JSL THE JOURNAL OF SCIENCE & LAW

COMMENTARY
published: October 11, 2016

The ethics of biomedical big data: Busting myths

Jacob M. Kolman[a,1]

[a]*Houston Methodist Hospital*

Biomedical research ethics has historically rested on cases of egregious harm and disrespect to subjects through direct experimentation on bodies. However, with the emergence of sophisticated health data and specimen analysis, a new type of research ethics case study has emerged to highlight the limitations of applying current research and privacy regulations to the study of Big Data. In this paper I challenge common myths about data protection and argue three points researchers must keep in mind: (1) De-identification does not always secure privacy in the manner intended, (2) Successful identification does not suffice to address all ethical concerns, and (3) any party that creates new health records should not presume that traditional regulatory restrictions have fully accounted for their own part in this vanguard of evolving responsibilities. To this last point, I argue that any researcher, including those operating in the largely unregulated domains of public data and citizen science, should seek ethics consultation to help them respect persons, avoid harms, and proceed justly beyond the legal minimums.

Keywords: big data, biomedical research, bioethics

Introduction

Biomedical research ethics tends to use a narrative in which historical case studies of shock and horror form the usual preface. I have some qualms with the reactionary tendencies this can inspire, but nevertheless I will use a similar tactic now with three recent high-profile cases, each of which involve biomedical "Big Data" research.[2] Afterward, I will review what makes these types of cases so problematic for ethicists and then showcase a few myths that I have seen biomedical researchers fall prey to – particularly those accustomed to interventional studies.

The myths I focus on here include:

(1) De-identifying data is a simple and unqualifiedly effective means of privacy protection;

(2) Because of the importance of privacy protection, effective de-identification is never, or at least rarely, an ethically problematic move to make; and

(3) More-traditional clinical researchers need not concern themselves with the unique problems of data research outside their field.

As Big Data integrates into more traditional interventional research methods (Angus, 2015) and opens new frontiers of citizen science (Hoffman, 2015), it is worth compiling and addressing a few key myths in one place (even if that list is not exhaustive). Although several of these have been discussed separately elsewhere, my purpose in highlighting them is instructive outreach rather than novelty.

[1] Jacob M. Kolman obtained an M.A. in Philosophy from Rice University in areas of study including biomedical research ethics and epistemology. He is currently a senior research assistant at Houston Methodist Hospital and coordinates work on an ongoing compendium of ethical and methodological standards for clinical trials. The views expressed herein are individual and do not represent Houston Methodist Hospital; no interests to declare.

[2] Big Data generally covers studies utilizing large databases of typically de-identified data, or even totally anonymous or public data. For reasons below, it is important to think of identifiable data as part of the same issue; I also concur with scholars who consider bio-*specimen* research to be relevantly similar in ethical respects to biomedical *data* research (Lynch, Bierer, & Cohen, 2016); projects such as *23andMe*, which start with samples but mainly operate with the derived data, also motivate the comparison (Drabiak, 2016). For sake of scope, I am not focusing on big data research outside of the biomedical context, such as the 2014 behavioral Facebook study controversy; similarities and differences may become evident, and I will return to that briefly in the conclusion.

OPEN ACCESS | Edited by: Pablo Ormachea, Center for Science and Law **Correspondence:** Jacob M. Kolman, Houston Methodist Hospital, Houston, TX, USA, jmkolman@houstonmethodist.org **Received:** May 20, 2016 **Accepted:** August 1, 2016 **Published:** October 11, 2016
Citation: Kolman, J. M. (2016). The ethics of biomedical big data: Busting myths. *Journal of Science and Law*, *2*(3): 18-23.

Three unsettling cases: When blood and data go where donors cannot follow

In 2009, families sued the Texas Department of State Health Services over the state keeping their newborns' blood samples. Samples had been taken for clinical screening purposes and then de-identified for secondary research use. The families won and the samples were destroyed; related cases in other jurisdictions also occurred (Lewis, 2015).

In a more publicized case, now commemorated in the off-Broadway musical *Informed Consent* (Tommy, 2015), the Havasupai Native American tribe agreed to diabetes research on their genetic specimens but later found unauthorized dissertation projects and publications naming their tribe (but not individuals). Topics ranged from incidence of incest to genetics-based migration claims which contradicted the tribe's religious narratives (in turn jeopardizing tribal territory claims based on these beliefs); the tribe won their samples back in an out-of-court settlement (Drabiak-Syed, 2010).

Chronologically last but not least, two related events demonstrate the use of healthcare data for research in the U.K.: First, a large-scale health research database called *care.data* attempted to launch in 2014, but after public outcry and large-scale opting-out, the project faltered (Carter, Laurie, & Dixon-Woods, 2015; Hall, 2016). Second, with concern over this databank still fresh, the National Health Service gave Google access to over one million patient records so that Google's "DeepMind" project could develop health-centered apps (Cabral-Isabedra, 2016). The creation of these databases and the use of coded medical records were within the U.K.'s research exceptions to privacy regulations, but both projects were critiqued on grounds of inadequate public consultation and notification prior to launch.[3]

Cases like these are discomforting for their novelty. Though controversial or outright inappropriate, all three examples are comparatively mild within the history of research ethics. They are not interventional experiments conducted in Nazi camps nor were any vulnerable subjects slipped LSD or given STDs, hepatitis, radiation, or cancer cells. Literally no *body* was experimented on at all.[4] Consequently, the intuitions of both researchers and ethics reviewers are not primed to think in terms of the unique forms of disrespect or danger these cases present. While the full restrictions associated with interventional experiments rightly seemed inapplicable, categories of regulatory leniency also failed to honor the legitimate concerns regarding public trust, scientific quality, and dual use (e.g., possible discriminatory projects based on research findings), *inter alia* (El Emam, Rodgers, & Malin, 2015; Hoffman, 2015, 2016). In order for data researchers to avoid similar controversy in the future, a few misconceptions must be addressed.

Myth #1: De-identifying data is a simple and unqualifiedly effective means of privacy protection

Currently, both HIPAA and the U.S. Common Rule[5] reward researchers for using de-identified data by eliminating regulatory hurdles, potentially exempting the researcher from ethics review (IRB approval) and/or waiving the need to obtain individual patient consent (DHHS, 2004; OHRP, 2016). As extra liability protection, research institutions frequently require their researchers to check with the IRB or Privacy Board at least briefly to verify that these exceptions apply, but on a straightforward reading of guidance, an activity does not "involve human subjects" unless individuals can be identified, and so is not regulated for human research protections (OHRP, 2016).

[3] Any comments regarding the U.K. databases and regulations are "pre-Brexit;" the future of British research regulations in relation to prior harmonization with the EU is, at the time of writing, unknown.

[4] Even for use of physical specimens, "secondary use" entails that the samples were already taken either for prior medical or primary research purposes, presumably with consent for the primary use. For brief summaries of the bleak cases referred to above, see: ("History of IRBs and the MWSU IRB," 2016; Macklin, 2013; Resnick, 2016), and *locus classicus* of the genre of American research ethics case study: (Beecher, 1966).

[5] The Common Rule, 45 CFR 46, is the key regulation governing federally-funded human subjects research. It is "Common" because the details have been copied over to the regulatory titles of multiple other U.S. Departments involved in research, not just Health and Human Services (Title 45) where it is most commonly cited. The first substantive revisions to this Rule since 1991 are currently under debate after a September 2015 Notice of Proposed Rule-Making release ("Federal policy for the protection of human subjects," 2015). The HIPAA Privacy Rule (implemented by 45 CFR 160 and 164), I trust is better recognized outside the biomedical research community for its notorious complexity and effect on the healthcare industry generally. If the data includes genetics, the Genetic Information Nondiscrimination Act of 2008 (GINA) can also restrict uses and disclosures.

These leniencies can imply that de-identification is a powerfully protective measure for privacy, but there are two important limitations. First, the meaning of this standard can be unclear. Both within and between countries, different rules equivocate on the relevant privacy terms (DHHS, 2004; Thorogood & Zawati, 2015). Lexicons in which "linked / unlinked," "coded/uncoded," "de-identify," "anonymize," and other cognate terms are used carry different contextual meaning and expectations for researchers and data stewards to honor. Even apart from direct equivocation, concepts can be subtle. For instance, "anonymiz*ed*" data has been stripped of identifiers which once existed, whereas the similarly sounding "anonym*ous*" data never contained identifiers to begin with (e.g., publicly available aggregate census data). It is doubtful that a patient not steeped in the regulatory language would appreciate the difference when skimming a data use policy; however, the subtlety matters because of the second limitation on de-identification: it might not work. *Re*-identification combines anonymized data, public records, and statistical analysis to guess the missing identifiers. Proactive statisticians have proven how re-identification can be accomplished with alarming precision on HIPAA-compliant de-identified data (Richardson, Milam, & Chrysler, 2015), a concern which does not apply to anonymous aggregate data.

The proposed U.S. Common Rule revisions have taken re-identification seriously enough, at least in the case of bio-specimens, to treat all specimen research as identifiable in principle ("Federal policy for the protection of human subjects," 2015). In practice, this closes the loophole which allowed secondary use of both the Texas infant blood spots and the Havasupai genetics to proceed without additional consent (that is unless researchers actively make the case that other exceptions apply).

Myth #2: Effective de-identification is rarely an ethically problematic move to make

For the sake of argument, let us suppose de-identification and careful data stewardship work well enough together to justify risks of re-identification, given the potential benefits of the research.[6] The next myth is that de-identification is always *ethically protective*, i.e. while it is not always necessary or perfect, removing identifiers is never ethically problematic in itself as a default option. This is false.

The Texas Blood Spot case did not revolve around identifiability but rather unauthorized future research. Parents cared about whether and for which research these specimens were kept. Similarly, even though the indirect, individual identifiability of the Havasupai tribespeople remains a commonly cited component of the case, tribal or community identifiability is not expressed in HIPAA protections. Only the naïve would think *indirect individual identifiability* was the main concern when the tribe discovered their (sacred) blood was not being used for the stated purpose – that somehow the social harms of implicating *the whole tribe* would be more acceptable than the danger of implicating any given part. This does not merely ignore a cultural difference between individualism and communalism but the inherently communal relevance of genetic data itself. An appeal to applicable privacy regulations might have been more legally strategic, as tort-based approaches to redress their cultural harms had failed in the past, but individual privacy was not the morally salient objection (Drabiak-Syed, 2010).

Surveys of patients and the public have also corroborated that the purpose of research matters to people, independent of data identifiability (Thorogood & Zawati, 2015; Tomlinson et al., 2015; Zarate et al., 2016).[7] An easily overlooked detail about these surveys is that responders do not always realize the full, and potentially objectionable, scope of a donation's uses. Unless the patient or data/specimen donor is primed to think about these possibilities or educated to know they exist, blanket consent for future research use can look permissive yet in fact be totally blind to the social implications (e.g., of genetics studies).[8] Thus, even if

[6] That is, of course, a substantial assumption compared to the technical complexity involved. While I argue here that privacy is not always the main issue, the sustained emphasis on privacy in the literature is not unfounded. Whenever privacy does become salient, technical problems are worth confronting with concentrated effort (Heatherly, 2016; Williams & Pigeot, 2016).

[7] Indeed, if de-identification is successful, researchers are actually *prevented* from knowing who to seek for permission prior to re-using records for controversial studies and are left either to forego the public benefits of conducting the research or risk damaging the public trust if donors discover and object to the research.

[8] One must also ask, "Who are we asking?" If these surveys are supposedly to reveal what the public wants to see in a data bank, systematic marginalization is a major issue (do minorities or localized tribes like the Havasupai take these surveys? Are they nevertheless affected by inferences made based on the results of public poll?).

the U.K. had more broadly advertised *care.data* and DeepMind, and even if they had gone one step further to an (arguably unwieldy) *opt-in* consent requirement instead of opt-out, the public may still have objected to some later research that their blanket consent had not anticipated; even without any chance of identifying the coded data, the objection would remain.

Related to this is the notion that deleting study data as soon as possible is ethically safe – a strategy reported to be part of the DeepMind terms of use (Cabral-Isabedra, 2016). This misconception is infrequent in regulated clinical trials (e.g., those done to satisfy drug marketing requirements) because all major national jurisdictions enforce explicit archiving periods for study data, some lasting over a decade. However, data researchers outside of this regulatory context should still recognize the reasons to archive. Data trails are audit trails, and audit trails are sources of accountability, confirmation, replicability, and (in certain circumstances) patient safety. Not so much of the data should be deleted that an authority cannot investigate whose data was touched and why. A large-scale and controversial project like DeepMind may see the offer to delete records as respectful in intent (seeking forgiveness, if not permission), but to a citizenry already skeptical after *care.data*, the image of an embezzler's shredding party may more likely be conjured.

Myth #3: More-traditional clinical researchers need not concern themselves with the unique problems of data research outside their field

The current regulatory distinction between interventional and data-driven research noted above encourages this assumption, and it remains likely that the relatively lower risks of Big Data to individuals will continue to motivate permissive legal policy. This myth is therefore addressed to colleagues conducting clinical trials, who may feel that rehearsing the case studies and issues above are irrelevant to their more tightly regulated industry.

Even though Big Data does not share all the dangers of experimental intervention, the converse is false: clinical trials will inherit the signature dangers of Big Data. There has been a well-publicized proposal by the International Committee of Medical Journal Editors (ICJME) to require all clinical trialists to share their data openly as a condition for publication (Taichman et al., 2016). This emerged not long after the U.S. Institute of Medicine encouraged exactly this development, along with risk-management recommendations (IOM, 2015; Mitka, 2015). The reasons for clinical trial data sharing are many, most notably to uphold scientific objectivity and accessibility of results as fundamental ethical imperatives for interventional research. Unpublished (or worse, *selectively* published) clinical trials do not produce the promised scientific gains, and without these benefits, there is no justification for putting research subjects under experimental risks and burdens. Even published trials could be delayed past the point of utility by the publication process itself.

Commentaries on the ICJME's drafted solution have been spirited. In particular, there are social ramifications of analyzing data sets out of context, prior to peer-review gatekeeping, and without clear controls to avoid circulating spurious conclusions to the public (Haug, 2016).[9] The responsible researcher, statistician, institution, and sponsor have new obligations to make technical and ethical provisions for secondary use of trial data (IOM, 2015). The very act of collecting data will give the dataset a life of its own, with the threat that trial stakeholders could become complicit partners in exactly the sort of cases discussed above without proper data stewardship policies.[10]

Conclusion

Traditional biomedical research ethics requires a

One might think in particular of the ethnically and religiously diverse minorities which exist in the UK, where *care.data* and DeepMind project custodians would presume to know how to respect all interests.

[9] Note again the force of this issue independent of privacy and confidentiality protection (or indeed complicated further insofar as privacy is protected, if the decontextualized data hides scientifically salient variables).

[10] It might be argued that all healthcare workers face this threat equally when medical records are anticipated to end up in research databanks; however, the treating physician and staff must make the record for the therapeutic sake of the patient regardless (how the physician responds to requests for or seizure of data is external to the therapeutic encounter). A clinical trialist conducts research for non-therapeutic purposes, and therefore is responsible for creating additional patient-subject data that would not have existed. Thus the researcher has the opportunity to consider secondary uses in the design stage prior to any data collection.

balance between respect for persons, beneficence, and justice; no single provision or safeguard accounts for these values perfectly and in the same way across research designs, which is why regulations include in-depth exceptions for every rule. Allowing Big Data research to proceed favors public beneficence and even justice (if directed toward neglected rather than privileged research goals), but risks of research to individuals and groups (including disproportionate risks to vulnerable, exploitable, or marginalized groups) should be analyzed. Respecting persons requires privacy and confidentiality protections and seeking out community comment on acceptable data use, often in lieu of individual specific consent (which may not be feasible or even desirable for larger or anonymous public data). This can be a good trade if handled with ethical and methodological care, but only if the inherent limitations of each measure are kept soundly in mind.

Although I have not strayed from the field of *biomedical* data research in the examples, there is nothing exclusively biomedical about respect for persons, beneficence, or justice, and some of the specific guidance derived from these concepts can be adapted. Big Data researchers in other academic fields (e.g., legal or sociological) and even citizen-scientists conducting their own research can consider what form these values take in their situation (and whether other more domain-specific or personal values must inform the balance).

In particular, a Big Data researcher should seek external advice from different perspectives, particularly those most attuned to any affected communities or parties, before starting a project. In biomedical and university settings, this practice is formalized into the institutional review board (IRB). This may seem alien to the unaffiliated citizen scientist, excited to donate to and access emerging big databanks, or to an academic from outside the disciplines where IRBs originated; however, as members of IRBs are quick to point out, they are not just a regulatory "entity" but an interdisciplinary sounding board of people, designed to see a project with fresh eyes, to protect the parties affected, and to consider the morality (not just legality) of the research plan. An IRB may not be available at the local university or hospital, or the IRB may see an outside project as beyond their authority or as an unjustified extension their workload (which is probably true); no matter – find ethics consults elsewhere; hire a private IRB; reconnect with colleagues and professors from other fields; announce the study on Kickstarter or Youtube to take public comment. Ethics, even applied research ethics, is a subset of philosophy, and philosophical reasoning always benefits from dialogue and perspectival diversity, which remains true for citizens with every (legal) right to access data, and even for studies on the presumably "safest" of anonymous datasets. Like de-identification and data stewardship, consultation is not fool-proof (none of our cases were challenged at the IRB / ethics committee or gatekeeper stages), but consultation is a good start that should not be neglected merely because the researcher works outside of domains where regulations compel it.

There are resources which offer additional concrete advice, worth recommending to any Big Data researcher, data steward, and even consenting donor (Drabiak, 2016; Hoffman, 2015; IOM, 2015).[11] Ethical Big Data research is not impossible, just more complicated than simply following the rules.

References

Angus, D. C. (2015). Fusing randomized trials with big data: The key to self-learning health care systems? *JAMA, 314*(8), 767-768.

Beecher, H. K. (1966). Ethics and clinical research. *NEJM, 274*, 1354-1360.

Cabral-Isabedra. (2016, May 1). Google strikes deal with NHS that gives AI unit access to 1.6 million patient records. *Tech times*. Retrieved from http://www.techtimes.com/articles/155059/20160501/googlestrikesdealwithnhsthatgivesaiunitaccessto16millionpatientrecords.htm.

Carter, P., Laurie, G. T., & Dixon-Woods, M. (2015). The social licence for research: why *care.data* ran into trouble. *J Med Ethics, 41*, 404-409. doi:10.1136/medethics-2014-102374

DHHS. (2004). *Protecting personal health information in research: Understanding the HIPAA Privacy Rule* Retrieved from https://privacyruleandresearch.nih.gov/pr_02.asp

[11] For those concerned with trans-national data sources, see the two-issue symposium in *J Law Med Ethics* Winter 2015 to Spring 2016 for commentaries on various privacy cultures and laws. McGill University also has a searchable database of policy documents governing genetics databanks and biobanks (http://www.popgen.info/).

Drabiak-Syed, K. (2010). Lessons from Havasupai Tribe v. Arizona State University Board of Regents: Recognizing Group, Cultural, and Dignitary Harms as Legitimate Risks Warranting Integration into Research Practice. *Journal of Health & Biomedical Law, VI*, 175-225.

Drabiak, K. (2016). Reponse to call for essays: Read the fine print before sending your spit to 23andMe. Blog Retrieved from http://www.thehastingscenter.org/response-to-call-for-essays-read-the-fine-print-before-sending-your-spit-to-23andme-r/

El Emam, K., Rodgers, S., & Malin, B. (2015). Anonymising and sharing invidivual patient data. *BMJ, 350*, h1139. doi:10.1136/bmj.h1139

Federal policy for the protection of human subjects. (2015). *Federal Register, 80*(173), 53933-54061.

Hall, K. (2016, 20 April 2016). One million patients have opted out of Care.data. *The register*. Retrieved from www.theregister.co.uk/2016/04/20/one_million_patients_have_opted_out_of_caredata/

Haug, C. J. (2016). From patient to patient - sharing the data from clinical trials. *NEJM, 374*(25), 2409-2411. doi:10.1056/NEJMp1605378

Heatherly, R. (2016). Privacy and security within biobanking: The role of information technology. *J Law Med Ethics, 44*(1), 156-160. doi:10.1177/1073110516644206

History of IRBs and the MWSU IRB. (2016). Retrieved from https://www.missouriwestern.edu/humansubsresearch/history-of-irbs-and-the-mwsu-irb/

Hoffman, S. (2015). Citizen science: The law and ethics of public access to medical big data. *Berkeley Technology Law Journal, 30*(3), 1744-1805. doi:10.15779/Z385Z78

Hoffman, S. (2016). The promise and perils of open medical data. *Hastings Center Report, 46*(1), 6-7. doi:10.1002/hast.529

IOM. (2015). *Sharing clinical trial data: Maximizing benefits, minimizing risks*. Retrieved from http://www.nationalacademies.org/hmd/Reports/2015/Sharing-Clinical-Trial-Data.aspx

Lewis, M. H. (2015). Lessons learned from the residual newborn screening dried blood sample litigation. *J Law Med Ethics, 43*(Supplement s1), 32-35. doi:10.1111/jlme.12211

Lynch, H. F., Bierer, B. E., & Cohen, I. G. (2016). Confronting biospecimen exceptionalism in proposed revisions to the Common Rule. *Hastings Center Report, 46*(1), 4-5. doi:10.1002/hast.528

Macklin, R. (2013). Ethical controversy in human subjects research. Retrieved from http://blogs.einstein.yu.edu/ethical-controversy-in-human-subjects-research/

Mitka, M. (2015). Clinical trial data: Share and share alike? *JAMA, 313*(9), 881-882.

OHRP. (2016, Feb 16). Human subjects regulations decision charts. Retrieved from http://www.hhs.gov/ohrp/regulations-and-policy/decision-trees/

Resnick, D. B. (2016, 7 July 2016). Research Ethics Timeline (1932-Present). Retrieved from http://www.niehs.nih.gov/research/resources/bioethics/timeline/

Richardson, V., Milam, S., & Chrysler, D. (2015). Is sharing de-identified data legal? The state of public health confidentiality laws and their interplay with statistical disclosure limitation techniques. *J Law Med Ethics, 43*(Supplement s1), 83-86. doi:10.1111/jlme.12224

Taichman, D. B., Backus, J., Baethge, C., Bauchner, H., de Leeuw, P. W., Drazen, J. M., . . . Wu, S. (2016). Sharing clinical trial data: A proposal from the International Committee of Medical Journal Editors. *JAMA, 315*(5), 467-468. doi:10.1001/jama.2015.18164

Thorogood, A., & Zawati, M. n. H. (2015). International guidelines for privacy in genomic biobanking (or the unexpected virtue of pluralism). *J Law Med Ethics, 43*(4), 690-702. doi:10.1111/jlme.12312

Tomlinson, T., De Vries, R., Ryan, K., Kim, H. M., Lehpamer, N., & Kim, S. Y. H. (2015). Moral concerns and the willingness to donate to a research biobank. *JAMA, 313*(4), 417-419.

Tommy, L. (Writer). (2015). Informed Consent [Play]. Duke on 42nd Street Theater.

Williams, G., & Pigeot, I. (2016). Consent and confidentiality in the light of recent demands for data sharing. *Biometrical Journal, 00*(0), 1-11. doi:10.1002/bimj.201500044

Zarate, O. A., Brody, J. G., Brown, P., Ramirez-Andreotta, M. D., Perovich, L., & Matz, J. (2016). Balancing the benefits and risks of immortal data: Participants' views of open consent in the Personal Genome Project. *Hastings Center Report, 46*(1), 36-45. doi:10.1002/hast.532