Ethical Issues of Open Medical Data

Sarah Hanna Fischer and Lisa Müllner

February 7, 2020

Abstract

Open data is becoming more and more popular, but with open medical data ethical aspects must always be considered. To be able to talk about open medical data firstly some terms like Open Data, Big Health Data and Data Ethics must be defined. Current and possible future sources for open medical data are found and discussed, such as HealthData.gov and GenBank. Open medical data can lead to an increase in scientific discoveries and a decrease in research cost. With access to open medical data small studies could find more significant results and this could lead to lesser known medical conditions being researched. It would also lead to more transparency as to what data is collected. Open medical studies could solve the reproducibility crisis and work against the publication bias. But with all of these benefits come ethical issues and considerations that must be made. Even with methods for de-identification privacy concerns arise, since medical data is so sensitive it could easily lead to discrimination if data is re-identified. Because of the nature of medical data re-identification is much more likely than with other data. Since even little knowledge about someone, for example a time at which they where ill, can lead to easy re-identification of medical data. Therefore, the goal of this paper is to give an overview of the current state in the area of open medical data as well as to discuss the major ethical issues. This should serve as a basis for further research in this area as well as for more awareness regarding this topic.

1 Introduction

In recent years the topics "open data" and "big data" have gained increasing attention in many different areas. There exists a whole movement that deals with opening up various kinds of data and one of the areas where their is a wish for opening up the data is the health care sector. [1]

There is a huge potential for the already existing amount of medical data that is gathered every day. For example, it could help to make new scientific discoveries, it would lower research costs, because no redundant research needs to be done and many more. However, most of this data is not publicly available due to the fact that medical data is very delicate, which makes it impossible to utilize the potential of this data. The main reasons for this are ethical issues that make it difficult to open up the data. Therefore, the goal of this paper is to give a short overview of the current state of open medical data and to primarily discuss the ethical issues that arise when opening up medical data.

We start our paper by providing first some general information and definitions of the topic related areas in Section 2. We will continue by presenting some publicly available data sources of medical data in Section 3. Afterward, the potential and importance of open medical data is shortly discussed in Section 4. Finally, some of the ethical issues that need to be considered when dealing with open medical data are presented in Section 5 and then we sum up our results in Section 6.

2 Definitions

In this section, we provide some general information and definitions of the topic related areas like Open Data, Data Ethics, Big Health Data, Open Medical Data, etc.

2.1 Open Data

> "Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike." [2]

Of course the words like "freely used" or "re-used and redistribution" can be understood differently by different people. Therefore the different aspects of the definition are specified in more detail to make sure that everyone knows exactly when data is open data and when not. The first aspect is "freely used" and this means that the data is of course available as a whole and it does not cost anything except for reproduction costs that are reasonable. The next aspect is "re-used and redistributed" which means that the data must be published under conditions that allows everyone to re-use and redistribute it and to combine it with other datasets. The last very important aspect is that the data needs to be accessible for everyone which means that there is no discrimination against specific groups or people or fields of endeavour. [2]

Something else important regarding "open data" is that a movement called "open data movement" exists, which has the goal to open up data and to provide tools that make the analysis of the data easier. Because sometimes even if data is freely available the data is so complex that special tools or software is needed to be able to analyse it. [1]

2.2 Big Health Data

Big health data means the sum of all the data that is produced, gathered and electronically stored during the daily routines of modern health care systems as well as health data that is collected by wearable devices regardless if it is open or not. There is a high potential for improving patient care with this huge amount of data but currently there is a lack of tools that are able to properly handle the data. [3, 4]

2.3 Open Medical Data

Open medical data, also called open health data, is as the name already says medical data that is openly available and conforms the definition of open data. The main focus of our paper is on this kind of data.

2.4 Data Ethics

Data ethics is based on computer and information ethics, but the emphasis lies on the data and the complex ethical challenges that come with it. [5]

The term data ethics is often used in the context of large datasets, for example big data but also in context of open data in general. Data ethics deals with the study and evaluation of moral problems that can occur while working with data to help to find morally acceptable solutions. Furthermore, it also evaluates algorithms and other practices like programming and hacking that are also a part of the usage of data. [5]

Examples for ethical challenges are presented in Section 5.

2.5 De-identification

> "De-identification of protected health information is an essential method for protecting patient privacy." [6, p. 1044]

The goal of the process of de-identifying health data is that the person that the data belongs to cannot be identified through this data. In other words, the data is being anonymized. There is a list of 18 identifiers that need to be removed and only if all of these identifiers are removed the data is considered "de-identified". Some examples for such identifiers are the name, the telephone number, the medical record numbers, biometric identifiers, etc. [6, 7]

2.6 Open Science

According to Kriegeskorte [8] open science consists of four pillars: open data, open code, open papers (open access) and open reviews (open evaluation). We already discussed the definition of the first pillar: open data. The next one is open code which means that the code should be shared for more transparency and for reuse purposes. One possible approach for achieving this would be the usage of a version control system. The next pillar is open papers/open access. This pillar deals with publishing papers that are openly available. This could be easily realised with using preprint servers because this ensures open access. [8]

The last pillar are open reviews. The goal of this pillar is to make the review process transparent.

> "Transparent review means (1) that reviews are public communications and (2) that many of them are signed by their authors." [8]

3 Open Medical Data Sources

In this section we provide a short overview of some existing medical data sources. Not all of them are fully open according to the open data definition but we will explain this further for each database.

3.1 HealthData.gov

The website HealthData.gov[1] is the official website of the United States Department of Health and Human Services which offers around 3800 datasets for the topics such as mental health, environmental health, community health, etc. They explain the purpose of the site as follows:

> "This site is dedicated to making data discoverable and making valuable government data available to the public in the hopes of better health outcomes for all." [9]

The data that is published their is free of charge and most of the datasets have a license that makes the data "open data" according to the definition (see Section 2). Only 11 datasets have a license that is not open which are mainly datasets that deal with patient data. Some of those have restricted access files which are accessible with a data use agreement and a brief online security training. [9]

3.2 GenBank

The GenBank [2] is a genetic sequence database from the National Institutes of Health. On their website they state that they do not add any restrictions to the use and distribution of the data which means all the data available their is "open data". However, they also

[1] www.healthdata.gov

[2] https://www.ncbi.nlm.nih.gov/genbank/

explain that some submitters may add any claims or copyrights to the data on which they do not have any influence. That means that everyone who is using data from their site needs to take care of this themselves. Furthermore, they have a short statement regarding privacy where they request people submitting data to the GenBank not to include any data that could identify the person the data belongs to.

3.3 European Data Portal (EDP)

The European Data Portal[3] is a website that provides metadata of public sector information from the public data portals of European countries. On the website thirteen categories are available, for example data about transport, education, environment and many others. Furthermore, the EDP also offers over 7.000 datasets from the health sector. Every dataset on the website of the EDP has a license and they state that everyone who wants to use the data needs to first check by themselves which license, attribution requirement and share-alike requirement the dataset has.

4 Importance of Open Medical Data

One major importance and advantage of open medical data is that it can potentially lead to new scientific discoveries. Since data that is collected by one group of scientists could be used by another for a different study. [7]

Open medical data can also reduce the cost for research, since a lot of the same research is done by different researchers. This research is expensive and if the researchers could freely inspect other researchers' work, the amount of redundant research could be reduced. Furthermore, observational studies that look at existing data could be done by different researchers. Small research project funded via crowd funding might also need a lot less funding to work, since it would not always be necessary to collect new data, when a lot of data is available. [7]

Open medical data is also a way to make governments more transparent and increase public education. Knowing what data the government collects and have insight into it could lead to debate about controversial data and therefore to change in the collection of this data. People could learn about the health care system as well as the costs of medical care. If some medical data where publicly available it could lead to improvements in substandard medical facilities. Especially in countries like the United States of America, where medical facilities are not as standardised as they are in Europa. The data would also be available to the media, which would lead to political change. [7]

4.1 Open Medical Studies

Medical studies, trials and experiments are a big part of modern medicine, but most of the data collected in these studies is never published. This can lead to misleading study results and problems when reproducing the studies. Some people say we are in a so called "reproducibility crisis" what this entails and more is discussed in this section. [10]

4.1.1 Publication Bias

The publication bias means that more studies that are successful are published than those which are not successful. Let us say that the same study is done ten times and never works and then be done an eleventh time and work and then only the eleventh study is published. This means that results that are achieved by just trying often enough, which can be considered coincidental results will be taken as actual science and significant results

[3] https://www.europeandataportal.eu/

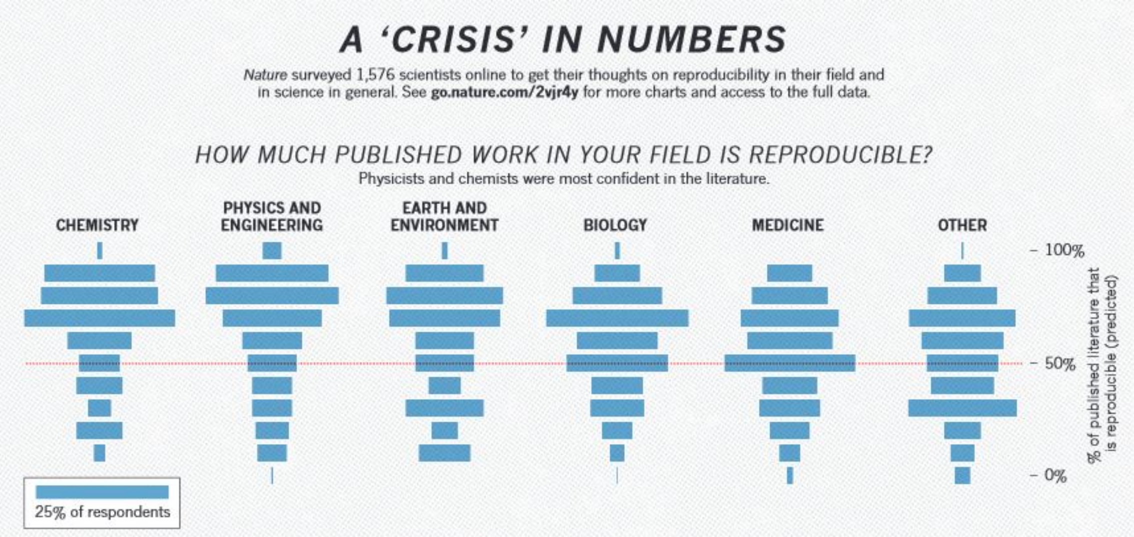

Figure 1: The reproducibility crisis shown in a graph as answer to the question of how much published work the researchers believed to be reproducible in different fields of science. [10]

because the studies that did not work are never published. This problem could be averted by having open science, this way all the studies would be available to other researchers and they could make up their own mind if the results are caused by coincident or have scientific significance. [11]

4.1.2 The Reproducibility Crisis

Reproducible Research is about researchers trying to replicate studies or experiments done by other researchers. Naturally this does not always work but one would expect it to work most of the time since experimental factors should mostly be tightly controlled and written down. The reproduction of results of medical studies or experiments is very important since it shows that the data is more likely to be correct and not the result of manipulation, error or plain coincidence. [12, 10]

More than 60% of medical researcher in Monya [10]'s survey have stated that they have failed to reproduce other people's research. Even more concerning is that more than 50% of them have failed to reproduce their own research. For the biggest reasons why research reproduction fails the researchers that participated in the survey named "Selective reporting" and "Pressure to publish". [10]

This is a major problem that could be tackled with open science. Because if all the research studies, even failed ones, were available for other researchers to look at reproducibility would most likely increase. The so called reproducibility crisis can be shown nicely with a graph of the survey done by Monya [10] as can be seen in Figure 1. In this graph one can see the results of the survey question "How much published work in your field is reproducible?". More than 1500 scientists took part in the survey and answered the questions for their respective field. It can be seen in the graph that while the subjects chemistry as well as physics and engineering seem to be rather confident in the reproducibility in their fields other subjects such as medicine are more on the sceptical side. Most scientists in the medical field stated that they thought that around half of the published literature was reproducible. This might not sound so bad at first, but considering this is the literature on which our medicine and medical treatments are based on having only 50% of reproducibility is quite concerning. Of course it must be remembered that this was only a question of researchers' opinions, not of facts. [10]

5 Ethical Issues of Open Medical Data

Big Data as well as open medical data in healthcare is becoming more and more popular, it has the opportunity to improve our medical knowledge and with that our ability to help people that need medical attention. But with all of these possible benefits come possible ethical problems. Questions arise, such as who owns this data and who has control over this data?

5.1 Big Data in Public Health

An example for the problems that can be encountered when medical data is not kept private is data.care, a system introduced on the 6th of July 2016 by the the United Kingdom's National Health Service (NHS). [4, 13]

The data.care system used data from hospitals and linked it with the general practice patient record system. The collected health data on the citizens of the United Kingdom was to be used for medical research. Data.care was an opt-out system, that means if a citizen did not want their data to be analysed by data.care they had to manually cancel the service. [14]

Data.care was canceled because of two independent reviews that called out major privacy concerns with it. Data.care was started 2013 but over one million people opted out of the service since then. When the program was finally closed 2016 it was done so with the promise of some other form of data sharing in the future. [13]

Another example is ELGA, a similar system as data.care by the Austrian government. In contrast to data.care the Austrian system is still in place and no attempt has been made to end it. As with data.care ELGA also has an opt-out system with two possible opt-outs. The first one is to refuse ELGA as a whole and the second option is a partially opt-out where only some parts of ELGA can be denied like electronical results for example. [15]

Vayena et al. [16] write in their article "Ethical Challenges of Big Data in Public Health" about how important big data is in digital epidemiology, the subject concerned with digitally detecting and analysing diseases. Big health data could potentially accelerate disease outbreak detection by assessing health behavior and attitudes. In their paper the main focus are the ethical challenges that come with big health data, which they form into tree categories.

The first is context sensitivity which is concerned with privacy and consent of patients. The second is connection of ethics and methodology which is about risk of harm and accountability. The last category includes trust, transparency and accountability and with that justice. For the first point they argue that instead of big companies that use citizen data for their own profits the digital epidemiology or digital disease detection (DDD) collects and analysis data for the common good. Concerning the second point they discuss how social network data is often used by DDD nowadays, this data is free but there are concerns that people using these networks do not know what they are agreeing to. [16]

This is a very interesting subject, since today a lot of people use social networks and give very private information on these platforms. They should know, of course that this means the data is available to anybody with access to the internet but still a lot of people are surprised when confronted with this reality. The last point must also be considered, trust and transparency are one of the main parts of open data. The researchers ague that a lot needs to be done still to ensure that ethical oversights are improved and eventually worked out. [16]

The problems with ethics and medical data are also discussed in Fairchild and Bayer [17]'s paper "Ethics and the Conduct of Public Health Surveillance". They conclude that while trying to protect the common welfare the protection of the individual might not always be possible in the same ways it normally is. They ague with the following:

> "It is inappropriate to regard ethical over-sight strictly as an impediment. In the context of public health surveillance, it can serve as a means of avoiding inadvertent breaches in confidentiality and stigma; it can help to ensure that the public understands that surveillance will occur and what purposes it serves; it can protect politically sensitive surveillance efforts. There is, after all, an ethical mandate to undertake surveillance that enhances the well-being of populations." [17, p. 632]

5.2 Privacy

One of the main concerns about open medical data is privacy. It is important that the privacy of the patient that the data belongs to is always ensured. One example of a project that has some privacy issues is mentioned by Hoffman [7] in her book. The Personal Genome Project for example has whole patient profiles on their website. [18]

Some of the profiles present not only the full name but also the gender, birthday, medical conditions and other private data. Anyone with access to the internet can potentially look at this data. But even the profiles that do not disclose names can probably be identified with relative ease, since a lot of personal data is given. Often the de-identification of data is done to not only uphold laws that prevent personal data of patients to be published but also to give people some privacy. But re-identification of data is often possible, which means there is no guarantee that de-identified data stays de-identified.

Re-identification is especially likely for medical data, since only very limited knowledge, like the time and place of an operation or accident can already lead to easier re-identification. This knowledge can be obtained easily for example by employers, more on this topic in the next section. [7, 18]

5.3 Discrimination

Another ethical issue that arises through opening up health data is that discrimination could happen. This could affect many different areas of life, for example in the world of employment or even in the financial or advertising area.

The first major area where discrimination could appear is the work place. It is no secret that employers do research about their job applicants beforehand on the internet, on social networks, etc. to find out more about the person. However, with open medical data it could be also possible to check the health condition of the employee. Of course, employers want to have healthy employees and it could be possible that if the employer finds out that the employee has some health issues that they do not want to employ them because it is a risk for them. Another example of discrimination in this area could be that someone finds out that a person has a sexually transmitted infection and even if the person is employed discrimination could appear at the workplace or the person could be bullied due to this knowledge. [19]

Another area where discrimination due to open medical data could appear is the financial sector. Lenders for example could refuse loaning money to a person from whom they know that they are not as healthy because then the risk of not getting back the loan exists. The person is maybe at some point in time not able to work anymore due to the health issues and then is not able to pay off the loan. This is a huge risk for the lenders and therefore such discrimination could be a real problem. [20]

Marketing and advertising is nowadays a huge topic and of course in this area open medical data can be also used in a discriminating way.

> "Marketers may also engage in discriminatory practices, offering promotions and discounts to some customers but not others, or advertising selectively so that they reach only certain consumers." [7, p. 1779]

With respect to this topic a popular case is often mentioned where a father figured out the pregnancy of his teen daughter through the mails his daughter got from the shop "Target". [21]

Although in this example open medical data was not directly used because the fact that the daughter was pregnant was maintained through many different data that was gathered from Target it is an important case. Because the pregnancy itself was a very sensitive medical information regardless how Target figured it out. Such information could then be much easier obtained.

These were only some examples of possible discrimination that could happen when opening up medical data. Of course, such discrimination is illegal but difficult to prove. The problem of possible re-identification as discussed previously shows that it is not unrealistic that the data, even though published de-identified might be re-identified. [7]

5.4 Consent Models

Another topic that is often discussed in literature when talking about ethical issues of open medical data is which model of consent should be used for the data of the participants. An overview of possible consent models for health data can be seen in Figure 2. The first option is the "specific consent" which means that the patient needs to give consent for each specific kind of information. The second possibility would be the concept of "broad consent" where the patient gives consent to using their data without knowing exactly for which purposes. [19]

These two consent options are of course some extremes. The specific consent for example leads on the one hand to more patient autonomy but on the other hand it is then very difficult to do research in the health area because it is not always possible to specify beforehand for which purposes the data is then actually used. [19]

Therefore, maybe a consent model which is something in between those two models would be the best one but this is something where more research needs to be done.

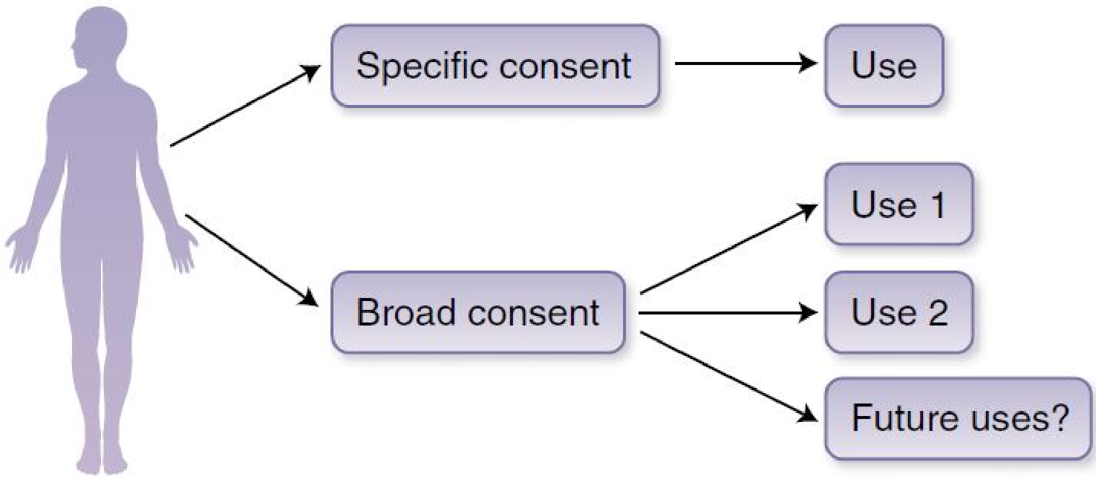

Figure 2: Overview of different consent models for health data. [19]

6 Conclusion

We have defined the necessary terms to talk about open medical data and pointed out some sources where to find such data. We listed only a few data sources because it is currently difficult to find sources for open medical data. This shows that there is really a lack in this area which is most likely due to the fact that many ethical issues can occur.

Therefore, we discussed some of the major ethical issues that need to be considered when dealing with open medical data. We conclude that ethical issues will always occur when medical data is concerned since it is very personal and private information. Especially nowadays where people are getting more and more concerned about their private data and try to be as careful as possible with it. It is important to discuss and evaluate those ethical issues openly to ensure transparency and trust. Because only if it is clearly defined for what purposes the data is used and how the privacy and security of the patient is ensured then the patients are getting more trust and maybe give their consent for using their data. This is something which takes a lot of time but it would have real benefits which we discussed in the paper. There is the potential of improving research in the health sector through open medical data.

However, we think that it is important to have a balance between the benefits of open medical data and the privacy of the patient. This is a difficult task that needs more research and more studies to find the best possible solution.

References

[1] Rob Kitchin. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

[2] Open Data Handbook. Open data handbook - what is open data? `http://opendatahandbook.org/guide/en/what-is-open-data/`, 2020. Accessed: 2020-01-20.

[3] Sebastian Schneeweiss. Learning from big health care data. *N Engl J Med*, 370(23): 2161–2163, 2014.

[4] Patty Kostkova, Helen Brewer, Simon de Lusignan, Edward Fottrell, Ben Goldacre, Graham Hart, Phil Koczan, Peter Knight, Corinne Marsolier, Rachel A McKendry, et al. Who owns the data? open data for healthcare. *Frontiers in public health*, 4:7, 2016.

[5] Luciano Floridi and Mariarosaria Taddeo. What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083):20160360, 2016. doi: 10.1098/rsta.2016.0360. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2016.0360`.

[6] Mehmet Kayaalp. Modes of de-identification. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1044. American Medical Informatics Association, 2017.

[7] Sharona Hoffman. Citizen science: The law and ethics of public access to medical big data. *Berkeley Technology Law Journal*, 30(3):1741–1805, 2015. ISSN 10863818, 23804742. URL `https://www.jstor.org/stable/26377581`.

[8] Nikolaus Kriegeskorte. The four pillars of open science. `https://nikokriegeskorte.org/2016/02/15/the-four-pillars-of-open-science/`, 2016. Accessed: 2020-02-03.

[9] U.S. Department of Health And Human Services. Healthdata.gov. https://healthdata.gov/, 2020. Accessed: 2020-01-30.

[10] Baker Monya. 1,500 scientists lift the lid on reproducibility. https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970, 2016. Accessed: 2020-02-03.

[11] K. Dickersin, S. Chan, T.C. Chalmersx, H.S. Sacks, and H. Smith. Publication bias and clinical trials. *Controlled Clinical Trials*, 8(4):343 – 353, 1987. ISSN 0197-2456. doi: https://doi.org/10.1016/0197-2456(87)90155-3. URL http://www.sciencedirect.com/science/article/pii/0197245687901553.

[12] Lorena A. Barba. Terminologies for reproducible research. *ArXiv*, abs/1802.03311, 2018.

[13] Department of Health and Social Care. Review of health and care data security and consent. https://www.gov.uk/government/speeches/review-of-health-and-care-data-security-and-consent, 2016. Accessed: 2020-02-04.

[14] Jon Hoeksma. The nhs's care.data scheme: what are the risks to privacy? *BMJ*, 348, 2014. doi: 10.1136/bmj.g1547. URL https://www.bmj.com/content/348/bmj.g1547.

[15] Pflege und Konsumentenschutz Bundesministerium für Soziales, Gesundheit. Review of health and care data security and consent. https://www.oesterreich.gv.at/themen/gesundheit_und_notfaelle/elga___elektronische_gesundheitsakte/Seite.3110002.html. Accessed: 2020-02-04.

[16] Effy Vayena, Marcel Salathé, Lawrence C. Madoff, and John S. Brownstein. Ethical challenges of big data in public health. *PLOS Computational Biology*, 11(2): 1–7, 02 2015. doi: 10.1371/journal.pcbi.1003904. URL https://doi.org/10.1371/journal.pcbi.1003904.

[17] Amy L. Fairchild and Ronald Bayer. Ethics and the conduct of public health surveillance. *Science*, 303(5658):631–632, 2004. ISSN 0036-8075. doi: 10.1126/science.1094038. URL https://science.sciencemag.org/content/303/5658/631.

[18] Personal Genome Project. Personal genome project, harv.med.sch. https://my.pgp-hms.org/users, 2020. Accessed: 2020-01-21.

[19] W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019.

[20] Sharona Hoffman. The promise and perils of open medical data. *Hastings Center Report*, 46(1):6–7, 2016.

[21] Kashmir Hill. How target figured out a teen girl was pregnant before her father did? https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did, 2012. Accessed: 2020-01-29.

About this document. This seminar paper was written as part of the lecture *Free and Open Technologies*, held by Christoph Derndorfer and Lukas F. Lang at TU Wien, Austria, during the winter term 2019/2020. All selected papers can be found online.[4]

[4] https://free-and-open-technologies.github.io