Research Ethics in Big Data

Marilyn J. Hammer, PhD, DC, RN

The ethical conduct of research includes, in part, patient agreement to participate in studies and the protection of health information. In the evolving world of data science and the accessibility of large quantities of web-based data created by millions of individuals, novel methodologic approaches to answering research questions are emerging. This article explores research ethics in the context of big data.

Hammer is the director of research and evidence-based practice in the Department of Nursing at Mount Sinai Hospital in New York, NY.

No financial relationships to disclose.

Hammer can be reached at marilyn.hammer@mountsinai.org, with copy to editor at ONFEditor@ons.org.

Keywords: big data; research guidelines; patient autonomy

ONF, 44(3), 293–295.

doi: 10.1188/17.ONF.293-295

George Orwell wrote the novel *1984*, in which every individual was under constant surveillance in a futuristic world. Now, 33 years past the setting of that novel, people are under considerable surveillance far exceeding that described in the story line of *1984*. Well beyond being captured on security cameras at every turn and occasional drones, we actively disclose where we are, what we are doing, what we like and dislike, and who we are connected to with virtually every tap of our e-devices. The amount of data created daily through technology has become so exponential (Mittelstadt & Floridi, 2015) that it is deemed an important commodity and has been termed *data capital* (MIT Technology Review Custom, 2016). We have essentially become our own best historians in real time, leaving a highly accessible data trail. These "digital footprints" are being used en masse as a rich source of research data (Bietz et al., 2015). All these data elements are freely available without our consent, so where does this leave us in terms of research ethics?

The Impetus for the Ethical Conduct of Research on Humans

Research ethics guidelines were initially based on the horrific treatment of human participants without their consent, such as in the Tuskegee experiment and in the atrocities of World War II. Historic information about the establishment of guidelines for the ethical conduct of research were detailed in a prior research ethics article (Hammer, 2016). Keeping in mind the historic infractions that inform the current research guidelines provides a solid, although sometimes narrow, framework with which research studies are conducted today. In brief, based on the Nuremberg trials, the 1947 Nuremberg Code, which introduced new research ethics principles, was written (National Institutes of Health, n.d.), followed by the Declaration of Geneva in 1948 (Fischer, 2005). Orwell's novel was published the following year but not with research ethics in mind. Current adherence to the *Belmont Report* holds researchers to the highest standards of ethics to ensure not only that study participants are treated respectfully and that studies provide more benefit than potential risk, but also that personal health information is protected (Office for Human Research Protections, 1979, 2016).

The protection of health information to prevent individuals from losing health insurance or employment, or incur other life losses because of personal health information disclosures became central with the 1996 Health Insurance Portability and Accountability Act (HIPAA) (Colorafi & Bailey, 2016). HIPAA was strongly embraced.

Technologic Advancements

A high level of protection can encumber research endeavors needed to understand diseases and how to manage them. Sometimes, research and clinical practice need to be concurrent in urgent or, more often, in emergent situations like epidemics or pandemics. The guidelines of conducting research to advance science or health, maintaining strong scientific rigor, and ensuring that the potential benefits of studies outweigh the risks (Gobat et al., 2015) are easily upheld in epidemic and pandemic situations. In such urgent or emergent events, maintaining patients' autonomy regarding informed consent becomes a challenge. In addition, in these critical situations, third-party, waived, and deferred consent have become acceptable to obtain the information through research that can quickly inform practice to optimize outcomes (Gobat et al., 2015). In addition, the use of large datasets or combinations of datasets can be tremendously beneficial in epidemic or pandemic situations. Beyond epidemics and pandemics, electronically captured health-related data can inform practice. One study revealed associations between Internet searches on symptoms that were often too subtle to warrant seeking health care and actual diagnoses—in this case, pancreatic cancer—finding that symptom searches often precede diagnoses (Paparrizos, White, & Horvitz, 2016). In that study, researchers used specific criteria to identify individuals who were likely recently diagnosed with pancreatic cancer and then compared those findings to the same individuals' searches related to symptoms. This method of detecting an aggressive and difficult-to-treat cancer early is an exciting advancement in oncology research, with the potential for improved outcomes. The evaluation of 9.2 million searches yielded a majority of negative findings, but with a number that large, positive findings were also significant (Paparrizos et al., 2016). Individuals were identified through codes and were included in the study without consent, which is ethically questionable. How this form of research could ultimately lead healthcare providers to prediagnose and solicit individuals based on symptom searches is unclear. In addition, the adoption of such a healthcare pathway is still unknown; however, based on current practices of sharing information, it will likely become acceptable.

Electronic Phenotyping

Tracking individual behaviors is relatively easy, and many for-profit companies have exploited personal information with great financial payoffs. In addition to posting about daily life events through multiple venues, many people self-track their own health (Bietz et al., 2015). With smartphones and specialized watches and wristbands, we can monitor every step and, for those so inclined, calculate calories and monitor blood pressure and blood sugar. The data are downloadable to personalized computers and can be shared with healthcare providers and even with the public at large.

Capturing such individual behaviors through technology has been termed *electronic phenotyping* (Cato, Bockting, & Larson, 2016). On a more analytical level, electronic phenotyping includes the use of information from electronic health records in predictive modeling. This innovative technology can accurately predict which patients have the greatest risks for certain adverse events during and following treatment, allowing for early interventions and, potentially, improved outcomes. Creating such models entails complex statistical analyses to assess a vast amount of variables per patient. This research has initially been conducted retrospectively with an exemption from institutional review boards. In clinical practice, research is not regarded as research until investigators start evaluating its effectiveness. Obtaining patient consent may be challenging depending on the study design and further complicated by the ability of the patients to fully understand the study.

Predictive modeling also provides information to the healthcare team that facilitates discussions with patients based on their health issues and/or demographics (Cato et al., 2016). Obtaining consent for the clinical usage of such data is also unclear and may even be considered profiling, based on beneficence (Cato et al., 2016), which is the principle of benefits outweighing harmful risks (Miracle, 2016). Profiling patients also has a negative connotation and can potentially lead to harm or the perception of being harmed. For example, if a man with a history of substance abuse seeks healthcare services from a new provider, and the provider brings up his history prior to him volunteering it, the patient may feel judged and as though his privacy has been invaded (Cato et al., 2016). Having maximum information to provide the best care while protecting people's privacy has been a challenge to the healthcare system for many years. Electronic phenotyping is both a contributor to improving health outcomes and a potential means of breaching privacy.

Conclusion

The term *big data* still has not been clearly defined (Mittelstadt & Floridi, 2015). Sources of large datasets are varied and include electronic health records, combined abstracted variable from multiple institutions, large insurance- or organization-based entities, and numerous websites. In addition, much information is stored in a virtual

cloud system with complex security (Suciu et al., 2015).

Because younger generations are born into a world of vast technology and are generations distant from the horrific infractions imposed on human beings under the guise of research, the impetus for the protection of private information may fade. The paradigm of research ethics may be shifting; keeping individuals truly de-identified is becoming increasingly difficult with genomic analyses and data gathering with every key stroke. The thought of Big Brother watching was abhorrent in *1984*. In 2017, however, many willingly contribute to surveillance. How this evolution is occurring and how research is conducted are being unveiled right before our eyes. How we maintain the ethical principles of research remains to be seen.

References

Bietz, M.J., Bloss, C.S., Calvert, S., Godino, J.G., Gregory, J., Claffey, M.P., . . . Patrick, K. (2015). Opportunities and challenges in the use of personal health data for health research. *Journal of the American Medical Informatics Association, 23*(E1), E42–E48. doi:10.1093/jamia/ocv118

Cato, K.D., Bockting, W., & Larson, E. (2016). Did I tell you that? Ethical issues related to using computational methods to discover non-disclosed patient characteristics. *Journal of Empirical Research on Human Research Ethics, 11,* 214–219. doi:10.1177/1556264616661611

Colorafi, K., & Bailey, B. (2016). It's time for innovation in the Health Insurance Portability and Accountability Act (HIPAA). *JMIR Medical Informatics, 4*(4), E34. doi:10.2196/medinform.6372

Fischer, B.A. (2005). A summary of important documents in the field of research ethics. *Schizophrenia Bulletin, 32,* 69–80. doi:10.1093/schbul/sbj005

Gobat, N.H., Gal, M., Francis, N.A., Hood, K., Watkins, A., Turner, J., . . . Nichol, A. (2015). Key stakeholder perceptions about consent to participate in acute illness research: A rapid, systematic review to inform epi/pandemic research preparedness. *Trials, 16,* 591.

Hammer, M. (2016). Informed consent in the changing landscape of research. *Oncology Nursing Forum, 43,* 558–560. doi:10.1188/16.ONF.558-560

Miracle, V.A. (2016). The *Belmont Report*: The triple crown of research ethics. *Dimensions of Critical Care Nursing, 35,* 223–228.

MIT Technology Review Custom. (2016). *The rise of data capital.* Retrieved from http://files.technologyreview.com/whitepapers/MIT_Oracle+Report-The_Rise_of_Data_Capital.pdf

Mittelstadt, B.D., & Floridi, L. (2015). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics, 22,* 303–341. doi:10.1007/s11948-015-9652-2

National Institutes of Health. (n.d.). *The Nuremberg Code.* Retrieved from https://history.nih.gov/research/downloads/nuremberg.pdf

Office for Human Research Protections. (1979). Protection of human subjects; *Belmont Report*: Notice of report for public comment. *Federal Register, 44,* 23191–23197.

Office for Human Research Protections. (2016). The *Belmont Report.* Retrieved from http://bit.ly/2pa49CB

Paparrizos, J., White, R.W., & Horvitz, E. (2016). Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice, 12,* 737–744. doi:10.1200/jop.2015.010504

Suciu, G., Suciu, V., Martian, A., Craciunescu, R., Vulpe, A., Marcu, I., . . . Fratu, O. (2015). Big data, internet of things and cloud convergence—An architecture for secure e-health applictions. *Journal of Medical Systems, 39*(11), 141. doi:10.1007/s10916-015-0327-y