A study of ethical issues in statistics and data science

Ksenia Lepikhina
University of Colorado, Boulder
STAT 5700: Philosophy of Statistics
Professor: Brian Zaharatos

December 18, 2019

Statistics and data are simultaneously powerful and dangerous tools. Data science can provide valuable inference for all areas of science, however, the misuse of data can have ethical consequences. This paper will go through two specific case studies of how handling data can raise ethical questions. The primary issues elaborated on in this paper will be the ethics behind data privacy and the ethics behind consent to data usage. In particular, this paper will discuss issues in data science using examples from the regulation of published science and medical research. For each area, intentions and consequences will be discussed in addition to ethical frameworks that attempt to find solutions to the quandaries.

The first case study this paper looks at pertains to a recent draft of a new Environmental Protection Agency (EPA) proposal [1]. The EPA is proposing an act titled "Strengthening Transparency in Regulatory Science" which would require all current, future, and previously published work to "disclose all of their raw data, including confidential medical records, before the agency could consider an academic study's conclusions" [2]. If the proposal passes, the consequences of this draft would impact research proposals enacting clean air or water rules and retroactively impact existing public health regulations.

The Environmental Protection Agency argues that they are "committed to the highest quality science" [2] and the overarching goal of the proposal is to reduce the quantity of "bad science". The EPA's intentions are to put up a fight against the replication crisis. The replication crisis is a scientific crisis pertaining to the ongoing issue that a multitude of scientific studies are difficult to or can not be reproduced [3]. The inability to reproduce results is due, in part, to the lack of data disclosure. A secondary reason is that with a new set of data, many of the statistical results can not be reproduced. Theoretically, if the data used for the analysis in a paper is published then the study can be verified as "good science" (meaning that it can be "independently validated... [and] hold up to scrutiny" [2]).

Requiring that all data be made public in order to produce better scientific results sounds like a brilliant idea on the surface, however the EPA proposal could have inadvertent effects. An example of previous research — that dictates current air-quality law — that could be

affected by the proposal is a 1993 Harvard project that draws the connection between poor air quality and premature deaths [2][4]. The scientists that worked on this study were asked to sign confidentiality agreements prior to collecting data on patients which they then linked to personal air quality data. Under the proposal, the EPA would be able to modify and potentially remove existing air quality regulations (such as ones implemented because of the 1993 study) due to the lack of data exposure.

While the proposal appears to be well intentioned, the request to make confidential data public raises ethical questions. Some questions that arise include: Should data scientists share confidential data to combat the replication crisis? Is it better to risk bad science to preserve patient privacy or to share patient data to assure better science? The EPA proposal appears to be forcing an answer upon this ethical question — it is best to make all data public. However well intentioned, the proposal seems to not consider the ethical implications.

There are a number of philosophical/ ethical theories that help study a proper solution to ethical questions such as data privacy ethics. One example is deontology. Deontology is the ethical theory that states that "people should adhere to their obligations and duties... to another individual or society" [5] when engaged in ethical decision making. The issue that arises with this ethical theory is that there is some confusion as to which duty a data scientist should adhere to. These ethical theorists may suggest that data scientist should keep confidential data private as it is their duty to keep the promise of privacy to individuals. Alternatively, the data scientist should adhere to their obligation to contribute to a scientific community that is committed to doing "good science".

Another ethical theory is utilitarianism. Utilitarians believe that the correct answer to an ethical question is the one that "yields the greatest benefit to the most people" [5]. Act utilitarians behave precisely as the definition states. Rule utilitarians behave within the bounds of the law and concern themselves with fairness in addition to maximizing benefits. As this ethical theory pertains to the EPA proposal, it raises the question of which solution provides the greatest benefit to the most amount of people while staying within the bounds

of the law. Exposing confidential data would lead to better reviewed science that could potentially impact the whole world. Maintaining the privacy of the data would ensure patient privacy and would continue to protect people individually. However, the exposure of confidential data may not actually lead to better science. Data and statistics can be manipulated to return misleading or incorrect results. It's possible publicizing the data will cause more harm than not. Under utilitarianism, the solution to this ethical query is that confidential data should remain private.

The second case study this paper looks at is the ethics behind using data collected without consent. In particular, this case study will examine Google's Nightingale project. The goal of the Nightingale project is to provide better tools for healthcare systems by making health records more searchable and accessible for doctors [6]. Google was provided lab results, doctor diagnoses, hospitalization records, and complete health records including patient names and birth dates [6].

Project Nightingale's overarching goal is to help doctors provide better healthcare. The project will take in medical records and return treatment plans, flag anomalies, suggestion different doctors for a specific patient, and monitor for narcotic policies [6]. Ascension (the healthcare organization providing the patient data) and Google both signed HIPAA (Health Insurance Portability and Accountability Act) business associate agreements [7] to assure that Google does not utilize the data for anything other than providing services to Ascension. Theoretically, if the data is handled properly and is only used for its intended purpose, then the project could provide patients with better healthcare.

Google appears to be well intentioned, however, soon after the Washington Post [6] broke the story on the project, it was discovered that Google never asked the doctors or the patients for consent to the use of the data. With more and more big data projects appearing, the question that is raised is whether or not consent for data is still relevant. Cheung, a faculty member of Law at the University of Hong Kong, writes that consent for big data projects may be obsolete [8]. Consent is typically requested *after* the research question has

been formulated and potential risks have been assessed. Cheung claims that with big data projects — such as Project Nightingale — "the risk and harm... may not be known at the time of data collection" [8] which implies that "the notion that one can fully specify the terms of participation through notice and consent has become a fallacy"[8].

Ethically speaking, there are a number of questions that arise from the statement that in an era of big data, consent might not be necessary. Because Project Nightingale is well intentioned, does it matter whether or not consent was granted? Would it be better if consent was granted to a project even though the full impact was unknown? The Nightingale Project appears to assume that Google's actions are ethical and that consent for big data is not necessary.

Like the first case study, Project Nightingale can be analyzed using a number of philosophical and ethical tools. As mentioned earlier in this paper, deontologists believe that the answer to ethical quandaries is to adhere to duties and obligations. In the case of the ethics behind using data without consent, a deontologist may argue that it is Google's duty to obtain the consent of the patients and the doctors prior to using their data. On the other hand, Google's intentions are to provide their powerful services to improve a healthcare system. If the duty of an individual is to help improve a community and potentially save lives, then Google is ethically obligated to use the data regardless of whether or not consent was obtained. The issue with this ethical ideology is that "there is no rationale or logical basis for deciding an individual's duties" [5]. This flaw makes it difficult to determine whether Google's use of non-consensual data is unethical due to the unknown logic that went into the decision (or whether there was any ethical thinking at all).

The second ethical ideology explored in this paper is utilitarianism. To repeat, utilitarians believe that the correct ethical choice is the one that maximizes the positive outcome. Arguably, this philosophical methodology can be the solution to the Project Nightingale ethical problem. Though Google did not ask for consent from patients or doctors, if they achieve their goal and use the data only as they stated they would (under HIPAA), then the

project's success could lead to better diagnoses for patients and improved tools for doctors. Alternatively, if Google requested consent from patients and doctors, then it is possible that Project Nightingale may not have begun. Without the innovation that Google could provide for the healthcare system through this project, it's possible that fewer people would benefit. With utilitarianism, the solution to the ethical quandary is that using data without the consent of individuals is okay as long as the overall effect is positive.

To recap — in this paper, the first case study looked into the ethics of publicizing confidential data in the name of "better science". The second case study discussed the ethics of using an individual's data without their consent. With each case study, this paper looked at how deontology and utilitarianism could potentially offer solutions to some of the ethical dilemmas in data science. Deontology did not provide a clear solution to the ethical quandaries because the duties in the ideology are personal and are therefore not clearly defined. Utilitarianism helped derive solutions by maximizing the positive outcome. For data privacy, the utilitarian argued that data should remain confidential (maintain privacy) but argued that in the era of big data, using personal data without consent (but bound under some legal agreement) is alright as long as the intentions and outcome are positive.

References

[1] *Strengthening Transparency in Regulatory Science*. Apr. 2018. URL: https://www.federalregister.gov/documents/2018/04/30/2018-09078/strengthening-transparency-in-regulatory-science.

[2] Lisa Friedman. *E.P.A. to Limit Science Used to Write Public Health Rules*. Nov. 2019. URL: https://www.nytimes.com/2019/11/11/climate/epa-science-trump.html.

[3] Ed Yong. *Psychology's Replication Crisis Is Running Out of Excuses*. Dec. 2018. URL: https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/.

[4] Douglas W Dockery, C Arden Pope, Xiping Xu, et al. "An association between air pollution and mortality in six US cities". In: *New England journal of medicine* 329.24 (1993), pp. 1753–1759.

[5] Larry Chonko. *Lecture notes in Ethical Theories*.

[6] Rob Copeland. *Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans*. Nov. 2019. URL: https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790.

[7] *House Committee Leaders Request Answers from Google and Ascension on Project Nightingale Partnership*. Nov. 2019. URL: https://www.hipaajournal.com/house-committee-leaders-demand-answers-from-google-and-ascension-on-project-nightingale-partnership/.

[8] Anne SY Cheung. "Moving beyond Consent for Citizen Science in Big Data Health and Medical Research". In: *Nw. J. Tech. & Intell. Prop.* 16 (2018), p. 15.