collision: modern genomics and medical privacy

Advances in data science and information technology are eroding old assumptions—and undermining researchers' promises—about the anonymity of DNA specimens and genetic data. The term 'de-identification' does not mean what the typical patient might expect: in fact, a 'de-identified' file with both genomic data and traditional medical data, including demographic information on the patient, increasingly can be 're-identified', either by connecting the genomic data to a source with identified genomic data or by connecting the medical data to an individual.[17] At best, the term 'de-identified' is a probabilistic statement about the perceived small likelihood of such re-identification.[18]

Databases of identified DNA sequences are proliferating in law enforcement, government agencies (eg the military, state health department newborn testing programs), genealogical databases, both commercial and public, and commercial direct-to consumer genetic testing enterprises, continually increasing the likelihood that a de-identified gene sequence could be re-identified (linked to a specific individual) if obtained by a person or entity with access to such 'reference' databases.[19] Substantial steps toward re-identification could be taken even by someone capable only of linking a file to identified genomic data of a first, second, or third degree relative of a data subject—relationships readily ascertainable from dense genomic information.

Beyond direct comparison to an identified DNA database, re-identification may also be possible when a third party defeats the de-identification measures used to protect the phenotype data (eg demographics and medical history,) typically linked to the genomic data used in research. Current de-identification practices for phenotype data generally involve removing specific data fields, such as names, addresses, and zip codes, but are not a guarantee of anonymity. Rare combinations of health and demographic data may leave specific individuals within a de-identified data set at a not insignificant risk of re-identification.[20] Ironically, this is particularly true among populations with a high incidence of a particular genetic disease for which research is needed.

And new re-identifications risks will emerge as scientists learn to profile individuals using information encoded in the genome itself, such as height, ethnicity, hair color, and eye color. This future is not mere theory or science fiction: authors of a 2014 study published in *PLOS Genetics* describe a method to use the genome and computerized rendering software to 'computationally predict' three-dimensional models of individual faces;

---

17 See eg Gina Kolata, *Web Hunt for DNA Sequences Leaves Privacy Compromised*, The New York Times, Jan. 17, 2013; but see Daniel Barth-Jones, *The Debate Over 'Re-Identification' Of Health Information: What Do We Risk*? Health Affairs Blog (2012), http://healthaffairs.org/blog/author/daniel-barth-jones/ (accessed Aug. 16, 2016) (concluding that re-identification fears are overblown for much de-identified data, but noting that risks vary with time and recommending legal and technical safeguards for de-identified data).

18 See Khaled El Emam et al., *De-identifying a Public Use Microdata File From the Canadian National Discharge Abstract Database*, 11 BMC Med. Inform. Dec. Making 53, 53 (2011) (describing process of calculating re-identification probability to determine level of de-identification for a health dataset).

19 See eg Catherine Heeney et al., *Assessing the Privacy Risks of Data Sharing in Genomics*, 14 Pub. Health Genomics 1 (2010) 17; see also Michael Grothaus, *How 23andMe Is Monetizing Your DNA Fast Company* (2015), http://www.fastcompany.com/3040356/what-23andme-is-doing-with-all-that-dna (accessed May 10, 2016).

20 See Heeny *et al.*, *supra* note 19, at 19; see also Daniel Barth-Jones, *NCVHS Hearing: De-identification and HIPAA National Committee on Vital and Health Statistics* (2016) http://www.ncvhs.hhs.gov/wp-content/uploads/2016/04/barth-jones.pdf (accessed Aug. 16, 2016).

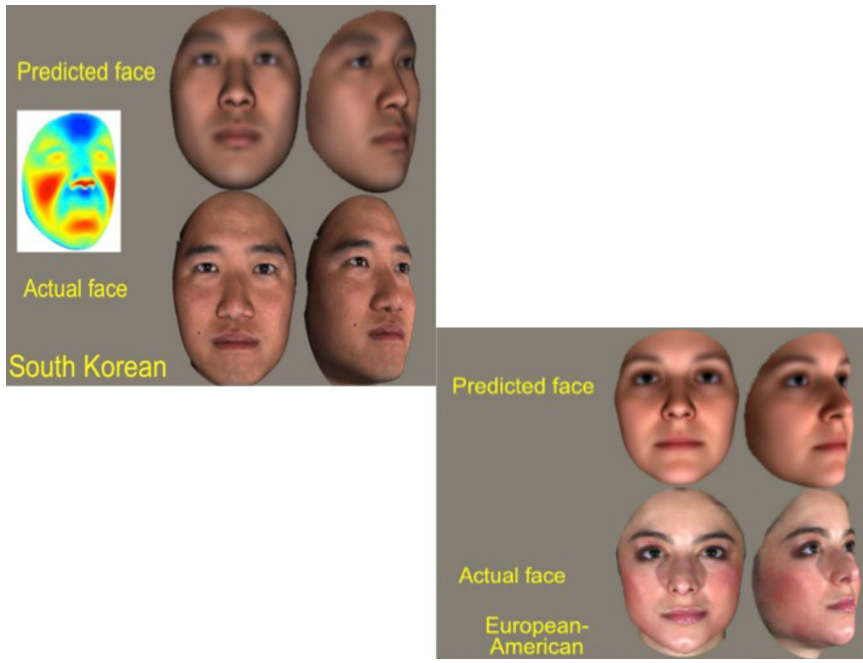

Figure 1. Comparison of computer rendered facial images predicted from genomic data to subject's actual photograph. Methodology described in Claes et al., *Modeling 3D Facial Shape from DNA*, 10 PLoS Genet (2014). Image provided by and used with the permission of Dr. Mark Shriver, Pennsylvania State University Department of Anthropology

the authors foresee widespread use of these techniques within a decade (See Fig. 1).[21] Physical attributes such as height, whose phenotypic expression is influenced by the environment and by multiple genes, may never be genetically profiled with precision, and 'gene photofitting', by itself, may never yield an absolute identification. These techniques will, however, be able to eliminate vast numbers of possible sources for genomic information and, in combination with the de-identified medical information routinely shared for genomic studies, could elevate the re-identification probability for gene sequence data.[22] Debates about re-identification often overlook this type of profiling risk, which is independent of the availability of any reference database.

Lastly, patients are, unwittingly, multiplying their own re-identification risk by transferring increasing amounts of their own identifiable health data to the web via Internet-based personal health records, genealogical tools, interactive medical devices, and even Google searches for disease sites and treatments. A typical de-identification scheme for health data never considers the cumulative identifiability of the health information an individual distributes across the Internet.

Today, medical ethicists, lawyers, and data scientists debate whether de-identification remains a reliable means of privacy protection. One camp maintains that

[21] See Peter Claes et al., *Modeling 3D Facial Shape from DNA*, 10 PLoS Genet. (2014), **DOI**: 10.1371/journal.pcbi.1002822.

[22] See *Building the Face of a Criminal from DNA*, BBC News, http://www.bbc.com/news/science-environment-33054762 (accessed Apr. 27, 2016).

the risks of re-identification are overstated, creating a climate that impedes research unnecessarily; another group of experts, the 're-identification scientists', counter by demonstrating repeatedly how they can re-identify supposedly anonymous subjects in genomic research databases.[23]

Yet to date, this debate has been largely academic, concerned primarily with the privacy of subjects in discrete research studies. Gene sequencing technology is only now maturing into clinical use, and the number of persons whose genomes have been sequenced for research in the USA is relatively small compared to the total patient population. Though many of these research subjects contributed DNA before the advent of sequencing technology and are almost certainly unaware that their genomes have been sequenced and shared, most did consent to participate in some form of medical research and provided DNA samples for this purpose. In theory, at least, these subjects all knew they were assuming new privacy risks arising from research.

This is about to change, and the consequences merit careful consideration.

The impetus for change will be the movement of gene sequencing from the research laboratory to the clinic. When the day arrives that most patients' genomes are sequenced routinely in the course of medical care, genomic data will be integrated in or linked to medical records.

The vehicle for change will be EMRs, which are rapidly replacing the traditional paper medical chart. EMRs that contain (or link to) gene sequence information will become a treasure trove for genomic research on a population-wide scale, allowing researches to forego recruiting DNA donors in favor of obtaining genomic data directly from the EMR.[24] Current accepted practices for records-based research, including waiver of HIPAA authorization and 'de-identification', could, if extended to include EMR genomic information, result in both genomes and health data distributed to networks of researchers throughout the country and, in some cases, around the world—all without the knowledge or permission of the patients themselves.[25] Calls to address privacy risk simply by penalizing re-identification attempts ignore the sad reality that data breaches, though illegal, are reported with increasing frequency for everything from financial records to political documents to health records, yet, while data custodians may be penalized, there are few reports of arrest, conviction, and punishment of the offenders who commit these breaches.

---

[23] See Ann Cavoukian & Khaled El Emam, *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy* (2011), https://www.ipc.on.ca/english/resources/discussion-papers/discussion-papers-summary/?id=1084 (accessed Apr. 27, 2016) (arguing for the effectiveness of de-identification); see also Michelle Meyer, *No, Donating Your Leftover Tissue is not Like Letting Someone Rifle Through Your Phone Forbes* (2015), http://www.forbes.com/sites/michellemeyer/2015/12/31/no-donating-your-leftover-tissue-to-research-is-not-like-letting-someone-rifle-through-your-phone/#121db8e019df (accessed May 10, 2016) (arguing that for de-identified data, 'it is generally not worth the effort and skill of a bad actor to re-identify research data'). For an example of the work of re-identification scientists, see Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321–324, 321–324 (2013).

[24] EMRs that exist in a form accessible across multiple providers or institutions are termed 'electronic health records', or EHRs. HealthIT.gov, *Definition and Benefits of Electronic Medical Records (EMR)*, http://healthit.gov/providers-professionals/electronic-medical-records-emr (accessed Apr. 27, 2016).

[25] The NIH, through its federally maintained database of genotype and phenotype data (dbGaP) obtained from grantee institutions, reports that it has disseminated genomic information [mainly SNP chip data] to over 2000 investigators in 41 countries. See Dina N. Paltoo et al., *Data Use Under the NIH GWAS Data Sharing Policy and Future Directions*, 46 NAT. GENET. 934–938, 934–938 (2014).

1. EMRs will transform records-based research

Electronic storage of clinical data is widespread: hospitals and health systems were early adopters of EMRs, and the National Center for Health Statistics reports that as of 2013, nearly 80 per cent of office-based physicians used some sort of electronic records system, many in response to multi-billion dollar federal incentive programs.[26] Though designed primarily to improve health care delivery and facilitate reimbursement, EMRs, with their large volumes of readily transmissible patient data, are becoming equally essential to medical research. Digital health data are so easily exported from clinical records that in a single project, a researcher using EMRs can study the health outcomes of thousands or even (in large health systems) millions of patients. By pooling data from the EMRs of multiple provider institutions, researchers have also begun to follow health trends and examine health outcomes in entire populations.

Virtually every American who receives health care has—or soon will have—an EMR combining health information with demographic data such as height, weight, birth date, and address. Already, an estimated 40 per cent of the American population has medical record information stored in an EMR manufactured by a single company Epic, a leading supplier of EMRs to academic medical centers and large health systems.[27]

The utility of a common electronic platform for data-driven patient care is already apparent. Epic has created an electronic health information exchange (HIE) among more than 200 institutions. Over a million records per month are shared across this exchange for patient care purposes, but this extensive network has also enabled novel research: a 2014 study pooled emergency department records across four Epic institutions and found that use of the Epic HIE avoided more than 560 duplicate diagnostic procedures during the 9-month study period.[28]

In short, EMRs permit research on a scale—and with a degree of predictive power—that was inconceivable in a world of paper medical charts. Because EMR-based research is so possible and so potentially powerful, most patients in large health systems are also becoming research subjects. The only apparent rate-limiting factors are persistent interoperability problems, particularly across platforms, and the variable quality of EMR data, which tends to be worst during the initial years of transition from paper-based systems.[29] Importantly, however, most EMR research happens outside the awareness of patients, under laws that facilitate the research use of health data.[30]

---

[26] See Chun-Ju Hsiao & Esther Hing, Use and characteristics of electronic health record systems among office-based physician practices: United States, 2001–2003 (2014).

[27] See Brandon Glenn, *Why Epic's Market Dominance Could Stifle EHR and Health IT Innovation Medical Economics* (2013), http://medicaleconomics.modernmedicine.com/medical-economics/content/tags/electronic-health-records/why-epics-market-dominance-could-stifle-ehr?page=full (accessed Apr. 27, 2016).

[28] See Epic, Inc., *Organizations on the Care Everywhere Network*, http://www.epic.com/careeverywhere (accessed Apr. 27, 2016); see also T. J. Winden et al., *Care Everywhere, a Point-to-Point HIE Tool*, 5 Appl. Clin. Inform. 388–401, 388–401 (2014).

[29] See Krister J. Kristianson, Henrik Ljunggren & Lars L Gustafsson, *Data Extraction from a Semi-structured Electronic Medical Record System for Outpatients: A Model to Facilitate the Access and Use of Data for Quality Control and Research*, 15 Health Inform. J. 305–19 (2009).

[30] These federal research and privacy laws and regulations, discussed later in this paper, provide various avenues for researchers to access and use medical records without patient consent (for example, by reducing the identifiability of records or obtaining a waiver from an Institutional Review Board).

WGS will become the clinical standard of care

Paralleling the expansion of EMR systems in medicine, a technological revolution in genomics has increased the speed and, to a remarkable degree, reduced the cost of decoding, or sequencing, an entire human genome. While at least a half billion dollars were spent to sequence the first human genome a decade ago, for a few thousand dollars it is now possible to sequence any patient's DNA and preserve all the sequence data for future use.

Today, at the request of a treating physician, a laboratory might sequence a single patient's genome to detect information relevant to that patient: namely, a small but growing number of genetic variants known to signal disease susceptibility or predict medication response. WGS is not yet common in medical practice because analytic and reporting techniques vary, and because for any given disease, insurers remain uncertain whether WGS is a medically necessary diagnostic service that merits reimbursement.[31] Studies also suggest that it is premature to use WGS to screen healthy adults because the reliability and clinical validity of many findings remains unclear.[32]

But these are short-term obstacles; consensus opinion holds that in the future, clinical demand for WGS will only increase. Similarly, other forms of genomic testing, such as whole exome sequencing (sequencing of the highly identifiable, protein coding regions of the genome) or sequencing of particular panels of genes or other significant genomic regions, may gain popularity as a more cost-effective alternative. In the next two decades, it is quite possible that some kind of genome sequencing will become standard clinical practice for newborn babies.[33]

To meet this demand, EMR vendors will be driven to solve what are, for the moment, daunting challenges: how to store very large gene sequence files (or allow the EMR to interrogate the databases where these data are stored); how to display genetic test results in standard format; how to create decision support tools to make the results meaningful to clinicians who are not genetic counselors.[34] To facilitate insurance coverage and claims processing, regulators, laboratories, and professional medical

---

31 See eg Anthem, Inc., *Medical Policy GENE.00043 Genetic Testing of an Individual's Genome for Inherited Diseases*, https://www.anthem.com/medicalpolicies/policies/mp_pw_c178373.htm (accessed May 3, 2016) (stating that the role of whole genome sequencing in clinical care 'has yet to be established'). More commonly in oncology, research laboratories will sequence the genome of a cancer patient's tumor; although the tumor's genome will be, in some places, different from the rest of the patient's genome, it will provide some information about the patient's non-tumor genome. See also Stephen F. Kingsmore & Carol J. Saunders, *Deep Sequencing of Patient Genomes for Disease Diagnosis: When Will It Become Routine?*, 3 Sci. Transl. Med. (2011), **DOI**: 1a1126/scitranslmed.3002695.

32 See Frederick E. Dewey et al., *Clinical Interpretation and Implications of Whole-Genome Sequencing*, 311 Jama 1035, 1035 (2014).

33 Francis Collins, the Director of the NIH, thinks so; see Francis S. Collins, The language of life: DNA and the revolution in personalized medicine (2010). The NIH is spending $5 million to fund four centers to experiment with newborn genome sequencing, looking at it from many different perspectives. See Anne Eisenberg, *The Path to Reading a Newborn's DNA Map*, New York Times (Feb. 9, 2014), at BU3. Moreover, Collins is not alone; Robert Green, who is leading one of the four research sites, agrees. See Rachel Fobar, *To Predict Future Diseases, Doctors Will Map Newborns' Genes*, Popular Science blog (Apr. 10, 2015) http://www.popsci.com/doctors-will-map-newborns-genes-test-diseases (accessed Dec. 5, 2016). So does Harvey L. Levy, *Newborn Screening: The Genomic Challenge*, 2 Mol. Genet. Mar 2014, at 81–84. And one of us (HTG) expects the forces of commerce, medicine, and 'hype' to lead to such a result, probably too soon.

34 Nancy Snider, Research Integration and Implementation Lead, Epic Systems Corporation., personal communication (Aug. 28, 2013).

societies will eventually develop common standards for reporting sequence data and coding sequencing services.[35]

EMRs will become a compelling tool for genomic research

While clinicians can use gene sequencing to diagnose known genetic conditions and predispositions, researchers are using this technology to identify new genetic factors in disease. Scientists combine WGS data (and similar subtypes, such as whole exome sequences), along with demographic and health data, to hunt for new genetic markers that correlate with health conditions. This research technique is one example of what is known as a genome-wide association study (GWAS).[36] Earlier GWAS efforts used data from inexpensive array technologies to study markers in the genome called single nucleotide polymorphisms (SNPs). SNP-based analysis almost always provided, at best, disappointingly weak associations between particular SNPs and diseases or traits. GWAS using WGS data should be much more powerful.

GWAS requires big data: GWAS researchers often assemble databases containing not only genomes, but information culled from the medical histories of thousands of patients. Such databases are expensive and time consuming to create in the traditional research model, where each DNA donor is recruited and consented as a study participant, each DNA sample is sequenced using research funds, and the relevant medical information must be extracted from each donor's medical chart.

Within the next decade, however, as gene sequencing becomes more common in clinical medicine, it is likely that the data necessary for more powerful, sequence-based GWAS will already exist in (or be linked to) EMR systems. When insurers begin to pay for sequencing in the course of routine care, this trend will accelerate.

As this happens, the totality of the sensitive information embedded in the genome—information about risk of future diseases or addictions, traits and susceptibilities shared with relatives and children, actual biological relationships and ancestral origins, and an unknown quantity of information, yet to be discovered, about the relationship between genes and health—will become an enduring part of EMRs.[37] This does not mean that everyone will have profoundly important or sensitive information in his or her genome, let alone a personal 'future diary'.[38] Still, a significant number of people will—and few if any will know in advance whether they are among those with such sensitive genomic information. This proliferation of clinical genomic data will occur just as the use of EMRs for research becomes commonplace, under norms that don't require patient consent.

Of course, to date GWAS has not been an unvarnished success, and as noted previously, the validity of EMR data can be variable.[39] Nonetheless, in the long run financial incentives strongly favor EMR-based genomic research, as scientists who make

35 See Steve Olson, Integrating large-scale genomic information into clinical practice: workshop summary (2012).

36 See William S. Bush & Jason H. Moore, *Chapter 11: Genome-Wide Association Studies*, 8 PLoS Comput. Biol. (2012), **DOI**: 10.13711/journal/pcbi.1002822.

37 See Kenneth Blum et al., *Genome Wide Sequencing Compared to Candidate Gene Association Studies for Predisposition to Substance Abuse a Subset of Reward Deficiency Syndrome (RDS): Are we throwing the Baby Out with the Bathwater?*, 4 Epidemiology: Open Access (2014) (describing potential GWAS approaches to the study of addiction risk).

38 George J. Annas, *Privacy Rules for DNA Databanks: Protecting Coded 'Future Diaries'*, 270 JAMA 2346 (1993).

39 See Kristianson, *supra* note 29, at 305.

secondary use of clinical genomic data bear neither the cost of gene sequencing nor the effort and expense of consenting individual patients and collecting project-specific phenotype data.[40]

De-identification is a moving target

For decades, medical ethicists have approved and regulators have allowed the non-consensual use of clinical records in research on the basis of one core assumption: that removing common identifiers such as names and Social Security numbers from the data nearly eliminates the risk of harm. This approach, once quaintly termed 'anonymization', is currently known as 'de-identification' (reflecting a growing understanding of the probabilistic nature of re-identification).[41] When data are de-identified, anonymity isn't, technically speaking, guaranteed: instead, identifiers are removed or masked to the point where the probability of re-identification appears at a given point in time to be (as specified in one federal regulation) very small.[42]

Yet, even if de-identification can protect many forms of health data by reducing the probability of re-identification, genomic data in their raw (non-transformed) format—or as a list of variants from a standard (reference) genome—may be unusually vulnerable to future changes in the level of re-identification risk. Unlike a blood type or a cholesterol test result, an individual's DNA sequence codes for unique combinations of physical traits that, collectively, may create a fully or partially identifying profile.[43] The more scientists learn about genetic profiling, the more this profiling re-identification risk will escalate. Meanwhile, the more commonly discussed possibility of re-identification via comparison of anonymous sequences with identified DNA databases in the public and private sector will also remain a growing risk.[44] In either case, to the extent genomic data are linked with 'de-identified' phenotype data, re-identifying a gene sequence will also mean re-identifying all of the EHR health and medical data associated with that sequence.

Skeptics might discount re-identification risk by arguing that no one would have much incentive to re-identify genomic information when other information stores, such as banking information, offer more low-hanging fruit. Apart from law

40 And, of course, to the extent that new information streams, such as data from the promising—and heavily promoted—field of 'mobile health' self-monitoring, become integrated with the EMR, EMRs with genomic data will become even more compelling sources of data for research.

41 See Committee on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine, SHARING CLINICAL TRIAL DATA: MAXIMIZING BENEFITS, MINIMIZING RISK. Washington (DC): National Academies Press (US); Apr. 20, 2015. *Appendix B, Concepts and Methods for De-identifying Clinical Trial Data*. http://www.ncbi.nlm.nih.gov/books/NBK285994/ (accessed Dec. 5, 2016) (describing measurement of the probability of re-identification risk for health data).

42 See Clete A. Kushida et al., *Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies*, 50 MED. CARE 82–101 (2012).

43 See Jen Wagner, *Re-Identification Is Not the Problem. The Delusion of De-Identification Is. (Re-Identification Symposium) — Bill of Health*, http://blogs.law.harvard.edu/billofhealth/2013/05/22/re-identification-is-not-the-problem-the-delusion-of-de-identification-is-re-identification-symposium/ (accessed Dec. 5, 2016) (arguing the a gene sequence is itself an identifier, and cannot be de-identified);see also Erika Check Hayden, *Privacy Protections: The Genome Hacker*, 497 NATURE 172–174, 172–174 (2013) (describing re-identification of participants in the international 1000 Genomes Project).

44 But see Bradley Malin et al., *Identifiability in Biobanks: Models, Measures, and Mitigation Strategies*, 130 HUM. GENET. 383–392, 383–392 (2011). (arguing, in 2011, that 'it is not yet possible to identify a person without an identified sample of DNA', and 're-identification is largely preventable').

enforcement and national security interests in genomic re-identification and profiling, however, one could easily foresee other motivations for genomic re-identification, from tabloid appetites for celebrities' medical information to sophisticated targeted marketing efforts, as well as profiling for life insurance and other purchases (unlike for health insurance, genomic profiling for life insurance or credit risk is not prohibited by federal law). The return on reidentification efforts will likely increase as technology improves and medicine can tell us more about the implications of genomic variation.

The obvious solution might seem to be technical innovations that might make genomes less identifiable. Although data scientists now proffer a variety of algorithms that purport to transform genomic data into less-identifiable forms, the genomic research community has not embraced these techniques or adopted any standard for data transformation. It is possible that the transformations necessary to reduce the re-identification risk degrade the informational value of a genomic sequence to an unacceptable degree; more likely, scientists may want to preserve their access to raw, untransformed sequence data for future use.[45] In either case, technology has yet to provide an attractive solution to the re-identification problem.

De-identification and its limits are more significant for records research as clinical data become electronic. The significant time and effort required to abstract data from paper medical charts manually have always constrained the size of research databases, limiting the aggregate privacy risk to patients. Electronic health data change this risk calculus in important ways: a typical EMR in a large health system contains tens of millions of records, and the effort required to export records is the same regardless of the number of records. By pooling data in multi-institutional studies and drawing upon multi-state electronic HIE systems, it is foreseeable that researchers might one day access the health data and the genomes of the majority of Americans on a continual basis. Indeed, this is the model envisioned by some policymakers and embodied in the concept, endorsed by the National Academies of Sciences Institute of Medicine, of a 'learning health system'.[46]

And why not, if privacy is protected? The conventional view, reflected in a 2012 report of the President's Commission for the Study of Bioethical Issues, is that the benefits of new knowledge substantially outweigh the privacy risks of genomic research, provided that researchers remove direct identifiers (eg names and addresses) from the data. The President's Commission analogized DNA to a fingerprint that does not

[45] The Privacy Rule's 'statistical expert' de-identification provision, an alternative to the safe harbor, that permits enumerated identifiers in a 'de-identified' dataset, requires finding a statistical 'expert' who will certify that re-identification risk is 'very low', (a term undefined in the regulation). 45 C.F.R. §164.514(a)-(b) (Dec. 5, 2016). See eg Deven Mcgraw, *Building Public Trust in Uses of Health Insurance Portability and Accountability Act De-Identified Data*, 20 JAMA 29–34, 29–34 (2013).

[46] In 2011, the National Academy of Sciences (NAS) published *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. The report advocated that researchers have real-time access to nationwide networks of clinical data, concluding that 'realizing the full promise of precision medicine, whose goal is to provide the best available care for each individual, requires that researchers and health-care providers have access to very large sets of health and disease-related data linked to individual patients'. See National Research Council (U.S.), TOWARD PRECISION MEDICINE: BUILDING A KNOWLEDGE NETWORK FOR BIOMEDICAL RESEARCH AND A NEW TAXONOMY OF DISEASE (2011).

encode identifying information and may only be identified if matched to a print from a known individual.[47]

This characterization is insufficiently forward looking: it neglects the rapid growth in the number of public and private reference databases of information that could be used to make a re-identifying match, whether those databases as genotypic, medical, or genealogical. It also fails to account for the re-identification risk stemming from the future prospect of genomic profiling, the compilation of an identifying list of physical features using only information encoded in the sequence data.[48] And, most fundamentally, it assumes that individuals' interests and rights in the use of personal information are disposable if some third party concludes overall benefits outweigh overall risks. Individual rights generally don't work that way.

The research community still maintains the perplexingly naive attitude that most data research, including genomic data research, should be considered 'minimal risk'. Compelled by government mandates, research institutions spend millions of dollars each year on compliance systems to reduce the statistically rare incidence of physical harm to research participants in clinical trials. Yet the same institutions often participate in large-scale secondary data use projects where hundreds of thousands or even millions of patient records are exported to third parties, sometimes with little effort, apart from a DUA, to ensure that data storage and access procedures meet security best practices. Recent massive commercial and government data breaches—and in particular, breaches of large health systems (the majority of which have now been compromised in some way)—demonstrate that few data systems are invulnerable, so it seems realistic to assume that breaches of large research databases are inevitable.[49] When this happens, the unaware participants may face real privacy and identity theft risks (medical identity theft is one of the fastest growing, and most expensive consequences of health care data breaches, imposing significant costs and burdens on patients and providers), and institutions themselves may be exposed to the very significant cost of providing credit monitoring, in addition to regulatory penalties and legal liability.[50]

Even perfect de-identification would not be enough

But assume, for the moment, that perfect de-identification—in essence, the elimination of re-identification risk—were possible. Would a reasonable patient still have grounds

47 The NAS report did acknowledge the shifting landscape of re-identification and recommended informed consent whenever DNA samples are obtained for sequencing. The report did not address the common IRB practice of waiving consent for the secondary use or sharing of sequence data. See *Id.*

48 See Manfred Kayser & Peter De Knijff, *Improving Human Forensics Through Advances in Genetics, Genomics and Molecular Biology*, 13 Nat. Rev. Genet. 753–753, 753–753 (2012) (discussing new techniques to infer ancestry and externally visible characteristics from genomic sequence data, known collectively as "DNA phenotyping").

49 See Jessica Davis, *7 Largest Data Breaches of 2015 The Healthcare Industry Lands Three Top Spots*, 2015, http://www.healthcareitnews.com/news/7-largest-data-breaches-2015 (accessed May 3, 1016) (describing the cumulative exposure of more than 100 million patient records).

50 The theft of medical information for the purpose of obtaining medical services, insurance reimbursement, and prescription drugs now accounts for more identity thefts than in the banking, finance, government, military, or education sectors, with breaches of nearly 70 million records, each a source of information that can be sold in underground markets. Genomic databases may become an attractive target because the unique nature of the genome might permit re-identification of an extensive record of otherwise de-identified medical information. See *The Rise Of Medical Identity Theft In Healthcare*, Kaiser Health News (2014), http://khn.org/news/rise-of-indentity-theft/ (accessed May 3, 2016).

to object to use of her health data and genome for research? Some commentators argue that patients would, and the available data seem to support this view.[51] Patients generally expect to exercise control over research uses of their information, and subgroups may actually object to certain uses.[52] Whatever researchers, lawyers, and ethicists think of patients' rights, to the extent that patients think they have such control, disregarding their understanding is unwise.

If, for example, data from members of one ethnic group were used, without the members' knowledge or consent, in an effort to demonstrate that group's inferiority or predisposition to stigmatizing diseases or conditions, it seems both reasonable and, indeed, predictable that those members might object, as they have in several cases involving biospecimens.[53] Causing distress in patients who learn only after the fact that they've become research subjects seems an ethical breach; it also seems likely to result in bad public relations and contentious politics for genomic science.

The ethics

Patients are not (automatically) research subjects

The 'patient' who passively places her health in the hands of a well-intentioned physician is a concept dating to antiquity. The 'human subject' who makes an informed affirmative choice to subjugate her own interests to those of science is a relatively modern construct. Not until the mid-twentieth century did organized bodies begin to define different ethical norms for medical care and human research, reflecting a growing understanding that research alters the physician-patient relationship (although the distinction between research and treatment can be blurred in areas such as oncology, where many patients are placed on protocols as a means to access investigational drugs).

The World Medical Association's Declaration of Helsinki, published in 1964, along with its predecessor, the Nuremberg Code (of 1948), changed the landscape of medical research profoundly, eventually informing new legal protections for human subjects in many countries, including the USA.[54] The Code is a widely cited appendix to the US military court's judgment in criminal trials of those responsible for horrific Nazi human experimentation; 16 years later, the Declaration expanded the Code's principles, making more explicit the obligations of physicians who conduct human research.

Both the Code and the Declaration assume that research introduces new risks and conflicts of interest to the physician–patient relationship; beyond informed consent, both documents also establish criteria for the research itself, such as societal value and

[51] See eg Fiona Riordan et al., *Patient and Public Attitudes Towards Informed Consent Models and Levels of Awareness of Electronic Health Records in the UK*, 84 Int'l J. Med. Inform. 237–247, 237–247 (2015).

[52] See eg Rebecca Dresser, *Public Preferences and the Challenge to Genetic Research Policy*, 1 J. L. & Biosci. 52–67, 52–67 (2014); see also Jill O. Robinson et al., *It Depends Whose Data are Being Shared: Considerations for Genomic Data Sharing Policies*, 2 J. L. & Biosci. 697–704 (2015). (stating there is an 'urgent need' for data sharing policies that accommodate variation in individual and group preferences).

[53] See eg Amy Harmon, *Havasupai Case Highlights Risks in DNA Research*, The New York Times, Apr. 21, 2010, http://www.nytimes.com/2010/04/22/us/22dnaside.html?ref=us (accessed May 3, 2016); see also Beth A. Tarini, *Storage and Use of Residual Newborn Screening Blood Spots: A Public Policy Emergency*, 13 Genet. Med. 619–620, 619–620 (2011) (describing acrimonious controversy resulting from non-consensual use of newborn blood samples for research, and to create a federal DNA database).

[54] See Robert V. Carlson, Kenneth M. Boyd & David J. Webb, *The Revision of the Declaration of Helsinki: Past, Present and Future*, 57 Brit. J. Clin. Pharmacol. 695–713, 695–713 (2004); See also WMA declaration of Helsinki: ethical principles for medical research involving human subjects 2013 .

risk minimization.[55] But the Code addresses human experimentation, not data privacy, while the Declaration, even in its most recent, seventh revision in 2013, mentions data research only in passing, concerning itself little with the circumstances under which patient records might become research data. (It does, however, require that '[f]or medical research using identifiable human material or data, such as research on material or data contained in biobanks or similar repositories, physicians must seek informed consent for its collection, storage and/or reuse'.[56]) The Code also assumes that data may be rendered 'anonymous'—an assumption that seems dangerous in our modern era of population-based genomic research.[57] Moreover, neither the Code nor the Declaration anticipates a world in which technology and big data make it possible to render every patient an involuntary subject of genomic research.

Medical research guidelines issued in the 1980s by the Council for International Organizations of Medical Sciences (CIOMS), in collaboration with the World Health Organization, further refined the ethical obligations of biomedical researchers.[58] Although these guidelines reflect the same overly sanguine assumptions about the effectiveness of de-identification and anonymization, as last revised in 2002 the CIOMS guidelines do distinguish sharply between patient and subject data, prescribing different standards for secondary research involving the records of consenting subjects and research involving the records of patients, where privacy expectations are greatest. The guidelines advise that when medical records will be disclosed for research without consent, providers should always notify patients, and should honor specific patient requests not to participate.[59]

As we discuss further below, US regulations pertaining to research and medical privacy also distinguish between patients and subjects, providing for IRB review, consent, and HIPAA authorization when researchers transform 'patients' into 'subjects' by using identifiable patient data for research. These regulatory schemes do permit waiver of patient consent and authorization when certain criteria are met, but arguably do not permit researchers to override the wishes of patients who express a desire to opt out of research use.[60]

Current practice affords less than full disclosure to data subjects

What do patients understand and believe about how clinical data is used and disclosed for research? Most probably don't have an informed opinion, because there is no legal requirement that patients be given specific information each time their providers disclose records for research—unless the patients themselves know enough to ask the

55 See eg *Id.*, Principle 17.

56 *Id.*, Principle 32.

57 See Carl Coleman, *How Should Ethics Be Incorporated into Public Health Policy and Practice?*, 85 BULL. WORLD HEALTH ORG. 504–504, 504–504 (2007); see also Carlson et al., *supra* note 54.

58 See COUNCIL FOR INTERNATIONAL ORGANIZATIONS OF MEDICAL SCIENCES & WORLD HEALTH ORGANIZATION, INTERNATIONAL ETHICAL GUIDELINES FOR BIOMEDICAL RESEARCH INVOLVING HUMAN SUBJECTS (2002). The 2020 Guidelines are currently in revision. SeeEmily A. Largent, *Recently Proposed Changes to Legal and Ethical Guidelines Governing Human Subjects Research*, 3 J. L. & BIOSCI. 206–216, 206–216 (2016).

59 See WMA DECLARATION OF HELSINKI, *supra* note 54 (Commentary to Guideline 18.)

60 These regulations grant subjects a perpetual right to withdraw from research, except where data have already been properly disclosed to third parties.

right questions.[61] The federal medical Privacy Rule does require providers to give patients a 'Notice of Privacy Practices' (NPP), but with respect to research, a provider can satisfy the regulation by simply stating in the NPP that the provider 'may use and share your information for health research', and then obtaining a waiver of the Privacy Rule's patient authorization requirement or using a DUA when disclosing data for specific projects.[62]

To the extent that any patient actually reads the NPP in its entirety, the required disclosures are quite vague and non-specific, and fall far short of conveying any sense of the sheer number of people, including third parties, who will be given access to patient information for records research—much less disclosing anything about the research itself.[63] The only way that a patient can learn which researchers are studying her medical records is to ask the provider for an 'accounting of disclosures'—and even such an accounting is limited in scope. Under the Privacy Rule, an accounting covers only the prior six years, is often not study specific, and includes only research involving 'individually identifiable health information' as defined by the Rule.[64]

This last limitation matters most, because a provider would not need to include disclosures of genomic data in a Privacy Rule accounting if such data are not considered identifiable health information. Typically researchers characterize genomic data as 'de-identified' information, and federal regulators have not objected. The research community has long operated as though a unique DNA sequence is not an identifier per se—unlike a fingerprint, driver's license number, or URL, each of which are enumerated identifiers under the Privacy Rule.[65] Genomic data reside in an identifiability gray zone: while most researchers and policymakers have acknowledged that gene sequences could in theory be *re*-identified, linking the data with the DNA source, they have maintained that the magnitude of this risk is small, so small that it doesn't warrant requiring informed consent for data use or oversight by federally regulated Institutional Review Boards (IRBs).[66]

We disagree, and we argue, as have other commentators, most notably George Church, leader of the Personal Genome Project, that it is no longer ethically defensible

[61] The Privacy Rule requires patient authorization for research disclosures, but permits an IRB or Privacy Board to waive such authorization, without specific notice to patients, under criteria that, as applied in practice, result in waivers for most big data health research.

[62] See Model Notice of Privacy Practices for Providers, www.hhs.gov/ocr/privacy/hipaa/npp_fullpage_hc_provider.pdf (accessed May 5, 2015) (indicating that to comply with the Privacy Rule, providers may tell patients simply 'We may use and disclose your health information for research').

[63] The number of researchers and study staff potentially accessing medical records for research is limited only by the provider's discretion; the provider who creates the records may also share them with external researchers, institutions, and companies—without patient consent—if the provider complies with regulatory requirements.

[64] HIPAA Privacy Rule, 45 C.F.R § 160.103 (Dec. 5, 2016) (definition of 'protected health information'), and 45 C.F.R §164.528 (Dec. 5, 2016) (Accounting of Disclosures provision requires covered entities to make available to an individual upon request an accounting of certain disclosures of the individual's PHI made during the six years prior to the request).

[65] See 45 C.F.R §164.514(a)-(b) (Dec. 5, 2016).

[66] For example, The National Cancer Institute, which funds population genomic studies in oncology, assembled a panel of experts in 2013 to discuss the topic, and appears to have concluded that the risk of re-identification is too remote to justify a more restrictive approach to the use of gene sequence data. See Carol J. Weil et al., *NCI Think Tank Concerning the Identifiability of Biospecimens and 'Omic' Data*, 15 GENET. MED. 997–1003, 997–1003 (2013).

or legally sound to maintain that gene sequence data are anything other than identifiable health information.[67] For WGS data, the research community should dispense with the hair-splitting nuances of federal regulatory schemes that attempt gradations of 'identifiability' in favor of a best practice that recognizes re-identification risk increases with time and patients are best protected if we treat their genomes as identifiers, now and in the future.

We are equally concerned about the transfer of gene sequence and medical data obtained in the course of clinical care to federal, commercial, and other third party academic medical center databases, without meaningful disclosure to the data subjects. We believe that at a minimum, patients and subjects should receive specific notice that this use of their genomes or other medical information can and does occur.

In the coming era of personalized genomics, we see patients' privacy expectations colliding with the growing demand in academia and industry for genomic data, and with the 'permission optional' culture of medical records research. Patients, conditioned by both deep cultural beliefs about doctor–patient confidentiality and the more recent federal Health Insurance Portability and Accountability Act (HIPAA) paperwork to believe that medical privacy is their right and their provider's obligation, will be worried—even angered—to learn how extensively their genomic information is used and shared for research without consent, and how variables are the current data privacy and security practices in research.

Of course, patients can only object to the research practices of which they are aware. We think such awareness is inevitable, and that it may come about in one of two ways: either the research community launches a frank and open dialogue with the public, explaining the benefits of genomic research and proposing uniform standards to protect privacy interests, or the issue will surface in an inflammatory context such as a major security breach, prompting restrictive policies that neglect the immense value of the new knowledge emerging from this work.[68]

Precisely because the research is too valuable to jeopardize by risking a public backlash and ill-considered legislative or regulatory measures, we hope to spark that open dialogue by proposing standards and norms for the research use of clinical gene sequence data in the EMR.

RECORDS-BASED RESEARCH TODAY

This section of the paper looks first at the relevant current legal rules. The paper then examines the current ethical and legal practices for secondary records research, noting where current practices may diverge in spirit or effect from the stated intent of the 'rules'.

Current rules

For the purpose of this paper, the three sets of current legal rules are important: medical record ownership, research subject protection, and health information privacy.

---

67 See Jeantine E. Lunshof et al., *From Genetic Privacy to Open Consent*, 9 NAT. REV. GENET. 406–411, 406–411 (2008); see also Jeantine E. Lunshof & Madeleine P. Ball, *Our Genomes Today: Time to Be Clear*, 5 GENOME MED. 52, 52 (2013).

68 See Misha Angrist, *Genetic Privacy needs a More Nuanced Approach*, 494 NATURE 7, 7 (2013).

Who owns the medical record?

Patients would be surprised to learn that they don't own the medical records that their providers maintain; whether paper or electronic, these records are generally viewed as a business asset owned by the patient's provider (or that provider's employer).[69] While federal and state medical privacy laws give patients certain rights of access to their providers' medical records, these laws don't confer ownership of the records, or even full, traditional 'privacy' rights, because they don't allow patients to control how such records are created, used, or shared, except under narrow circumstances.[70]

Instead of a basis in true privacy or property rights, the 'privacy' regime in health care comprises series of state and federal statutes and regulations offering what could more accurately be described as confidentiality protection: covered providers (and their vendors and contractors) are required by these laws to preserve patient confidentiality by maintaining medical records securely and disclosing identifiable information only for legitimate purposes, and subject to certain controls.[71] The protection regime focuses on the provider's record: once information from this record is no longer under the control of the covered provider—for example, once it is in the hands of third party researchers—it is largely beyond the reach of most medical records privacy regulations.[72]

Legal protections for human subjects

In the USA, two federal regulations are the primary source of protection for human research subjects, but each regulation is limited in scope, and only one, the Federal Policy for the Protection of Human Subjects ('Common Rule'), addresses the secondary use of clinical data.[73] The Common Rule dates to 1991, prior to the significant use of EMRs, and extends only to research (a) funded or(b) conducted by the DHHS (which includes NIH-funded research) or by other federal agencies that have adopted the Rule by regulation or (c) by federally funded entities that elect to extend the Common Rule to all of their research). (Eighteen federal agencies follow the Common Rule.)[74] Common Rule agencies require institutions receiving federal grant money (eg research universities) to file a Federalwide Assurance certifying that the grantee complies with

---

69 A provider's ownership of the legal medical record is distinct from ownership of the underlying information held in other forms; once disclosed to patients or other providers, such information may be subject to competing ownership claims. Also, in isolated cases, state courts have recognized a patient ownership interest in medical records, and legislatures in a handful of states have granted patients vaguely defined 'property' rights in genetic information. See Barbara J. Evans, *Much Ado About Data Ownership*, 25 HARV. J. L. & TECHNOL. 69, 73 (2011).

70 See Mark A. Hall, *Ownership of Medical Information*, 301 JAMA 1282, 1282 (2009); see also Daniel J. Solove, *Conceptualizing Privacy*, 90 CAL. L. REV. 1087 (2002). http://scholarship.law.berkeley.edu/californialawreview/vol90/iss4/2 (accessed Dec. 5, 2016) (describing traditional legal concept of privacy as control of access to the self or information).

71 For example, the HIPAA Privacy Rule incorporates mechanisms such as DUAs and business associate agreements when providers release protected information without patient consent.

72 An exception exists when the covered entity includes researchers within its designated workforce (assuming responsibility for their compliance) and elects to treat its own research as a HIPAA-covered activity.

73 See 45 C.F.R. §46.102 (f) (2) (Dec. 5, 2016).

74 U.S. DHHS, Federal Policy for Protection of Human Subjects (The Common Rule), http://www.hhs.gov/ohrp/humansubjects/commonrule/ (accessed Dec. 5, 2016).

federal human subjects protection policies; grantees whose assurance is suspended or revoked for non-compliance may no longer spend federal grant funds.[75]

The Common Rule generally requires, among other safeguards, that grantees obtain IRB review and seek participants' informed consent when research will require an intervention with a subject or, importantly for this article, will involve the investigator obtaining what the Rule defines as 'identifiable private information' about living individuals, unless the research qualifies for one of several categories of exemption.[76] The Common Rule exempts data research from these protections if the investigator otherwise has legitimate access to the data (eg is a physician studying her own patients), and will not record identifiers for the research.[77] The regulation defines private information as 'identifiable' if an investigator may 'readily ascertain' the identities of the data subjects.[78]

Most relevant to data research, the Common Rule permits an IRB to waive participants' consent if the research risks are minimal, the waiver would not adversely affect subjects' rights and welfare, and the research could not practicably be carried out without the waiver.[79] When an IRB approves an EMR-based research study (which typically involves many patient records), that IRB will almost invariably waive subjects' consent and authorization as being impracticably expensive and time consuming to obtain. (The version of the Common Rule adopted by the Federal Food and Drug Administration targets research involving FDA-regulated products and does not contemplate this kind of records-only research.)[80]

In July 2011, the DHHS issued an 'ANPR, signaling an intent to make extensive changes to the federal Common Rule for the Protection of Human Subjects.[81] The subsequent Notice of Proposed Rulemaking, published on September 8, 2015, generated a flood of comments, with many academic medical institutions focusing on the practical implications of a proposed consent mandate for biospecimen research.[82] The future, and eventual terms, of these proposed amendments remains uncertain, but one provision, if adopted, could have a significant, if largely unnoted, effect. Few commenters, whether from academic medical centers or elsewhere, paid attention to the proposal to exclude from human subject protection regulations entirely any research use of identifiable health information governed by the HIPAA Privacy Rule.[83]

---

75 45 C.F.R. § 46.103 (Dec. 5, 2016).

76 45 C.F.R. § 46.109; 116 (Dec. 5, 2016).

77 45 C.F.R. § 46.103 (Dec. 5, 2016).

78 45 C.F.R. § 46.101(b) (Dec. 5, 2016).

79 45 C.F.R. § 46.116(c) (Dec. 5, 2016).

80 The U.S. Food and Drug Administration (FDA) regulates human research involving 'clinical investigations' of FDA-regulated products. 21 C.F.R. § 50.1(a) (Dec. 5, 2016). FDA's human subjects regulations contemplate clinical studies in which an investigator intervenes directly with subjects, administering an investigational product or test; these regulations do not permit waiver of informed consent except in very limited circumstances (eg certain research involving emergency situations). 21 C.F.R. § 50.23 (Dec. 5, 2016).

81 Notice of Proposed Rulemaking, Federal Policy for the Protection of Human Subjects, 80 Fed. Reg. 53931 (Sept. 8, 2016) (ANPRM).

82 AAMC Submits Comments to HHS on Common Rule NPRM–2016 (2016), https://www.aamc.org/advocacy/washhigh/highlights2016/451934/010816aamcsubmitscommentstohhsoncommonrulenprm.html (accessed May 3, 2016).

83 *Id.* § 101 (b)(2)(iv) NPRM, *supra* note 81.

With this one provision, barely referenced in the preamble to the Proposed Rule (which noted simply that the researcher and the provider must both be covered by the rule), DHHS would effectively deregulate and remove from IRB review almost all EMR-based research conducted by covered entities. Note that researchers who are not otherwise covered by the Privacy Rule (and would therefore remain subject to the Common Rule) could become 'covered entities' for the purpose of accessing a covered entity's EMR, simply by providing some service (such as abstracting data from the EMR for the researcher's own study) to the HIPAA-covered entity and signing a HIPAA 'business associate agreement' with that entity.[84]

The proposed exclusion from the Common Rule of secondary records research leads us to conclude that the NPRM, if it became a final rule, would not impose any significant regulations relevant to our topic. We recognize, however, that until the DHHS publishes a final rule we cannot be certain—for example, it could adopt the suggestion (acknowledged in the NPRM text but not proposed as a change to regulatory language) that gene sequence data be defined to be identifiable under both the Common Rule and HIPAA.

Legal protections for medical data privacy and security

The federal medical Privacy Rule, promulgated under the Health Insurance Portability and Accountability Act of 1996, restricts how 'covered entities' (eg most providers and insurers) may use and disclose 'individually identifiable health information' for research. In comparison to the Common Rule, the Privacy Rule might appear to broaden privacy protections in data research. The Privacy Rule extends beyond federal grantee institutions to all US entities that transmit health data electronically for a covered purpose (almost all providers and health care institutions, as well as insurers). The Privacy Rule also defines 'identifiable' more broadly than the Common Rule, which protects the subjects of private information only when an investigator can readily ascertain the identity of those data subjects. The Privacy Rule, by comparison, protects all health information held by a covered entity when there is a reasonable basis to believe such information can be used to identify an individual, even if not 'readily'.[85] Before a covered entity may use or disclose this protected health information (PHI) for research, the entity must obtain each data subject's written authorization.[86]

Importantly, however, the Privacy Rule contains its own waiver provisions, with criteria resembling those in the Common Rule, and in addition to waiver provides several other routes for a covered entity to use or disclose information to researchers without any form of patient permission. The first is a 'de-identification' regulatory safe harbor, under which the covered entity may treat health information as completely outside the scope of the Privacy Rule's protections if the entity removes 18 enumerated identifiers

[84] *Id.* at 53954. Potentially any researcher, even one not employed by a HIPAA-covered entity, could become HIPAA covered for this purpose, in compliance with the Rule's requirements for third party contractors, by agreeing to abstract the data for her own project or to provide minimal analytic services to a provider under an HIPAA business associate agreement.

[85] 45 C.F.R. § 164.514(a) (Dec. 5, 2016) Privacy Rule De-Identification Standard Under the Privacy Rule, a healthcare provider who conducts research may elect to exclude its own research from its HIPAA-covered entity; choosing whether to do so involves a complicated assessment of the administrative burden of segregating research activities for HIPAA purposes.

[86] 45 C.F.R. § 164.508 (Dec. 5, 2016) (Privacy Rule) Requirements: Research Authorizations for Use or Disclosure of Protected Health Information.

(ranging from name and zip code to URLs and biometric identifiers) and has no 'actual knowledge' that the remaining data could be re-identified; alternatively, the entity must obtain certification from a 'statistical expert' that for the combination of elements in a given data set, the probability of re-identification is 'very low'.[87] The covered entity also may elect to create a 'limited data set' by removing specific 'direct identifiers', such as name and Social Security Number, and may then disclose the remaining data to researchers under a 'data use agreement' that contains terms specified in the Privacy Rule.[88]

The HIPAA Security Rule, a companion regulation, applies to electronic PHI (ePHI), and requires covered entities to adopt administrative, physical, and technical safeguards to protect ePHI from unauthorized access and maintain the integrity and availability of ePHI.[89] The Security Rule's standards govern how a covered entity stores and transmits any ePHI that entity maintains for any purpose, including research. Both the HIPAA Privacy and Security Rules apply in addition to any existing state laws pertaining to medical records privacy. Importantly, genomic information, if not deemed identifiable, need not be maintained in an electronic form that meets Security Rule standards.

Federal and some state regulations also require 'breach notification' in the event of certain data breaches, mandating covered entities to notify consumers and the government of large, unauthorized disclosures of identifiable personal information that have the potential to cause harm (eg disclosures of unencrypted data containing identifiers.)[90] The HIPAA breach notification regulations apply to defined breaches of all ePHI, but typically state breach laws define covered information more narrowly, limiting notification to disclosures of breaches of information associated with a direct identifier, such as a name or Social Security Number. (State regulators may, however, have the power to impose substantial fines for certain data breaches).[91]

Lastly, an evolving landscape of class action litigation has created liability-related incentives for hospitals and physician practices to maintain the privacy and security of clinical data. The litigation climate for security breaches is unsettled, with plaintiffs pursuing new theories of liability in the wake of large, highly publicized data breaches. While the elements of a successful claim are not yet clear, it is evident that large providers in states with more consumer-friendly breach statutes have begun to enter multi-million dollar settlements in class action cases.[92]

Each of these legal protections, whether for data research, data security, or data breach, is available only to data meeting various standards for identifiability. Unless

[87] The 'statistical expert' provision is less commonly used in research, given the uncertainty around the meaning and application of the terms. 45 C.F.R. 164.514(a-c) (Dec. 5, 2016) (De-Identification Safe Harbor); 45 C.F.R. 514(e) (Dec. 5, 2016) (Disclosure of Limited Data Set).

[88] *Id.*

[89] 45 C.F.R. Part 160 and Subparts A and C of Part 164 (Dec. 5, 2016).

[90] See examples at the National Conference of State Legislatures Website, www.ncsl.org/research/telecommunications-and-information-technology/seuciryt-breach-notification-laws.aspx (accessed Dec. 5, 2016).

[91] 45 C.F.R. §§ 160.103, 164.400 (Dec. 5, 2016). See also CA Civil Code §§1798.25-1798.29 (Dec. 5, 2016) (describing notice requirements for state agency databases).

[92] See Jason Green, *Settlement-Possible-Stanford-Medical-Information-Breach*, THE MERCURY NEWS, Mar. 22, 2014, http://www.mercurynews.com/ci_25398083/4-1m-settlement-possible-stanford-medical-information-breach (accessed May 3, 2016).

genomic data receive this designation, unauthorized uses and disclosures of patient genomes will not incur legal penalties or civil liability.

Current practices

As a general matter, a given element of personal data is protected only to the extent that a given law or rule defines the term 'identifiable' to include that element, but inconsistent legal definitions of identifiable—and inconsistent, sometimes equivocal guidance from federal agencies—cloud the status of genomic data. Moreover, perhaps due in part to the absence of any private right to sue under the Common Rule or the HIPAA Privacy Rule, there has been little if any judicial interpretation of 'identifiable' in these contexts.

The multiple meanings of 'identifiable'

The federal Common Rule, drafted in the 1980s in an era of paper medical charts, deems information individually identifiable only if the identity of the subject may be *readily* ascertained by the investigator or associated with the information.[93] Re-identification science and electronic data mining were not anticipated by regulators of the Common Rule era. Federal regulators have attempted in guidance documents to define the circumstances under which information is considered 'Common Rule' identifiable, but in so doing have highlighted the different standard for identifiability under the HIPAA Privacy Rule.[94] In the preamble to the recent NPRM for the Common Rule, regulators considered but appear to have rejected the possibility of harmonizing these regulations by adopting the Privacy Rule standard for identifiable information.

The HIPAA Privacy Rule, written after the advent of electronic HIE and as a result of legislation expressly focusing on such records, extends the definition of identifiable to any health information where there is a *reasonable basis to believe* it *can* be used to identify an individual. Among the data elements that the Privacy Rule specifies as de facto identifiers are any 'biometric identifier' and any 'unique, identifying number, characteristic, or code'.[95] This more sweeping definition would seem, on its face, to include the genome, the ultimate biometric, a truly unique (but for identical twins) identifying characteristic and code. DHHS has not taken a position either way on that argument.

If there were any room for doubt, the HIPAA Privacy Rule, in its 'safe harbor' provisions, states that even after removing all 18 of the enumerated identifiers, a covered entity must treat the remaining data elements as protected information if that entity has 'actual knowledge' that a recipient *could* re-identify the information. What this means is at the heart of the genomic privacy debate in research.

The National Institutes of Health, most plainly in a recent NIH policy document on genomic data sharing, consistently states that genomes are de-identified information.[96]

---

93 45 C.F.R. § 46.102(f) (Dec. 5, 2016) (Definition of 'human subject').

94 See U.S. DHHS, Office of Human Research Protections, *Guidance on Research Using Coded Private Information or Specimens (2008)*, http://www.hhs.gov/ohrp/policy/cdebiol.html (accessed May 5, 2016).

95 45 C.F.R. § 164.514(b)(2) (Dec. 5, 2016).

96 See The National Institutes of Health, NOT-OD-14-124: NIH Genomic Data Sharing Policy, NOT-OD-14-124: NIH Genomic Data Sharing Policy, http://grants.nih.gov/grants/guide/notice-files/not-od-14-124.html (accessed May 3, 2016), at SectionII (c)1, stating '[T] data in the NIH database of Genotypes and Phenotypes (dbGaP) are de-identified by both the HHS Regulations for Protection of Human Subjects and HIPAA Privacy Rule standards . . .'.

NIH continues to hold this position even though it both has imposed increasingly stringent security precautions for access to the genomic data that it collects and maintains, now treating these data as *potentially* identifiable by requiring investigators who access the dbGaP genome repository to sign DUAs containing confidentiality, security, and access restrictions.[97]

The federal regulators who oversee Common Rule compliance for the DHHS at the Office for Human Research Protections (OHRP) have not challenged the scientific practice of assuming genomes are de-identified and conducting secondary genomic research without consent or IRB review. Thus far, OHRP has not publicly questioned the NIH position that whole genome sequence data are de-identified, and therefore their use does not constitute 'human subjects research' under the Common Rule. This is true even though the NIH reportedly has, in resisting compulsory disclosure of dbGAP data under the Freedom of Information Act (FOIA), argued that disclosure of such data would be an invasion of subjects' personal privacy.[98]

The federal DHHS's Office of Civil Rights (OCR), which interprets and enforces the HIPAA Privacy Rule, has been somewhat equivocal, if not cryptic on this topic. OCR has stated in its guidance on de-identification that an 'identifying characteristic or code' is one that would *currently* allow for re-identification.[99] With respect to the question of when a provider has 'actual knowledge' that data may be re-identified (thereby negating the safe harbor), ORC guidance states that the mere publication of re-identification techniques is not sufficient to meet this standard—leaving open the question of what kind of knowledge would suffice.[100]

DISTINGUISHING RESEARCH FROM OTHER SECONDARY USES OF MEDICAL RECORDS

Research is not the only—or even the most common—use of EMR data beyond the direct provision of care to patients. Healthcare providers routinely use and disclose information from their medical records, often in identifiable form, and without patient consent, for purposes that include a broad category of health care business activities such as billing, accounting, finance, strategic planning, and quality improvement (collectively termed 'healthcare operations' by federal privacy regulations);[101] as well as for state and federal public health activities and to satisfy document demands from

97 See Michael Krawczak, Jürgen W Goebel & David N Cooper, *Is the NIH Policy for Sharing GWAS Data Running the Risk of Being Counterproductive?*, 1 Invest. Genet. 3, 3 (2010).

98 See Amy L. Mcguire & Laura M. Beskow, *Informed Consent in Genomics and Genetic Research*, 11 Annu. Rev. Genom. Human Genet. 361–381, 361–381 (2010).

99 See U.S. DHHS, Office of Civil Rights, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#uniquenumber (Dec. 5, 2015).

100 See *Id.*, stating as follows: 'A covered entity may be aware of studies about methods to identify remaining information or using de-identified information alone or in combination with other information to identify an individual. However, a covered entity's mere knowledge of these studies and methods, by itself, does not mean it has "actual knowledge" that these methods would be used with the data it is disclosing'.

101 The model NPP developed for providers by the federal Office of Civil Rights states simply that the provider will use the patient's information to, among other things, "treat you," "run our organization," "bill for your services," "help with public health and safety issues," "do research," and "comply with the law." http://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/npp_fullpage_hc_provider.pdf (accessed Dec 5, 2016). See *Id.*

regulators, law enforcement, and litigants. We describe those other disclosures briefly, largely in order to distinguish the issues they raise from those involved in research.

Healthcare operations

Under the rubric of 'health care operations', the federal HIPAA Privacy Rule permits entities covered by the Privacy Rule (most health care providers, insurers, and pharmacies) to use or share identifiable patient information without consent to provide treatment. Covered entities may also use identifiable patient information internally, without consent, as necessary to conduct normal business operations, such as to obtain payment, process claims, or assess the quality of care. And finally, the Privacy Rule also permits 'covered entities' to disclose or share patient information, also without consent, with other covered entities for treatment or reimbursement purposes, and with vendors and contractors who sign an agreement (known as a HIPAA Business Associate Agreement) containing certain privacy and security obligations.[102]

One example that illustrates the scope of these 'TPO' (treatment, payment, and health care operations) disclosures is the electronic HIE. Through an HIE, providers across a state or region can network much of their clinical data for purposes that initially were treatment focused, but are now expanding to include research. HIEs may be public or private, though many were initially funded and facilitated by the federal HITECH Act. Some states have established HIEs on the state level, but third party operator entities have also moved to aggregate providers and health systems in multi-state HIE consortia.[103] No specific notice to patients is required when a provider or facility participates in an HIE, although a minority of participants do seek prior patient consent. Most HIEs provide an opt-out for those patients who somehow learn of the HIE, object to data sharing, and contact the HIE operator directly.[104]

Beyond HIPAA, some state and other, narrower federal laws further restrict a provider's ability to use and disclose sensitive information, such as records of HIV or substance abuse treatment, without patient consent. To the extent they offer greater privacy protection, these laws are not preempted by HIPAA.[105]

Public health activities

State and federal government agencies, from the CDC to state health departments to the US Food and Drug administration, also routinely collect identifiable patient information abstracted from medical records for what are termed 'public health activities'. These uses range from tracking disease incidence to evaluating prevention programs or investigating adverse events related to drugs or medical devices. The Privacy Rule specifically permits these disclosures without patient consent, although, as mentioned

[102] 45 C.F.R. §§ 164.502(e), 164.504(e), 164.532(d) and (e) (Dec. 5, 2016).

[103] The Health Information Technology for Economic and Clinical Health (HITECH) Act is part of the American Recovery and Reinvestment Act of 2009, Pub. L. No. 111-5, 123 Stat. 115, 516 (Feb. 19, 2009). HITECH includes several billion dollars of funding to be distributed to states for the creation of HIE infrastructure and services.

[104] For example, Maryland's major hospitals, clinical laboratories, and radiology facilities participate in the state-sanctioned CRISP HIE, not all providers inform patients of this fact, and only patients who learn of the HIE and visit its website have an opportunity to opt out. See CRISP > FOR PROVIDERS > PARTICIPATING ORGANIZATIONS, www.crisphealth.org (accessed Dec. 5, 2016).

[105] See 45 C.F.R. Part 160, Subpart B (HIPAA preemption provisions).

above, patients may request that a provider or other covered entity provide an 'accounting' of instances in which that patient's identifiable health information was shared for certain purposes, including public health, during the past six years (though anecdotal evidence suggests few patients are aware of or exercise this right).[106]

Compliance and law enforcement

With certain procedural restrictions, the federal medical Privacy Rule also provides a pathway for providers and other covered entities to release identifiable information to federal agencies such as the Center for Medicare and Medicaid Services (for Medicare-related billing, quality, audit, and other purposes); to other federal agencies performing audit or investigation functions; to federal, state, and local law enforcement; and to private litigants.[107] This could include, for example, releasing information to the police as part of a criminal investigation or to counsel in personal injury case who is seeking information to use against a party in that case. The Privacy Rule sometimes requires legal process, such as a subpoena, before a covered entity may make compliance and law enforcement disclosures.

Why research is different

Does research differ in any material way from these myriad other uses of medical records of which most patients are unaware, and over which patients exercise may little or no control? In many respects, the answer is no, but there are two important exceptions. The first caveat is that researchers who obtain information from patient records may operate outside the governance and regulatory and contractual confidentiality obligations that apply to providers and insurers (and to their contractors), to federal agencies, and (to some extent) to state law enforcement and civil litigants.

While these legal requirements help to raise the bar for data security among operational and government users of EMR data, they aren't a fail-safe; a recent report estimates that one in 10 US citizens has been affected by a breach of medical records security involving a provider or its contractors.[108] But in the absence of mandated safeguards or even agreed-upon standards, data privacy and security in research turn on whether individual investigators understand and implement encryption, access controls, firewalls, and other basic electronic data safety measures. As a commentator noted in the journal *Nature,* the genomic information collected for research 'is supposed to be highly protected [but] it is disseminated to various institutions that have inconsistent security and privacy standards … data protection often comes down to individual scientists.. [o]nce leaked, these data would be virtually impossible to contain'.[109] It is important to note that, as discussed above, even if adopted the proposed changes to federal research regulations would likely not change this analysis. The proposal, which include unspecified security standards, would not apply to most secondary research

[106] Carol Richardson, Privacy Officer, Johns Hopkins Medical Institutions, personal communication (estimating annual accounting requests in the single digits.)

[107] See Steven E. Brenner, *Be Prepared for the Big Genome Leak*, 498 NATURE 139–139, 139–139 (2013).

[108] SeeKatie Wike, *HHS: Data Breaches Affect 1 In 10 HHS Data Breaches Affect 1 In 10* (2014) http://www.healthitoutcomes.com/doc/hhs-data-breaches-affect-in-0001 (accessed May 3, 2016).

[109] See Brenner, *supra* note 107.

using EMR clinical data, because the revisions would largely remove the secondary use of HIPAA-covered data from the Common Rule.[110]

The second and more important caveat is that research is not something the patient is required, either legally or practically, to participate in. A patient must accept some uses and disclosures of information for a health care provider to operate a health care business or to respond to governmental demands. Each medical records research project conducted without consent, however, could be viewed as an elective intrusion upon patient privacy. Even though these intrusions may ultimately benefit this patient, or other patients, they are different from the trade a patient must make when giving up some privacy to access health care.

This elective aspect distinguishes research ethically from some of the other routine uses of medical records. The weight that we give this distinguishing factor may vary, but we can't disregard it. And, especially for research involving clinical genomic data, before we assume that patients will support this use unquestioningly, we must honor that ethical distinction, conveying the full scope of the privacy intrusion and explaining the limits of any assurances we make about confidentiality.

It could be argued that permitting the broad use of medical records for research should be a public duty, like providing evidence, mandatory vaccinations, compulsory education, or paying taxes. Effectively, such legally authorized conscription of medical data for research already exists, to the extent that the federal Common Rule and Privacy Rule permit waiver of individual consent and authorization for medical records research without prior specific notice to patients. Although we, as coauthors, may disagree about the proper (and practical) scope of waiver in the research context, we both believe EMR research involving genomic data implicates privacy and security practices that exceed current norms.

ESSENTIAL RESEARCH VERSUS INDIVIDUAL RIGHTS

The potential for widespread EMR research using genomic data threatens a direct confrontation between the needs of research and the rights and interests of patients. In this section, we contend that this impending conflict requires special attention and possibly exceptional responses.

EMR research using genomic data is important

It is easy to see where the interests of all parties to EMR research align: Patients, providers, payors, and researchers all benefit when well-designed, ethically conducted studies produce useful new knowledge. From diabetes to cancer to infectious disease, much of what we are now learning about population disease risk and health outcomes—knowledge that currently improves care for millions of patients—results from researchers mining clinical data in EMRs.[111] Adding genomes to this data

[110] Although HIPAA-covered entities are subject to the relatively undefined standards of the HIPAA Security Rule (45 C.F.R. Part 160 and 164, Subparts A and C) research itself is not a 'covered function' under HIPAA, so it is unclear whether these standards apply to secondary research use of clinical data.

[111] See T. A. Manolio, *Genomewide Association Studies and Assessment of Risk of Disease*, 363 New Eng. J. Med. 2076–2077, 2076–2077 (2010) (describing the state of GWAS research); see also Robert H. Shelton, *Electronic Consent Channels: Preserving Patient Privacy Without Handcuffing Researchers*, 3 Sci. Transl. Med. 4 (2011).

mining effort creates a potent scientific tool that should lead to a better understanding of disease, and, ultimately, more effective, efficient treatments.[112]

Researchers themselves have a direct and substantial interest in maintaining their access to immense volume of valuable clinical information stored in the EMRs of providers and health systems. Providers use the findings of EMR-based research to set practice standards and make evidence-based treatment decisions. Payors now use EMR-based research to decide whether treatments work and are cost-effective. Taxpayers, who subsidize federal payors such as Medicare, Medicaid, and the Veteran's Administration, have a decided economic interest in supporting the kind of EMR research that creates a sound evidence base for reimbursement decisions.

EMR research using genomic data requires higher standards

But do these interests, in the aggregate, outweigh the individual patient's autonomy interests—interests that traditionally we attempt to honor in research? Some bioethicists have argued that they do, proposing that when medical records research involves minimal risks, everyone who is a patient has an ethical obligation to participate.[113] Whether or not that is a compelling ethical argument, no express legal obligation currently exists (though, as noted previously, current laws permit the conscription of much patient information for research through a waiver process).[114] But even if there were such an obligation, we can ask whether research involving clinical genomic information is different in ways that justify an exception to this obligation principle.

The policy argument over genetic exceptionalism reflects conflicting views about whether genetic information differs in important ways from other clinical information, and deserves special protections. Some states have endorsed this view, singling out genetic testing in confidentiality statutes and non-discrimination statutes. In contrast, federal regulators rejected this approach when drafting the HIPAA Privacy and Security rules in 2000, refusing to declare genomes to be categorically different from other health information. This latter view is not uniform across federal legislation: The federal GINA, though not a confidentiality statute, specifies that genetic information is a special category of health data that health insurers and employers may not use in coverage or hiring decisions.[115]

We think that among the many types of health information, several characteristics make genomic data especially, if not uniquely, sensitive. Like biometric identifiers, dense genomic datasets are unusually subject to re-identification, they can reveal sensitive family and ancestry information, and they predict current and future health concerns to an extent that is, at least currently, collectively unclear and almost completely

---

[112] See Paltoo et al., *supra* note 25.

[113] See Ruth R. Faden et al., *An Ethics Framework for a Learning Health Care System: A Departure from Traditional Research Ethics and Clinical Ethics*, 43 HASTINGS CENTER REP. (2013), **DOI**: 10.1002/hast.134.

[114] One might argue that the waiver provisions of the Common Rule and Privacy Rule create an implicit norm for mandatory research participation Michelle Meyer, J.D., Ph.D. Assistant Professor of Bioethics and Director of Bioethics Policy in the Union Graduate College-Icahn School of Medicine at Mount Sinai Bioethics Program, personal communication, Oct. 2015 DHHS, however, suggests in its NPRM that, at least with respect to biospecimen research, participants who refuse to consent could not be compelled to participate via the waiver process.

[115] Genetic Information Nondiscrimination Act of 2008 (2008 - H.R. 493). *GovTrack.us*, https://www.govtrack.us/congress/bills/110/hr493 (accessed Dec. 5, 2016).

unknown to any individual. Perhaps most importantly, people *believe* genomic data are sensitive, and at least in some contexts (eg FOIA, as noted above), government entities appear to agree. By recognizing the dynamic, uncertain quality of re-identification risk and the near consensus that genomic data have some special sensitivity, we can address the tension between individual and collective interests by focusing, not on patient obligations, but on the obligations that should accompany the use and disclosure of clinical genomic data for research.

This does not mean, however, that we dismiss the privacy risks associated with the secondary use of other types of clinical data. For example, we think the widespread sharing of three-dimensional cranial MRI and CT datasets for research with few (if any) controls on data use poses a current and not insignificant risk to the privacy of the patients whose images are shared. Very little skill is required to use open source software to render a facial image from such a dataset (one could do this on a home computer); recent work suggests that with the help of facial recognition software, such renderings can be matched correctly to subjects' photographs in nearly one third of comparisons.[116] Though beyond the scope of this paper, the identifiability of imaging data is a research privacy problem that providers and imaging researchers should take seriously.

EMR genomic research may be a special case

Much medical records research will never be conducted with patient consent; many argue that for practical and scientific reasons, it can't be. In fact, regulators and commentators are entertaining proposals to eliminate consent for EMR research, or to simply deem such uses of EMR data 'healthcare operations' and therefore not research, thus removing them altogether from requirements for research oversight.[117]

But use of patients' genomic sequences, which we believe to be identifiable within the plain meaning of that term, should be a special case.[118] We do not believe claims that *all* secondary-use genomic research involves minimal risk to the data subjects, although the existing practices in effect treat all of it as such. The heightened potential for re-identification of genomic data and the inherent sensitivity of such data are compelling reasons to distinguish WGS data studies from other medical records research, and to afford patients' autonomy and privacy interests greater respect than is the current practice.

We also argue that providers and researchers have an equally compelling, if less-often noted, interest in prioritizing patient autonomy and choice in especially sensitive areas of research. For economic reasons, providers must be concerned about meeting patient expectations and reducing liability exposure. Honesty and transparency about records disclosures should make good business sense—at least to the extent that such practices become industry norms.

---

[116] See eg Jan C. Mazura et al., *Facial Recognition Software Success Rates for the Identification of 3D Surface Reconstructed Facial Images: Implications for Patient Privacy and Security*, 25 J. DIGIT. IMAGING 347 (June 2012);see also Fred W. Prior et al., *Facial Recognition From Volume-Rendered Magnetic Resonance Data*, 13 IEEE TRANS. INF. TECHNOL. BIOMED. 5–9 (2009).

[117] See Devin McGraw, *Paving the Regulatory Road to the 'Learning Health Care System'*, 64 STAN. L. REV. 75 (2012).

[118] The Oxford English Dictionary defines 'identifiable' as 'able to be recognized'. See Definition of identifiable in English: IDENTIFIABLE: DEFINITION OF IDENTIFIABLE IN OXFORD DICTIONARY (AMERICAN ENGLISH) (US), http://www.oxforddictionaries.com/us/definition/american_english/identifiable (accessed May 3, 2016).

Perhaps most pragmatically, genomics researchers will need continued public support, both for funding and for access to medical records. For EMR-based research, scientists' access to data will depend on providers' willingness to open their records; a change in public sentiment, prompted by revelations that genomes are disclosed to researchers without consent or IRB oversight, could affect that willingness dramatically. We have seen recent examples of popular backlash against unconsented and unknown research, from the Havasupai lawsuit against unexpected uses of health information and DNA samples given for diabetes research to lawsuits by parents in Texas and Minnesota over undisclosed research using their children's neonatal blood spots.[119] And in response to such revelations, would researchers really argue to patients that, although their fingerprints and even their URLs are identifiers under federal law, their genomes are not?

V. FINDING A BALANCE: WORKABLE PRACTICES THAT RESPECT PATIENT RIGHTS

Legal mandates for privacy protection can usefully set a floor for conduct and enable the government to single out extraordinarily bad or negligent behavior for sanction. But as a means to establish best practices, laws and regulations have significant limitations: in genome-related research, just as in the financial services industry, the law—especially the protracted rule-making process of regulation—can never keep pace with innovation. Legislative responses can be backward looking and inflexible. Laws, regulations, and legal precedent are, for the most part, jurisdiction specific, while today's genomic research can involve international collaborations and multi-national corporations.

The better, more nimble, and more far-reaching approach is voluntary, but normative. Although the Privacy Rule (and proposed changes to the Common Rule) gives them the latitude to do otherwise, health care providers and the research community should adopt a common set of best practices to govern use and disclosure of genomic information created for clinical purposes. Professional societies, academic institutions, and major provider entities who publicly endorse consensus best practices have the power to create a de facto standard of conduct that evolves, flexibly and organically, with advances in science and technology.

There is precedent for such an approach in the embryonic stem cell research committee (ESCRO) structure first proposed in 2005 by the National Research Council of the Institute of Medicine, with the goal of addressing emerging controversies in the largely unregulated area of human embryonic stem cell research.[120] Many institutions have altered the NRC's procedural recommendations in favor of a more efficient review process, but the core proposals still garner praise as an example of successful scientific self-governance.[121]

Voluntary standards have also been proposed for international genomic database research: in 2009, the Organization for Economic Cooperation and Development, a

[119] See Beleno v. Lakey, No. SA-09-CA-188-FB (W.D. Tex., Sept. 17, 2009); see also Bearder v. State, 788 N.W.2d 144 (Minn. Ct. App. 2010).

[120] See National Research Council (U.S.) & Institute of Medicine (U.S.). Guidelines for human embryonic stem cell research. (The National Academies Press, 2005).

[121] See Henry T. Greely, *Assessing ESCROs: Yesterday and Tomorrow*, 13 Am. J. Bioethics 44–52 (2013); Mary Devereaux & Michael Kalichman, *ESCRO Committees—Not Dead Yet*, 13 Am. J. Bioethics 59–60, 59–60 (2013).

member organization comprising 34 countries (including the United States), published guidelines to govern research involving biobanks and databases of genomic information.[122] These standards are stated in broad terms, but include IRB (or, internationally, ethics committee) review of most secondary uses of genomic data, and would require data sharing agreements and specific protocols for data access and protection. Similarly, Knoppers et al., on behalf of three international genomics research organizations, have published a data sharing Code of Conduct for international research collaborations.[123] Neither of these standard sets specifically addresses the secondary use of genomic data from medical records, although both guidelines recognize the need for policies that extend beyond the use of data collected in the context of a research protocol.

With these models in mind, and building upon this prior work, we offer the following proposed standards to govern research use of clinical genomic data.

First principle: avoid surprises

In 2006, the United Kingdom's Academy of Medical Sciences studied how British researchers use National Health Service medical records. The AMS concluded that the existing NHS goal—to seek patient consent whenever records could not be anonymized—'will never be feasible for much research using patient data'.[124]

Arguing that anonymized data simply isn't useful for much research, and further, that British law allows researchers to use identifiable medical records without consent under defined circumstances, the AMS also endorsed what one commentator called the 'no surprises' principle: don't assume that the public understands and agrees; instead, reach out to inform patients and then study their attitudes and preferences, using what you learn to inform policy decisions.[125]

But asking the question means risking an unfavorable response. A recent UK study of patients in NHS outpatient clinics found that when asked, only 14 per cent supported use of their identifiable records for research, while 18 per cent would not permit research use even if their records were de-identified.[126] In the USA, although studies suggest that patients do support the general concept of medical records research, when it comes to their own data, patients expect to be informed; many also want the opportunity to consent.

One of the best studies of US patient expectations, conducted in 2010, found that more than two thirds of patients who had donated DNA for genetic research did not want their genomic data shared with the federal dbGaP database without their express consent.[127] This large survey involved elderly patients who had already joined a

[122] See Organization for Economic Cooperation and Development. *Guidelines for Human Biobanks and Genetic Research Databases (HGBRDs)* (2008). www.oecd.org/sti/biotechnology/hbgrd (accessed Dec. 5, 2016).

[123] See Bartha Knoppers et al., *Towards a Data Sharing Code of Conduct for International Genomic Research*, 3 GENOME MED. 46, 46 (2011).

[124] See ROBERT L. SOUHAMI, PERSONAL DATA FOR PUBLIC GOOD: USING HEALTH INFORMATION IN MEDICAL RESEARCH (2006).

[125] See *Id.*

[126] See Serena A Luchenski et al., *Patient and Public Views on Electronic Health Records and Their Uses in the United Kingdom: Cross-Sectional Survey*, 15 J. MED. INTERNET RES. (2013). **DOI**: 10.2196/jmir.2701.

[127] See Evette J. Ludman et al., *Glad You Asked: Participants' Opinions Of Re-Consent for dbGap Data Submission*, 5 J. EMPIRICAL RES. HUMAN RES. ETHICS 9–16, 9–16 (2010). The authors note that most other studies of participant attitudes toward reconsent in data studies have relied upon hypothetical scenarios.

longitudinal, NIH-funded dementia study and had a lengthy relationship with the investigators; the authors note that a younger, more diverse sample of patients who have never participated in research might feel even more strongly about consent.

Quite possibly, despite glancing at the HIPAA NPP in their physician's office, few patients realize that their identifiable data, much less their genomes, could be disclosed outside their local clinic or hospital and used by researchers other than their own providers. On the basis of the few studies of attitudes and preferences conducted to date, however, it seems clear that patients do want to know. Whether one sees that fact as ethically important, practically important, or both, it clearly should be important.

So, in terms that they can understand, providers must tell patients that this happening, and explain why. The fact that IRBs routinely waive patient consent (and HIPAA authorization requirements) for EMR research, on the grounds that seeking consent is impracticable for large samples, does not justify the failure of researchers and providers to give patients any meaningful notice of how (and how often) identifiable medical information —particularly genomic information—is used and disclosed for research.

Nor can we justify this failure to inform by resorting to the argument, advanced by some bioethicists, that patients have an ethical obligation to participate in medical records research—even if, under some circumstances, we agree.[128] Such an obligation, even if it exists, would not be an obligation to participate blindly, with no awareness of scope of the privacy risk or the scale of the potential benefits of the research. Unlike the HIPAA NPP, notice to patients about EHR genomic research should be informative.

Meaningful notice to patients could take many forms, but the simplest approach might be an electronic roster, maintained at the provider's website, of all studies to which the provider has disclosed genomic data, along with each investigator's contact information. Such a notice would contain information that patients are already entitled by to receive under federal law, but which few actually do receive unless they are aware of and exercise their right to request a HIPAA 'accounting of disclosures' from each of their providers. An electronic roster of data studies might also help patients and institutions to hold investigators accountable for data security, by making public information about which third party researchers are holding genome sequence data initially created for clinical purposes. The existence of the roster could be disclosed to patients in person or by email, mail, or the telephone, in addition to being present on the institution's website.

Provide more information, not less

We know that genome sequence data in the EMR will likely be used for research one day, even if it isn't possible to know by whom, or for which studies. What, then, should a physician who orders genomic sequencing for a diagnostic purpose tell his or her patient about this eventuality?

Patients deserve more than a generic statement that their medical information may be used for research. But the providers who are ordering gene sequence tests may have little or no information to share about particular studies involving EMR records. Increasingly, decisions about which data to export from the EMR and for what purpose are handled centrally within large medical centers and health systems, so providers in

[128] See Faden *et al., supra* note 113.

those environments may not even know when data about their patients are released to researchers.

What providers do know, and can tell patients at the point of care, is that research using patient records is now common and can be an important tool for discovering new relationships between genes and health. Physicians can tell patients that they share records because new research findings can improve the quality and cost-effectiveness of medical care. Providers must tell patients that it will not always be possible to ask for consent, but they can reassure patients that through DUAs and other legal means, any researcher receiving genomic or other potentially identifiable information from your medical records will obligated to protect the security and confidentiality of those data.

Yet providers should not promise absolute confidentiality. Patients should know that their genomes are unique and can't be made anonymous. Providers should also help patients understand that research findings developed through the use of EMR data may be too new and uncertain for medical use and therefore individual results will not, in most cases, be returned to patients.

Consider asking for permission and offering patients control

Consent is one of the biggest ethical challenges for EMR-based genomic research. Once the research community stops insisting that it is reasonable and ethical to treat genomic data as either 'de-identified', 'anonymous', or 'not readily identifiable', then in most cases federal regulations (and some state laws) dictate that researchers must obtain patients' consent—or an IRB waiver of consent—before using these data for research.[129]

When medical records research involves identifiable information, IRBs often agree, consistent with regulatory criteria, that it would be impracticable to contact thousands of subjects to ask permission, and further, that the potential for response bias (differences between the health or demographic characteristics of those who consent and those who refuse) might compromise the validity of the study. The power of these justifications has made consent waiver routine in records research, to the point where millions of patients are currently the subjects of such research without having any idea that this is the case.

Innovations in the EMR space could disrupt the consent waiver paradigm by undermining the impracticability argument. One feature of many EMR systems is a 'patient portal', through which patient and provider may exchange information in a secure, encrypted communication. Portals are also a means for providers to push information to their patient population, and for patients to respond to satisfaction surveys or indicate preferences related to their care.

The patient portal could also be a way for patients to record their preferences about participating in genomic research. Because patient portals interface with the EMR, researchers using the EMR can identify those patients who have either given global consent or opted out of research participation, without the need to contact any patient directly. Several patient advocacy groups such as the Genetic Alliance and Autism Speaks are constructing a similar form of patient portal, with the goal of giving their members

[129] Adoption of the proposed changes to the Common Rule transferring control over research with identifiable data to HIPAA would, of course, change this by leaving in place only the HIPAA authorization requirement, which a Privacy Board may (and typically does) waive using criteria nearly identical to those use to waive consent under the Common Rule.

more control over the use of their samples and genomic information; conceivably such existing systems might be programmed to interface with Epic and the other major EMR systems.

We believe that providers who are using EMR systems should move toward allowing patients to document their willingness to participate in genomic research via use of a portal-based general permission form. We argue that this documented permission should not be equated to consent under federal research regulations, or to HIPAA authorization, because the point-of-care based process will be too prospective and attenuated to meet these strict regulatory standards. There are also clear ethical limitations to seeking general consent for unspecified future research: most obviously, that patients can't make a fully informed decision about uses and risks not yet identified by investigators or IRBs.

And the challenges of implementing an EMR-based permission system are greater than they might first appear. When patients receive a preference form through an EMR patient portal, the burden is likely to fall on the primary care provider—the point of contact with the patient—to answer questions about risks and benefits of unspecified future research. The time constraints of the primary care setting dictate that any preference form be short and easy to read, so it is unlikely that the process or documents could meet the extensive consent requirements of federal research regulations. It may well make sense for institutions to set up alternative contacts for questions about this research permission.

Despite these limitations, and recognizing that the process may not meet all regulatory standards for research consent and HIPAA authorization, we still believe that asking permission for research use at the time of clinical testing demonstrates respect for patient rights and autonomy.

Importantly, however, an ethical permission process must inform patients that there are circumstances when permission cannot or will not be sought.

Be honest when permission isn't possible

Even if it were possible to give every patient the ability to log into a portal and record his or her preferences about research use of the EMR, there will still be instances in which patients' clinical genomic data are used and shared for research without permission.

It simply isn't possible for a provider to apply patient preferences retroactively when patients' DNA and data have already left the control of the provider's institution. Further, providers' pathology departments and clinical laboratories still share 'de-identified' clinical specimens for research, especially in academic medicine, and not infrequently for gene sequencing studies, without any requirement to document which specimens were shared, or with whom.

We can advocate against this practice, and cite a 2012 proposal by the federal OHRP that all biospecimen research be, at a minimum, conducted in a traceable, secure manner (ie be registered with an IRB and subject to data security standards), but it will take time to change long-standing attitudes and expectations about the free exchange of 'de-identified' biospecimens. As a result, a patient who has ever had pathology or clinical laboratory testing can't be sure that her biological materials—or her medical information derived from them, including her genomic information—won't be used for research. Nor can her provider.

The provider might offer the patient choices and some degree of control over the use of EMR genomic data that the provider has not yet disclosed, but any permission form must explain the circumstances under which patient preferences will not or cannot be respected.[130] For example, if the provider participates in an HIE whose laboratory test data, including gene sequencing data, may be used for research without patient consent, that provider should inform the patient of this possibility and of the availability of any opt-out.

Be scrupulous about data security

The research community's long practice of treating genomic information as de-identified or describing such data as 'anonymized' has impeded the development of community norms for data privacy and security in genomic research.

Providers, whether individuals or institutions, should only release EMR genomic data into secure environments. Release should be subject to a DUA between provider and recipient that contains enforceable indemnification provisions (supported by proof of insurance coverage) and is signed by a person with the authority to bind the researcher's employer entity. Terms of the DUA should include the following:

I. A minimum set of security standards that include encryption, storage only on secure servers behind institutional firewalls, and appropriate access and authentication protocols
II. A designated list of approved recipients, with a prohibition on access by, or transfer to, unapproved third parties without the provider's written permission.
III. A requirement to provide the data source an annual accounting of all copies and all users of the data set.
IV. A prohibition on attempts to re-identify our contact data sources, or to create new, identifiable information through joinder with other available datasets.

Yet DUAs are not sufficient to truly protect privacy. DUAs provide no direct protection to data subjects, who are not parties to these agreements and whose information is already compromised in the event of any breach. The effectiveness of a DUA depends upon the data recipient's compliance; meaningful penalties for breach are difficult to enforce, especially in foreign jurisdictions.[131] To achieve a more forward-looking security solution, federal policymakers should put a high priority on the development of secure, central data enclaves where researchers can access and analyse genomic data without creating and downloading new copies of the data.[132] DbGaP, which currently

[130] See eg Djims Milius et al., *The International Cancer Genome Consortium's Evolving Data-Protection Policies*, 32 NAT. BIOTECHNOL. 519–523, 519–523 (2014). (describing the tightening of the ICGE access policy, and the importance of explaining to patients that 'promises of absolute privacy and data protection are unrealistic'.

[131] There is a reason to question whether DUAs between US providers and researchers in some foreign countries (eg China) offer any enforceable protections for data. See eg Dan Harris, *Why Suing Chinese Companies In The US Is Usually A Waste Of Time*, CHINA LAW BLOG (2009) http://www.chinalawblog.com/2009/09/why_suing_chinese_companies_in.html (accessed May 3, 2016). (Post by international law expert, noting that agreements written in English and applying US law are not enforceable in China.)

[132] See Robert H. Shelton, *Electronic Consent Channels: Preserving Patient Privacy Without Handcuffing Researchers*, 3 SCI. TRANSL. MED. 69cm4, 69cm4 (2011).

distributes copies of the genomic data it warehouses, would seem the obvious starting point for such a project.

Build reciprocity: help patients see and share the benefits of EMR research

In many aspects of their lives, people give up privacy in exchange for something—often ease and convenience in using, for example, credit cards, websites that require cookies, or automated systems for paying road or bridge tolls. In medical research, with few exceptions, participation is often for the promise of future societal versus individual benefit. Telling those whose data is part of research about the concrete outcomes of that research—and doing it in language that they can understand—is one small but important way to try to 'give back' to those whose data was used in research, and perhaps to build support for research more broadly.

Yet genomic data provides more than just grist for a researcher's mill. Some things can be learned that individual patients or subjects might find valuable, or at least interesting. The question of 'incidental findings' has been a controversial one in the world of research ethics, but in some contexts, accurate and useful information about incidental findings can confer real benefits on research participants. People may also find interesting some general information about their genetic backgrounds. For example, ancestry information is not always benign, but, particularly at a relatively high level of abstraction, it usually will be. Similarly, some trait or even disease risk susceptibility information might, if sufficiently accurate, be interesting or even useful to research participants. Often, to preserve privacy, researchers agree not to attempt direct contact with subjects, and are not provided any contact information—making return of individual results impracticable, if not impossible. Even so, researchers should think hard about safe and useful ways in which, in summary form, the genomic information they are analyzing might somehow provide a nice little 'thank you' gift, a 'lagniappe', to the people whose data made their research possible.

CONCLUSION

The promise of big data and the appetite of researchers for access to information are enormous, to the point where, in pursuit of new knowledge, we've all but abandoned participant consent in records-based research, relying instead upon various degrees of de-identification to satisfy ethical concerns and meet regulatory requirements. There is very little in the way of transparency in most records-based research: apart from blanket reassurances in the HIPAA privacy notice that 'your privacy is protected', providers don't offer patients specifics about who will receive what information. Nor is any disclosure to patients likely to convey the uncertainty that lies beneath any categorical statements about privacy protection in research.

We can debate whether this preemption of individual choice is defensible. In the era of EMRs, the new knowledge obtained from population-scale, records-based research is immensely valuable; it may seem unfair to allow patients to benefit from these research findings without sharing in the privacy risk of the research itself. Proponents of choice preemption argue that where risk is minimal and benefits substantial, we should

not allow dissenting patients to impose the response biases and process burdens that an opt-out would entail.[133]

We can even ask whether privacy and consent still matter, both as individual rights and as protections for human subjects, if de-identification strategies can effectively minimize the re-identification risk associated with a given set of EMR data. For many projects involving medical data this may be true. But when it comes to genomic research, scientists should not - and we believe, ethically cannot - promise anonymity, label gene sequences as de-identified information, or fail to tell patients, in specific terms, who is studying their EMR genomic data and where copies of those data reside.

Unrealistic and even deceptive promises of anonymity are all too common throughout the online world, where website privacy policies promise that their corporate sponsors collect only 'anonymous' user data, even as these same sites track and aggregate browsing habits to form highly detailed profiles of the shopping, reading, and religious preferences of individual consumers.

But while 'browser beware' may be the norm in the commercial internet space, patients expect, and have the *right* to expect, that their medical providers—and the researchers who use patient data that obtained from those providers—will hold themselves to a higher standard. For EMR-based genomic research, the starting point should be meaningful, *specific* notice to patients of all research uses and disclosures of genomic information. As technology permits, providers should also strive to offer patients some degree of control over discretionary uses such as research.

The standard 'notice and choice' model of privacy protection has limitations, especially in EMR research, where the future uses of patient data—and the future privacy risks—are unknown.[134] Patients might give permission, but they can never provide fully informed consent at the point of care for all future genomic research. We should insist upon other protections for clinical genomic data used in research, such as data security measures that are no less rigorous than the standard for electronically maintained clinical information.

Given how rapidly the landscape of re-identification risk is evolving in genomic research, neither IRBs nor researchers can predict future risk with confidence. Geneticist George Church, who heads his own genomic sequencing project, argues that we should simply admit there is no reliable, enduring technical solution to privacy, and then work to convince DNA donors that the consequences of a research privacy breach are acceptable.[135]

We disagree. Researchers, providers, and regulators can—indeed must—do more than aim to convince patients to accept the privacy risks of EMR-based genomic research as an inescapable cost of receiving medical care. From an ethical standpoint, there is little meaningful difference between a research subject asked to contribute her blood specimen for gene sequencing—and afforded the right to say 'no', a right reaching back to the Nuremburg Code and other foundational statements of research

[133] See Faden et al., *supra* note 113. See also Rosamond Rhodes, *Rethinking Research Ethics*, 10 Am. J. Bioethics 19–36, 19–36 (2010).

[134] See Helen Nissenbaum, *A Contextual Approach to Privacy Online*, 140 Daedalus 32–48, 32–48 (2011).

[135] See Eryn Brown, *Geneticist on DNA Privacy: Make It So People Don't Care*, Los Angeles Times, Jan. 18, 2013, http://articles.latimes.com/2013/jan/18/science/la-sci-sn-george-church-dna-genome-privacy-20130118 (accessed May 3, 2016).

ethics—and a patient whose genome is sequenced in the course of clinical care. Why isn't the patient entitled to know when her genome is shared with researchers? Why shouldn't she have a say in the matter?

Notice and a degree of control will produce one additional benefit: sunshine. If patients must be told which researchers, institutions, commercial entities, and federal research institutes receive their genomic data, patients can hold those recipients to account for data security, or even request that genomes be removed from research databases. This new scrutiny, though it might be uncomfortable at times, could actually prompt a greater level of patient engagement in genomic research. Researchers who can explain why EMR genomic research is valuable and how privacy is protected may find that patients, the ultimate beneficiaries, become vocal champions and enthusiastic participants. This paper is an effort both to point out the way the status quo impedes such a result and to describe a set of practices that are more likely to lead to it. We do not expect that we have said—or that anyone else will think we have said—the last word on this issue, but we hope we have opened, and moved forward, this crucial discussion.

ACKNOWLEDGEMENTS

This paper was funded by a grant from the Greenwall Foundation. The authors gratefully acknowledge the contributions of Debra Mathews, Ph.D., Michelle Meyer, J.D., Ph.D., and Mark Rothstein, J.D., each of whom commented on earlier drafts of the manuscript. The views expressed in the paper are our own and do not necessarily represent those of the reviewers or our employers.