Volume 16, Number 1 (2018)
Northwestern Journal of Technology and Intellectual Property

MOVING BEYOND CONSENT FOR CITIZEN SCIENCE IN BIG DATA HEALTH AND MEDICAL RESEARCH

Anne S.Y. Cheung[*]

ABSTRACT—Consent has been the cornerstone of the personal data privacy regime. This notion is premised on the liberal tenets of individual autonomy, freedom of choice, and rationality. The above concern is particularly pertinent to citizen science in health and medical research, in which the nature of research is often data intensive with serious implications for individual privacy and other interests. Although there is no standard definition for citizen science, it includes generally the gathering and volunteering of data by non-professionals, the participation of non-experts in analysis and scientific experimentation, and public input into research and projects. Consent from citizen scientists determines the responsibility and accountability of data users. Yet with the advancement of data mining and big data technologies, risks and harm of subsequent data use may not be known at the time of data collection. Progress of research often extends beyond the existing data. In other words, consent becomes problematic in citizen science in the big data era. The notion that one can fully specify the terms of participation through notice and consent has become a fallacy.

Is consent still valid? Should it still be one of the critical criteria in citizen science health and medical research which is collaborative and contributory by nature? With a focus on the issue of consent and privacy protection, this study analyzes not only the traditional informed consent model but also the alternative models. Facing the challenges that big data and citizen science pose to personal data protection and privacy, this article explores the legal, social, and ethical concerns behind the concept of consent. It argues that we need to move beyond the consent paradigm and take into account the much broader context of harm and risk assessment, focusing on the values behind consent – autonomy, fairness and propriety in the name of research.

I. INTRODUCTION

Consent has been the cornerstone of the personal data privacy regime.[1] It authorizes the collection, use, and processing of personal data. When it comes to health and medical research, consent is a prerequisite for the intervention in one's body, the collection of bio-specimens, and the use of personal data.[2] The doctrine of consent is premised on the liberal tenets of individual autonomy, dignity, and integrity, rooted in the fundamental respect to a person, and intertwined with the right to respect for privacy.[3] More importantly, consent is only meaningful if it is freely given (voluntary), specific, and informed.[4]

The above concern is particularly pertinent to citizen science in health and medical research, in which the nature of research is often data-intensive and has serious implications for an individual's privacy and other interests.[5] Although there is no standard definition for citizen science, the European Commission has highlighted its general features to be the gathering and volunteering of data by nonprofessionals and the participation of nonexperts in analysis and scientific experimentation, with public input into research and projects.[6] Citizens become experimenters, stakeholders, purveyors of data, research participants, or even partners in the process.[7] Consent from citizen scientists is indispensable as it is a constitutive element for participants' self-determination and self-empowerment. Furthermore, consent from participants as data subjects determines the responsibility and

---

* Anne S.Y. Cheung, Professor of Law, the University of Hong Kong. An earlier version of this article was presented at the "Disciplining or Empowering the Citizenry Through Citizen Science" Conference, organized by the Institutum Iurisprudentiae Academia Sinica, Taiwan in December 2016. The author benefits from the valuable comments and suggestions of the participants. The author is grateful for the research assistance of Jason C.P. So and Michael M.K. Cheung.

1 *See generally* ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, THE OECD PRIVACY FRAMEWORK (2013), http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf [http://perma.cc/U9BZ-KHSB].

2 MARCUS DÜWELL, BIOETHICS: METHODS, THEORIES, DOMAINS 193 (2013).

3 TOM L. BEAUCHAMP & JAMES F. CHILDRESS, PRINCIPLES OF BIOMEDICAL ETHICS 107 (7th ed., 2012).

4 Article 29 Data Protection Working Party, *Opinion 15/2011 on the Definition of Consent*, WP187, at 34–35 (July 13, 2011), http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2011/wp187_en.pdf [http://perma.cc/EJE8-9YGF].

5 *Opinion No. 29 of the European Group on Ethics in Science and New Technologies to the Europeon Comission*: *The Ethical Implications of New Health Technologies and Citizen Participation*, at 20 (Oct. 13, 2015) [hereinafter *Opinion No. 29 of the EGE*], http://ec.europa.eu/research/ege/pdf/opinion-29_ege.pdf [http://perma.cc/4VDE-J5MY].

6 *Id.* at 23.

7 *Id.* By citizens as "experimenters," this refers to patients participating in various degrees in experimentation. *Id.* at 25. "Stakeholders" refers to patient expert groups. *Id.* "Purveyors of data" refers to citizens or patients sending data through digital devices, mobile devices, and other information communication technology. *Id.*

accountability of data users. Under orthodox understanding, consent should be given on a one-on-one basis, for a "single study at a single institution for a specific purpose."[8]

While this may sound sensible and reasonable, technology always produces new directions and challenges for research. With advances in information technology and data analytics, health and medical research have become data-intensive, global, and virtual.[9] Biobanks and virtual research repositories are gaining prominence and significance.[10] At the same time, the risk inherent in health and medical research and big data technology has often extended beyond the existing data. In addition, the use and transfer of data for other unforeseen purposes is often outside the control of the original research team. Plus, the risks and harm of subsequent data use may not be known at the time of data collection. Hence, consent becomes problematic because the traditional understanding of consent, that one can fully specify the terms of agreement in advance, becomes questionable in big data science and citizen science.

We cannot help but ask: is consent still valid? Should it still be one of the critical criteria in citizen science health research which is collaborative and contributory by nature? While the big data challenge is not unique to citizen science, the inherent sensitivity of health and medical data exacerbates the problem, which calls for close scrutiny of the doctrine of consent. With a focus on citizen science in health and medical research, this study examines the doctrine of consent and its inadequacies. It then analyzes the alternative and adaptive models of consent: open, broad, dynamic, portable, and meta consent.[11] Facing the challenges that big data and citizen science pose to personal data protection and privacy, this article explores the legal, social, and ethical concerns behind consent. It argues that navigating one's way through different models of consent and the varied choices in consent forms can be a legal minefield. We need to move beyond the consent paradigm into a broader framework of accountability, taking into consideration harm and risk assessment. Ultimately, what lies behind consent are the entailing values of autonomy, fairness, and propriety in the name of research.

---

[8] Bridget M. Kuehn, *Groups Experiment with Digital Tools for Patient Consent*, 310 JAMA 678, 679 (2013).

[9] BETTINA SCHMIETOW, *Ethical Dimensions of Dynamic Consent in Data-Intense Biomedical Research—Paradigm Shift, or Red Herring?*, *in* ETHICS AND GOVERNANCE OF BIOMEDICAL RESEARCH, RESEARCH ETHICS FORUM 197 (Daniel Strech & Marcel Mertz eds., 2016).

[10] *See generally* Charles Auffray et al., *Making Sense of Big Data in Health Research: Towards an EU Action Plan*, 8 GENOME MED. (2016), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4919856/ [http://perma.cc/V45C-7EJ2].

[11] *See infra* Part IV.

II. CITIZEN SCIENCE AND BIG DATA HEALTH RESEARCH

We begin our discussion by looking at the nature of citizen science. In the mid-1990s, Alan Irwin popularized and prompted the discussion on citizen science.[12] Irwin examined the relationship between citizens and science, including science that assists citizens' needs and concerns and science developed and enacted by citizens themselves.[13] Irwin flagged up the need to face up to the challenges posed by risks in scientific research and sustainable development.[14]

By the 21st century, citizen science has developed into different forms of participation by nonprofessional scientists, exhibiting various dimensions in the cooperation between professionals and nonprofessionals and opening up multiple levels of engagement in the health sector. At one end of the spectrum, there are citizen-led, patient-owned initiatives of sharing quantitative information, exchanging experiences on treatment, and searching for the right clinical trials on online platforms. Prominent examples include PatientsLikeMe,[15] CureLauncher[16] and CureTogether.[17] At the other end of the spectrum, there is commercial or government-led research. For instance, Sage Bionetworks took advantage of smartphone-based health technology to study the lifestyle of 17,000 Parkinson's disease patients.[18] It also paired up with Apple ResearchKit in 2015 to study the quality of life of breast cancer survivors.[19] In the same year, President Obama announced the nationwide $215 million Precision Medicine Initiative to build a large-scale research enterprise between public and private sectors, calling for one million volunteers to contribute their health data so as to extend precision medicine to all diseases.[20] In between the two models, there

---

[12] Barbara Prainsack, *Understanding Participation: The "Citizen Science" of Genetics*, *in* GENETICS AS SOCIAL PRACTICE: TRANSDISCIPLINARY VIEWS ON SCIENCE AND CULTURE 147 (Barbara Prainsack et al. eds., 2014).

[13] ALAN IRWIN, CITIZEN SCIENCE: A STUDY OF PEOPLE, EXPERTISE AND SUSTAINABLE DEVELOPMENT xi (1995).

[14] *Id.* at x. Irwin's case study was mainly on environmental development. *Id.* at xii.

[15] PATIENTSLIKEME, https://www.patientslikeme.com (last visited May 9, 2018).

[16] CURELAUNCHER, http://curelauncher.com [https://web.archive.org/web/20180103112551/http://www.curelauncher.com/] (last visited Aug. 8, 2017).

[17] CURETOGETHER, http://curetogether.com (last visited Aug. 8, 2017).

[18] Max Little, *Crowdsourced Parkinson's Research: Engaging People, Opening Up Science*, THE GUARDIAN (Apr. 7, 2014, 8:00 AM), https://www.theguardian.com/media-network/media-network-blog/2014/apr/07/parkinsons-disease-research-science-health.

[19] Aditi Pai, *Apple's ResearchKit Now Available to Medical Researchers*, MOBIHEALTHNEWS (Apr. 14, 2015), http://www.mobihealthnews.com/42370/apples-researchkit-now-available-to-medical-researchers.

[20] The White House: Office of the Press Secretary, *Fact Sheet: President Obama's Precision Medicine Initiative* (Jan. 30, 2015), https://obamawhitehouse.archives.gov/the-press-

is a third catering to joint collaboration between citizens and health professionals in creating knowledge. A prominent example includes the Sarroch Bioteca Foundation founded in 2012 in pursuit of a "citizen veillance on health" project in Italy.[21] The project was launched by the Sarroch municipality in 2006 to gather biological samples donated by citizens to monitor genetic changes as health indicators in relation to the environment.[22] All citizens of the municipality could become members of the project.[23] The aim was to use science to inform both health regulations and institutional implementation policy.[24] The Sarroch example can be seen as a joint effort for collective governance and a model for democratic health choice.[25]

Regardless of the level of citizen involvement, participation issues related to the data privacy and security of research subjects (or the form of cooperation in the above models as contractual, contributory, collaborative, co-creative, or collegial[26]) will be triggered whenever citizens have contributed their data or bio-specimens to the projects. Following the new wave of citizen science research in big data is a whole new set of legal and ethical concerns. First, we are witnessing an unprecedented scale of online crowdsourcing, with researchers pooling data together using big data capture strategies and data analytics.[27] As the progression of research often extends beyond the existing data, big data technology use and the transfer of data for other unforeseen purposes is often outside the control of the original research

office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative [https://perma.cc/TSG8-D7NA].

[21] Mariachiara Tallacchini, Philip Boucher & Susana Nascimento, European Comm'n, Joint Res. Center Sci. and Policy Reports, *Emerging ICT for Citizens' Veillance*, at 30, EUR 26809 EN (2014), http://publications.jrc.ec.europa.eu/repository/bitstream/JRC90334/civ%20-%20final%20draft.pdf [https://perma.cc/L2UD-VAWH].

[22] *Opinion No. 29 of the EGE*, *supra* note 5, at 28.

[23] *Id.*

[24] *Id.*

[25] *Id.*

[26] Jennifer L. Shirk et al., *Public Participation in Scientific Research: A Framework for Deliberate Design*, 17 ECOLOGY & SOC. 29 (2012), http://www.ecologyandsociety.org/vol17/iss2/art29/ (last visited Mar. 27, 2018). The typology of the five project models is formulated by Shirk and her colleagues. Contractual projects refer to communities asking professional researchers to conduct a specific investigation and report. *Id.* Contributory projects refer to those designed by scientists, with citizens contributing data. *Id.* Collaborative projects are similar to contributory project except with citizens helping to refine project design, analyze data, or disseminate findings. *Id.* Co-created projects are jointly designed by scientists and citizens, in which some citizens are actively involved in most or all aspects of the research process. *Id.* Lastly, collegial contributions refer to non-credential individuals conducting research independently with varying degrees of expected recognition by professional scientists. *Id.*

[27] Mark A. Rothstein, John T. Wilbanks & Kyle B. Brothers, *Citizen Science on Your Smartphone: An ELSI Research Agenda*, 43 J.L. MED. & ETHICS 897, 897 (2015).

team.[28] Second, despite the promise of data de-identification, third parties can match data sets to reidentify individuals.[29] Third, with the advancement of data mining and big data technologies, the risks and harm associated with subsequent data use may not be known at the time of data collection and use. For instance, large-scale harvesting of health data can reveal unnoticed correlations between lifestyle and medical conditions of individuals. These correlations are important information for insurance companies.[30] The fear is that insurance companies may use big data analytics to draw conclusions on consumers' health care use and thus increase premiums in unprecedented ways.[31] In other words, consent becomes problematic in health and medical research, especially in the big data era. The notion that one can fully disclose and specify the terms of notice and consent at the outset has become illusory.

III. RETHINKING CONSENT

Consent has been a cardinal doctrine in clinical treatment and research. It is premised on the respect for individual autonomy, which embodies the principle of self-rule that is free from "controlling interference by others and limitations that prevent meaningful choice." [32] It is enshrined in numerous international treaties, legal guidelines, and codes.[33] Namely, on the protection of human rights, the International Covenant on Civil and Political Rights (ICCPR) stipulates that "free consent" is a prerequisite for medical and scientific experimentation.[34] On personal data, the European Union's General Data Protection Regulations (GDPR) state that "explicit consent" is

---

[28] *See* Bartha Maria Knoppers & Adrian Mark Thorogood, *Ethics and Big Data in Health*, 4 CURRENT OPINION IN SYSTEMS BIOLOGY 53, 54 (2017); UNESCO, *Report of the IBC on Big Data and Health*, at 11–12 (2017), http://unesdoc.unesco.org/images/0024/002487/248724E.pdf.

[29] Lataya Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, 25(2&3) J.L. MED. & ETHICS 98, 98 (1997).

[30] Aaron Stanley, *Tech Companies See Market Opportunity in Healthcare Innovation*, FINANCIAL TIMES (May 5, 2015), https://www.ft.com/content/709aa784-efd4-11e4-ab73-00144feab7de.

[31] With big data analytics, insurance companies can collect large amounts of personalized data about individuals for evaluating individual habits and lifestyles to create new predictive and risk models. Cathy O'Neil, *Big Data Is Coming to Take Your Health Insurance*, BLOOMBERG (Aug. 4, 2017), https://www.bloomberg.com/view/articles/2017-08-04/big-data-is-coming-to-take-your-health-insurance. Different jurisdictions have laws that prohibit the use of genetic data in health insurance. Knoppers & Thorogood, *supra* note 28, at 53–54.

[32] BEAUCHAMP & CHILDRESS, *supra* note 3, at 101.

[33] For an historical overview, see Benjamin M. Meier, *International Protection of Persons Undergoing Medical Experimentation: Protecting the Right of Informed Consent*, 20 BERKELEY J. INT'L. L. 513 514–33 (2002). For further comparison of global guidelines on consent and informed consent, see Zulfiqar A. Bhutta, *Beyond Informed Consent*, 82 BULL. WHO 771, 771–77 (2004).

[34] International Covenant on Civil and Political Rights art. 7, *opened for signature* Dec. 16, 1966, 999 U.N.T.S. 171 (entered into force Mar. 23, 1976) ("No one shall be subjected to torture or to cruel, inhuman or degrading treatment or punishment. In particular, no one shall be subjected without his free consent to medical or scientific experimentation.").

necessary for the processing of genetic, biometric, and health data.[35] Under article 4 of the GDPR, consent means "any freely given, specific, informed and unambiguous indication of the data subject" by a clear affirmative action signifying agreement to the processing of personal data. On experiments done by physicians, the Nuremberg Code sets the standards to which physicians must conform when carrying out experiments on human subjects, including obtaining consent and ascertaining competence from human subjects in experiments.[36] On medical research, the World Medical Association's Declaration of Helsinki calls for "informed consent, preferably in writing" from physicians.[37] Again, on biomedicine research, the Convention on Human Rights and Biomedicine calls for "free and informed" consent.[38] The *International Ethical Guidelines for Biomedical Research Involving Human Subjects* insists that investigators must obtain the "voluntary informed consent of the prospective subject. . . . Waiver of informed consent is to be regarded as uncommon and exceptional, and must in all cases be approved by an ethical review committee."[39] On research involving human subjects in general, the U.S. Code of Federal Regulations specifies that informed consent from participants in research must involve discussion of the nature of the involved procedure, its risks and benefits, and alternative treatments available.[40] Finally, participants must also give free assent.[41]

Regardless of whether it is free, explicit, informed, and voluntary, the four essential elements of valid consent are comprehension or understanding,

---

[35] Commission Regulation 2016/679, art. 9, 2016 O.J. (L 119) 38 [hereinafter GDPR], http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN. In addition to article 9, article 7 requires that the request for data subjects' consent must be clearly distinguishable, intelligible, easily accessible, and expressed in clear and plain language. *See id.* at art. 7. The GDPR came into force on May 25, 2018. *See id.* at art. 99.

[36] 2 TRIALS OF WAR CRIMINALS BEFORE THE NUERNBERG MILITARY TRIBUNALS UNDER CONTROL COUNCIL LAW NO. 10, at 181–82 (1949) (stating that "[t]he voluntary consent of the human subject is absolutely essential . . . [and includes] legal capacity . . . free power of choice . . . sufficient knowledge and comprehension of the [nature, duration, and purpose of the experiment] . . . to make an understanding and enlightened decision.").

[37] *WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects*, World Med. Ass'n (Mar. 19, 2018), https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/. The declaration was promulgated in 1964 and revised nine times since. *Id.*

[38] Convention for the Protection of Human Rights and Dignity of the Human Being with Regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine art. 5, *opened for signature* Apr. 4, 1997, E.T.S. No. 164 (entered into force Jan. 12, 1999).

[39] COUNCIL FOR INT'L ORG. OF MED. SCI., INTERNATIONAL ETHICAL GUIDELINES FOR BIOMEDICAL RESEARCH INVOLVING HUMAN SUBJECTS 22 (2002).

[40] 45 C.F.R. § 46.116 (2017).

[41] *Id.*

voluntary participation, competence, and disclosure.[42] While the first three refer to the duty of doctors or researchers to obtain the voluntary agreement of human subjects before participation, the last one refers to their duty to disclose adequate information to the subjects.[43] Both limbs are integrated requirements of consent as a single legal and moral doctrine.[44] Seemingly, the above frameworks in international and national law, codes, and guidelines have provided the necessary and sufficient legal basis for informed consent and for the use of one's data or bio-specimens. But how detailed descriptions of the research should be and how much disclosure would be required as adequate remains controversial.

In the context of citizen science in health and medical research, participants must face the additional uncertainty and unpredictability of research progress. For example, in the 1980s, a group of Canavan disease-affected families developed a disease registry and tissue bank to encourage research in the area.[45] They provided tissue for research on the disease and aided in the identification of other affected families.[46] With three nonprofit organizations, they developed a confidential database and Canavan disease registry, attracting financial sponsorship.[47] However, when one of the chosen physician-researchers decided to isolate and patent the Canavan gene sequence and develop genetic screening tests for it, the families sued the researcher and his institution.[48] The bitter legal battle ended only in half-victory. The U.S. District Court for the Southern District of Florida dismissed several of the plaintiffs' claims, including lack of informed consent, breach of fiduciary duty, fraudulent concealment of the patent, and misappropriation of trade secrets.[49] Nevertheless, the court upheld the claim of unjust enrichment made by the tissue donors on the grounds that "the facts

[42] Gopal Sreenivasan, *Does Informed Consent to Research Require Comprehension?*, 362 LANCET 2016, 2016 (2003).

[43] 45 C.F.R. § 46.116.

[44] *Id.*

[45] *About Canavan Disease*, CANAVAN FOUNDATION, http://www.canavanfoundation.org/about_canavan_disease (last visited Aug. 8, 2017). Canavan disease is a progressive, fatal neurological disorder that begins in infancy. *Id.* It is caused by an inherited genetic abnormality resulting in improper transmission of nerve signals. *Id.*; *see* Barbara J. Evans, *Barbarians at the Gate: Consumer-Driven Health Data Commons and the Transformation of Citizen Science*, 4 Am. J.L. & Med. 651, 652 (2016) (discussing the dispute).

[46] Greenberg v. Miami Children's Hosp. Research Inst., Inc., 264 F. Supp. 2d 1064, 1067 (S.D. Fla. 2003).

[47] *Id.*

[48] *Id.* at 1067–68.

[49] *Id.* at 1077.

paint a picture of a continued research collaboration that involved Plaintiffs also investing time and significant resources."[50]

In more recent times, another notorious example is the 23andMe project. It became known in late 2007 as a company offering genetic testing at a very low price of $299, giving out saliva collection kits and asking for saliva samples.[51] The testing was perceived as a "fun way" to learn about one's genetics.[52] In 2008, 23andMe added a new research feature named 23andWe which played up citizen science rhetoric and community-driven research, emphasizing strong participatory features.[53] It invited customers and participants to vote on of a list of diseases which the company promised it would then prioritize in its research.[54] In exchange, the customers and participants were asked to disclose details about their lifestyles and other relevant information for research purposes.[55] By that time, the company had offered free saliva collection kits to people who had been diagnosed with the types of diseases that the company wanted to focus on and research.[56] Additionally, the company lowered the price of saliva collection kit to $99.[57] By 2012, 23andMe had about 150,000 users.[58] The business was operating as a commercial company drawing heavily on the contributions of citizen science participants .

Some patients and members of patient support groups joined 23andMe under the impression that they were contributing their genetic and personal data for the development of treatment and long-term research.[59] However, they soon woke up to reality when the company filed a number of patent applications in 2012.[60] People realized that 23andMe was sharing aggregate data about its customers and participants with third parties, and that Google

---

[50] *Id.* at 1072–73.

[51] Charles Seife, *23andMe Is Terrifying, but Not for the Reasons the FDA Thinks*, SCI. AM. (Nov. 27, 2013), https://www.scientificamerican.com/article/23andme-is-terrifying-but-not-for-the-reasons-the-fda-thinks/ [https://perma.cc/W6KT-W26P].

[52] *Id*.

[53] Prainsack, *supra* note 12, at 156; *see also* Dan Vorhaus, *Genomic Research Goes DTC*, ROBINSON BRADSHAW HINSON GENOMICS L. REP. (July 9, 2009), https://www.genomicslawreport.com/index.php/2009/07/09/genomic-research-goes-dtc/ [https://perma.cc/22F5-4SYD].

[54] Prainsack, *supra* note 12, at 156.

[55] *Id*.

[56] *Id*.

[57] Seife, *supra* note 51.

[58] Prainsack, *supra* note 12, at 156.

[59] Katherine Drabiak, *Caveat Emptor: How the Intersection of Big Data and Consumer Genomics Exponentially Increases Informational Privacy Risks*, 27 HEALTH MATRIX 143, 154 (2017).

[60] Prainsack, *supra* note 12, at 156.

had invested in the company.[61] Consequently, 23andMe was severely criticized, but defended itself by arguing it had informed the customers and participants all along in its Terms of Service and consent forms.[62] Although 23andMe's actions were technically lawful, Barbara Prainsack pointed out that it was dishonest and immoral for the company to capitalize on the "free labour" and data capital of its customers, patients, and participants for profit under the grand name of research.[63] In particular, Prainsack observed that 23andMe's business model was "continually evolving."[64] It was highly unlikely that participants could keep up with the frequent modifications to the Terms of Service and the fast-changing, constantly-updated terms in small print on the website.[65]

These two incidents illustrate that there is a misalignment of orientations between citizen science participants and expert researchers. The former group was motivated by a genuine commitment to facilitate disease research, to contribute to health knowledge, and to create collective benefits. In contrast, the latter was motivated by profit and individual or corporate success. Yet, this mismatch might not be present at the research project's outset. Rather, it is due to the fluid and flexible nature of health and medical research and citizen science that the projects soon spin out of control of citizen science participants and evolve beyond their own meanings.

Additionally, expert researchers' motivations can gradually grow apart from the motivations of their research subjects. The relation between expert researchers and participants is not on a traditional one-to-one model. Instead, it rests on an elaborate network backed by complex organizational structures and staffed by different experts at various levels. Citizen scientists or participants, motivated by altruism to share their personal data, can mistakenly think they can retain some form of control in a collaborative or cooperative manner.[66] Their solidarity, sadly, is later exploited by researchers or commercial groups in both public and private spheres.

Furthermore, if informed consent requires disclosure by researchers and comprehension by participants, then full disclosure of information will become neither definable nor achievable at the outset of the research due to the fast-changing nature of research. Writing on clinical treatment and research, Onora O'Neill has remarked on the inherent deficiency of informed

---

[61] In fact, the founder of the company, Anne Wojcicki, was the wife of Google boss Sergey Brin at that time. *Id.* at 158; Seife, *supra* note 51.

[62] Prainsack, *supra* note 12, at 156.

[63] *Id.* at 156–57.

[64] *Id.* at 157.

[65] *Id.* at 156–57.

[66] *Opinion No. 29 of the EGE*, *supra* note 5, at 52.

consent as a doctrine.[67] She explains that this is not due to any procedural deficiencies ensuring that informed consent has been fulfilled, but rather that consent is a "propositional attitude."[68] It is a "*description* of a *proposal*" for treatment or research.[69] One can only consent to the specific descriptions of a proposition but may not be aware of the foreseeable consequences.

Does O'Neill's conclusion mean that informed consent is no longer valid? Alternatively, does it suggest that it is high time for an urgent refinement of the requirement of consent? Brent Mittelstadt and Luciano Floridi have argued that the traditional framework on informed consent "does not cleanly transfer" to research involving biomedical big data.[70] They point out this is because the doctrine of informed consent is formulated for single, specific research or treatment but not for the sharing, aggregating, or repurposing of data that may reveal unforeseen information.[71] As a result, until we have found a satisfactory alternative model, the pressing concerns on obtaining informed consent for citizen science research in health data remain: deciding why the data are collected and how long the data will be kept, identifying who is permitted to have access to the data and who is processing the data for what purposes, and determining what to do in case the data are misused.[72]

IV. RENEGOTIATING CONSENT

To tackle the above, researchers in this area have formulated different models of informed consent. Here, we evaluate the common forms of open, broad, dynamic, portable, and meta consent.

A. Open Consent

In light of the inherent uncertainty and unpredictability of big data health research, some scholars have advocated for veracity or "radical honesty"[73] in the model of "open consent," which deliberately excludes any promises about privacy and requires participants to demonstrate

---

[67] ONORA O'NEILL, AUTONOMY AND TRUST IN BIOETHICS 42–43 (2002).

[68] *Id.* at 43.

[69] *Id*.

[70] Brent D. Mittelstadt & Luciano Floridi, *The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts*, 22 SCI. ENG. ETHICS 303, 311 (2016).

[71] *Id.* at 312.

[72] *Opinion No. 29 of the EGE*, *supra* note 5, at 51–54.

[73] JOHN T. WILBANKS, *Portable Approaches to Informed Consent and Open Data*, *in* PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT 234, 234–35 (Julia Lane et al. eds., 2014).

comprehension of the nature of the research and the risks involved prior to enrollment.[74]

Open consent has been used in the famous Personal Genome Project (PGP) by Harvard University since 2005.[75] The aim of the project is to test DNA sequencing technologies on human subjects by building a database of human genomes and traits, with the ambition to be a global network project.[76] The nature of the database is open source, open access, participatory, and collaborative. The target is to collect the genomes of 100,000 individuals and to make the information public with no serious effort at de-identification.[77] Since DNA is the ultimate digital identifier of an individual but de-identification of samples would impoverish the data,[78] the PGP research team has decided to be forthright and honest with the participants, aiming for them to be "truly informed" about the nature of the research.[79] Participation of the public is encouraged and volunteers are asked to give open consent to ensure that they understand the scientific nature of the experiment and that they also understand that privacy and confidentiality cannot be guaranteed.[80]

Misha Angrist, one of the original ten participants of the PGP back in 2006, shared his experience and reflections.[81] According to him, participants had to first go through an eligibility screening process which included filling out a questionnaire regarding family circumstances and privacy preferences.[82] Second, they would review a study guide that covered the potential risks of participating.[83] Third, they then took an "entrance exam" that covered the areas of how PGP worked, knowledge of genetics, ethical principles governing human subjects research, and their comfort level with having their genome and health records in the public domain.[84] They had to score 100% on the exam before they could be enrolled in the project.[85] Finally, they had to sign a consent form, which stated the possibility of re-

---

[74] Madeleine P. Ball et al., *Harvard Personal Genome Project: Lessons from Participatory Public Research*, 6 GENOME MED. 10 (2014). Other examples of open consent model include the Omics project, The Human Microbiome Project, and the American Gut Project. *See Opinion No. 29 of the EGE*, *supra* note 5, at 14–16.

[75] THE PERSONAL GENOME PROJECT, http://www.personalgenomes.org/ (last visited May 9, 2018).

[76] At the time of writing, UK, Canada, and Austria have joined the network. *Id.*

[77] Misha Angrist, *Eyes Wide Open: The Personal Genome Project, Citizen Science and Veracity in Informed Consent*, 6 PERSONALIZED MED. 691, 694 (2009).

[78] *Id.* at 693.

[79] *Id.* at 693–94.

[80] *Id.* at 694–95.

[81] *Id.* at 695.

[82] *Id.* at 694.

[83] *Id.*

[84] *Id.*

[85] *Id.* at 695.

identification, disclosure of non-paternity, and loss of insurance, as well as other risks of embarrassment, discrimination, data loss, or any unforeseen problems.[86]

Participants were described as "co-drivers" of the project.[87] They were expected to have solid knowledge about the field. Take Angrist as an example; he himself is a scientist who worked for an established institute for genome science and policy.[88] He completed an early version of the entrance exam and suggested changes to certain questions, and questioned the rationale for specific analyses.[89] He was also one of the three initial ten participants who served on the PGP Board of Directors.[90] He was careful enough to carry out a certain test on his genotype and make sure the result was negative before deciding to make his cell line available to the public.[91] However, it is doubtful how many other citizen scientists or participants could achieve such thorough understanding of the research and its implications to privacy.

B. Broad Consent

Rather than asking participants to take a leap of faith into uncertainty, a slightly refined model of broad consent has been proposed. While one gives consent to a framework for future research of certain types, ethical review of each specific research project by an independent ethics committee is required.[92] In addition, researchers must provide strategies on how to regularly update the participants and how to enable ongoing withdrawal opportunities for the participants.[93] Examples of broad consent model research are the UK Biobank project and the Norwegian Mother Child Cohort Study.[94] Nevertheless, regular updates to participants in ongoing research are seen as "extras" in this model.[95] Thus, legal and ethical concerns have been raised as to whether broad consent is a form of genuine informed

---

[86] *Id.*

[87] *Id.* at 693.

[88] *Id.* at 691; *see also* FACULTY: MISHA ANGRIST, https://sanford.duke.edu/people/faculty/angrist-misha (last visited Apr. 13, 2018).

[89] Angrist, *supra* note 77, at 693.

[90] *Id.*

[91] *Id.* at 695–96.

[92] Kristin Solum Steinsbekk et al., *Broad Consent Versus Dynamic Consent in Biobank Research: Is Passive Participation an Ethical Problem?*, 21 EUR. J. HUM. GENETICS 897, 897–900 (2013).

[93] *Id.* at 897.

[94] Jane Kaye et al., *Dynamic Consent: A Patient Interface for Twenty-first Century Research Networks*, 23 EUR. J. HUM. GENETICS 142, 142 (2015).

[95] *Id.*

consent when participants are reduced to passive subjects rather than research partners.[96]

C. Dynamic Consent

The alternative model of dynamic consent approaches consent from a unique perspective. It is a model tailor-made to the need of participants by utilizing an online interface and information technology-based platform. Information about the specific use of personal data and tissue as well as requests for consent for such use are put to the participants through the online platform.[97] Participants are allowed to engage in an interactive personalized interface as much or as little as they choose and to alter their consent choices in real time.[98] Consent is seen as a process, an ongoing interaction between researchers and participants. Hence, consent becomes dynamic because it allows participants to interact with the researchers over time, to consent to new projects, and to alter their consent choices in light of any new circumstances. This model was first designed for the EnCoRe project of three biobanks in Oxford from 2008 to 2012.[99] Another example is the Registries for All (Reg4All) project run by Genetic Alliance in partnership with the technology company Private Access.[100] Reg4All allows participants to decide how their data are being used and shared with particular researchers, institutions, or people studying a specific disease.[101] Participants can track who has used their data and how.[102]

Scholars praise the model of dynamic consent as providing a "personalised communication interface for interacting with patients, participants and citizens,"[103] implementing engagement 2.0 in the era of Web 2.0.[104] Apparently, it enables consent to be given to multiple researchers and

---

[96] Harriet JA Teare et al., *Towards 'Engagement 2.0': Insights from a Study of Dynamic Consent with Biobank Participants*, DIG. HEALTH 1, 2 (2015), http://journals.sagepub.com/doi/pdf/10.1177/2055207615605644.

[97] Kaye et al., *supra* note 94, at 145.

[98] *Id.* at 142.

[99] *Id.* at 145; Marco C. Mont et al., *EnCoRe: Dynamic Consent, Policy Enforcement and Accountable Information Sharing Within and Across Organisations*, HEWLETT-PACKARD DEV. COMPANY, L.P. 1, 1–4 (2012), http://www.hpl.hp.com/techreports/2012/HPL-2012-36.pdf.

[100] Debra J. H. Mathews & Leila Jamal, *Revisiting Respect for Persons in Genomic Research*, 5 GENES 1, 7–8 (2014), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3978508/pdf/genes-05-00001.pdf; Courtney Humphries, *New Disease Registry Gives Patients Some Privacy*, MIT TECH. REV. (Mar. 14, 2013), https://www.technologyreview.com/s/512456/new-disease-registry-gives-patients-some-privacy.

[101] *Id.*

[102] *Id.*

[103] Kaye et al., *supra* note 94, at 141.

[104] Teare et al., *supra* note 96, at 1.

projects, to open-ended and ongoing research, and to the use of secondary research or downstreaming of data use. Besides, dynamic consent overcomes the problem of locked-in consent confined to one experimental procedure for granting autonomy, choice, and control to individuals. At the same time, researchers can also manage the necessity to re-contact and to seek re-consent from participants much more easily.

Understandably, dynamic consent as a participant-centric initiative has its special appeal. Refining the model of dynamic consent, there are further variations on it.

D. Portable Legal Consent (PLC)

Another model is "portable legal consent" proposed by John Wilbanks of Sage Bionetworks.[105] This model recognizes that individuals have rights with respect to the data generated from their bodies.[106] They therefore will decide the kind of data that they would like to donate and share. For example, in the Sage Bionetworks, the suggested five categories of data are genetic sequence, clinical information, medical record, patient reported outcomes, and personal sensor data.[107] Consent is not tied to any particular study but carried around by the participants like organ donation status.[108] In that sense, consent becomes portable and controllable. Obtaining consent is done through an online interactive consent system.[109] Participants can share their own data broadly in the public domain to serve scientific research regardless of the particular institution involved.[110] In turn, the database of genomic information being collected through portable legal consent will be available to anyone who agrees to its terms.[111] These include a guarantee not to use the data to harm anyone or to identify the participants.[112] Users also agree to publish their work based on an open-access policy.[113]

---

[105] WILBANKS, *supra* note 73, at 245.

[106] *Id*.

[107] *Id.* at 249.

[108] Kuehn, *supra* note 8, at 679.

[109] WILBANKS, *supra* note 73, at 246.

[110] *Id*. at 245.

[111] *Synapse Terms and Conditions of Use: Summary of Key Provisions*, SAGE BIONETWORKS (Oct. 26, 2015), https://s3.amazonaws.com/static.synapse.org/governance/SageBionetworksSynapseTermsandConditionsofUse.pdf?v=4. Synapse operates data governance for Sage Bionetworks. SYNAPSE, https://www.synapse.org/ (last visited Apr. 13, 2018).

[112] *Id*.

[113] *Id*.

E. Meta Consent

Rather than focusing on the distinct categories of personal data, the meta consent model allows participants to express a preference for how and when to provide consent at a meta level, i.e. how and when they would like to be presented with a request for consent to the use of their personal health data and biological material.[114] In this model proposed by Ploug and Holm, participants must be provided with a predefined set of types of consent,[115] data,[116] and research contexts to choose from.[117]

While acknowledging similarities with dynamic consent, Ploug and Holm argue that meta consent is different in that dynamic consent was originally designed for biobanks.[118] In contrast, meta consent is developed with the aim to handle and configure consent preferences for the entire population for all kinds of data and biological samples, with a vision that every citizen is a potential participant in big data research—especially in medical research.[119] The meta consent model is designed "to provide a definitive answer by letting individuals design future consent requests on the basis of predefined types of consent, data, and contexts."[120]

V. LIMITATIONS OF CONSENT

Regardless of which variation or refinement of consent one chooses, problems remain. Other than the fact that the nature of research and open consent require an advanced level of comprehension from citizen science participants, open consent is far from *true* consent. First, open consent does not allow participants to act meaningfully on their continuing interest in their own health data. It does not include recontact of subjects, the subject's right to withdrawal, and the setting of time limits on the use of data.[121] Further, it does not have restrictions on how information or materials are going to be

---

[114] Thomas Ploug & Søren Holm, *Meta Consent – A Flexible Solution to the Problem of Secondary Use of Health Data*, 30 BIOETHICS 721, 724 (2016).

[115] *Id.* at 725–26. Ploug and Holm mention three types of consent. The first is specific consent referring to consent request for each new specific project using data, but not for each and every use of data. *Id.* at 725. The second is broad consent for "broader categories of research." *Id.* The third type is blanket consent and blanket refusal for one-off decisions concerning participation or non-participation in research. *Id.* at 725–26.

[116] *Id.* at 726. This includes data from electronic patient records, "[t]issue/[g]enomic data," health databases, and linkage to non-health data. *Id.*

[117] *Id.* Research context refers to private versus public, commercial versus non-commercial, and national versus international. *Id.*

[118] *Id.* at 732 n.29.

[119] *Id.*

[120] *Id.*

[121] Timothy Caulfield et al., *DNA Databanks and Consent: A Suggested Policy Option Involving an Authorization Model*, 4 BMC MED. ETHICS 1, 3 (2003).

shared to third parties, which is a potential cause for concern especially if information is used later for commercial purpose.[122] Besides, the way that it operates does not take prevention of harm, such as discrimination and other problems, into account. Altogether, so-called veracity has become an excuse to absolve researchers from accountability and responsibility. Although open consent may be legally valid, its practice remains ethically vague and questionable.

In a similar vein, the alternative forms of consent are in essence information governance models, which are useful only for well-informed, engaged, and e-health literate participants.[123] Concerns of digital divide and social exclusion have yet to be addressed in the dynamic consent model.[124] Participants will be asked for consent continuously because each new project requires fresh consent to be given. Arguably a person may potentially receive hundreds of consent requests each year.[125] This is likely to cause routinization of consent behavior, resulting in people not reading the information and not reflecting on the choice, but simply choosing habitually to consent or refuse to consent.[126] Although the refined model of meta consent allows opt-out or broad consent for future use of data, there is skepticism over whether meta consent is considered a form of valid informed consent under the new European Union regime of GDPR, which requires "explicit consent" for each and every data processing.[127] The common thread that runs through the various models is that control has seemingly been passed to individual participants. Yet at the same time, responsibility has also shifted to them without ensuring that they have the required knowledge and competence to

---

[122] *See* Henry T. Greely, *Informed Consent and Other Ethical Issues in Human Population Genetics*, 35 ANNU. REV. OF GENETICS 785 (2001).

[123] *See* WILBANKS, *supra* note 73.

[124] Steinsbekk et al., *supra* note 92, at 898–99.

[125] Ploug & Holm, *supra* note 114, at 723. Ploug and Holm discuss the hypothetical experience of a resident of Denmark if specific consent is required for every secondary use of data. *Id.*

[126] *Id.* at 723–24.

[127] *See supra* Part III. Article 4 of the GDPR has strengthened the requirements for consent to be freely given, specific, informed, unambiguous, clear, and affirmative. Gauthier Chassang, *The Impact of the EU General Data Protection Regulation on Scientific Research*, 11 ECANCERMEDICALSCIENCE (Jan. 3, 2017), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5243137/. Recital 33 allows a certain degree of flexibility for scientific research. *Id.* Data subjects are allowed to "give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose." *Id.* Hence, scientific research projects can only include personal data on the basis of consent if the purpose is clearly defined and well-described. Where purposes are unclear at the start of a scientific research program, stricter control and scrutiny may apply. Chassang points out that personal data impact assessment or compatibility tests should be carried out for data use in the context of research. *Id.*

make informed decisions.[128] Steinsbekk et al. have even argued that based on the above point, broad consent may be better than dynamic consent, as at least independent review from a research ethics committee is required.[129]

Overall, the above consent models may have enabled better participation of participants, but they remain largely information strategies. The mere passing of more information to participants, and the seeking of their indication at different stages of research, does not necessarily amount to building a democratic and participatory model of health and medical research. Steinsbekk et al. point out that the participation envisaged is limited as it is "participation *inside* an already established research arena where only minor changes of policy are up for discussion."[130] Furthermore, at best, we have filled only part of the knowledge gap (mentioned at the end of Part III of this article) in enabling participants to find out more about the purposes of data collection, the persons who have been accessing their data, and empowering those participants to have more control on how their data are being used down the stream of data reuse. The success of the remaining alternative models is highly dependent on how informed, competent, knowledgeable, and reflective the participants are. However, we have not addressed the nightmare scenario of what to do in case things go wrong.

VI. BEYOND CONSENT: THE MODEL OF ACCOUNTABILITY

Indisputably, consent plays an important role in health and medical research, but mere information disclosure and seeking participants' indication of choices do not necessarily guarantee the respect and self-determination of individuals. At most, the above-suggested alternative consent frameworks have fulfilled the contractual ritual required by law.[131] They may have "managed" the legal concerns,[132] but they have not resolved the problems of risks and harms not mentioned in the terms of agreement. Rather than shifting across different modes of consent and putting participants through a strenuous exercise of choices and forms, a consent model should be complemented with an accountability model.

Big data technology has opened up undreamed-of capacities to gain a sophisticated understanding about the way we can process and use data to organize our society and our lives. Those insights, unfortunately, can be pitfalls at the same time. Governments in different jurisdictions are eager to

---

[128] Steinsbekk et al., *supra* note 92, at 900.

[129] *Id*.

[130] *Id*.

[131] Ulrike Felt et al., *Refusing the Information Paradigm: Informed Consent, Medical Research, and Patient Participation*, 13(1) HEALTH 87, 101 (2009).

[132] *Id*. at 102.

capture the benefits of big data but also to weed out its harms.[133] In the present discussion, the medical sectors have used big data to monitor disease and assist in clinical decision-making. Yet the potential harms of big data technology should not be overlooked, especially when individuals' personal lives are being affected significantly. When big data is used to define and construct identity—as in defining who a healthy citizen or employee is—issues of privacy and personal data protection, discrimination and exclusion, as well as procedural fairness are inevitably involved.[134] One common fear related to identification from health and medical data is insurance discrimination based on disease susceptibility.[135] Another fear is group-level harm from analysis of aggregated data, including the risks of stigmatization.[136] This is considered to be more problematic as all members of the community will be affected, not only those who have given consent for their data to be used.

The risk and harm of stigmatization cannot be underestimated. For instance, it has been reported that a "warrior gene" is found to be prevalent in New Zealand Maori, which some scientists have suggested might explain why violence is common in the Maori community.[137] This conclusion carries a potentially stigmatic effect beyond genetic research when one considers that police officers and jurors may be influenced by the finding.[138] Similar claims that a particular population has a high level of a genetic variation associated with alcoholism, diabetes, or obesity could be stigmatic and lead to victim blaming.[139] Anonymized data subjects may be grouped according to geographical, socio-economic, ethnic, or other characteristics.[140] Indigenous groups have raised concerns about the risk that they will be singled out for discrimination in big data health research.[141] Furthermore, there is the concern of cultural harm which poses threats to the group in an unforeseen and unintended manner. For instance, data subjects give blood to researchers believing it is for diagnosis without realizing that researchers are

---

133 *See* Bart van der Sloot & Sascha van Schendel, *Ten Questions for Future Regulation of Big Data: A Comparative and Empirical Legal Study*, 7 J. INTELL. PROP. INFO. TECH. & ELEC. COM. 110, 116–17 (2016), http://www.jipitec.eu/issues/jipitec-7-2-2016/4438.

134 *See* SERGE GUTWIRTH & MIREILLE HILDEBRANDT, *Some Caveats on Profiling*, *in* DATA PROTECTION IN A PROFILED WORLD 31, 35–36 (Serge Gutwirth et al. eds., 2010).

135 Mittelstadt & Floridi, *supra* note 70, at 316.

136 Greely, *supra* note 122, at 789–791.

137 Jason Grant Allen, *Group Consent and the Nature of Group Belonging: Genomics, Race and Indigenous Rights*, 20(2) J.L. INF. & SCI. 28, 32 (2009).

138 *Id*.

139 *Id*.; Greely, *supra* note 122, at 790.

140 Mittelstadt & Floridi, *supra* note 70, at 318.

141 Greely, *supra* note 122, at 794–95.

taking their human DNA and patenting its products.[142] Yet, the practice of patenting a human gene sequence is deeply offensive and considered to be fundamentally immoral to certain native tribes.[143] The above reveals some controversial legal and ethical issues of informed consent, group level harm, and the control of research uses and materials.

Scholars have advocated for the incorporation of risk and harm assessment to tackle the problems of re-identification and discrimination in data privacy protection.[144] Although they are writing in the larger context of cloud computing and big data technology, their proposed models on data-driven accountability are equally applicable in our context of big data health research.[145]

A. Risk Assessment of the Disclosure and Reuse of Data

To ensure accountability, regulation of disclosure, and reuse of personal data, it is necessary to include de-identified data[146] because third parties may identify the individuals concerned through data combination. This may lead to profiling, and the risks and adverse effects of profiling through data mining and data combination are well-recognized.[147] Data brokers have been collecting, analyzing, selling, and linking individual identities without our knowledge for some time.[148] For example, Acxiom, the largest data broker in

---

[142] Pauline Lane, *Blood Money*, THE GUARDIAN, Jan. 21, 1998, http://www.theguardian.com/science/1998/jan/21/genetics.

[143] Allen, *supra* note 137, at 35. Other types of cultural harm include research findings that contradict the traditional beliefs and knowledge system of the group studied, leading to the loss of political or legal claims to certain territories or upsetting a group cohesion or social identity when certain members are discovered to be "genetic outsiders." *Id.* at 33–35.

[144] *See supra* Part IV § A.

[145] *See* Anne S.Y. Cheung, *Re-personalizing Personal Data in the Cloud*, *in* PRIVACY AND LEGAL ISSUES IN CLOUD COMPUTING 69 (Anne S.Y. Cheung & Rolf H. Weber eds., 2015) (including parts of the discussion remaining in this paper).

[146] *Id.* at 78–79. De-identified data includes anonymized and pseudonymous data. The former refers to "personal data that has been collected, altered or otherwise processed in such a way that it can no longer be attributed to a data subject." *Id.* at 78. The later refers to "explicit identifiers being replaced with codes." *Id.* at 79.

[147] *See* FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? UNDERSTANDING THE ISSUES 8–12 (2016), https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf; *see also* Eur. Data Prot. Supervisor, *Opinion 7/2015 Meeting the Challenges of Big Data: A Call for Transparency, User Control, Data Protection by Design and Accountability* 7–9 (2015), https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-11-19_Big_Data_EN.pdf.

[148] The FTC uses the term "data broker" to refer to those that "collect and aggregate consumers' personal information from a wide range of sources and resell it for an array of purposes, such as marketing, verifying an individual's identity, and preventing financial fraud." FED. TRADE COMM'N, WHAT INFORMATION DO DATA BROKERS HAVE ON CONSUMERS, AND HOW DO THEY USE IT 2 (2013), https://www.ftc.gov/sites/default/files/documents/public_statements/prepared-statement-federal-trade-

the U.S. and a marketing giant, holds an average of 1,500 pieces of information on each of more than 200 million Americans.[149] Also, it is estimated that each piece of information that users post on Facebook is worth five cents and that each Facebook user is worth $100 as a source of information.[150] Presently, there is limited regulation of the secondary use of data in most jurisdictions, particularly when they take the ostensible form of de-identified, nonpersonal data.[151] Ultimately, this is an issue of data security, relating to the obligations of data controllers to protect against unauthorized data access, use, and disclosure by third parties.

I am not advocating for a complete ban on the use of de-identified data. Indeed, there are legitimate reasons to reuse de-identified (pseudonymous) data,[152] such as in pharmaceutical trials and medical data research or for other legitimate purposes that serve the public interest. In such cases, scholars have recommended that clear guidelines be set, with minimum standards established for the de-identification of datasets and independent reviews of the risk of re-identification before data disclosure.[153] Many have advocated that a specific model be used to measure the continuum of risk involved. For example, Hon et al. use the "realistic risk of identification" as a benchmark,[154] whereas Schwartz and Solove suggest the "substantial risk of identification."[155] More concretely, Ohm recommends that any risk

commission-entitled-what-information-do-data-brokers-have-consumers/131218databrokerstestimony.pdf.

[149] Steve Kroft, *The Data Brokers: Selling Your Personal Information*, (CBS News Mar. 9, 2014), http://www.cbsnews.com/news/data-brokers-selling-personal-information-60-minutes/.

[150] VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 119 (2013).

[151] The new regime under the EU will regulate pseudonymised data under article 4 section 5 of the GDPR. GDPR, *supra* note 35, at art. 4. Both the US and Australia do not regulate de-identified data as the data are not subject to privacy laws. *How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule?*, U.S. DEP'T OF HEALTH & HUM. SERVICES NAT'L INSTITUTES OF HEALTH, https://privacyruleandresearch.nih.gov/pr_08.asp (last visited May 18, 2018); *De-identification and the Privacy Act*, AUSTL. GOV'T: OFFICE OF THE AUSTL. INFO. COMM'R, https://www.oaic.gov.au/resources/agencies-and-organisations/guides/de-identification-and-the-privacy-act.pdf (last visited May 24, 2018). *See generally International Review: Secondary Use of Health and Social Care Data and Applicable Legislation*, DELOITTE (Sept. 5, 2016), https://media.sitra.fi/2017/02/28142605/International_review_secondary_use_health_data.pdf.

[152] Article 4 section 5 of the GDPR defines "pseudonymisation" to be "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person." GDPR, *supra* note 35, at art. 4.

[153] W. Kuan Hon et al., *The Problem of 'Personal Data' in Cloud Computing: What Information Is Regulated?—The Cloud of Unknowing*, 1 INT'L. DATA PRIVACY L. 211, 215 (2011).

[154] *Id.* at 226.

[155] Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814, 1882 (2011).

assessment should take account of (1) the data-handling techniques used by database owners, (2) the nature of information release, with the public disclosure of data being subject to stricter scrutiny, (3) the quantity of data involved, (4) the likely motives and economic incentives for anyone to re-identify the data, and (5) the trust culture in a particular industry or sector—that is, the existing standard of fiduciary duty or duty of confidentiality in that sector.[156] Furthermore, as data identification and combination technologies are advancing at a rapid pace, I contend that any risk assessment concerned should be carried out on a regular basis with citizen scientists, rather than only at the stages of data collection and de-identification and disclosure.

B. Re-identification: Data Quality and Size

When considering threats of re-identification external to the original research team or organizations, data quality and size also need to be taken into account. Data quality refers to the nature, sensitivity, and linkability of data to individuals.[157] Linkability refers to the different degrees of data identifiability or the levels of effort required to identify an individual.[158] An example of good quality data is the information presented in Google Flu Trends. Regardless of whether its predictions are accurate,[159] the information that Google gathers from the online web search queries submitted by millions of individuals is abstracted at a high level and safely aggregated.[160]

Another important element of data quality is data size. The size of a database is determinative of how easy it is to link the information therein to an individual. The larger the database, the easier that link is to make.[161] However, the law seems to be silent regarding data controllers and how much data they may collect, how long they may retain data, and whether stricter security measures are needed for large databases.[162] Ohm argues that new quantitative limits and guidelines should be enacted to address these

---

[156] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1765–68 (2010).

[157] *Id.* at 1766.

[158] Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymisation Technique*, WP216, at 11 (Apr. 10, 2014), http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf ("*Linkability*, which is the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases).").

[159] Charles Arthur, *Google Flu Trends Is No Longer Good at Predicting Flu, Scientists Find*, THE GUARDIAN (Mar. 27, 2014, 6:27 AM), https://www.theguardian.com/technology/2014/mar/27/google-flu-trends-predicting-flu.

[160] Schwartz & Solove, *supra* note 155, at 1882.

[161] Ohm, *supra* note 156, at 1766–67.

[162] *Id*.

issues.[163] Such limits and guidelines would undoubtedly have an impact on bio-banks given the vast quantity of data stored in them, but they are still certainly deserving of further consideration.

In sum, in determining the likelihood of re-identification, we must also consider the quality and quantity of the data in question. Ultimately, the core issue in personal data protection is identity protection.

C. Sensitive Data and Recombination of Data

Many of the foregoing measures are dependent on the compliance framework of the data controllers and the organizations or companies concerned. Participants often have no idea that their data are being reused and processed or that they have been re-identified. It is therefore important to formulate an alternative privacy framework that is based less on consent and more on holding data controllers accountable for the particular reuse of data based on risk and the likely adverse impact on data subjects when the unauthorized disclosure takes place. Namely, the European Union has a higher standard for the use of sensitive personal data, which includes genetic, biometric, and health data under the new GDPR, while the U.S. regulates the combination of data.

The EU affords sensitive personal data special protection. Article 9 of the GDPR requires explicit consent from data subjects before any processing of sensitive data.[164] In addition, recital 51 of the GDPR specifies that personal data "which are, by their nature, particularly sensitive in relation to fundamental rights and freedoms merit specific protection as the context of their processing could create significant risks to the fundamental rights and freedoms."[165] Examples of such include personal data revealing racial or ethnic origin. Although the categories of sensitive data are likely to be controversial in different contexts and cultures, the "sensitive" nature of certain data reveals the underlying values and harm concerned. For example, data related to an individual's health (particularly sensitive health information such as HIV status) may lead to discrimination against that individual.[166] Bearing in mind the threat of harm arising from the re-

---

[163] *Id.* at 1767.

[164] Article 9 of the EU GDPR prohibits the processing of personal data revealing "racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade-union membership, and the processing of genetic data, biometric data for the purposes of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation," unless data subjects provide their explicit consent or other conditions under article 9(2) are satisfied. GDPR, *supra* note 35, at art. 9.

[165] GDPR, *supra* note 35.

[166] It was recently reported that the U.S. Federal Bureau of Investigation (FBI) has been "collecting racial and ethnic information and 'mapping' American communities around the country based on crude

identification of certain data, here I would argue that organizations need to ensure that sensitive data, which may perhaps be better described as critical data, are stored separately from the general network. They also need to ensure that access to such data is carefully monitored and that combination with other data cannot easily take place. Public disclosure must be impossible.[167]

Rather than imposing a high standard on a discrete category of sensitive data, there is special restriction on the combination of data in the U.S. For instance, in 2013, California amended its law on personal information to include regulation of the practice of data combination by imposing new requirements on the operators of commercial websites or online services that collect the personal information of Californian consumers.[168] The relevant provisions were further amended in 2016.[169] Under the amended section 1798.29 of the California Civil Code, the definition of personal information has been expanded to include "[a]n individual's first name or first initial and last name in combination with any one or more" of five stated categories of data fields (if any is unencrypted): (1) social security number, (2) driver's licence number or California identification card number, (3) bank account number or credit or debit card number in combination with any required "security code, access code, or password" that would permit access to an individual's financial records, (4) medical information, (5) health insurance information, and (6) "information or data collected through the use or operation of an automated license plate recognition system."[170] The definition also now includes "a user name or email address, in combination with a password or security question and answer that would permit access to an online account."[171] All of this information is subject to a specific duty of notice of breach and security requirement.[172] California's approach to regulating the combination of certain categories of unencrypted information constitutes a move in the right direction.

---

stereotypes about which groups commit different types of crimes." Seeta Peña Gangadharan & Sean Vitka, *Knowing Is Half the Battle: Combating Big Data's Dark Side Through Data Literacy*, SLATE (Apr. 2, 2014, 10:13 AM), http://www.slate.com/blogs/future_tense/2014/04/02/white_house_big_data_and_privacy_review_we_need_federal_policy_about_digital.html [https://perma.cc/6BS6-65QN].

[167] *See* Ohm, *supra* note 156, at 1768.

[168] S. 46, 2013–2014 Leg., Reg. Sess. (Cal. 2013).

[169] Assemb. 2828, 2015–2016 Leg., Reg. Sess. (Cal. 2016).

[170] CAL. CIV. CODE § 1798.29(g)(1) (West 2017). An "[a]utomated license plate recognition system" is defined under section 1798.90.5 of the California Civil Code to mean a "searchable computerized database resulting from the operation of one or more mobile or fixed cameras combined with computer algorithms to read and convert images of registration plates and the characters they contain into computer-readable data." CAL. CIV. CODE § 1798.90.5 (West 2016).

[171] CAL. CIV. CODE § 1798.29(g)(2) (West 2017).

[172] CAL. CIV. CODE § 1798.29 (West 2017). *See* CAL. CIV. CODE § 1798.82 (West 2017).

The emphasis is rightly on data being unattributable to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organizational measures to ensure anonymity.[173] In addition, the U.S. Federal Trade Commission (FTC) has recommended a more robust system of de-identification and accountability.[174] Rather than toiling with various concepts of de-identified data (anonymous, anonymized, and pseudonymous data), the FTC acknowledges that the de-identification of data is not foolproof, and thus there is always a possibility that individuals will be re-identified.[175] Accordingly, it recommends that companies should adopt a three-prong approach: (1) robustly de-identify personal data; (2) publicly make a commitment not to re-identify data, and (3) contractually require the same public commitment from any downstream users with which they share data with.[176] Such requirements should extend to the sharing of data with third-parties owing to the possibility of subsequent attribution by later parties.[177]

VII. CONCLUSION

Big data and information communication technologies hold great promise for health and medical citizen science. Citizen scientists can connect and exchange data with one another and with researchers. This has led to growing expectations to access and reuse the data in bio-banks and repositories. In grappling with the shifting nature of data and ever-evolving technology, various notions of consent have been formulated to resolve the tension between researchers' need for data and subjects' will for privacy and self-determination. Yet all the attempts to refine and redefine consent have proved to be futile conquests to preserve an individual's full autonomy.

Embedded in big data analytics is the use of both data and personal data and the matching of data sets. Arguably, one does not have enough data and medical science literacy to give meaningful consent to research involving such technology. To most participants, their consent may have reiterated their dependency on expert researchers, medical professionals, or state authorities. Regardless whether it is open or dynamic or a variation in-between, consent only gives an illusion of control in the big data age. Despite

---

[173] Recital 29 of the GDPR governs the measures on the use of pseudonymous data. GDPR, *supra* note 35.

[174] FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS 21–22 (2012), https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

[175] *Id.* at 22.

[176] *Id.* at 21–22, 37.

[177] *Id.* at 21.

the fact that the formulations of consent may be compatible with existing legal standards, they may be a far cry from ethical imperatives such as personal dignity, equality, and democratic accountability. The layered meanings of consent often come with a broader shift of unsolicited responsibility from public healthcare authorities, commercial actors, or institutional researchers to individual participants.[178] Regulators have warned that the transfer of risk and regulation should not "signal a reduction in the standards and quality of healthcare provision."[179] What lies behind the seeming empowerment of citizens should not be a disguised exploitation or extraction and sale of personal data leading to discrimination against individuals or groups.

While consent is still essential in medical and health research, it must be assured by a complementary system of data-driven accountability. Consent alone is not enough to restore autonomy to individual and citizen scientists in fast-evolving, data-intensive research. As there are different dimensions and forms of citizen science, so should the participation of citizen scientists at various stages of research follow the life cycle of data usage including risk and harm assessment, re-identification, and combination of data. The solution to attain autonomy must come through a comprehensive set of citizen science practice involving data research.

[178] Eline M. Bunnik et al., *A Tiered-Layered-Staged Model for Informed Consent in Personal Genome Testing*, 21 EUR. J. HUM. GENETICS 596, 598 (2013). The layered meanings of consent refer to the different layers or levels of information that is required. *Id.* at 598. The first basic level is directed at fundamental information essential for informed consent, which is explicitly offered to data subjects and is often kept minimal to enable easy communication. *Id.* The second or further layers of consent is based on extendable information, accessible for data subjects who actively seek for it or who have signed up for ongoing research, re-contact, or extras in the models of dynamic consent, portable legal consent, or meta consent. *Id.*

[179] *Opinion No. 29 of the EGE*, *supra* note 5, at 62.