DOI:10.1145/2935882

Jacob Metcalf

Computing Ethics

Big Data Analytics and Revision of the Common Rule

Reconsidering traditional research ethics given the emergence of big data analytics.

"Big data" is a major technical advance in terms of computing expense, speed, and capacity. But it is also an epistemic shift wherein data is seen as infinitely networkable, indefinitely reusable, and significantly divorced from the context of collection.[1,7] The statutory definitions of "human subjects" and "research" are not easily applicable to big data research involving sensitive human data. Many of the familiar norms and regulations of research ethics formulated to prior paradigms of research risks and harms, and thus the formal triggers for ethics review are miscalibrated. We need to reevaluate longstanding assumptions of research ethics in light of the emergence of "big data" analytics.[6,10,13]

The U.S. Department of Health and Human Services (HHS) released a Notice of Proposed Rule-Making (NPRM) in September 2015 regarding proposed major revisions (the first in three decades) to the research ethics regulations known as the Common Rule.[a] The proposed changes grapple with the consequences of big data, such as informed consent for biobanking and universal standards for privacy protection. The Common Rule does not apply to industry research, and some big data science in universities might not fall under its purview, but the Common Rule addresses the burgeoning uses of big data by setting the tone and agenda for research ethics in many spheres.

The NSF-supported Council for Big Data, Ethics and Society[b] has focused on the consequences these proposed changes for big data, including data science and analytics.[9] There is reason for concern that the rules as drafted in NPRM may muddle attempts to identify and promulgate responsible data science research practices.

Is Biomedicine the Ethical Baseline?

The Common Rule was instituted in 1981. It mandates federally funded research projects involving human subjects to receive prior, independent ethics review before commencing. Most projects go through Institutional Review Boards (IRB)[3] responsible for

a So named for its common application across signatory federal agencies.

b See http://bdes.datasociety.net

IMAGE BY ANDREW OSTROVSKY

researchers' due diligence in identifying and ameliorating potential physiological, psychological, and informational harms to human subjects. The Common Rule grew out of a regulatory process initiated by the 1974 National Research Act, a response to public scandals in medical and psychological research, including the Nuremberg Doctors Trial, the Tuskegee syphilis study, and Milgram experiment on obedience to authority figures. The Act led to a commission on human-subjects research ethics that produced the Belmont Report (1979). The Belmont authors insisted that certain core philosophical principles must guide research involving human subjects: respect for persons, beneficence, and justice. The HHS developed the specific regulations in the Common Rule as an instantiation of those principles.[12]

Importantly, the Belmont authors understood that not all activities that produce knowledge or intervene in human lives are "research," and not all research about humans is sensitive or personal enough to be about "human subjects." To delimit "human-subjects research" within biomedicine, the Belmont commission considered "the boundaries between biomedical and behavioral research and the accepted and routine practice of medicine."[12] This boundary reflects the ethical difficulties posed by unique social roles of physician-researchers who are responsible for *both* patient health and societal well-being fostered by research knowledge. This unique role creates ethical dilemmas that are often not reflected in other disciplines. *Research* defined by the Belmont Report is, "an activity designed to test an hypothesis, permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge." *Practice* is, "interventions that are designed solely to enhance the well-being of an individual patient or client and that have a reasonable expectation of success."[12]

Not surprisingly, the first draft of the Common Rule came under attack from social scientists for lumping together all forms of human-subjects research under a single set of regulations that reflect the peculiarities of biomedical research.[2] Not all research has the same risks and norms as biomedicine. A single set of rules might snuff out legitimate lines of inquiry, even those dedicated to social justice ends. The HHS responded by creating an "Exempt" category that allowed human-subjects research with minimal risk to receive expedited ethics review. Nevertheless, there has remained a low-simmering conflict between social scientists and IRBs. This sets the stage for debates over regulating research involving big data. For example, in her analysis of the Facebook emotional contagion controversy, Michelle Meyer argues that big data research, especially algorithmic A/B testing without clear temporal boundaries or hypotheses, clouds the distinction between practice and research.[8,11] Jacob Metcalf and Kate Crawford agree this mismatch exists, but argue that core norms of human-subjects research regulations can still be applied to big data research.[10]

Big Data and the Common Rule Revisions

The Common Rule has typically not been applied to the core disciplines of big data (computing, mathematics, and statistics) because these disciplines are assumed to be conducting research on systems, not people. Yet big data has brought these disciplines into much closer intellectual and economic contact with sensitive human data, opening discussion about how the Common Rule applies. The assumptions behind NPRM leaving big data science out of its purview are empirically suspect.

Excluded—A New Category

Complaints about inconsistent application of the *exempt* category have prompted HHS to propose a new category of *excluded* that would automatically receive no ethical review due to inherently "*low risk*" to human subjects (§___.101(b)(2)). Of particular interest is exclusion of:

- research involving the collection or study of information that has been or will be acquired solely for non-research activities, **or**
- was acquired for research studies other than the proposed research study when the sources are publicly available, **or**
- the information is recorded by the investigator in such a manner that human subjects cannot be identified, directly or through identifiers linked to

The contentious history of the Common Rule is due in part to its influence on the tone and agenda of research ethics even outside of its formal purview.

the subjects, the investigator does not contact the subjects, and the investigator will not re-identify subjects or otherwise conduct an analysis that could lead to creating individually identifiable private information. (§__.101(b)(2)(ii))[4]

These types of research in the context of big data present different risk profiles depending on the *contents* and *what is done* with the dataset. Yet they are *excluded* based on the assumption that their *status* (public, private, pre-existing, de-identified, and so forth) is an adequate proxy for risk. The proposal to create an excluded category is driven by frustrations of social and other scientists who use data already in the public sphere or in the hands of corporations to whom users turn over mountains of useful data. Notably, social scientists have pushed to define "public datasets" such that it includes datasets that can be purchased.[2] The power and peril of big data research is that large datasets can theoretically be correlated with other large datasets in novel contexts to produce unforeseeable insights. Algorithms might find unexpected correlations and generate predictions as a possible source of poorly understood harms. Exclusion would eliminate ethical review to address such risks.

Public and *private* are used in the NPRM in ways that leave this regulatory gap open. "Public" modifies "datasets," describing access or availability. "Private" modifies "information" or "data" describing a reasonable subject's expectations about sensitivity. Yet *publicly available datasets containing private data* are among the most interesting to researchers and most risky to subjects.

For example, a recent study by Hauge et al.[5] used geographic profiling techniques and public datasets to (allegedly) identify the pseudonymous artist Banksy. The study underwent ethics review, and was (likely) permitted because it used public datasets, despite its intense focus on the private information of individual subjects.[5] This discrepancy is made possible by the anachronistic assumption that any informational harm has already been done by a public dataset. That the NPRM explicitly cites this assumption as a justification to a priori exclude increasingly prominent big data research methods is highly problematic.

Perhaps academic researchers *should* have relaxed access to maintain parity with industry or further scientific knowledge. But the Common Rule should not allow that de facto under the guise of empirically weak claims about the risks posed by public datasets. The Common Rule might rightfully exclude big data research methods from its purview, but it should do so explicitly and not muddle attempts to moderate the risks posed by declaring public data inherently low risk.

Exempt—An Expanded Category

The NPRM also proposes to expand the Exempt category (minimal review largely conducted through an online portal) to include secondary research using datasets containing identifiable information collected for non-research purposes. All such research would be exempt as long as subjects were given prior notice and the datasets are to be used only in the fashion identified by the requestor (§__.104(e)(2)). The NPRM does not propose to set a minimum bar for adequate notice. This can be reasonable given the high standard of informed consent is intended primarily for medical research, and can be an unreasonable burden in the social sciences. However, to default to end user license agreements (EULA) poses too low a bar. Setting new rules for the exempt category should not be a de facto settlement of this open debate. Explicit guidelines and processes for future inquiry and revised regulations are warranted.

Conclusion

The NPRM improves the Common Rule's application to big data research, but portions of the NPRM with consequences for big data research rest on dated assumptions. The contentious history of the Common Rule is due in part to its influence on the tone and agenda of research ethics even outside of its formal purview. This rare opportunity for significant revisions should not cement problematic assumptions into the discourse of ethics in big data research.

References

1. boyd, d. and Crawford, K. Critical questions for big data. *Information, Communication & Society 15*, 5 (2012), 662–679.
2. Committee on Revisions to the Common Rule for the Protection of, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, et al. Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences, 2014; http://www.nap.edu/read/18614/chapter/1.
3. Department of Health and Human Services Code of Federal Regulations Title 45—Public Welfare, Part 46—Protection of Human Subjects. 45 Code of Federal Regulations 46, 2009; http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html.
4. Department of Health and Human Services. Notice of Proposed Rule Making: Federal Policy for the Protection of Human Subjects. Federal Register, 2015; http://www.gpo.gov/fdsys/pkg/FR-2015-09-08/pdf/2015-21756.pdf.
5. Hauge, M.V. et al. Tagging Banksy: Using geographic profiling to investigate a modern art mystery. *Journal of Spatial Science* (2016): 1–6.
6. King, J.L. Humans in computing: Growing responsibilities for researchers. *Commun. 58*, 3 (Mar. 2015), 31–33.
7. Kitchin, R. Big data, new epistemologies and paradigm shifts. *Big Data & Society 1*, 1 (2014).
8. Kramer, A., Guillory, J., and Hancock, J. Experimental evidence of massive-scale emotional contagion through social networks. In *Proceedings of the National Academy of Sciences 111*, 24 (2014), 8788–8790.
9. Metcalf, J. Letter on Proposed Changes to the Common Rule. Council for Big Data, Ethics, and Society (2016); http://bdes.datasociety.net/council-output/letter-on-proposed-changes-to-the-common-rule/.
10. Metcalf, J. and Crawford, K. Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society 3*, 1 (2016), 1–14.
11. Meyer, M.N. Two cheers for corporate experimentation: The a/b illusion and the virtues of data-driven innovation. *Colorado Technology Law Journal 13*, 273 (2015).
12. National Commission for the Protection of Human Subjects, of Biomedical and Behavioral Research and The National Commission for the Protection of Human Subjects (1979) The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research; http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.
13. Zwitter, A. Big data ethics. *Big Data & Society 1*, 2 (2014).

Jacob Metcalf (jake.metcalf@gmail.com) is a Researcher at the Data & Society Research Institute, and Founding Partner at the ethics consulting firm Ethical Resolve.

This work is supported in part by National Science Foundation award #1413864. See J. Metcalf "Letter on Proposed Changes to the Common Rule. Council for Big Data, Ethics, and Society (2016)"[9] for the public comment on revisions to the Common Rule published collectively by the Council for Big Data, Ethics and Society. This column represents only the author's opinion.