

The Ethics of Using De-Identified Medical Data for Research without Informed Consent

Prag Batra, BS and Arthur L. Caplan, *

Keywords: de-identification, privacy, medical data, bioethics, confidentiality

INTRODUCTION

A recent publication by Ludvigsson et al¹ attempted to explain and justify the nature of health registries in Nordic countries. These registries contain de-identified medical information from each of the individuals who interact with the nationally-run healthcare system and are used for research and quality improvement purposes. According to current laws in these countries, individual informed consent is generally not required for large-scale, registry-based studies that are deemed ethical by an ethics committee (EC). In Nordic countries, these regional ethics committees (RECs) essentially perform the role of IRBs for human subjects research including clinical research but have less formal regulation to guide their practice.² As this holds similarities to laws in other countries such as the United States, this is an issue of global relevance. Ludvigsson et al. made several philosophical arguments to justify the option of informed consent, including the social contract theory of government, the ability of a representative ethics committee to serve as a proxy for individual interests, the principle of social justice, and the principle of utilitarianism.

Here, we present an alternative point of view to these arguments made in support of the use of de-identified medical data in the form of health registries for research purposes without informed consent.³ Though the argument Ludvigsson et al. puts forth specifically focuses on national health registries in Nordic countries, regulations in the US also permit the use of de-identified health information in research without informed consent. As such, the ethicality of using de-identified medical data for research studies without informed consent is relevant both in the US and abroad.

Argument and Response

The first argument posited by Ludvigsson et al involves the social contract theory of government. This theory states that individuals who are part of a given society have a set of moral obligations according to a contract, or agreement, between the individuals of that society which allows the society as an entity to exist. By choosing to remain part of a society, individuals are therefore agreeing to these obligations. In the case of Ludvigsson et al's argument, individuals have an assumed agreement to contribute their personal data for research in exchange for receiving free or reduced-cost healthcare from the state. Consequently, participation is compulsory and informed consent is not necessary. This parallels other government

*Prag Batra and Art Caplan, PhD, New York University School of Medicine, New York, NY

requirements that aim to support society in exchange for the benefits of living in the society, such as drafting individuals to serve in the armed forces and requiring individuals to pay taxes.

One practical caveat to this argument involves individuals who obtain private health insurance in the US. In this case, an argument could be made that because these individuals' healthcare is not paid for by government-funded insurance such as Medicare or Medicaid, these individuals are not subject to the same agreement to contribute their data as those whose healthcare is provided by the state. A similar argument could be made for relatives of individuals who reside in a different country. In this case, certain medical data about an individual who resides in a given country may be used to infer medical information about the individual's family members or relatives. Examples of this may include hereditary medical conditions or individual DNA sequence data. If the individual's information is then used in a research study, this could also lead to indirect inclusion of relatives' medical data in that research study, yielding potential violations of the relatives because they are not subject to the same social contract theory. Furthermore, their information is being used in research studies without their informed consent.

In each of these cases, however, Ludvigsson et al's argument falls short if we consider individual privacy as a fundamental human right. In this case, a government or other agency's use of personal data without permission represents a violation of this right, whereas requiring individuals to serve in the armed forces or pay taxes does not. A counterargument may posit that because the medical data stored in these registries is "de-identified" and thus cannot be used to uniquely identify an individual, storage and use of this data without consent does not violate an individual's right to privacy. While this argument appears sound, there are two important technical and philosophical considerations: (1) whether an individual's medical data can be truly de-identified, and (2) whether new technologies in the future might enable re-identification of data that was believed to be de-identified. For instance, given that DNA sequence data is unique for each individual, creating a registry of this data on its own could conceivably be linked back to specific individuals in the future, even if this cannot be done at present. Furthermore, because there is greater similarity between genome sequence data of closely related family members, family trees could conceivably be created using the genome sequence data of individuals in a population-level registry, leading to potential identification of the genome sequence data of specific families that are outliers in some way. A similar technique could be applied to identify minority groups with known patterns in their genome sequence data. Combining these data analyses with other information such as census reports or demographic surveys could enable researchers to further identify specific groups and individuals from "de-identified" data, leading to privacy violations. From a philosophical standpoint, if we cannot truly de-identify medical data while retaining sufficient accuracy to conduct scientifically valid research studies, any use of data for research studies could potentially be used to personally identify individuals and thus would require informed consent in order to protect privacy. As a result, Ludvigsson et al's argument that individuals have an assumed agreement to contribute their personal data for research in exchange for free or reduced-cost government healthcare falls short if we value privacy rights and acknowledge that individual medical data cannot be truly de-identified.

A second argument posed by Ludvigsson et al. stipulates that research study approval from an ethics committee may replace individual informed consent regarding the use of medical data in a national registry because ethics committees are believed to represent the general public. While this argument is plausible in cases where there is unanimous public agreement regarding use of registry medical data for clinical research, it falls short when there is disagreement. In such cases, an ethics committee's blanket statement concerning medical data in a registry that includes data from individuals who do not agree with the ethics committee's position would violate their autonomy. Therefore, preserving individual autonomy would require us to ask each individual whether he/she approves of the use of his/her data for a particular research

study, which is essentially informed consent. As such, Ludvigsson et al's argument that ethics committee approval may replace individual informed consent for the use of medical registry data in the context of clinical research only holds weight if public opinion surrounding a particular study is unanimous.

A third argument posed by Ludvigsson et al. involves the principle of social justice. Here, social justice connotes a fair distribution of wealth, opportunity, and social privileges among the members of a society. The argument is as follows: Informed consent may be especially challenging to obtain from high-risk populations. Informed consent protocol for a given research study would therefore exclude these populations, resulting in selection bias and publication of research findings that may not be relevant to these populations. The unequal publishing of research findings relevant to these populations could lead to further marginalization of these populations, further perpetuating social inequalities. This represents a violation of the principle of social justice. A potential counterargument involves violation of individual autonomy in research studies performed without informed consent. This may represent a violation of social justice if informed consent is obtained from some individuals but not others if we consider the opportunity to exercise one's individual autonomy through the informed consent process to be a social privilege. However, this counterargument does not apply if informed consent is not obtained.

A fourth argument posed by Ludvigsson et al is a utilitarian argument. In this case, utilitarianism refers to the principle of maximizing utility, or producing the greatest benefit to the greatest number of individuals. Specifically, requiring informed consent would drastically reduce the number of study participants and thus statistical power of national population-level studies. Furthermore, the cost required to obtain consent from millions of individuals as part of a national health data registry would be exorbitant. Because this research has the potential to benefit many of the individuals whose data are being used, requiring informed consent would conflict with the principle of utilitarianism because it would prevent such research from being conducted. One counterargument here involves the nature of research. Given the inherent uncertainty of research studies, it is possible that the results of large-scale research studies using national registries would not yield results that benefit many of the individuals whose data was used. Additionally, while conducting large-scale research studies may maximize benefits to society, it may not maximize benefits to the individuals whose data was used to conduct these studies. If there is a greater risk of harm than benefit to some individuals and we choose to conduct the research without allowing these individuals to choose whether their data is used, we set a dangerous precedent that justifies harming a few individuals in order to benefit society. This also represents a potential violation of the research ethics principle of favorable risk-benefit ratio, which states that the potential benefits of a research study should be proportional to or outweigh the potential risks, if we assume that this principle should apply to each individual participant as well as to society as a whole. One possible response to this objection is that if each research study must be approved by an ethics committee, it would be this committee's responsibility to ensure that the research would not result in harm to a few individuals in order to benefit the many. Thus, Ludvigsson et al's argument is sound if we value the principle of utilitarianism over individual autonomy and assume that large-scale research studies are likely to confer the maximum benefits to society. Indeed, from a utilitarian point of view the potential harms to the people whose data are used are quite low, particularly when protections are in place, while the potential benefits to society are quite large. This tends to tip the scale in favor of utilitarian-based violations of privacy rights.⁴

Discussion and Concluding Remarks

Ludvigsson et al's arguments regarding social justice and utilitarianism appear to be sound. However, Ludvigsson et al's arguments regarding the social contract theory of government and ethics committees

serving as a substitute for individual informed consent appear to be flawed. In order to determine whether use of de-identified medical data for research purposes is ethical in the absence of informed consent, we must consider whether the principles of social justice and utilitarianism supplant individual autonomy.

Several additional considerations are warranted in the context of de-identified individual medical data for research studies. These include potential benefits to data subjects as a result of de-identification of data, additional tools that can be used to protect de-identified data from the risk of individual privacy violations in addition to or instead of informed consent, and protection of both de-identified and personally identifiable research data from unauthorized access and use. Specific additional tools include technical tools to assess the level of de-identification of data and statistical risk of re-identification, additional legal regulations on the use of individual medical data to prevent potential consequences of privacy violations such as employment discrimination or insurance discrimination, more granular access controls for researchers and research groups working with potentially re-identifiable data sets, and increased education and awareness of best practices for research personnel working with potentially re-identifiable data. Each of these considerations will now be discussed.

The first additional consideration includes potential benefits to research subjects as a result of de-identification of data. Even if there is a risk that the data cannot be completely de-identified, using de-identified data as opposed to personally identifiable data in research studies, whether these studies involve informed consent or not, reduces the risk of individual privacy violations if there is a data breach. While data may need to remain individually identifiable in order to perform certain analyses in some research studies, individuals should be made aware of the level of de-identification of their data as part of the informed consent process and the data should be kept as de-identified as possible to minimize participant risk. In particular, given that there are multiple ways to de-identify data and de-identification exists along a spectrum,⁵ explaining what steps, if any, will be taken to de-identify a participant's data and the associated risks of re-identification as part of the informed consent process for a given research study will enable participants to make a truly informed decision regarding use of their data in a given research study.

The second important consideration involves whether informed consent is the only tool that can be used to protect "de-identified" data from the risk of individual privacy violations. Clearly, requiring a participant to consent to use of his/her data in a research study protects against potential privacy violations because the participant has given permission for his/her data to be used in this manner. If the participant has not given informed consent, however, whether or not a given use of the participant's data as part of a research study constitutes a privacy violation depends on whether the data may be used to re-identify the participant. As such, developing technical tools to better assess the re-identifiability of a given set of data, both alone and in conjunction with other information, may help assess the risk to individual privacy of using a given set of data as part of a research study. Legal regulations could serve as another potential tool to protect against privacy violations. For instance, laws preventing discrimination for employment or insurance coverage on the basis of individual health data such as genetic information could protect individuals from unauthorized use of their health data by employers or insurance companies in the event that their "de-identified" data is able to be re-identified. Current laws such as the Genetic Information Nondiscrimination Act of 2008 (GINA) already exist to protect against genetic discrimination from employers and for health insurance coverage.⁶ Another potential tool involves restricting access of specific research groups or research personnel to different data sets to reduce the likelihood of a single group or researcher obtaining complementary data sets that could be used to re-identify individual study participants. This may be analogous to conventional checks and balances used across large organizations and in government to prevent any single individual from wielding too much power. Finally, increased education of research personnel regarding the risks of working with de-

identified data and a set of best practices for storing and analyzing this data may also reduce the risk of inadvertent data breaches or re-identification. Thus, while informed consent is a valuable tool to protect de-identified data from posing a risk to individual privacy, it is not the only tool that can be used for this purpose.

The third consideration involves protection of research data, both “de-identified” and personally identifiable, from unauthorized access. Given that de-identified data can be re-identified and the likelihood of re-identification depends on the ethics of the entity analyzing the data and the other data the entity has access to, maintaining controlled access to de-identified data is crucial. Determining who is able to access this data requires both technical safeguards and approval from an ethics committee regarding which research studies and research groups are permitted to access and use a given set of data, along with how the data may be used as part of a given research study or by a particular research group.

¹ Jonas F Ludvigsson et al., “Ethical Aspects of Registry-Based Research in the Nordic Countries,” *Clinical Epidemiology* 7 (November 23, 2015): 491–508, <https://doi.org/10.2147/CLEP.S90589>.

² R. Froud et al., “Research Ethics Oversight in Norway: Structure, Function, and Challenges,” *BMC Health Services Research* 19 (January 10, 2019), <https://doi.org/10.1186/s12913-018-3816-0>.

³ “Privacy and Security Conditions Required for Research Using EHR and Other Electronic Health Data,” ASPE, November 23, 2015, <https://aspe.hhs.gov/report/feasibility-using-electronic-health-data-research-small-populations/privacy-and-security-conditions-required-research-using-ehr-and-other-electronic-health-data>.

⁴ Luke Gelinias, Alan Wertheimer, and Franklin G. Miller, “When and Why Is Research without Consent Permissible?,” *The Hastings Center Report* 46, no. 2 (April 2016): 35–43, <https://doi.org/10.1002/hast.548>.

⁵ Mark A. Rothstein, “Is Deidentification Sufficient to Protect Health Privacy in Research?,” *The American Journal of Bioethics : AJOB* 10, no. 9 (September 2010): 3–11, <https://doi.org/10.1080/15265161.2010.494215>.

⁶ “The Genetic Information Nondiscrimination Act of 2008,” National Human Genome Research Institute (NHGRI), accessed December 16, 2018, <https://www.genome.gov/27568492/the-genetic-information-nondiscrimination-act-of-2008/>.