Research Ethics Issues Raised in Collecting and Maintaining Large Scale, Sensitive Online Data (White Paper)

Michelle N. Meyer, PhD JD

This white paper discusses the collection, maintenance, and analysis of passively acquired "big data." The ethics of interventional research on online consumer platforms is beyond the scope of this paper.[1]

I. Respect for persons: Consent and other mechanisms for manifesting respect

Although commonly forgotten, in the U.S. tradition of research ethics that developed in the late 1970s, the default requirement of study-specific, voluntary informed consent emerged from—and is merely one way of honoring—the principle of respect for persons. Both the Belmont Report and the federal Common Rule it animated recognize that such consent is not always feasible or ethically (or legally) required for all research.[2] Individual consent will also be a poor fit in cases where data are inherently relational, as at least some data of interest to ISSOD will be (e.g., data pertaining to Facebook "friending" behavior or Twitter follower behavior).[3] Nevertheless, this section provides a brief overview of the kinds of consent that are possible, the kinds of data and research that may not legally[4] or ethically require any form of consent, and the spectrum of other mechanisms beyond consent that are available for respecting data subjects as persons.

A. Varieties of consent

1. *Study-specific consent*. Traditional research consent is study-specific, and discloses a great deal of information about the nature and purpose of a particular study, prior to data collection, in order to enable a prospective participant to decide whether to participate or not. This kind of consent will rarely be feasible in collecting large scale data from online platforms. Even when partnering with a platform that has the ability to push informed consent mechanisms (e.g., pop-up dialogue boxes with forced tutorials[5]) out to users, platforms will (understandably) be unwilling to interrupt the user experience every time a researcher wishes to collect or analyze user data.

2. *Broad and blanket consent*. Such a mechanism might be deployed for particular studies on rare occasions, but ordinarily, consent will have to be "broad," in which a user consents to

[1] For an overview, see Meyer, M. N. (2018). Ethical considerations when companies study — and fail to study — their customers. In E. Selinger, J. Polonetsky & O. Tene (Eds.), *Cambridge handbook of consumer privacy* (pp. 207–231). Cambridge, UK: Cambridge University Press.

[2] Meyer, M. N. (2015). Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colorado Technology Law Journal, 13*(2), 273–331.

[3] This is similarly true of genetic/omic research, since genetic/omic information is inherently relational, pertaining to families and broader kin groups. Although this problem is well-recognized, no satisfactory solution for consent that honors this relationality has been identified. With some exceptions for identical twins, individuals invited to participate in genetic/omic research are merely advised to discuss their choice with their blood relatives.

[4] This white paper address research regulation under the U.S. federal Common Rule only; application of the GDPR, other non-U.S. laws, and U.S. state laws are beyond the scope of the paper.

[5] 23andMe uses such a mechanism not for research but to secure user informed consent before they access sensitive health results (those pertaining to breast cancer, Parkinson's, and Alzheimer's risk).

unspecified future research within a broad scope (e.g., "research about information consumption" or "research that can help us better understand civil engagement"), or "blanket," in which the scope of research questions is boundless. Forced choice mechanisms (with or without tutorials and/or competence quizzes to ensure informedness[6]) could be used as one-time means to secure such broad or blanket consent (they could also be deployed on a regular basis, e.g., annually, to ensure that the user's preferences have not changed[7]). Whatever their legal status, platform Terms of Service (TOS) essentially never provide ethically meaningful informed consent. Virtually no one reads TOS, nor—rationally speaking—should they take the time and effort to do so, since they are largely incomprehensible contracts of adhesion.[8]

Although broad consent is not without its critics among ethicists,[9] the dominant view among ethics scholars appears to be that (well-executed) broad consent is ethically permissible. Biobank researchers have long used broad and blanket consent by availing themselves of the Common Rule's provisions for waiver or alteration of consent, and the revised Common Rule explicitly provides for broad consent as an option for the collection, maintenance, and secondary use of identifiable biospecimens or identifiable, private information.[10] Broad consent is a policy compromise that recognizes two sets of realities. On the one hand, human tissue is expensive to collect and maintain but a critically important resource in advancing human welfare. Every possible appropriate research use of such tissue or the information derived from it cannot be foreseen at the time of collection, nor is it feasible to recontact and reconsent all tissue or data sources when new research uses are contemplated. There is also considerable value in allowing as many researchers as possible to access this valuable resource. On the other hand, broad consent preserves the ability of individuals to exercise control over their tissue and/or data by declining participation.

Broad consent should be familiar to citizens of representative democracies; in both cases, subjects agree to allow a governance body to make decisions on their behalf. When data resulting from tissue collected under broad or blanket consent are deposited into repositories for broader access by other researchers, a Data Access Committee (DAC) reviews data access requests, asking whether the proposed secondary analysis falls within the scope of the broad consent or runs afoul of any other Data Use Limitations (DUL) provided in the consent (e.g., a promise that researchers will not attempt to re-identify subjects or a restriction to use by affiliates of non-

---

[6] Harvard's Personal Genome Project requires prospective participants to score 100% on a quiz testing their knowledge of genetics and the risks of posting their whole genome sequence and health data on the open Internet before they are accepted into the study. NIH's *All of Us* Research Program (formerly the Precision Medicine Initiative) uses an e-Consent platform developed by Sage Bionetworks that uses formative assessments of comprehension during the consent process to reinforce learning but does not require any score to join.

[7] A variant of broad consent is "dynamic consent," which involves a two-way technological interface between participant and researcher that allows unanticipated research uses that might stretch the bounds of the original consent to be communicated to participants and also allows participants with evolving preferences to make different decisions about their data. Kaye, J., Whitley, E. A., Lund, D., Morrison, M., Teare, H. & Melham, K. (2015). Dynamic consent: a patient interface for twenty-first century research networks. *European Journal of Human Genetics, 23*, 141–146.

[8] Meyer, M. N. (2015). Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colorado Technology Law Journal, 13*(2), 273–331.

[9] See, e.g., Caufield, T. & Kaye, J. (2009). Broad consent in biobanking: Reflections on seemingly insurmountable dilemmas. *Medical Law International, 10*(2), 85–100.

[10] 45 C.F.R. § 46.116(d). These provisions are limited to identifiable biospecimens and data because the Common Rule does not apply to secondary research on non-identifiable biospecimens or data, nor does it apply to secondary analysis of identified or identifiable but non-"private" information. *Id.* § 46.102(e).

profit research organizations). Care should be taken when drafting broad consent language to ensure that laypersons and researchers share an understanding of the terms used. For instance, many researchers would understand "research related to health" to encompass the social determinants of health (e.g., educational attainment, income, zip code), but most laypersons are not familiar with that concept, leading to secondary research that does not align with user expectations.

3. *Tacit consent.* A final form of consent[11] is tacit consent, which rests on an individual's failure to actively opt out of passive research participation. For such consent to be ethically meaningful, it must satisfy the publicity principle: individuals must know that they have the ability to opt out, they must appreciate the consequences of not opting out, and opting out must be fairly easy. Like other nudges, opt-out consent can be an effective way of enrolling more participants in research who are not especially opposed to it by leveraging inertia or indifference, while still allowing those with clear views to easily make a different choice. Platforms could enable opt-out consent. Again, however, inviting an opt-out process in the TOS will not suffice; the invitation must be much more prominent than that. Fiesler and Proferes recommend that researchers using online data find ways to allow users to opt in or out of particular kinds of research: "This could be, for example, a flag set in the user profile or a black/white list included as part of the API. Another potential design would be to build a system that could provide public notices when data collection begins from a specific hashtag, informs users when their tweets are included in a dataset, and/or links those who have had their tweets used back to a published paper based on the results."[12]

B. *When consent is not legally and (arguably) not ethically required*

Several categories of data or research of relevance to ISSOD fall outside the scope of the Common Rule—and hence its default requirement of informed consent.

1. *Research with non-identifiable data*. First, research with non-"identifiable" data that were collected for a purpose other than a particular study (i.e., the researcher does not intervene or interact with subjects to obtain the data) does not involve "human subjects" as the Common Rule defines that term. Data are "identifiable" if "the identity of the subject is or may readily be ascertained by the investigator or associated with the information."[13] Note that, depending on how one interprets "readily ascertainable," this definition of "identifiable" is quite weak. Federal regulators are aware of improvements in reidentification techniques and the revised regulations require Common Rule agencies, within one year and at least every four years thereafter, to a) reconsider this definition and b) consider whether any analytic technologies or techniques should be considered to necessarily produce identifiable data.[14] For present purposes, the point is that

[11] Two other recently articulated kinds of consent are worth noting: meta consent, which invites people to determine how and when they would like to be presented with a request for consent, see Ploug, T. & Holm, S. (2016). *Bioethics, 30*(9), 721–732, and dynamic consent, which entails ongoing communication between researchers and data subjects about evolving research questions and user preferences, see Budin-Ljøsne, I., et. al. (2017). *BMC Medical Ethics, 18*(4), 1–10.
[12] Fiesler, C. & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media + Society, Jan.–March*, 1–14.
[13] 45 C.F.R. § 46.102(e)(5).
[14] *Id.* § 46.102(e)(7).

federal research regulation codifies a reasonable policy choice that treats research with non-identifiable data (however defined) quite differently than research with identifiable data, and excludes the former from regulation.[15]

Moreover, even if data aren't likely to be associated with individual identities—or were already publicly available in identified or identifiable form—people may still have autonomy interests in controlling those data. For instance, data subjects may have an interest in avoiding complicity in research with which they disapprove by being able to prevent their data from contributing to such a project. The classic case is that of the Havasupai Tribe of Arizona, many of whose members entered into an oral agreement with academic researchers to give blood in order to study the tribe's epidemic of diabetes. A written consent form, however, provided for broad use of the blood for behavioral or health research, and the data were eventually used to study schizophrenia, consanguinity, and tribal migration patterns, all of which were highly objectionable to both the participants and to the other members of the tribe who felt equally stigmatized even without have provided blood.

In the case of ISSOD, it is not difficult to imagine that some data subjects might object to what they perceive to be partisan research on U.S. politics. Some data repositories (e.g., NIH's *All of Us* Research Program) have attempted to define "sensitive research" and develop processes for reviewing proposals to conduct it. The challenges of defining in advance categories of research that some (a majority?) of data subjects will find substantively objectionable, however, are formidable.

2. *Research with public data*. Under the Common Rule, "human subjects" are similarly not involved if researchers study identifiable but non-"private" information. "Private" information is "information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information that has been provided for specific purposes by an individual and that the individual can reasonably expect will not be made public (e.g., a medical record)."[16] A great deal of data of interest to ISSOD falls into this category, including non-protected tweets, Facebook posts set to "public," public Twitter and Facebook profile information, and comments on sites such as Reddit and 4/8Chan. Although, as discussed below, some users and some ethicists believe that what constitutes meaningfully "public" data and spaces is or should be contested, these data are not "private" under any plausible interpretation of the Common Rule, and hence research with them—even with identities fully attached—does not involve "human subjects" and therefore is not subject to the Common Rule, including to its requirements of consent or IRB review.

There are some important complications to the claim that research with "public" data falls outside the scope of the Common Rule and does not require consent. First, sometimes a platform's TOS will either deem platform data to be "private" or, more commonly, will prohibit research use of the data. One prominent advisory body has suggested that in both of these cases, seemingly public data should be considered private for purposes of the Common Rule.[17] Others,

---

[15] HIPAA makes the same distinction, with the same consequences for (lack of) consent ("authorization") for research use of de-identified data, although HIPAA sets a higher bar for when data are "de-identified" than does the Common Rule with respect to non-"identifiable" data.

[16] *Id.* § 46.102(e)(4).

[17] Secretary's Advisory Committee on Human Research Protections. (2013). Attachment B: Considerations and recommendations concerning Internet research and human subjects research regulations, with revisions. https://www.hhs.gov/ohrp/sachrp-committee/recommendations/2013-may-20-letter-attachment-b/index.html#backfn2.

however, have argued that violating TOS in order to conduct research must be ethically justified, but is possible.[18]

A second complication is that although the Common Rule deems data to be private only if a data subject's expectations of privacy are "reasonable," many data users clearly do misunderstand the public nature of the data they produce. These misunderstandings may indeed be unreasonable, but from an ethical perspective, it is not obvious in all cases that researchers should take advantage of the user's mistake, whether reasonable or not. For instance, a recent exploratory survey[19] of Amazon Mechanical Turk workers who reported being current or former Twitter users found that 61% had not previously realized that public tweets might be used in research, 43% (incorrectly) believed that researchers are not permitted to use tweets in research without user permission (because doing so would violate Twitter TOS, copyright law, or—most commonly—research ethics rules). Facebook users might misunderstand the platform's privacy settings, which for a long time were rather non-intuitive. In some cases, users would not have contributed sensitive data had they realized how public it was and what kinds of things might be done with it.

Some scholars go further and believe that producers of "public" tweets, public Facebook posts, and the like in fact have reasonable expectations that their data will not be used in research without their explicit permission.[20] Many overlapping reasons have been offered for this. For instance, many users seem to adhere to the principle of "privacy by obscurity":[21] although an unprotected tweet can technically be accessed by anyone with an Internet connection, ordinarily, the user with only a handful of followers can reasonably expect that it will in fact reach only a small audience. Similarly, although a user voluntarily provides each public datum, the fact that it is generally infeasible for others to aggregate and analyze her entire platform profile (perhaps discovering new, additional information about the user in the process of analysis that the user didn't intend to share and may even have been unaware of herself) provides a sort of informational privacy. Data repositories that aggregate and organize data for easy searching eliminate or at least reduce that privacy-protecting obscurity. Similarly, depending on how public data are stored and reported in published research, research can have the effect of greatly amplifying attention to what were public-but-obscure data. Other scholars and users have, under the capacious label of "privacy," expressed concerns about data decontextualization: studying data out of context (e.g., a tweet written in the heat of the moment, or one that is best understood

---

[18] Bruckman, A. (2016). Do researchers have to abide by Terms of Service (TOS)? The Next Bison: Social Computing and Culture. https://nextbison.wordpress.com/2016/02/26/tos/.

[19] Fiesler, C. & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, Jan.–March: 1–14. Although the total sample size for this survey was 368, the *n* for most questions discussed here was only 268. One limitation of this study is that MTurkers are likely habituated to written informed consent disclosures and may incorrectly assume that all human subjects research requires such explicit, study-specific informed consent, which may have affected some of their responses about the appropriateness of researcher use of public tweets and profile information. Another limitation is that the survey did not elicit respondents' overall preferences for research use of public tweets and Twitter histories in light of the tradeoffs involved if individual, study-specific consent for each research use of a public tweet were required.

[20] Michael Zimmer. (Feb. 12, 2010). "Is it ethical to harvest public Twitter accounts without consent?," https://www.michaelzimmer.org/2010/02/12/is-it-ethical-to-harvest-public-twitter-accounts-without-consent/. *See also* Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology, 51*(6), 1149–1168.

[21] Hartzong, W. N., & Stutzman, F. (2013). The case for online obscurity. *California Law Review, 101*(1), 1–49.

in the context of a thread or of an unnamed exogenous event, e.g., political event) can portray the data subject in a poor or unfair light.[22]

A final complication is that it is unclear when a space is public or private and people do not share the same view about this. For instance, if one needs to register or create an account to join a platform, but anyone can do so (i.e., the registration process is entirely undiscriminating), is that a public or a private space? Sometimes, platforms created for particular purposes will generate expectations about how and why information is shared. For instance, someone who shares sexual preferences on an online dating site for the purpose of meeting a compatible companion is not necessarily thereby saying that that information is free for the taking, to be used by those outside that community for other purposes, simply because they registered.[23]

3. *Exempt research*. Two exemptions from the Common Rule mirror the two ways, described above, that research can be deemed not to involve human subjects (and so are subject to the same caveats described above). First, human subjects research involving "observation of public behavior (including visual or auditory recording)" is nevertheless exempt from the Common Rule (and, hence, its consent default requirement) if any one of the following criteria is met: the data are recorded by researchers in such a way that the subject's identify "cannot readily be ascertained, directly or through identifiers linked to the subjects;" the data are not sensitive; or the data are identifiable and sensitive but an IRB conducts a limited review of the data security plan.[24] Notably, with respect to the first prong, OHRP guidance advises that if the researcher and the holder of a key to coded private information enter into an agreement under which the researcher will never be given access to the key (or will not be given access until the subjects are deceased, at which point they are no longer "human subjects" under the Common Rule), then this meets the requirement that subjects' identities cannot be readily ascertained.[25] A platform could serve as this data broker, providing coded data to ISSOD (and, through it, researchers). A separate exemption is available permitting nonconsensual secondary research with identifiable, private data if the data are either "publicly available" or are recorded in a non-identifiable manner, as per the previous exemption.[26]

4. *Minimal risk research in the public good*. The Common Rule permits research that is no more than minimal risk to be conducted without informed consent if: the research "could not practicably be carried out" without the waiver or (when identifiable data are used) with non-identifiable data; if waiving consent would not adversely affect the rights or welfare of subjects; and if, whenever appropriate, subjects are debriefed after the research.[27] Although IRBs do not always agree about when research with consent would not be "practicable," two common qualifying reasons are if the researchers have no feasible means of interacting with subjects (typically the case with big data research) and if informing subjects about a specific study in advance would bias the results (often the case when researchers wish to study behavior).

---

[22] Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, 79, 101–139.
[23] See, e.g., discussion of the OkCupid scraping in Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science, 1*(1), 1–14.
[24] *Id.* § 46.104(d)(2).
[25] Office for Human Research Protections. (2008). Coded private information or specimens used in research, guidance. https://www.hhs.gov/ohrp/regulations-and-policy/guidance/research-involving-coded-private-information/index.html.
[26] 45 C.F.R. § 46.104(d)(4).
[27] *Id.* § 46.116(f).

Some ethicists have argued compellingly that certain minimal risk research should be permitted without consent—whether or not obtaining consent is practicable. These scholars point out that although individuals and groups have legitimate interests in privacy, they also have legitimate interests in public goods such as better quality health care, and that research policy must better balance these sets of interests than it currently does. Some have even argued that citizens have an ethical obligation to participate in low-risk, passive research such as health data research[28] and digital epidemiology.[29]

C. Other ways of respecting data subjects

When consent is infeasible or otherwise inappropriate, other ways of respecting data subjects as persons should be pursued. For instance, researchers can often provide notice of the research, either before or after the fact, and either study-specific or broad. Notice can also include return of aggregate results of research to data subjects. Some empirical studies of public perceptions of research have found that many (though not all) people would find notice of the purpose, nature, and/or results of research to be nearly as satisfying as consent for minimal risk research. For instance, in the aforementioned survey of MTurk Twitter users, although most thought that researchers should ask permission before collecting or using public tweets, "a number of respondents specifically framed their desire to both *know about* the research and to see it when it's finished as an issue of respect. Informed consent can be seen as both *informing* and *consenting*, and for many respondents, the former would be sufficient."[30] A national survey of public perspective on pragmatic randomized trials found that although most respondents preferred traditional, written informed consent (despite the majority recognizing that the trials depicted did not pose additional risk on participants), a substantial minority—nearly 40% in one arm—not only tolerated but *preferred* general notification.[31]

ISSOD could require all publications based on ISSOD data to be made publicly available on the ISSOD site, accompanied by accessible lay summaries. A more ambitious version would involve a technology-enabled platform that tells data subjects exactly which publications their data contributed to.[32] ISSOD could also host videos of lightning talks that explain to the public the major discoveries enabled by their data. Finally, in addition to returning aggregate research results, it can be valuable to return individual research results. The National Information Study that ISSOD envisions might, for instance, show each survey respondent how their answers compare to others.[33]

---

[28] Ballantyne, A. & Schaefer, G. O. (2018). Consent and the ethical duty to participate in health data research. *Journal of Medical Ethics, 44*(6), 1–5.

[29] Mittelstadt, B., Benzler, J., Engelmann, L., Prainsack, B. & Vayena, E. (2018). Is there a duty to participate in digital epidemiology? *Life Sciences, Society and Policy, 14*(9), 1–24.

[30] Fiesler, C. & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, Jan.–March: 1–14.

[31] Nayak, R. K., Wendler, D., Miller, F. G. & Kim, S. Y. H. (2015). Pragmatic randomized trials without standard informed consent? A national survey. *Annals of Internal Medicine, 163*(5), 356–364.

[32] 23andMe has such a feature: when logged into the platform, the user sees a message such as, "Michelle, you contributed to 17 published discoveries!" followed by a list of those publications with links to the journal (unfortunately, often not open access).

[33] The Facebook app Genes for Good offers something like this. Participants are asked to complete numerous phenotype surveys, but when asked, e.g., how frequently they smoke, they are rewarded through a simple data visualization showing how their answer compares to those of others, e.g., the respondent is in the top quintile for smoking activity.

Another way of respecting data subjects and their contribution to research is by ensuring that ISSOD enables reproducible science. Reproducible science is an ethical (not merely scientific) issue because whatever risks data subjects bear are for naught if the resulting science is not reproducible. Social Science One, for instance, will require researchers who access their platform to adhere to the "replication standard": all funded research must archive replication data files, including code, methods and metadata.[34] In addition, ISSOD could, similar to NIH and clinicaltrials.gov, require preregistration of hypotheses (if any) and data analysis plans and require public posting of results, including null results, within a specified time period (subject to extension) on an ISSOD site (whether or not the results are also reported in a peer-reviewed journal).

II. Other ethical issues raised

A. *Group harm, vulnerable populations, and participant engagement*

Certain populations may both perceive that they are at increased risk by having their data aggregated and shared and may in fact be at increased risk of harm. For instance, one survey study found that "[l]esbian, gay and bisexual (LGB) respondents were more likely to express concern over their Twitter posts being used in government (odds increase of 2.12) and commercial settings (odds increase of 1.92), compared to heterosexual respondents," perhaps owing to historical online abuse of LGBT persons and/or antagonistic relationships between these communities and some governments.[35] If ISSOD expects to enable research on vulnerable populations, it should consider engaging those communities early in the process of developing the platform to see whether risks can be mitigated or community representatives can be included in ISSOD oversight structures. More generally, it is a good idea to engage the public (whether vulnerable or not) in development and oversight of the research platform.

B. *Access to data*

Data repositories typically calibrate data access according to the sensitivity and/or identifiability of the data, providing different access mechanisms for different kinds of data. At one end the spectrum, the most sensitive data might only be accessible through virtual or actual data enclaves. At the other end of the spectrum, metadata and innocuous, individual-level raw data might be made publicly accessible. In between are myriad options (see Table 1 in Meyer, 2018[36] and Appendix) that involve trade-offs between data security, on the one hand, and transparency and fair access, on the other. For instance, it is common to restrict data access to "qualified researchers," usually permanent faculty (i.e., those with "PI privileges") affiliated with a research institution.[37] The purpose of this restriction is that the institution can be required to be a party to the data use agreement, thereby lending that agreement some teeth. But this practice

[34] King, G. & Persily, N. (Aug. 10, 2018). A new model for industry-academic partnerships. https://gking.harvard.edu/files/gking/files/partnerships.pdf.
[35] Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology, 51*(6), 1149–1168.
[36] Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science, 1*(1), 1–14.
[37] E.g., Social Science One.

excludes citizen scientists, independent academics, and journalists, who may have legitimate interests in the data and who are likely to be more representative, politically and in other ways, of the general population whose data comprises the repository. A data repository that purports to serve the public good but is accessible only by academics may be viewed with skepticism by citizens on the political right, who may associate academia with biased, ideologically-driven research.

Data use agreements should prohibit attempts by researchers to re-identify or contact data subjects without the explicit permission of ISSOD (which can review proposals for, e.g., re-identification research).

C. *Commercialization of research*

It is essentially unheard of for data sources to be compensated for the research use of their data or to share in any profits derived from research, whether those data derive from human tissue or online behavior. Nevertheless, some people feel differently about the issue of their data in commercial as opposed to non-profit purposes, and a persistent minority of data subjects believe they should be financially compensated for use of what many regard as their "property."[38] For instance, one of several objections by some to the case of Henrietta Lacks (in which physician-investigators collected and used leftover, quasi-pseudonomized clinical tissue for research without consent, as was then and remains today legal) is that, although those physician-investigators did not profit from the research, others downstream from the initial research did—and handsomely. Some ethicists have argued cogently that sources of passively collected research materials that the source would have produced anyway and which therefore entail no extra effort by or inconvenience to the data subject (e.g., leftover clinical tissue or, in the present context, digital data already produced for the user's own purposes) are not morally owed compensation.[39] Still, in response to this minority public sentiment, the revised Common Rule newly requires research consent to include, where appropriate, a "statement that the subject's biospecimens (even if identifiers are removed) may be used for commercial profit and whether the subject will or will not share in this commercial profit."[40] The regulations do not, however, provide that tissue sources must, or even ought, to share in profits.

Although compensating each data subject for their contribution is unreasonable and infeasible, it does make sense to constrain researchers from commercializing ISSOD-enabled research in ways that would threaten public access to the benefits of that research. This is especially important if, as seems likely, the public will be passively contributing data to ISSOD and bearing some (minimal) risk for doing so. Social Science One, for instance, precludes

---

[38] Fiesler, C. & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media + Society, Jan.–March*, 1–14; Federal Policy for the Protection of Human Subjects, 82 Fed. Reg. 12 (Jan. 19, 2017), at 7164–7165 (reporting qualitative data from survey respondents suggesting that public Tweets are the user's property and researchers should compensate them accordingly); *Federal Register: The Daily Journal of the United States*. Web. 19 Jan. 2017 (describing public comments in response to the proposed revised Common Rule arguing that biospecimens, regardless of identifiability, are "property" and that tissue sources should share in any resulting profits from research with these tissues).

[39] Truog, R. D., Kesselheim, A. S. & Joffe, S. (2012). Paying patients for their tissue: The legacy of Henrietta Lacks. *Science, 337*(6090), 37–38.

[40] 45 C.F.R. § 46.116(c)(7) (2017). Note that because the Common Rule does not apply to research with non-identifiable biospecimens collected for a purpose other than the instant research project, these new requirements technically only apply to research in which researchers intervene or interact with tissue sources to newly collect identifiable or non-identifiable biospecimens.

researchers who are funded by the platform from patenting their results, but it does not preclude other for-profit uses of the data (e.g., writing a profitable trade book about the research results).[41]

D. *Partnering with, versus scraping, platforms*

In general, partnering with platforms provides several potential advantages. First, platforms can incorporate notice of (if not consent to) data sharing with ISSOD into their user-facing materials. Similarly, platforms may be able to and help transmit aggregate or individual results as a gesture of respect. Third, the standard rule in biobank research is that samples from participants who withdraw are destroyed and no longer used in analyses going forward, but that their data is not removed from completed analyses. By analogy, partnering with a platform could entail a mirrored research database that refreshes in real time, enabling deleted tweets to be automatically disappeared from the research platform. (The evolving content of the database would have to be accounted for somehow to enable reproducibility of the analyses.) Conversely, scraping is often frowned upon by both platforms and users, especially if the platform is quasi-closed and community- or purpose-specific, where researchers may be viewed as interlopers. This can undermine trust, setting back the goal of large scale data sharing. The primary risk of partnering with platforms is that it would be important to insulate the research from influence by the platform.

III. Governance

There is not currently consensus on the ethical issues raised in this white paper, and that is unlikely to be achieved in the near future, if ever.[42] Moreover, as always, different research studies will raise different concerns, significantly challenging the feasibility of one-size-fits-all ethics rules. For both reasons, it is wise to focus on process.

What should that process look like? Most university IRBs are unfamiliar with both data sharing or research with online data, leading to both Type 1 and 2 errors in reviewing proposals such as this. Moreover, the vast majority of secondary research with data housed with ISSOD will either be non-human subjects research or exempt from IRB review. Most IRBs will not conduct a substantive review that falls outside of their jurisdiction. Moreover, many aspects of the Common Rule, which IRBs are trained to apply, are a poor fit for ISSOD. The act of data collection and maintenance itself does not meet the regulations' definition of "research,"[43] and as discussed above, most secondary analyses of these data will be exempt from the Common Rule or fall outside of it completely. The Common Rule is further hampered by a weak definition of "identifiable" and a focus on risks to individual data subjects only, as opposed to third parties, groups, or society.

---

[41] Social Science One. (n.d.). FAQ, https://socialscience.one/faq?page=1

[42] *Accord* King, G. & Persily, N. (Aug. 10, 2018). A new model for industry-academic partnerships. https://gking.harvard.edu/files/gking/files/partnerships.pdf; Vitak, J., Shilton, K., & Ashktorab, Z. (2016). Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)* (pp. 941–953). ACM; Michael Zimmer. (Feb. 12, 2010). "Is it ethical to harvest public Twitter accounts without consent?," https://www.michaelzimmer.org/2010/02/12/is-it-ethical-to-harvest-public-twitter-accounts-without-consent/.

[43] Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science, 1*(1), 1–14.

Still, some sort of prospective group ethics review is desirable, at least for some categories of data collection (e.g., prior to scraping a platform) and some categories of secondary research with collected data (e.g., research with highly sensitive and/or highly identifiable data, research investigating a highly sensitive or controversial question, or research that targets vulnerable populations). Social Science One requires university IRB approval or (more likely) determination of non-human subjects or exempt status;[44] prospective peer review that includes a review of the proposal's scientific merit and potential benefits, but also of the "ethical track record" of the PI and the potential costs to data subjects and others;[45] and, assuming the proposal passes peer review, separate ethics review by ethicists appointed by Social Science One with specific expertise in online research ethics. Finally, Social Science One collaborates with an NSF-funded team of information scientists, PERVADE, to provide continuous ethics feedback about the platform's decisions. Something like this body might be engaged to review at least some subset of data collection and data analysis activities. Members of that committee should include not only those who specialize in online communities and digital data but also those who are broadly trained in moral reasoning. ISSOD should also consider include laypersons on the committee (ideally, more than the single community member most IRBs retain). For more on oversight of research when IRB review and the Common Rule are poor fits, see Meyer 2018, pp. 224–227.[46]

IV. Public trust and the need for additional research about public preferences

Whether behavior is morally right or wrong is not dictated by public opinion. However, data sharing initiatives like ISSOD will not be successful if they do not earn the public's trust. This is likely to be one of the biggest obstacles to success. In the U.K., for instance, the National Health Service (NHS) attempted to invoke a social license to justify the extraction of data from medical records, partly for research, unless patients opted out. Public and even professional opposition was strong enough that the program, *care.data*, was shelved. Some scholars have argued that the program failed to secure all necessary aspects of a social license for research, which require that

[44] King, G. & Persily, N. (Aug. 10, 2018). A new model for industry-academic partnerships. https://gking.harvard.edu/files/gking/files/partnerships.pdf incorrectly states that federal research regulations require IRBs to make exempt determinations. In fact, the Common Rule is silent on this question. Although OHRP has historically recommended that this not be left to investigators, https://www.hhs.gov/ohrp/regulations-and-policy/guidance/faq/exempt-research-determination/index.html, during the Common Rule revision process, HHS/OHRP proposed a "decision tool" that would allow investigators (or others, as the institution prefers) to make their own exempt determinations, with impunity, so long as the inputs they enter are accurate. That did not make it into the Final Rule, largely because the agencies ran out of time to develop the tool, but HHS has indicated that it intends to introduce this in the future. Perhaps anticipating this, some IRBs have developed online decision tools, of a sort, through which they permit their investigators to make their own exempt determinations. The proposed practice of investigators making their own exemption determinations is controversial, and until the dust settles, it would be best if IRBs continue to make these determinations for ISSOD projects.

[45] It is unclear how peer reviewers are meant to investigate the "ethical track record" of the PI. Nor is it clear that peer reviewers are sufficiently knowledgeable about the risks of this work to be helpful (much as IRBs are often insufficiently knowledgeable about the scientific merits of proposed research).

[46] Meyer, M. N. (2018). Ethical considerations when companies study — and fail to study — their customers. In E. Selinger, J. Polonetsky & O. Tene (Eds.), *Cambridge handbook of consumer privacy* (pp. 207–231). Cambridge, UK: Cambridge University Press.

data subjects perceive participation to be voluntary and governed by values of reciprocity, non-exploitation, and service of the public good.[47]

There is some existing research investigating perceptions of research use of digital data, and it suggests the challenges ahead. In the aforementioned survey of Twitter users, 65% believed that researchers should not be able to use even *public* tweets without explicit user permission. When asked whether they themselves would agree to allowing a university researcher to use their tweet, 53% said yes, 14% said no, and 33% said it would depend on contextual factors. When asked if they would opt out of having their tweets used in all academic research, 29% said yes and another 25% again said it would depend. When asked which factors would influence their comfort with "a tweet" of theirs being used in research, respondents (n = 268) were most likely to indicate being somewhat or very uncomfortable when: the tweet was from their protected account (75%), no consent was sought (67%), it was a public tweet they had later deleted (64%), the tweet was quoted in published research and attributed to their Twitter handle versus quoted anonymously (56% vs. 27%),[48] researchers also analyzed their public profile information (e.g., username and location) versus researchers not having such information (55% vs. 20%), they were informed after the fact (50%), and if their tweet was one of only a few dozen being analyzed versus one of millions (47% vs. 21%), and a human read their tweet to analyze it versus a computer program doing so (37% vs. 17%). When asked about their overall comfort level with tweets being used in research, only 21% to 27% of respondents said they were somewhat or very uncomfortable. But when asked about their comfort if "your entire Twitter history was used," that number shifted to 49%. When asked if they would want to know that a tweet of theirs was used in a university study, 80% of respondents said yes.

Notably, this survey did not elicit the strength of respondent preferences for consent if such a requirement would hamper or preclude research. A study of patient preferences regarding consent to medical record review found that although most respondents preferred an in-person consent session with their physician, only 13.8% would prefer such research not to occur if written or verbal consent would make the research too difficult to conduct.[49]

Several barriers to public acceptance of large scale, nonconsensual data collection are likely, including an ineffable sense of "creepiness" (especially, for instance, in the case of data collected from platforms that present themselves as private, such as WhatsApp and Facebook Messenger), fear of research, and lack of appreciation of how little is known about important social phenomena (and, hence, the importance of research with big data). Additional research on lay perceptions and preferences is needed for an initiative such as ISSOD to be successful, including: a) research that progresses beyond opinion poll-like surveys and takes an incentive compatible and/or experimental approach to measuring preferences and otherwise elicits preferences in ways that require respondents to acknowledge the trade-offs of data privacy, b) research that investigates perceptions and preferences of users on platforms other than Twitter (e.g., Facebook, Reddit, 4/8chan, Instagram, WhatsApp), and c) research that goes beyond baseline perceptions and preferences and investigates how to communicate initiatives like ISSOD to the public in ways that engage, rather than alienate, the public.

---

[47] Carter, P., Laurie, G. T. & Dixon-Woods, M. (2014). The social license for research: why care/data ran into trouble. *Journal of Medical Ethics, 41*(5), 404–409.

[48] The authors note that respondents may not have realized how easy it is to reidentify a user whose tweet is quoted verbatim, even if the username is omitted.

[49] Kraft, S. A. et al. (2016). A comparison of IRB and patient views on consent for research on medical practices. *Clinical Trials, 13*(5), 555–565.

Appendix: Data Access Provisions of Major Data Repositories

NHANES

Participants

- Nationally representative sample of 5,000 American children & adults/year
- Oversamples people over 60 years, Hispanics & African Americans

Data collected

- Survey (demographic, socioeconomic, dietary, health-related questions)
- Exam (medical, dental, physiological measurements, lab tests, genomic)
- Smoking, alcohol consumption, sexual practices, drug use, physical fitness & activity, weight, dietary intake, reproductive health (e.g., use of oral contraceptives & breastfeeding practices)
- SNP data (after 2003)

Consent language

- Survey consent:
  - "When we allow other collaborators & researchers to use survey data, we protect your privacy. We assign code numbers in place of names & never reveal other facts that could directly identify you."
  - "What you tell us, your exam findings, & samples you give are a rich resource for health science. Many Federal agencies, universities, & other public & private groups use NHANES data. They use it to help find new cures & treatments for diseases & disabilities. The aim is to enhance the health of all people."
- Biospecimens consent:
  - "Researchers from Federal agencies, universities, & other scientific centers can submit proposals to use the stored specimens. These proposals will be reviewed for scientific merit and then by a separate board that determines if the study proposed is ethical. The NHANES program will always know which samples belong to you, but we will not give other researchers any information that could identify you."
  - "What genetic studies will be done with the samples? Science and medicine are continually advancing. New tests and new ways of looking at results will be developed in the future. We can't predict what test will be done or what the results will mean for your health."
  - Re: Risks: "There may also be a risk that some people may use the information from the genetic studies to exaggerate or downplay differences among people. The ethics board that will review all studies using these samples will attempt to prevent any misuse of the information gained from the NHANES DNA samples."

Data tiers

1. Open/public/unrestricted
2. Some small anonymized genomic datasets not linkable to any other datasets are available on request w/o IRB review (because no human subjects involved w/non-identifiable data) through Data Use Agreement/release form
3. Restricted: Data that could compromise confidentiality of survey respondents or institutions or is "sensitive by nature"
   - All geographic data below national level
   - Exact interview & exam dates
   - Most genomic data

Processes for accessing restricted data

1. NHANES Data Support Agreements: initiated by NHANES w/identified experts under signed agreement to assist in data collection or processing

2. NHANES QA/QC Collaborator datasets: Inter-agency QA/QC dataset agreement w/current NHANES collaborators 3 mos prior to public release
3. NHANES Special Use Data Agreements: Under special circumstances NCHS enters into agreement w/Collaborators, CDC employees, or any researcher to provide limited non-public special dataset; request reviewed by Director & Confidentiality Officer
4. NCHS RDC applications: Requests by "any researcher" to match NHANES data to external data sources; to analyze lower level geography or indirect identifiers; for access to non-public release data which are the basis of published analyses, e.g., published analyses based on one year of data:

    - Application (example) submitted to Research Data Center (RDC), judged by: Well-defined research question addressing public health concern (consistent w/consent scope), explanation of why restricted variables are necessary, technical feasibility, disclosure risk (based on variables requested, remote vs. on-site access, analytic plan including stats methods)
    - Review Committee (including Analyst, Data System Rep(s) & Confidentiality Officer) may approve, disapprove, or (often) R&R
    - Avg review: 6-8 weeks
    - Approval doesn't guarantee all output generated by analysis will be released; output is reviewed for disclosure risk & will be suppressed if necessary
    - Completion of online Confidentiality Orientation & 100% score on quiz
    - Signed Confidentiality Agreement (e.g., use data only for approved purpose; no attempt to re-ID or discover suppressed cells; no attempt to introduce any additional data through statistical programming or otherwise; don't use data in way that poses additional risk to respondents; if you can inadvertently deduce small cells (<5) or an individual–level-information, don't share that information with anyone or in any publication and immediately notify RDC)
    - Signed Designated Agent affidavit
    - Review of Disclosure Manual
    - Submission of fee
    - Manuscript must be submitted to RDC Analyst prior to submitting for publication
    - Appears that NCHS Ethics Review Board (ERB) reviews all proposals to analyze restricted (i.e., identifiable) data after RDC approves (see here); biospecimens consent refers to such ERB review but survey consent doesn't, even though some survey data are restricted/identifiable. No local IRB review appears to be required.

National Longitudinal Survey of Youth (NLSY)

Participants

- Nationally representative birth cohorts
- Incidentally includes prisoners; their data is not specially protected or restricted but questions about illegal activity are skipped in interviews if prisoner cannot enter answers directly into laptop (i.e., if prisoner would have to answer out loud

Data collected

- Education, training, cognitive tests; age, gender, geographic residence & neighborhood composition; household composition; race, ethnicity & immigration; computer & Internet access
- Sensitive questions: income & assets, religion, relationships w/parents & family, sexual experiences, abortion, drug & alcohol use, criminal activities, homelessness, runaway episodes
- Geocode data: county, metropolitan statistical area, ZIP Code, census tract and block, & latitude & longitude of residence

Consent language

- BLS "will hold your responses in confidence & will not disclose them in identifiable form w/o your

informed consent"

- "Some of your answers will be made available to researchers at the BLS & other government agencies, universities, & private research organizations through publicly available data files. These publicly available files contain no personal identifiers…"
- "Some researchers are granted special access to data files that include geographic information, but only after those researchers go through a thorough application process at the BLS. Those authorized researchers must sign a written agreement making them official agents of the BLS & requiring them to protect the confidentiality of survey participants. Those researchers are never provided w/the personal identities of participants."
- Respondents are advised at start of interview they can choose not to answer any questions that they prefer not to answer; respondents may also enter answers to sensitive questions into laptop directly rather than telling interviewer

Data tiers

1. Open/public/unrestricted
2. Restricted data
    - Geographic data: state, county, metropolitan statistical areas of residence, country or state/county of birth, state of residence, state, country, or region of world in which respondent's parents & grandparents were born
    - Name & locations of colleges & universities attended
    - Dates of birth, marriage, divorce, death, school attendance, etc. are month-year only in public files; specific dates restricted
    - Income & assets variables are public but topcoded (see p. 157 here)

Processes for accessing restricted data

- Project-specific application (new use of data requires new application):
    - Clear statement for general audience of project (max 4 paragraphs) & explanation why geographic variables are necessary
    - Project must further mission of BLS & the NLS program "to conduct sound, legitimate research in the social sciences" (see p. 157 here)
- Researcher must work/study at U.S. institution (needn't be U.S. citizen)
- Non-negotiable Letter of Agreement pledging to adhere to BLS confidentiality policy, signed by BLS office & official at requestor's institution w/authority to enter into legal agreements on institution's behalf
- Each individual authorized to access data under Letter of Agreement signs non-negotiable BLS Agent Agreement designating him as unpaid agent of BLS & requiring him to take certain security & confidentiality measures
- Geocode agreements last 1 year for students, 3 years for most faculty; term for adjuncts, visiting faculty, postdocs, etc. depends on how long they'll stay at institution; extensions granted on case-by-case basis; if researcher leaves institution before end of agreement, his BLS agent agreement terminated & if no one else is authorized at institution, Letter of Agreement also terminated
- Average 6-8 weeks after application submitted until legal docs signed at BLS
- After letters executed, can order appropriate Geocode CD
- Research outputs subject to review by BLS to ensure compliance w/confidentiality requirements
- Facilities where Geocode data used subject to BLS inspection to ensure compliance w/Letter of Agreement
- May not link geocode data w/individually identifiable records from any other dataset
- Penalty for misuse
- Process to access original cohorts geocode data slightly different

Framingham Heart Study (FHS)

Participants

- Three generations comprising 15K clinically & genetically well-characterized participants
- Since 1994, two groups from minority populations added

Data collected

- Medical records, specimens (DNA, urine, blood & blood products), physical examination, blood tests, electrocardiogram, genetic data

Consent language

- "Data & DNA will be distributed to . . . other qualified researchers interested in the genetics of heart, lung & blood diseases & other diseases & health conditions. The scientists from these laboratories will be given the DNA without any potentially identifying information."

Data tiers (doesn't seem to be any public data)

1. Much FHS data is available through other repositories according to their access policies: geno/phenol data via dbGaP; phenol data via BioLINCC
2. Authorization required: studies of existing data, collection of new data (from participants or existing samples), images or medical records

Processes for accessing restricted data

- Applicant first requests & receives an account
- Application: background & rationale, specific aims, methods, data requested
- Application routed to appropriate committee(s):
    - Executive Committee (requests for participant contact)
      Lab Committee (requests for Framingham bio-specimens for non-genetic research)
    - DNA Committee (requests for genomic data not included in dbGaP; requests for Framingham DNA or other bio-specimens for genetic research)
    - Research Committee (requests for clinical data not available in BioLINCC)
- Application review criteria:
    - Does proposal complement Framingham's research scope?
    - Is collaboration w/a Framingham investigator planned?
    - Does proposal require unique characteristics of FHS cohort(s)?
    - Does proposal put minimal demand on FHS resources?
    - Does proposal show proof of resources for conducting project?
    - Investigators strongly encouraged to use Omni data/biospecimens
    - Local (i.e., investigator) IRB approval required of all approved data &/or material distributions. ("Although Framingham data is de-identified, FHS is a study of a single community & hence one's identify can be more easily ascertained, even if traditional identifiers are removed.")
- If proposal involves new participant contact or additional specimen collection, Observational Studies Monitoring Board (external to FHS) also reviews proposal.
- Proposals eligible for expedited review (w/in 2 weeks) if request only existing data or new phenotypic data w/o participant contact
- Among application questions: Will this project generate new individual level data on Framingham participants? For example, sets of analyzable data from individual level measurements, images or lab specimens.
- Fee from $3-$10,000.
- Framingham Data & Materials Distribution Agreement (DMDA) required of all approved data &/or material distributions:
    - No attempted re-ID

- Data & materials not used for any purpose contrary to consent; must consult w/Study Investigators re: consent terms & conditions
- No use beyond approved research project
- No further sharing
- Advance notification of publication, etc.

Wisconsin Longitudinal Study (WLS)

Participants

- NIA-sponsored longitudinal study of a random sample of 10,317 1957 WI H.S. graduates
- Broadly representative of white, non-Hispanic Americans w/at least high school education; minorities not well represented

Data collected

- Genotypic data from Illumina HumanOmniExpress array w/ quality metrics (supplied by U of Washington Genetic Analysis Center)
- Genotype imputation data w/genotypes imputed to 1000 Genomes Project phase 3 reference panel
- Life course, intergenerational transfers & relationships, family functioning, physical & mental health & well-being, morbidity & mortality from late adolescence through 2011, social background, youthful aspirations, schooling, military service, labor market experiences, family characteristics & events, social participation, psychological characteristics & retirement, attractiveness rating, relative BMI

Data tiers & processes for access

1. Most WLS data are publicly downloadable & free w/some variables removed (geography, birth & death months, data re: friends & relationship w/other participants, names of colleges) or top- or downcoded to prevent ID of outliers (many monetary values, height, weight, BMI). They only require downloaders of the public data (Level One) to provide them with their name, a valid email address, geographic location, and academic area of specialty.
2. Subject to approval: email WLS staff, explain why level-1 data insufficient, sign confidentiality agreement There is a smaller subset of our data that is defined as either sensitive or has a marginally higher risk of identifiability (Level Two). Access to that data is granted after the researchers provide us with a statement on why the public data is not sufficient to answer their research question, a copy of their CV, and proof of human subjects training.'
3. Accessible thru secure server: extremely sensitive data (e.g., genetic data, audio recordings) available w/research plan, local IRB approval, signed DUA, analyses conducted on WLS secure server, genetic data also approved by WLS Genetic Advisory Board. Finally our most sensitive data, including the DNA data and some geographic codes, is a Level Three request. These data require a fully-executed Data Use Agreement (DUA), and proof of IRB (or equivalent) approval from the researcher's home institution. Once the data are licensed through the DUA we provide users with a copy of the data to analyze at their home institution.
4. Accessible only in physical coldroom: SS earnings & benefits

Health and Retirement Study (HRS)

Participants

- Longitudinal panel study of representative sample of 20K Americans supported by NIA & SSA

Data collected

- In-depth interviews, health data, genetic data

Data tiers & processes for access

1. Most survey data is unrestricted & publicly available w/
    a. Registration (name, valid email, phone, organization, state; type of organization, including none or other; primary role in org—faculty, students, staff, other; highest degree; whether you're working alone, w/collaborator w/name or under supervision w/name; primary research area, whether you've published using HRS data)
    b. Agreement (via registration) to conditions of use (e.g., no attempted ID, no data transfer to 3rd parties)
2. Sensitive health data (e.g., biomarkers, prescription drug data, diabetes study, cognition & behavior phenotypic data, telomere data, memory): available from public portal w/
    a. Sensitive DUA (no re-ID, store & use data in secure environment, cite HRS & provide copies of publications to HRS, guidelines for frequency & magnitude tabulations, publish only aggregate stats)
    b. Verification of identity & institutional affiliation (unclear if independent/citizen sci ok)
3. Restricted data (SSA admin data, VA health data, national death index cross-year cause of death, CMS cross-ref file, geographic info, pension estimation program & database, industry & occupation data, cancer site, Part D Plan info, interview date, date of death, cross-wave race & ethnicity, college tuition imputations) available in 2 ways w/different data security plans & confidentiality agreements (flow chart; text narrative):
    a. MiCDA Enclave Virtual Desktop Infrastructure (VDI): submit application (for each participating institution) including:
        - Letter from Dept. Chair (for students)
        - Proof of local IRB review (exempt, expedited, or full)
        - 1-3 p. Research proposal (what restricted variables you need & why, study team details, project goals)
        - Data order form
        - MiCDA VDI Data Security Plan
        - MiCDA Data Enclave Acceptable Use Policy
        - ISR Pledge to Safeguard Respondent Privacy
        - Confidentiality Agreement
        - [Certain data merges can only be performed by visiting Enclave in person w/additional signed Confidentiality Agreement Restricting Disclosure & Use of Data from the MiCDA Enclave]
    b. Traditional licensing agreement (required for SSA, CMS linkages):
        - Proof of IRB approval (expedited or full—apparently NOT exempt): once your application complete & data security plan is acceptable, you submit proposal to your IRB &/or your institutional Contracting Authority for review; completed reviews are submitted to HRS
        - 1-3 p. Research proposal (what restricted variables you need & why, study team details, project goals)
        - Data order form
        - Data security plan (see also this checklist)
        - CV(s)
        - Institutional Federalwide Assurance (IRB registered w/OHRP)
        - Proof of current federal funding (primary penalty for breach is notification via NIA to your funding agency)
        - Confidentiality Agreement w/institutional countersignature
4. Access to SSA administrative data & CMS research data is "more involved"; contact HRS for details (seems to involve encrypted physical media to qualified researchers?)
5. Genetic data: genetic data products (candidate gene & SNP files, genptype data, exome data) from 20K genotyped respondents: available after applying first for dbGaP access to controlled data & then submitting to HRS a Genetic Data Access Use Agreement & Genetic Data Order Form
6. Additional restrictions on merging restricted data files w/each other (e.g., restricted SSA admin records may not be merged w/geographic data)

Completed application goes for final review by full Data Confidentiality Committee (DCC). If approved, HRS PI signed the Confidentiality Agreement & data is made available. Restricted data agreement must be renewed annually. Any change in circumstances requires modification to agreement. Periodic inspection of site by HRS possible. On termination. Licensing agreement users mist return of destroy data (if latter, must provide counter-signed certification of destruction).

Notes about IRB review

- HRS considers public data to pose a sufficiently low risk of re-identification that it is exempt from IRB review
- This memorandum from the HRS PI to IRBs makes clear that restricted data poses a sufficiently high risk of re-identification that exemption is inappropriate, that the risk for the IRB to consider is that of re-identification, & that IRB review should therefore center on the "Restricted Data Protection Plan, and those aspects of the Research Plan that deal with issues of respondent anonymity and data security, if any. By the time they reach you, HRS will have approved these Plans. But we ask for your review because you will be better able to judge the extent to which, in your institution's physical and computing environment, whether the Plans are adequate to ensure participant anonymity and limitation of access to the restricted data to the persons specified in the agreement." [NB: This partially contradicts what is said about the VDI pathway, which is that your IRB may indeed find your research exempt.]
- Certification of IRB review form: Certify that 1) your IRB has an active FWA and 2) "Our Institutional Review Board/Human Subjects Review Committee has reviewed, according to its standards and procedures for live human subjects, and approved, the Restricted Data Protection Plan (and those portions of the Research Plan that deal with respondent anonymity and data security, if any), approved by the Health and Retirement Study, of the Restricted Data Investigator above; and has approved those plans."

Possible penalties for breach of Agreement for Use of Restricted Data

- Denial of future access to HRS data
- Notification to your institution's scientific integrity office & request for sanctions
- Notification to your current funding agency w/recommendation that all current funds be terminated & all future funds be denied
- Other remedies available at law

General Social Survey (GSS)

Data collected

- Demographic, behavioral, & attitudinal questions, plus topics of special interest (civil liberties, crime & violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, stress & traumatic events)

Data tiers (individually identifying info, e.g., name, address, never provided)

1. Public use files (include no geocoded data)
2. GSS geographic identification code files (state, primary sampling unit, county, & Census tract) available to researchers under special contract w/NORC

"Sensitive Data": any data that might compromise anonymity or privacy of respondents. Specifically, any data file that, for either individuals, or families, includes:

a) Identification numbers or demographic information (such as month & year of birth, age, ethnicity, occupation, industry, gender, etc.);
b) Geographic identification of areas smaller than Census Division, including, but not limited to state, county, minor civil division, primary sampling unit (PSU), segment, city, place, zip code, tract, block numbering area, enumeration district, block group, or block;

c) Any variables or fields derived from the data mentioned in items a)-b) above, including data linked to a GSS dataset using the data mentioned in items a) & b) above as linking or matching variables.

Process for access

- Research Plan: describe which datasets & variables you want; must be project-specific
- CV for each researcher
- Sensitive Data Protection Plan (data security plan)
- Human Subjects Review from your Institution, using Sensitive Data Protection Plan as part of the application for approval; may result in approval *or waiver*
- Contract for Use of Sensitive Data:
    - If Investigator isn't fulltime permanent faculty at institution, requires co-I who is fulltime, PhD-level faculty member
    - Data must be returned or destroyed
    - Investigator(s) & institution must assume liability up to 100K for any violations by any person at the institution
    - Signed by representative who can enter into contracts on behalf of institution
    - If investigator leaves institution, agreement terminates
- Process can take several months
- Fee: $750

Criteria for review

- GSS takes its promise of anonymity to its respondents very seriously and this is the basis for the contract process
- GSS aims to make high-quality data easily accessible to scholars, students, policy makers, and others, with minimal cost and waiting

dbGaP

Participants & data collected

- Genotype & phenotype data collected via many studies under a wide range of consent terms

Data tiers

1. Open: available to anyone w/no restrictions
2. Controlled: allows download of individual-level genotype & phenotype data that have been de-identified (i.e., no personal identifiers, such as name)

Process for access

- PI (must be registered by your institution as a PI in your eRA account) & institutional Signing Official (both w/NIH eRA Commons accounts) co-sign request for data access
- Statement summarizing proposed research use for the requested data
- List of collaborating investigators at same institution (collaborators at other institutions must submit own requests)
- Submission of request constitutes agreement to Data Use Certification (e.g., use limited to proposed project, no re-ID or re-contact, no further distribution of data)
- Agree to Code of Conduct
- Adhere to data security measures

Criteria for access

- Data access & use must be consistent w/participants' intent as reflected in consents; datasets are placed in different "consent groups," e.g.:
    - General research use: use limited only by model Data Use Certification

    - Health/medical/biomedical research only: no ancestry, no possibly stigmatizing research, no non-health research
    - Non-profit use only
- Some datasets require local IRB approval; others (e.g., often, general research use) don't
- Access to all controlled datasets requires approval of an NIH Data Access Committee (DAC), which looks to ensure proposed plan matches limitations (if any) of consent group; no additional ethics review is involved