CHAPTER 4

Big Data in Biomedical Research

Biomedical research comprises basic science/bench research as well as clinical research. It involves disciplines such as epidemiology, diagnostics, clinical trials, therapy development and pathogenesis (Nederbragt 2000). Studies in these fields aim to enhance the scientific knowledge and understanding of (public) health and diseases. Key objectives are the development of effective treatments and thus the improvement of healthcare.

Biomedical research has for a long time involved large datasets. However, big data and novel analytics approaches have been increasingly emphasised as significant trends (see also Parry and Greenhough 2018, 107ff.). Big data-driven research projects draw on data retrieved from, for instance, social networking sites, health and fitness apps, search engines or news aggregators. Critical factors for this biomedical 'big data revolution' are technological innovation, the popularisation of personal, mobile computing devices, and increasingly ubiquitous datafication (Margolis et al. 2014; Costa 2014; Howe et al. 2008).

In this chapter, I outline the discursive conditions for such biomedical big data-driven research, especially in the field of digital public health surveillance. To recapitulate, I derived two main, analytic questions from previous research in critical data studies (CDS), pragmatist ethics, and Habermas' deliberations on discourse ethics in particular:

1. What are the broader discursive conditions for big data-driven public health research?
   a. Which actors are affected by and involved in such research?
   b. Which factors may shape the views of affected actors and their engagement in public discourse?
2. Which ethical arguments have been discussed; which validity claims have been brought forward?

How to cite this book chapter:
Richterich, A. 2018. *The Big Data Agenda: Data Ethics and Critical Data Studies*. Pp. 53–69. London: University of Westminster Press.
DOI: https://doi.org/10.16997/book14.d.

The first question, including the two sub-questions, is predominantly examined in this chapter. Chapter 5 responds mainly to question 2, by analysing ethico-methodological developments, justifications and tensions concerning specific big data-driven research projects. However, I also come back to some of the issues explored below when discussing ethical arguments and specific project constellations.

The following sub-chapter starts with a reflection on what commonly classifies as biomedical data. This is followed by an overview of stakeholders affected by big data-driven public health research. Subsequently, I elaborate on some of these stakeholders in more detail, specifically those that have a notably powerful role in setting a discursive agenda for big data-driven research. Specifically, I highlight the role of (inter-)national grant schemes and corporate interests, as well as (financial) support for biomedical and big data-driven research. This focus takes into account that certain (f)actors may a priori bias the discursive conditions for public opinion formation and debate.

Strictly Biomedical?

With regards to big data developments in biomedical research, one can differentiate, very broadly speaking, between two categories of relevant data Certain data are generated from biological sources such as human tissue and body fluids. In addition, observational data, for instance patient diagnoses, are provided by clinicians and other medical professionals, and documented in medical records. Parry and Greenhough (2018) describe these types of data as *derivative* and *descriptive* bioinformation (5ff.).

Vayena and Gasser (2016) argue that such data should be considered 'strictly biomedical', referring , among others, to 'clinical care data, laboratory data, genomic sequencing data' (20). In these cases, biological material (derivative) or observations (descriptive) are transferred into digital data. However, there is another category of 'digitally-born' data that are not extracted from encounters with patients or analyses of biomedical material. Instead, these data are generated by documenting individuals' interactions with computing devices and online platforms. While often created without being intended primarily as a contribution to understanding (public) health issues, these data have shown to carry 'serious biomedical relevance' (Vayena and Gasser 2016, 17).

According to Vayena and Gasser (2016), the category 'strictly biomedical' applies to genomics. This interdisciplinary science is concerned with sequencing and analysing genetic information, i.e. the DNA in an organism's genome. While the samples and methods of data collection may be considered more 'traditional' (even though, of course, highly advanced on a technological and scientific level), developments in sequencing technologies have led to new challenges of data storage and management.

Since the finalisation of the Human Genome Project in 2003, with its complete mapping and examination of all human genes, the amount of biological sequence data has dramatically increased. One of the main reasons is that '[s]equencing a human genome has decreased in cost from $1 million in 2007 to $1 thousand in 2012' (O'Driscoll, Daugelaite and Sleator 2013, 774). In turn, this has created a heightened need for data storage options, computing tools and analytics. At the same time, it has facilitated a commercialisation of genetics and related services such as 23andMe for which regulations were only enforced with some delay (see e.g. Harris, Wyatt, and Kelly 2013a, 2013b, 2016).

The use of 'digitally-born data' is being explored in various fields of biomedical research. For example, it has been asserted that data retrieved from social media such as Twitter may contribute to detecting adverse medication reactions (Freifeld 2014) or content which may indicate depression (Nambisan et al. 2013), as well as the geo-mapping of epidemics (Chunara 2012). The significance of such data as biomedical information is context-dependant, even more so than in the case of derivative and descriptive bioinformation. Content exchanged on social media – such as, for example, posts and status updates indicating meals or eating habits – may enable health-related insights. However, these data were collected without individuals' intention and mostly without their awareness that they may be used for biomedical research (see also Chapter 3 on 'Informed Consent'). In the first place, they were created to interact with friends, peers, or broader audiences: e.g. to display or discuss experiences, opinions, achievements etc.

In this context, Vayena and Gasser (2016) pointedly stress the need for new ethical frameworks regarding the largely unregulated use of such digitally-born data (28ff.). The authors refrain, however, from calling these data 'biomedical', since they do not regard it as bioinformation in a strict sense. Instead, they describe such data as 'non-biomedical big data of great biomedical value' (Vayena and Gasser 2016, 23). In contrast, I also speak of biomedical (big) data with regards to digitally-born data. A main reason for doing so is to account for the comparable epistemic value and significance of those data. This is also acknowledged by Vayena and Gasser when they state that '[…] although biomedical data are categorized on the basis of their source and content, big data from non-biomedical sources can be used for biomedical purposes' (2016, 26). But while the authors still make a differentiation based on biological or physical observations versus digital sources, I propose not to distinguish in this case, since this may also suggest that a priori different, potentially less strict, ethics guidelines should apply.[42]

In this chapter as well as in Chapter 5, I focus on those digitally-born data whose significance for biomedical research is currently being explored. I mainly investigate research aimed at using big data for public health surveillance/epidemiological surveillance. There are two main reasons for this choice: First, this is a crucial field for which digital health data have been employed so far. Second, due to the fast-paced technological and institutional developments

in collecting and analysing health-relevant data, the ethical debate is only successively catching up with big data-driven research in this domain.

Who is Affected, Who is Involved?

A first step towards assessing the formation of social norms, according to Habermasian discourse ethics, is to identify: who is affected by certain developments, who *has* a say in related debates and/or who *should have* a say. Additionally, it is relevant which stakeholders play a part in shaping the respective development in the first place. This also gives some indication of interests that these actors may discursively pursue.

The big data ecosystem of public health research is complex, and an overview of stakeholders is inevitably a simplification. That said, Zwitters' (2014) classification of big data stakeholders, into *generators*, *collectors* and *utilisers*, is a useful starting point. The author differentiates between: a) natural/artificial actors, or natural phenomena that *generate* data, voluntarily or involuntarily, knowingly or unknowingly; b) actors and entities that define and control the *collection*, storage and analysis of data; and c) those *utilising* the collected data, i.e. actors and entities which may receive data from collectors for further, potentially redefined utilisation (Zwitter 2014, 3). These broader categories also apply to the field of big data-driven health research, although it appears useful to add another, potentially crosscutting category: d) entities incentivising and promoting the use of big data in research, for example by providing financial support.

Biomedical big data have implications for a broad range of professions, domains and actors. For example, during a workshop on 'Big data in health research: an EU action plan', organised by the EC's Health Directorate[43] (Directorate-General for Research and Innovation) in 2015, a long list of international experts participated. The list included '[…] bioinformaticians, computational biologists, genome scientists, drug developers, biobanking experts, experimental biologists, biostatisticians, information and communication technology (ICT) experts, public health researchers, clinicians, public policy experts, representatives of health services, patient advocacy groups, the pharmaceutical industry, and ICT companies' (Auffray 2016). One extremely heterogeneous group is notably absent, though: those individuals generating the digital data that are now complementing biomedical research (see also Metcalf and Crawford 2016).

Users who contribute to digital platforms and generate big data of biomedical relevance are not necessarily doing so in their role as patients. In contrast to most derivative and descriptive bioinformation, big data are also retrieved from users who are not consciously part of a certain health or research measure. Accordingly, those individuals whose data are fed into big data-driven research are key stakeholders. They enable big data approaches, since they are the source

of the data in question. However, they rarely contribute actively to the decisions made with regards to if and how personal data are retrieved, analysed, sold, and so on. Their 'involvement' is commonly limited to the opt-in or opt-out options enforced by corporate terms of services and usage conditions. As well as those users whose data are included in retrieved data sets, non-users of respective platforms should also be considered as relevant stakeholders. Non-users may be systematically excluded from benefits that other, participating users may receive (see the example of fitness trackers in Chapter 2); or they may experience pressure to participate in the generation of digital health data as these dynamics become more common.

One should not mistake 'being affected' with consciously noticing the effects of a development. This is one of the main problems that much of big data-driven research is hesitant to foreground: the ethical and practical implications of such research are largely unclear. At the very least, individuals are exposed to uncertainties regarding how the data are used and what this might mean for them as stakeholders now and in the future (see also Zwitter 2014). As personal data are automatically retrieved on an immense scale, the implications of such approaches for users' autonomy, dignity and right to privacy need to be considered. However, this is an extremely heterogeneous group of stakeholders. It needs to be seen on a case by case basis (see chapter 5), in which specific, potentially vulnerable groups, may be affected by big data-driven research projects more concretely. This also includes how they may relate to the outcome and results of big data-driven health research, for example as beneficiary or harmed party.

In their paper on the US 'Big Data to Knowledge Initiative', which I introduce in more detail below, Margolis et al. (2014) propose that '[k]ey stakeholders in the coming biomedical big data ecosystem include data providers and users (e.g., biomedical researchers, clinicians, and citizens), data scientists, funders, publishers, and libraries' (957). Here, researchers are labelled as 'users'. The wording is telling, and points towards Zwitter's (2014) category c. In big data-driven studies, researchers tend to act as data *utilisers*. They are affected by big data developments, since they are faced with what is promoted by e.g. peers or funders as novel research opportunities. Big data in this context may be perceived or portrayed as an opportunity for innovation. But, for scientists, it might also turn into a requirement to engage with this phenomenon or into a competitive trend, channelling biomedical funding into big data-driven studies. As big data utilisers, biomedical researchers are repurposing data retrieved from social networking sites and other sources. At the same time, they shape normative discourses on why and how these data may be used in biomedical research. This may further incentivise biomedical research involving big data. The ethical discourses articulated by scientists involved in big data-driven research, as well as counterarguments where applicable, are considered in Chapter 5.

Apart from scientists encouraging or discouraging specific normative discourses, also more authoritative institutions come into play in this respect.

Stakeholders representing (inter-)governmental funding programmes and grant schemes, such as Horizon 2020 for the EU or the US National Institutes of Health (NHI) programmes, have also taken an interest in big data-driven research. Big data are not only a development promising research innovation and improved healthcare, but also a way to reduce (healthcare) costs. Funding bodies and institutions are important stakeholders to consider, because they are decisive for the discursive governance of research. They set broader research agendas and appear as expressly influential stakeholders shaping discursive conditions. Therefore, this point will be covered more extensively in the next sub-chapter.

Instead of or besides derivative and descriptive bioinformation, biomedical researchers in big data-driven projects draw on data collected by stakeholders such as global internet and tech corporations. As big data collectors, the latter are key stakeholders, since they have come to be decisive gatekeepers for data access and analytics expertise. Corporate data collectors and scientific data utilisers are both discursively powerful groups. Yet (inter-)dependencies between these two may notably affect researchers' agency, in their role as big data utilisers, and their integrity and expert authority.

Researchers' big data practices and related ethical discourses are often inevitably linked to the data collection approaches of internet and tech corporations such as Alphabet and Google or Facebook. Such big data collectors define which data are retrieved, how these are processed and stored, and with whom they are shared. Moreover, these corporations progressively fund and support biomedical research. In this role, they add to (inter-)national grant schemes and funding provided by other industries, such as pharmaceutical companies. This engagement simultaneously incentivises research involving big data, a development which appears to be of corporate interest for multiple reasons.

Health data analytics as corporate services are an important development in this respect too. Being data-rich actors, internet and tech corporations have developed leading expertise in this field. This applies to the expertise of individuals employed at such companies, as well as data analytics and storage infrastructures. In this domain, one can observe two, interrelated trends: one is that researchers and/or public health agencies are acting explicitly as customers of tech corporations. They do not only draw on the data collected by tech corporations as outlined above, but may also make use of their data analytics services. The other trend is that tech corporations have shown an interest in biomedical data from public sources, since these can support them in developing and maintaining health related services.

The triple role of data collector, service provider and funding body is a defining feature of internet/tech corporations. It puts these stakeholders in a powerful position, with regards to biomedical big data generators and utilisers alike. Therefore, this aspect will be covered in greater detail in the sub-chapter after next. First, though, I expand on the role of (inter-)governmental funding schemes raised above.[44]

Funding Big Data-Driven Health Research

Due to the rising size and complexity of biomedical datasets, as well as the digital origins of certain data, computer/data science expertise has become more and more important for biomedical research. Emerging technosciences such as *bioinformatics* and *biocomputing* refer to interdisciplinary research approaches. They merge data science, computing and biomedical expertise. Scholars in the interdisciplinary research field of bioinformatics, for example, create platforms, software and algorithms for biomedical data analytics, knowledge production and utilisation (Luscombe, Greenbaum, and Gerstein 2001).

The emergence of such intersections between life/health sciences and computing is also linked to the tendency that contemporary funding schemes require technology development and private-public partnerships (see e.g. 'Information and Communication Technologies in Horizon 2020' 2015). Technological output such as software or hardware prototypes and applications is increasingly decisive for various national and transnational grants. This applies also and particularly to research on and with biomedical big data.

In 2012, the United States National Institutes of Health (NIH) launched a major data science initiative, called 'Data Science at NHI'. This involved creating a new position called Associate Director for Data Science, currently [January 2018] held by Philip Bourne, a computer scientists specialising in health research. Moreover, it established a new funding scheme called 'Big Data to Knowledge' (BD2K). The programme's main aim is to explore how biomedical big data may contribute to understanding and improving human health and fighting diseases (Data Science at NIH 2016).[45] The programme is divided into four main clusters: centres of excellence for big data computing (11 centres in 2017); resource indexing; enhancing training; and targeted software development. The latter framework provides funding for projects working towards software solutions for big data applications in health research.

The European Commission (EC) too displays a clear interest and mounting investments in big data developments. In 2014, the EC published an initial communication document titled 'Towards a Thriving Data-Driven Economy' (COM 442 final 2014). The document highlights the economic potential of big data in areas such as health, food security, climate, resource efficiency, energy, intelligent transport systems and smart cities. Stating that 'Europe cannot afford to miss' (COM 442 final 2014, 2) these opportunities, the document warns that European big data utilisation and related technologies lag behind projects established in the US. Three years later, in January 2017, a follow-up communication was released: 'Building a European Data Economy' (COM 9 final 2017) One of the aims declared in this document is to '[…] develop enabling technologies, underlying infrastructures and skills, particularly to the benefit of SMEs [small and medium enterprises]' (COM 9 final 2017, 3).

On the EC website, this big data strategy is also presented by posing questions such as: 'What can big data do for you?' Under this point/question, the first aspect mentioned is 'Healthcare: enhancing diagnosis and treatment while preserving privacy'. This emphasis indicates that big data are seen as important development in healthcare, but also that healthcare is showcased as an example of how individuals can benefit from big data. Building on these focal points, the EC provides targeted funding possibilities such as the call 'Big data supporting Public Health Policies' (SC1-PM-18. 2016) which is part of the programme Health, demographic change and well-being.

Projects like Big Data Europe, which involves a big data health pilot, also received funding from grant schemes such as 'Content technologies and information management: ICT for digital content, cultural and creative industries' (BigDataEurope 2016). Such trends relate back to the EC's *Digital Agenda for Europe* (*DAE*) (a 10-year strategy development running from 2010 until 2020) and its priority 'eHealth and Ageing'. The *DAE* aims at enhancing the EU's economic growth by investing in digital technologies. Complementing national and EU-wide efforts, it also entails endeavours for enhanced global cooperation concerning digital health data and related technologies ('EU and US strengthen collaboration' 2016). Moreover, biomedical big data funding initiatives have been set up by various governments in Europe (see e.g. Research Councils UK n.d.; Bundesministerium für Bildung und Forschung n.d.).

The World Health Organisation (WHO), as a United Nations (UN) agency, likewise takes an interest in the use of big data for health research, disease monitoring and prevention. Stressing that this development opens up new possibilities and challenges, the WHO's eHealth programme states: 'Beyond traditional sources of data generated from health care and public health activities, we now have the ability to capture data for health through sensors, wearables and monitors of all kinds' ('The health data ecosystem' n.d.). With regards to big data utilisation for public health and humanitarian action, the *WHO* collaborates closely with the UN Global Pulse initiative (see also chapter 3 on data philanthropy).

Global Pulse's main objectives are the promotion and practical exploration of big data use for humanitarian action and developments, notably through public-private partnerships (see 'United Nations Global Pulse: About' n.d.). It is organised as a network of so-called 'innovation labs': with a headquarter in New York and two centres in Jakarta (Indonesia) and Kampala (Uganda). These labs develop big data-driven research projects, applications and platforms which are closely connected to local communities in the respective area and country. Among other factors, Global Pulse was inspired by NGO research initiatives such as Global Viral (which is linked to the commercial epidemic risks analytics services offered by Metabiota Inc.), the Ushaidi crisis mapping platform, and Google Flu Trends (see UN Global Pulse 2012, 2).

This overview indicates that the 'big data agenda' (Parry and Greenhough 2018, 108), in these cases the promotion of big data's use for health research, is

not simply a bottom-up development stirred by individual researchers. Instead, the trend towards big data-driven health research is incentivised by authoritative institutions and actors, also in the role of funding bodies. It could be argued of course that most of these initiatives claim to go back to democratic processes, consulting experts and other stakeholders (Auffray et al. 2016). However, these consultations tend to privilege renowned experts and, to a lesser extent, patient advocacy groups, rather than directly involving actors who are affected by big data practices because they are made part of the data generation process.

Discursively, what is accentuated in (inter-)national funding schemes and policy documents is big data's impact on economic competitiveness, innovation and societal wellbeing. Considerably less emphasis is put on potential risks and uncertainties, although some improvement has been noticeable during the last two years. Thus, as stakeholders, these institutions also contribute to establishing big data as a field of interest for scientific research. The economic advantages, innovation potential and health benefits, alleged in respective grant schemes or policy documents, are authoritatively promoted as research rationales.

The Role of Tech Philanthrocapitalism

Apart from national and intergovernmental initiatives, private and corporate funding opportunities also play a role. Historically, this is of course by no means a new development in (biomedical) research. For example, in the US it was only in the 1940s that '[t]he national shift from primarily philanthropic to governmental funding took place as the National Institutes of Health (NIH) became the main vehicle for research' (Brandt and Gardner 2013, 27; see also Cooter and Pickstone 2013). In Europe, philanthropic organisations such as the (American) Rockefeller Foundation were very influential, notably in the context of World Wars I and II (Weindling 1993).[46] What is new however, is the peculiar role of internet and tech corporations. These companies have very specific interests and agendas, especially with regards to how their products may feature in contemporary research and in relation to public policies. Moreover, they invest in the development of health technologies considered auspicious additions to their product portfolio. In 2016 and 2017, for example, increasing venture capitalist and private equity funding was reported for digital health technologies (see e.g. Silicon Valley Bank 2017; Mercom 2016).

It has been noted that tech corporations increasingly receive public funding. Regarding privately held or mediated databases, Sharon (2016) observes that '[…] public money is channelled, indirectly or directly, to their development, as has been the case with 23andMe, which recently secured a US$1.4 million research grant from the NIH to expand its database, and with recent National Cancer Institute funding of Google and Amazon run genome clouds' (Sharon 2016, 569). These developments are part of the emerging data, analytics, skills

and infrastructure asymmetries depicted in Chapter 3. It is important to be aware of money and data not only flowing from tech corporations to (public) research institutions, but also vice versa. Since I mainly focus on studies conducted by academics at universities, however, the following sections describe investments and funding provided by internet/tech corporations for such research projects.

More generally, it has been argued that '[…] a transition from public to private sector funding has already taken place in some domains of the sciences' (Inverso, Boualam and Mahoney 2017, 54). One of these domains is biomedical research. A report by the American Association for the Advancement of Science shows that while federal government funding is still the main source for research, 'industry has caught up' (Hourihan and Parkes 2016, 6). A well-known issue in this context is that private funding tends to privilege research that promises to deliver short-term results and product development (ibid.). While private companies spend 80% of their research and development investments on development, only 20% go into basic and applied research, a ratio which is reversed for federal nondefense agencies in the US.

Even before the big data hype, in the early 2000s scholars observed that in the US, industry influence on biomedical research had dramatically risen within two decades (Bekelman, Li and Gross 2003). Based on an analysis of articles examining 1140 biomedical studies, Bekelman, Li and Gross (2003) showed that statistically '[…] industry-sponsored studies were significantly more likely to reach conclusions that were favourable to the sponsor than were nonindustry studies' (463). From an ethical perspective, the authors problematise conflicts of interests emerging from entanglements between researchers and industry sponsors.

These entanglements have a bearing on the results that certain research may generate. Furthermore, considering industry's tendency to sponsor development-driven research, this sways the type of studies being conducted. Given such earlier insights, we should carefully scrutinise how internet and tech corporations support and fund scientific research. Financial or in-kind support is commonly made in domains that are relevant to their economic, tech-political interests and their favourable public perception.

With regards to Google, a 2017 report published by the Google Transparency Project, an initiative of the US Campaign for Accountability, comes to the conclusion that: 'Google has exercised an increasingly pernicious influence on academic research, paying millions of dollars each year to academics and scholars who produce papers that support its business and policy goals' (Google Transparency Project 2017). The report highlights among other things that between 2005 and 2017, 329 research papers dealing with public policy issues in the interest of Google were funded by the corporation. Moreover, corporations such as Alphabet, as Google's parent company, are heavily investing in biotechnology start-ups.

In 2009, Alphabet launched Google Venture (GV) as its venture capital arm. Since then, GV has invested, for instance, in 23andMe[47], Doctor on Demand,

and Flatiron, a company developing cloud-based services for oncological (cancer research and care) data. Four years earlier, in 2005, Google started its charitable offshoot Google.org. In 2017, it was stated on the website of this Google branch that it annually donates '$100,000,000 in grants, 200,000 hours, $1 billion in products'. Investments and grants are particularly targeted at projects exploring how new technologies and digital data can be used to tackle societal and ecological challenges. Various Google-sponsored tech challenges/competitions worldwide complement these efforts.

Since 2016, 'Crisis Response' has been one of Google's declared focal points, next to 'Disabilities', 'Education and Digital Skills', and 'Racial Justice'. The crisis response team was already formed in 2010, in reaction to the 2010 Haiti earthquake and the ensuing humanitarian crisis. It provides services such as Google Public Alerts, Google Person Finder, and Google Crisis Map.[48] In February 2017, Google.org specifically highlighted its efforts in 'Fighting the Zika Virus' and 'Fighting Ebola'. From 2006 until 2009, Google.org was led by Larry Brilliant. Before his appointment, the physician and epidemiologist had been involved in various enterprises, ranging from research for the WHO to co-creating the early online community The Well as well as the health-focused Seva Foundation.

After leaving Google.org in 2009, Brilliant joined the Skoll Global Threats Fund (SGTF) as managing director. The SGTF is part of the Skoll Foundation (SF), an NGO initiated by eBay founder Jeff Skoll in support of 'social entrepreneurship'. It maintains the website endingpandemics.org which describes itself as a 'community of practice' aimed at accelerating the detection, verification, and reporting of disease outbreaks globally. Similarly to the SF, the Bill and Melinda Gates Foundation, with an endowment of $44.3 billion, proposes that '[w]e can save lives by delivering the latest in science and technology to those with the greatest needs'.[49]

Not only technologies, but also the funding enabled by profitable tech corporations has been styled as an important contribution to research and healthcare. In 2016, a philanthropic investment of Mark Zuckerberg and his wife Priscilla Chan was however rather controversially discussed, at least in San Francisco. After receiving a donation of $75 million from the couple, the San Francisco General Hospital and Trauma Center (where Chan was trained as paediatrician) was renamed into the 'Priscilla and Mark Zuckerberg San Francisco General Hospital and Trauma Center' ('Mark Zuckerberg and Priscilla Chan give $75 million' 2015). The decision to rename the hospital triggered criticism from some, because it was said to ignore the continuous input of taxpayers, as well as the alarming impact of Silicon Valley on San Francisco (Heilig 2015; Cuttler 2015).

Apart from such donations, less is known about Facebook's role and interest in health research applications. Information on this has been largely speculative, partly because only few official statements are provided on the company's interests in this domain. In 2013, a report by *Reuters* suggested that the company

was interested in establishing patient support websites such as PatientsLikeMe, as well as health and lifestyle monitoring applications involving wearable technologies (Farr and Oreskovic 2013). This initiative has not, however, materialised so far. Yet, Facebook often highlights its relevance as catalyst and enabler of health- relevant and humanitarian initiatives. This applies, for instance, to a status feature through which users can identify themselves as organ donors, and to 'Community Help' and 'Safety Check'. The latter are features allowing users to ask for support from others or indicate that they are safe, for example in areas hit by natural disasters.

Chan and Zuckerberg recently revealed the new health focus of The Chan Zuckerberg Initiative. This limited liability company (LLC) was founded in December 2015. After initially mainly investing in education and software training, The Chan Zuckerberg Initiative launched its science programme in September 2016. On behalf of Chan and Zuckerberg, it was declared that the programme would help 'cure, prevent or manage all diseases in our [Chan and Zuckerberg's] children's lifetime' (see also Heath 2016).

An important part of this science programme is the Chan Zuckerberg Biohub. The programme provides funds for this centre, which comprises (medical) researchers and engineers from Berkeley, University of California; University of California San Francisco; and Stanford University. In February 2017, the two main research projects were the 'Infectious Disease Initiative' and the 'Cell Atlas'. The Chan Zuckerberg Biohub, its funding structure, and its involvement of researchers are an example for emerging entanglements between university research on (public) health and tech corporations. The funding available to the 47 researchers part of the hub is unrestricted.

Zuckerberg is not the only Facebook founder investing in philanthrocapitalism. Also in 2017, the venture capital firm B Capital Group, co-initiated by Eduardo Saverin (co-founder of Facebook), invested in the technology start-up CXA group. Its declared aim was to '[t]ransform your current healthcare spending into a benefits and wellness program where your employees choose their own path to good health'. Already in 2011, another Facebook co-founder, Dustin Moskovitz, initiated the private foundation Good Ventures, together with his wife Cari Tuna. Good Ventures invests in domains such as biosecurity and pandemic preparedness, as well as global health and development.

While this is not an all-encompassing overview of corporate, philanthropically framed investments in the public health sector, it allows for initial insights into entanglements between internet and tech giants such as Alphabet and Facebook and contemporary research. More generally, since the 'Giving Pledge Campaign' was initiated by Bill Gates and Warren Buffett in June 2010[50], there has been an increase in diverse, tech philanthrocapitalist initiatives. While one may intuitively deem that philanthropic investments as such should not be seen as a problematic development, these practices raise considerable economic and ethical issues and contradictions. The Chan Zuckerberg Initiative

has been described as a poster child of *philanthrocapitalism* (Cassidy 2015), a term which has turned out to be an effective euphemism for a form of 'disruptive philanthropy' (Horvath and Powell 2016, 89).

Horvath and Powell (2016) argue that disruptive, corporate philanthropy bypasses democratic control over spending in domains significant to societies' wellbeing and public good. Relating this back to Habermas' deliberations on discourse ethics, this also implies that critical public debate on such issues is largely irrelevant for these corporate decision-making processes that are not overseen by institutions embedded in democratic processes. Three main, interrelated problems should be considered here: first, emerging dependencies between corporate actors, health researchers and public health institutions; second, the tendency that large sums of otherwise taxable money are invested into philanthropically framed projects; third, the influence which corporate actors exert on content choices and developments concerning health relevant research.

With regards to Google funding, it was observed that '[t]he company benefits from good PR while redirecting money into charitable investments of its choice when, if that money were taxed, it would go toward government programs that, in theory at least, were arrived at democratically' (Alba 2016). The work of Horvath and Powell (2016) is highly insightful in this regard, since they examine how the rise of corporate, philanthropic activity is linked to the decline of democracy (89; see also Reich, Cordelli, and Bernholz 2016). According to the authors, approaches to destructive philanthropy are characterised by three key features: 1) They attempt to change the conversation and influence how societies evaluate the relevance of current challenges and possible solutions. 2) They are built on competitive values. 3) They explore new models for funding public goods. With regards to the intersection of public health research and corporate big data, these are relevant considerations. Horvath and Powell (2016) illustrate aptly how efforts in destructive philanthropy shape what is seen as societal issues, and which methods are considered appropriate for addressing respective problems (see 89ff.).

These strategies stand in stark contrast to Habermasian principles for valid social norms, notably the requirement that persons should make assessments and decisions based on the force of the better argument. Given that powerful stakeholders such as leading internet and tech corporations are shaping relevant discourses, the basis for public debate appears troubled. It is also of concern that such corporate shaping of discourses occurs conspicuously by mobilising the credibility of scientific research. Tech/internet corporations' discursive and financial engagement at the intersection of technology and biomedical research raises the question how this may shape the public perception of big data.

Furthermore, notably in the US, novel, corporate funding mechanisms influence ethics review procedures and requirements. Rothstein (2015) depicts some of the practical consequences for big data-driven health research:

> 'Of immediate concern is that the use of personal information linked to health or, even worse, the intentional manipulation of behavior, is not subject to traditional, federal research oversight. The reason is that these studies are not federally funded, not undertaken by an entity that has signed a federal-wide assurance, and not performed in in contemplation of an FDA [US Food and Drug Administration] submission.' (425)

As the author implies, this raises the question whether adjustments in regulations for research are needed. It also begs the question of the responsibility and capacity of corporations to ensure that funded projects are equipped with and incentivised to address ethical issues.

Tech and internet corporations take great interest in maintaining and fostering a view of (their) technologies as beneficial to scientific advancements and societal wellbeing. As part of this broader agenda, they have also come to play an influential role in heralding the benefits of big data for public health. By providing funding, data, analytics and other support, they set incentives for researchers to engage in related technoscientific explorations. In doing so, they act as important gatekeepers in defining research choices as well as implementations. This seems all the more important, since internet/tech corporations often act as crucial data and analytics providers, a tendency which is highly salient for the field of digital public health surveillance.

Digital Public Health Surveillance

> 'I envision a kid (in Africa) getting online and finding that there is an outbreak of cholera down the street. I envision someone in Cambodia finding out that there is leprosy across the street.' (Larry Brilliant, in Zetter 2006)

Envisioning the benefits of new technological developments is a common practice. In competitive contexts – be it for start-ups competing for venture capital or researchers competing for funding – persuasive promises emphasising the need for and benefits of a product/service/technology are indispensable. It is therefore not surprising that projects involving biomedical big data have made bold promises. As Rip observes:

> '[P]romises about an emerging technology are often inflated to get a hearing. Such exaggerated promises are like confidence tricks and can be condemned on bordering at the fraudulent. But then there is the argument that because of how science and innovation are organised in our societies, scientists are almost forced to exaggerate the promise of their envisaged work in order to compete for funding and other resources.' (2013, 192/193)

This mechanism does not only apply to research. It likewise applies to corporations and their promotion of new technological developments and services, as illustrated with the above comment by Larry Brilliant. Google.org's former director ambitiously pushed and promoted its engagement in infectious disease prediction.

Epidemiology, and its sub-discipline of epidemiological/public health surveillance, has undergone significant changes since the 1980s. [51] Most recently, these are related to technological developments such as the popularisation of digital media and emerging possibilities to access and analyse vast amounts of global online user data. Epidemiological surveillance involves systematic, continuous data collection, documentation and analysis of information which reflects the current health status of a population. It aims at providing reliable information for governments, public health institutions and professionals to react adequately and quickly to potential health threats. Ideally, epidemiological surveillance enables the establishment of early warning systems for epidemic outbreaks in a geographic region or even multinational or global pandemics.

The main sources relevant to 'traditional' public health surveillance are mortality data, morbidity data (case reporting), epidemic reporting, laboratory reporting, individual case reports and epidemic field investigation. The data sources may vary however, depending on the development and standards of a country's public health services and medical facilities. Since the 1980s at the latest, computer technology and digital networks have become increasingly influential factors, not merely with regards to archiving and data analysis, but in terms of communication and exchange between relevant actors and institutions. Envisioning the 'epidemiologist of the future', Dean et al. suggested that she/he '[…] will have a computer and communications system capable of providing management information on all these phases and also capable of being connected to individual households and medical facilities to obtain additional information' (1994, 246).

The French Communicable Disease Network, with its *Réseau Sentinelles*, was a decisive pioneer in computer-aided approaches. It was one of the first systematic attempts to build a system for public health/epidemiological surveillance based on computer networks. Meanwhile, it may seem rather self-evident that the retrieved data are available online. Weekly and annual reports present intensities (ranging from 'minimal – very high activity') for 14 diseases, including 11 infectious diseases such as influenza.[52]

Similar (public) services are provided by the World Health Organisation's (WHO) 'Disease Outbreak News',[53] the 'Epidemiological Updates'[54] of the European Centre for Disease Prevention and Control (ECDC) and (only for influenza cases in Germany and during the winter season) by the Robert Koch Institute's 'Consortium Influenza'. With its *Project Global Alert and Response (GAR),* the WHO additionally establishes a transnational surveillance and early-warning system. It aims at creating an 'integrated global alert and response system for epidemics and other public health emergencies based on

strong national public health systems and capacity and an effective international system for coordinated response'.[55]

In this sense, computerisation and digitalisation have significantly affected approaches in epidemiological surveillance for decades. However, one aspect remained unchanged until the early 2000s: these were still relying on descriptive and derivative bioinformation, for example data from diagnostics or mortality rate statistics. In contrast, more recent strategies for epidemiological surveillance have utilised 'digitally-born' biomedical big data. Various terms have been coined to name these developments and linguistically 'claim' the field: infodemiology, infoveillance (Eysenbach 2002, 2006, 2009), epimining (Breton et al. 2013) and digital disease detection (Brownstein, Freifeld and Madoff. 2009).

Approaches to digital, big data-driven public health surveillance can be broadly categorised according to how the used data have been retrieved. Especially in the early 2000s, digital disease detection particularly focused on publicly available online sources and monitoring. For example, news websites were scanned for information relevant to public health developments (Zhang et al. 2009; Eysenbach 2009). With the popularisation of social media, it seemed that epidemiologists no longer had to wait for news media to publish information about potential outbreaks. Instead, they could harness digital data generated by decentralised submissions from millions of social media users worldwide (Velasco et al. 2014; Eke 2011).

Platforms like Twitter, which allow for access to (most) users' tweets through an open application programming interface, have been considered especially useful indicators of digital disease developments (Stoové and Pedrana 2014; Signorini et al. 2011). Moreover, attempts were made at combining social media and news media as sources (Chunara et al. 2012; Hay 2013). Other projects used search engine queries in order to monitor and potentially even predict infectious disease developments. The platforms *EpiSPIDER*[56] (Tolentino et al. 2007; Keller et al. 2009) and *BioCaster* (Collier et al. 2008) combined data retrieved from various online sources, such as the European Media Monitor Alerts, Twitter, reports from the US Centers for Disease Control and Prevention and the WHO. The selected information was then presented in Google Maps mashups. However, these pioneer projects seem to have been discontinued, whilst the *HealthMap* platform is still active (see Lyon et al. 2012 for a comparison of the three systems).[57]

Big data produced by queries entered into search engines have also been utilised for public health surveillance projects. In particular, studies by Eysenbach (2006), Polgreen et al. (2008) and Ginsberg et al. (2008) have explored potential approaches. The authors demonstrated that Google and Yahoo search engine queries may indicate public health developments, while they likewise point to methodological uncertainties caused by changes in users' search behaviour. Such approaches using search engine data have been described as problematic, since they are based on very selective institutional conditions for data access,

and have raised questions concerning users' privacy and consent (Richterich 2016; Lupton 2014b, Chapter 5).

In this context it also seems significant that a project such as Google Flu Trends, which was initially perceived as 'poster child of big data', was discontinued as a public service after repeated criticism (Lazer et al. 2014; 2015). The platform predicted influenza intensities by analysing users' search queries and relating them to influenza surveillance data provided by bodies such as the ECDC and the US CDC. The search query data are still being collected and exchanged with selected scientists, but the project is not available as a now-casting service anymore. Instead, some indications of the data are published in Google's 'Public Data Explorer'. In light of such developments and public concerns regarding big data utilisation (Science and Technology Committee 2015; Tene and Polonetsky 2012, 2012a), ethical considerations have gradually received more attention (Mittelstadt and Floridi 2016; Vayena et al. 2015; Zimmer 2010).

While it has been discontinued as a public service, 'Google Flu Trends' is still an illustrative example which highlights how collaboration between epidemiologists and data/computer scientists facilitated research leading to a concrete technological development and public service. Some of the aforementioned authors, such as Brownstein, Freifeld, and Chunara, have also been involved in research aimed at developing digital tools and applications in digital epidemiology. For example, they created the websites and mobile applications HealthMap (which also receives funding and support from Google, Twitter, SGTF, the Bill and Melinda Gates Foundation, and Amazon) as well as FluNearYou. HealthMap draws on multiple big data sources, for example, tweets and Google News content, while FluNearYou is an example of 'participatory epidemiology' and presents submissions from registered community members.

Considering such entanglements between big data collectors and data utilisers, an analysis of individual research projects appears insightful and necessary. This chapter explored how relevant stakeholders are involved in shaping the discursive conditions for big data-driven health research. But which ethical discourses have in fact evolved under the described discursive conditions? In response, the following chapter examines which ethical arguments have been mobilised in research projects and big data-driven approaches to public health surveillance. It shows which validity claims have been brought forward. Particular attention is paid to validity claims to normative rightness, although it appears characteristic for big data-driven research discourses to interlink ethical arguments with validity claims to truth.