Article

Theory Culture & Society

Theory, Culture & Society
30(4) 131–150

DOI: 10.1177/0263276413485900
tcs.sagepub.com

SAGE

Illness Online: Self-reported Data and Questions of Trust in Medical and Social Research

Sally Wyatt
Maastricht University, The Netherlands; Royal Netherlands Academy of Arts and Sciences,The Netherlands

Anna Harris
Maastricht University, The Netherlands

Samantha Adams
Erasmus University Rotterdam, The Netherlands

Susan E Kelly
University of Exeter, UK

Abstract
Self-reported data are regarded by medical researchers as invalid and less reliable than data produced by experts in clinical settings, yet individuals can increasingly contribute personal information to medical research through a variety of online platforms. In this article we examine this 'participatory turn' in healthcare research, which claims to challenge conventional delineations of what is valid and reliable for medical practice, by using aggregated self-reported experiences from patients and 'pre-patients' via the internet. We focus on *23andMe*, a genetic testing company that collects genetic material and self-reported information about disease from its customers. Integral to this research method are relations of trust embedded in the information exchange: trust in customers' data; trust between researchers/company and research subjects; trust in genetics; trust in the machine. We examine the performative dimension of these trust relations, drawing on Shapin and Schaffer's (1985) discussion of how material, literary and social technologies are used in research in order to establish trust.

Our scepticism of the company's motives for building trust with the self-reporting consumer forces us to consider our own motives. How does the use of customer data for research purposes by *23andMe* differ from the research practices of social

Corresponding author:
Sally Wyatt, e-Humanities Group, Royal Netherlands Academy of Arts and Sciences, PO Box 94264, Amsterdam, 1090 GG, The Netherlands.
Email: sally.wyatt@ehumanities.knaw.nl

scientists, especially those who also study digital traces? By interrogating the use of self-reported data in the genetic testing context, we examine our ethical responsibilities in studying the digital selves of others using internet methods. How researchers trust data, how participants trust researchers, and how technologies are trusted are all important considerations in studying the social life of digital data.

Keywords
data, digital, genetics, methodology, technology

Since their earliest days, computers and networked computers have been used in academic, government and commercial research settings (Abbate, 1999; Agar, 2006) because they are extremely good at storing and processing structured data, and powerful computational tools can help researchers to analyse and represent complex data. Since what we now call the internet became publicly and commercially available in the early 1990s (Thomas and Wyatt, 1999), the digitally mediated communication of ordinary people has proved to be a fascinating source of (self-reported) data, which has only intensified with the rise of so-called 'web 2.0' or social media applications. The use of transactional data generated by billions of people going about their everyday (online) activities is of enormous value not only to market researchers and intelligence services but also to scholars in the humanities and social sciences (Savage and Burrows, 2007).

This is particularly evident in fields such as healthcare, where information about people's states of health and illness is now easily amassed digitally, for research and other purposes, through a range of technological artefacts and practices (Fearnley, 2006; Ginsberg et al., 2009; Oudshoorn, 2011; Foster and Young, 2012; Pols, 2012). Digital technologies facilitate not only more detailed measurement and monitoring (Lupton, 2012) but also the diffusion of responsibility for providing data from medical researchers to individual citizens, who are considered to be particularly helpful in generating 'zettabytes of medical data' (Swan et al., 2010). Increasingly, individuals are encouraged to become more actively involved in data collection because they are considered to be the best (yet often 'untapped') resource for information about their own states of health and illness.

While we do not wish to exaggerate the novelty of self-reporting in research, medical research *is* changing in that it increasingly relies on networks of data and collections of tissues stored by research institutions (Lipworth et al., 2011), with online platforms, large datasets and computational abilities allowing new kinds of research. The use of self-reported data is particularly controversial, as it blurs and contests the boundaries of expertise previously established in the medical research world. Self-reported data acknowledge lay knowledge about one's body and healthcare experiences, yet are often criticized by scientific

and medical researchers who believe they are unreliable (Arnquist, 2009; Prainsack, 2011). By showing trust in self-reported data, in the individual as a source of information about their own health behaviour, rather than the medical record produced by clinicians, these practices foster new ways of thinking about the relations of trust between research participants and researchers.

This article therefore focuses on a context where individuals are actively encouraged to collect and communicate data about their health and illness for medical research purposes. Rather than examining self-reporting for clinical purposes, we are interested in intentional reporting of data outside of the clinical encounter that are nonetheless used for medical research purposes. Specifically, we focus on how data are obtained and what kinds of trust issues are raised by self-reporting. We work on the premise that issues of trust are brought into relief wherever there are perceived to be exchanges – or the possibility of appropriation – of information and money. In their analysis of the successful rise of the scientific experiment in the 17th century dispute between Hobbes and Boyle, Shapin and Schaffer (1985) show how 'matters of fact' became established using material, literary and social technologies. Similarly, we show how trust, another problem of social order, can be built using material, literary and social technologies, all three of which are now also affected by the digital. This is especially true when the science (genetics) and the means of data collection (self-reported data via the internet) are themselves emerging. Thus the trust relationships between actors from science, industry and the public, themselves often mediated via the internet, are precarious and require constant attention.

How does a company that encourages self-reporting by individuals represent and establish trust in its products and services and in its means for interacting not only with its customers but also with its investors and research partners? We answer this question by examining the case of *23andMe*, a company that sells genetic tests directly to customers and uses genetic and phenotypic data provided by the customers to conduct scientific research. We begin with a short review of the developing relationship between the 'participatory turn' in medicine and the role of reporting data through technology. We then explore what kinds of trust relationships are entailed in participation in this online genetic research, in particular how the commercial enterprise for genetic research, *23andMe*, promotes trustworthiness in the company, genetics, research and technology in order to encourage participation in what it describes as a research revolution. We show how the company attempts to use material, literary and social technologies (Shapin and Schaffer, 1985) in order to perform trust between itself and its customers, its investors and the scientific community. We examine how the company itself comes to trust its customers/medical subjects/citizen researchers/biosubjects/guinea pigs and the mutual (sometimes fragile) trust relationships that must be

established and sustained. Before concluding, we reflect on the implications of our analysis for our own position as researchers in the online environment of user-generated data.

Self-Reported Data and 'the Participatory Turn'

The self-reporting of personal health and illness states forms an important element of the social life of digital data in research relationships of the online personal genomics industry. These relationships are rhetorically constructed as moving personal health outside the realm of medical expertise and drawing on the embodied personal knowledge of the individual participant/consumer, aligning the sharing of personal data with empowerment enabled by access via digital media. However, we argue that trust in self-reported data, in line with other forms of online production, should be considered as performatively constructed by social actors in particular contexts.

Self-reporting for health research can take numerous forms, such as reporting experience about drug reactions, reporting symptoms of infectious disease on a health map or providing details of one's medical history. Self-reported data are used by an increasing number of online non-profit health organizations such as PatientsLikeMe (Allison, 2009; Wicks et al., 2011) and others (e.g. the Personal Genome Project, Genomera, TuAnalyze and LAMsight), as well as in state-run databases (e.g. the Icelandic Biobank) and commercial enterprises (e.g. *23andMe*).

Self-reported data seldom stand alone or unmediated in the disciplines to which they are integral, such as medicine and psychology. Self-reports of bodily states and health-related activities form a central component of the traditional medical interview; however, they are 'translated' or mediated into 'objective' signs and symptoms through the 'clinical gaze' (Foucault, 1963) of the physician. This mediation constitutes work, the work of re-positioning self-reported information within expert realms of 'seeing and knowing'. Mol (2002) emphasizes how medicine 'enacts' the body and disease that are the object of its practice, thus moving away from epistemological questions about accuracy and underlying truths that would conceptualize self-reports in terms of the extent to which they represent reality. Mol points out that a patient and a physician are not talking about exactly the same thing when they discuss a disease state. The practices of coordination involved in communicating and acting across this ontological plurality are less apparent when self-reported data are communicated digitally. It is not apparent what coordinating work is done – and how, by what agents, etc. – in bringing together multiple possibilities of representation and of context.

Thinking of self-reported data online as a form of representation calls attention to the nature of actors, the forms of activity in which they are engaging, and the contexts of data production (see, for example,

White, 2002). Although patients, citizens and 'experience-based experts' (Collins and Evans, 2002) have been participating in scientific and medical research endeavours for centuries (e.g. Star and Griesemer, 1989; Lawrence, 2006; McCray, 2006; Bruyninckx, 2013), the recent 'participatory turn' (Tutton and Prainsack, 2011) in healthcare research claims to challenge the conventional delineation of medical expertise, by using the aggregated self-reporting of experience from patients and 'pre-patients' through the internet. Patient advocacy group participation in particular has been discussed in relation to myopathies (Callon and Rabeharisoa, 2003, 2008), human immunodeficiency virus (Epstein, 2008), stem cell research (Langstrup, 2011), autism and Tourette's syndrome (Panofsky, 2011). Tutton and Prainsack (2011) compare direct-to-consumer genetic testing (DTC GT) research practices to those of population biobanks, and identify the 'entrepreneurial' subjectivities of DTC GT participants. Within the healthcare context, broadly defined, online participation covers a variety of practices, including the counterparts to those mentioned above, such as online patient advocacy (Akrich, 2010), sharing of patient experiences (Adams, 2010), as well as newer forms of participation facilitated by social media such as health hacking and providing data for medical research (Harris et al., 2013).

Filled with 'aspirations, promises, expectations, hopes, desires and imaginings' (Brown, 2003: 4), this participatory movement in medical research has been interpreted by many commentators as a move towards more empowered patients (Arnquist, 2009), and part of more democratic healthcare practices already thought to be enabled by the internet more broadly (Piras and Zanutto, 2010: 586). The general argument behind these promises and expectations is that while people were primarily recipients of online health information in the early days of the internet, with the rise of 'web 2.0', people also produce it, by sharing personal healthcare experiences (e.g. in the form of blogs and fora comments) as well as by contributing to various research enterprises, in practices described as crowd-sourcing or open-source research (Arnquist, 2009). In many cases, participating as an individual is linked to benefits for the community, whereby the idea of contributing data and information becomes intertwined with ideals of good citizenship (Adams, 2010) and improved science. Active health citizenship therefore entails not only responsible self-monitoring, self-care (Piras and Zanutto, 2010: 586) and self-tracking,[1] but also self-reporting.

The move toward reliance on self-reported data that this participatory turn requires brings longer-standing issues of trust to the fore. Self-assessment of symptoms, health and illness, and the validity of self-reported data, are not taken on their own in the medical context; they are contextualized in various ways by the medical professional, using instruments, practices and the application of expert knowledge. This is an inherent tension and one that is at the heart of diagnostic practices,

understood as social (Jutel, 2010). Further, medical interviews are structured by power relationships (e.g. Waitzkin, 1991) and some self-reports in the medical context are given more credibility than others (West, 1984). The use of self-reported data in medical surveillance and in medical research has been called into question (Gordon et al., 1993; Smith et al., 2008), particularly in the online, personal genomics context (Hall et al., 2009). This means that trust becomes actively constructed through materials and discourses in the interactions between various actors. In the next section we examine how the company enrols the consumer into self-reporting data online, and the trust relationships enmeshed in this exchange of information.

Initial Enrolment of the Consumer-Research Participant

23andMe is one of the largest and best-known companies offering genetic testing online. Upon entering the company's website, the potential consumer is greeted by clean, simple, slick web pages, with text promoting the empowering potential of personal genetics, images of healthy-looking people and hyperlinks to other internet platforms such as the company blog, Facebook and Twitter. Genetic tests are offered for health (our focus) and ancestry. The visitor to the site can read about people engaging in feel-good genetic discovery, connecting to each other through their genetic results, and the latest of the company's research endeavours.

These material and literary technologies contribute towards the company's attempts to build a personal, trusting relationship with the consumer. This engagement is personalized through language use, such as the use of 'You' for example, the company slogan of 'genetics just got personal' highlighting this further. *23andMe* also promotes links between the company and trustworthy organizations through highlighting funding it received from the National Institutes of Health or providing profiles of its scientific advisory board members with details of their university affiliations. It is important for the company to foster trust in itself, not only to enrol people into becoming consumers, but also subsequently to enrol these consumers into becoming research participants. Trust is not an inherent property of the site and isn't automatic on the part of the person visiting the site. These relations need to be built into the internet-mediated transactions, and draw upon broader trust relations.

A trust relationship must be established between the consumer, the company and the product it is selling. People who choose to enter this genetic testing marketplace pay to provide a saliva sample and in exchange are provided with information about their genetic make-up. Because the product is a genetic test, the potential consumer needs to have some element of trust in genetics. *23andMe* fosters the

geneticization of health and behaviour on its website, although acknowledging in relatively 'small print' the 'influence' of environmental and other factors.

The potential consumer must also have some trust in the internet. On the *23andMe* website, emphasis is given to the security of the site as a place to share information: genetic, financial and health-related. In the privacy statement, there are technical details about firewalls, secure online payment systems, genetic information and health information being kept separate from account information, encryption of data and connections, monitoring of employees' use of databases and restricted access to internal servers and the data centre. *23andMe* has also obtained a Certificate of Confidentiality from the US Department of Health and Human Services that gives it legal bounds to protect consented research participants' data from subpoenas.[2] *23andMe* presents the internet as a relatively risk-free place in which to find out genetic information about oneself, information that is not necessarily tied to one's medical record.

Keeping the Consumer-Research Participant Active

Once enrolled as a customer, individuals are invited to engage in further company-related participatory practices. *23andMe*, for example, invites its customers to share genetic information with other users, to find 'relatives' and to post on community fora.[3] The company promotes 'sharing' and attempts to normalize this activity, as a social and material technology. Since 2008, the company has been asking its customers to share even more information: data about their health and other traits, or self-reported phenotypic data. The data are combined with customers' genetic information to create a research database in order to conduct genome-wide association studies (GWAS).

Conducting genetic research has always been part of the business profile of this company, a move strongly backed by the husband of one of the co-founders, Sergey Brin, a founder of Google and a major funder of *23andMe*, described by *WIRED* magazine as a man who wants to 'bypass centuries of epistemology in favor of a more Googley kind of science' (Goetz, 2010). *23andMe* claim to be 'revolutionizing' research by building 'an entirely new model for conducting research' which sets 'the standard for web-based genetic studies' (Eriksson et al., 2010: 17).

Participating in *23andMe* research involves filling in surveys about health and other traits on the website. Surveys such as 'Ten Things About You', 'Health Habits' and 'Ten More Things About You' are posted on the company blog, sent by email, pasted into forum threads, advertised through Twitter and greet consumers when they log in. The surveys ask questions about pulse rates, cholesterol levels, eye colour or family history of disease. As with similar surveys on 'taste' websites such as *Hunch*, or *Good Reads*, these surveys are designed to be simple,

enticing, fun and addictive in order to encourage participation. The data from these surveys are combined with the genetic information analysed from the consumers' saliva samples, and mined for various genetic associations.

The company highlights that the self-reported method of data collection offers a speedy, 'web-based' alternative to 'traditional methods' of medical research (Eriksson et al., 2010: 2) such as collecting information from medical records or directly by medical researchers. Recognizing that there are some 'errors' in self-reported data the company states that the large numbers that it can generate using these methods outweigh these limitations. In order to help keep response rates high, partially completed self-reported surveys are also used. Thus, the databases are created using not only a controversial method, self-reporting, but also incomplete data, something that many genetic researchers would disregard (Kotz, 2012: 3).

Three years after starting the research programme, the company reported that more than three-quarters of its 100,000 consumers had agreed to take part in research activities, with 60 per cent having taken surveys and hundreds submitting research topics. As the customer base has increased to at least a reported 150,000[4] in 2012, it can only be assumed that the number of research participants has also increased. As of late 2012, *23andMe* has published three research articles based on what it describes as a 'participant-led', 'patient-driven', 'consumer-enabled', or 'consumer-driven' research methodology (Eriksson et al., 2010; Do et al., 2011; Tung et al., 2011). Each article has been authored by company researchers and company founders. Recently the company also applied, successfully, for a patent from one of its novel genetic findings, and has made its first acquisition, of another self-reporting website called *Cure Together*.[5] We focus on the company's own research, but it is worth noting that many consumers have access to their raw *23andMe* genetic data, and may contribute to other research projects or 'participatory' ventures (see Swan et al., 2010 for example).

In order to enrol research participants, and to encourage them to share further personal information by answering surveys, the company must build upon the trust relations it has already established with the consumers. From our analysis of the *23andMe* website material, we found that they do so performatively, in a number of ways. They put 'a face' to the research, personally introducing different members of the research team on the website and in the blog. Lab certifications are also displayed on the site, and links are provided to scientific studies, not only those with genetic findings upon which they base their analyses, but also those using similar research methodologies (genome-wide association studies, or GWAS) to their own techniques.

Public engagement and trust are further built through the early feedback of research results to participants (and would-be participants),

using various platforms such as blogs, fora and Twitter. This is something currently encouraged in medical research, but as yet under-realized. Results are fed back to participants at various stages of the research, from immediately after completing the surveys, when participants can see how they compared to others, to blog posts about recently published research articles in open-source journals. The company states in a research article that this is one of its research strategies, that 'a platform like this one that maintains an on-going relationship with participants, including sharing data with them, may motivate individuals to participate and stay active in research' (Tung et al., 2011: e9). A reliable cohort of individuals willing to supply self-reported data is an incredibly valuable resource to such an enterprise.

Establishing Trust with Investors

Establishing relations of trust with consumers/participants is an important aspect of the *23andMe* research project; however, in order to have the infrastructural arrangements to conduct this research, the company relies on venture capital and a trusting relationship with investors, combining material and social technologies. We have already mentioned one of the major funders of this company, Google, and companies in biotechnology and other sectors are financially supporting *23andMe*. Funders want to see a return for their investment, and it increasingly appears as if profits will not be obtained from the sale of genetic tests, but rather from the potential of the research database to generate revenue from pharmaceutical companies, from other biotechnology firms, and through the development of patents.[6]

In order to secure these profits, however, the company potentially jeopardizes its trust relations with consumers, as highlighted when the company announced its first novel genetic association patent on its blog. While there were several positive comments about this development, the reaction from consumers was largely hostile, many considering that they had been 'duped' into participating in this revenue-generating venture disguised as a participatory patient-led initiative. Lawyer and commentator on genetic testing, Daniel Vorhaus observed that it was a surprising move for the company to apply for this patent, which is of secondary importance to its most valuable asset: 'an engaged, enthusiastic and growing community of customers-qua-research participants who, provided 23andMe can keep from alienating too many of them, represent something much more unique, and inventive, than US Patent number 8,187,811'.[7] The controversy that the patent application evoked highlights the fragility of trust relations in this context. As Sterckx et al. (2012: 5) have suggested, 'what undermined trust was not so much the profit motive but rather the fact that the company did not provide any clear indication to consumers that it was seeking patents on its

discoveries'. As internet users become increasingly aware of the business practices behind sites (for example, in the highly-publicized dispute about properly informing customers about privacy policies and settings on the social networking site Facebook), maintaining trust amid growing consumer scepticism (e.g. through information strategies that demonstrate transparency in practice) will remain important.

Establishing Trust from the Scientific Community

While *23andMe* does much to distinguish itself from 'traditional science', it also relies extensively on science in order to establish its legitimacy as a research organization. The company draws on 'traditional' medical research to promote the GWAS method, for example. Its own submission of articles to peer-reviewed science journals[8] shows that it wants to position itself as a legitimate research organization within the scientific community. Obtaining ethics approval is one way to do this, and the company sought and obtained ethics approval from an (albeit commercial) Institutional Review Board for its research even though this was not formally required for the kind of research it was conducting.[9] In order to publish its research, *23andMe* must obtain the trust of journal editors, reviewers and readers, despite its 'unorthodox' means for collecting data (that is, unorthodox in the context of modern scientific methods and how the validity and reliability of research is assessed).

In order to establish itself as a legitimate scientific research organization, and present its material to research communities and journal editors, the company must itself demonstrate trust in the self-reported data provided by its consumers. The company advertises its trust in self-reported data through rhetorical statements on the website; however, the claims in scientific papers are somewhat more modest, recognizing that some data being used are 'incomplete' (Do et al., 2011). It also admits that there may be data that are more trustworthy than others, stating in one study that 'some classes of diseases were likely not well phenotyped … through some combination of misdiagnosis and misreport' (Tung et al., 2011: e7). Examples included autoimmune diseases which have a low prevalence and non-specific symptoms, or psychiatric disorders for which diagnosis requires a subjective clinical evaluation, and where it may 'make more sense to have a family member, friend, or caregiver provide information for an individual' (Tung et al., 2011: e7).

This points to the importance of verified (and the verifiability of) data in a research context, an issue that is related to the reliability and representativeness of the sample. Evidence-based medicine through Randomized Controlled Trials (RCTs) is the dominant research standard in Western medicine. Research findings are confirmed through replicated trials and theory testing. For a site such as *23andMe*, having the research

results be taken seriously in, and picked up (i.e. trusted) by, the medical community means providing data that can be checked (in terms of origins, correctness, comparability, etc.) and confirmed as correct. Verifiability is currently an aspect that lies on the fringes of the research, with the company claiming that it attempts to 'verify' medical diagnoses to the best of its ability in the surveys, and that other data may need to be 'verified' by additional interviews.

Representativeness of the sample is trickier to establish when data are collected via the web. There has been some question about whether the contributions to websites reflect the readership or (potential) audience of a website, which is already limited in relation to the population at large (Genes, 2006). In a recent analysis of the availability of large quantities of data and what this means for social research, boyd and Crawford (2012), however, suggest that while some scholars are careful to point to the limitations of e.g. Twitter data in their published work, the public discourse often conflates 'people' and 'Twitter users'. They explain that Twitter makes only a fraction of its material available, that Twitter users are not representative of the global population, that Twitter accounts and users are not equivalent, and that some users are much more active in producing content than others. Furthermore, some users have multiple accounts, and some accounts are 'bots' (software applications that run automated tasks on the internet). Similar observations have been made about Wikipedians (Pentzold, 2011), and they also apply to the claims being made by *23andMe*.

At the same time, it is in the interest of *23andMe* to demonstrate that it works with 'representative' samples, in accordance with the established scientific norms for medical research. In 2011, *23andMe* offered its tests to African Americans for free, an implicit recognition that its database under-represented this group. This distinguishes the case from e.g. Twitter because the kinds of knowledge produced by and for individual (always human) users are very different from the kinds of knowledge that can be produced using advanced computational techniques to harvest and analyse millions of tweets. The complicated issue of representativeness of the *23andMe* database highlights the work/performance required to establish and maintain various types of trust relations that are integral to the overall project, and the various material, social and literary technologies used to do so.

Using Self-Reported Data in Social Science Research

Interrogating the use of self-reported data in the genetic testing context gives us cause to reflect upon how issues related to establishing trust influence the positions and research of social scientists who themselves conduct web-based research. We need to remain aware of the likely mismatch between users of a particular application and the population as a

whole, as boyd and Crawford (2012) remind us in their study of Twitter. We also need to remain aware that technology is not a neutral tool that gives researchers unmediated access to data, references, ideas and people. Technology mediates and structures researchers' interactions at all stages of the research process, just as it structures people's efforts to find and share information about their own health. Computational tools, including algorithms, linked data and databases, can be used to harvest data from multiple and diverse sources in order to recombine them for other purposes. While such tools offer many possibilities and may well reduce the time and effort associated with some tasks, we need to remain alert to the ways such tools may render invisible some literature, information, data, categories, institutions, or people (Bowker and Star, 1999). Digital technologies can affect medical and scholarly research in a variety of ways, all of which may raise ethical and normative questions at any moment in the process and may affect researchers' relationships not only with research participants but also with colleagues, funders and users of research. In order to understand the full dimensions of these changes, it is important to take technology seriously, and consider carefully how the qualities of digital devices demand the rethinking of many assumptions in social science (Ruppert et al., this issue).

In the preceding pages we have focused on how a private company makes use of both genetic and phenotypic data provided by people who pay to do so. Similar developments (though usually without asking people to pay) can be observed in the social sciences where the availability of vast quantities of online data generated by people as they shop, travel and visit websites and social media may, some claim, seem to obviate the need for traditional social science methods such as the survey, the interview or the focus group (Savage and Burrows, 2007). Social scientists have long grappled with the disjunction between what people say and what people do. The combination of utterances captured when people use social media and their traces when engaging in computer-mediated exchanges means that social scientists now have access to both what people say and what they do. In the using of data generated as people go about their normal activities, data-intensive science may even obviate the need for theory. Levallois et al. (2013) examine these debates in sociology and economics, and conclude that what is taking place is not so much a shift from one paradigm to another but a realignment, both within and between disciplines.

This implies that researchers are confronted with various ethical dilemmas when dealing with the vast quantities of data online, especially data generated by people going about their daily lives, either as traces of transactions and movements or as more-or-less considered reflections via social media. Are they simply to be treated as any other publicly available information? If so, then appropriate citation and acknowledgement of sources solves many problems. For much information found online,

that is indeed more than adequate. However, in informal settings, including many patient groups, such an approach would not meet the basic ethical principle, of protecting people, sometimes from themselves.

Debates around questions of the privacy and anonymity of respondents (and researchers) have been well-rehearsed (e.g. Ess and AoIR Ethics Working Committee, 2002; Ess, 2009; Wyatt, 2012). Problems arise because protecting the human subject is seen as the primary obligation of individual researchers, professional disciplinary associations and ethical review committees. It is assumed there is a simple, clear relationship between data found online and individuals in the real world. This is what Carusi (2008) calls 'thin' identity, and what Beaulieu and Estalella (2011) refer to as 'traceability'. Given that individuals may reveal a great deal of personal information online, researchers have to be concerned to protect the anonymity and privacy of research subjects. The danger facing social researchers and their respondents is that individuals could be easily identified and traced if, for example, their words, avatars, or nicknames are mentioned in academic texts. Even when attempts are made to anonymize individuals, other details (in combination with search engines with increasingly attuned algorithms) could more or less inadvertently enable their identification. For example, in his critique of the 'Tastes, Ties and Time (T3)' research project based on the Facebook accounts of a cohort of university students, Zimmer (2010) points out that despite measures taken to protect the anonymity of the students, it did not take long to identify the university and even individual students, based on analysis of the course offerings and other demographic data. Elsewhere, he questions whether users of social media fully understand the trade-offs that they are making when posting personal information online (Zimmer, 2008).

Individuals going about their everyday online lives are not obliged to be part of research. Even if they are voluntarily providing extensive details about their calorie intake, drug reactions or mental health status, it does not necessarily mean that this information is fair game for researchers. It is not always adequate for researchers to say this information is in the public domain. Nissenbaum (2010) discusses this in terms of 'contextual integrity' and draws attention to the expectations of privacy people may have in particular contexts, online and offline. She highlights the right to privacy 'neither as a right to secrecy nor a right to control but a right to appropriate flow of personal information' (Nissenbaum, 2010: 127). Similarly, Bakardjieva and Feenberg (2000) point to the dangers of alienation arising from indiscriminate use of material found online, alienation experienced by people who have provided information in one context who may understandably not be happy to find it taken up in another. These concerns have become more acute with the rise of integrated data and more powerful search techniques, working precisely to cross contexts. Furthermore, healthcare

practitioners, policy makers and researchers may sometimes have good reasons to integrate data about individuals from different domains in order, for example, to examine the relationship between income inequality and life expectancy.

Others point to the dangers associated with assuming an isomorphic relation between individuals and some of their online utterances, and argue for treating online material as forms of representation. By doing so, other ethical issues and responsibilities emerge. For White (2002), it is important to recognize the constructed nature of online material, so that researchers can challenge the abundance of hate speech that is easily found. She argues that by recognizing the highly mediated and representational character of online material and by considering the ethical codes of literary studies or of art history and visual culture, different sorts of research questions could be addressed, opening up different forms of analysis. Bishop (2009) bemoans the fact that a focus on the privacy of respondents makes it more difficult to consider the ethical obligations researchers may have to other actors and stakeholders.

Opening up the possibility that online data may be constructed by social actors for particular audiences in particular contexts raises challenges, especially for medical researchers working within more realist research frameworks dominated by the RCT, and for social scientists studying both everyday life and medical science. In many ways, these issues are neither new nor specific to web 2.0 and/or the health domain. Issues related to ethics, trust, representativeness, online identity, and so on have been raised since researchers began studying the web in the mid-1990s. But the web and associated research opportunities continue to grow and change, which may lead to new issues, new iterations of old issues or changes in the nature and scale of existing issues. This demands continued methodological reflexivity on the part of social science researchers, and continued attention to the relationships of trust with respondents, peers, funders and other stakeholders with whom they are involved.

Conclusion

The internet and web 2.0 have opened up myriad possibilities allowing people not only to access information and data but also to generate it themselves, in the form of numbers, text and images. The data themselves undergo changes and transformation as they travel from one location to another, from one form to another, and from one research context to another. In the case described above, people are not only entering phenotypic data about their health on their computer screens, they are also providing samples of their DNA. All of this information is converted into digital form so that it can be stored in databases and subsequently analysed both for the individual and at a collective level by researchers.

The information provided by the company and by people contributing to its fora and blogs has been analysed by us, as we conduct social science research about the emerging phenomena of DTC GT and user-generated content. All sorts of researchers thus have interesting new possibilities for collecting, storing and analysing data on multiple levels for a variety of different purposes.

We described how this works in the case of *23andMe*, a DTC GT company that offers genetic testing to its customers usually for a fee. Drawing on insights from medical sociology and new media studies, as well as Shapin and Schaffer's (1985) work on trust technologies in scientific research, we examined how *23andMe* seeks to establish trust with its customers. We showed how the company attempts to develop trust relationships at different moments in the process and in different ways, and that these relations are performed for various purposes. First, visitors to the site have to be enrolled as customers, willing to part with their money and their saliva. Subsequently, customers have to be transformed into research participants, willing to share information about themselves, not just once, but regularly and often. The company also needs to initiate and maintain relationships with its funders and with the scientific community. We described how the internet is deployed in these multiple practices, and demonstrated how fragile this relationship can be, as when the 'customers' reacted with anger and disappointment upon learning that the company had filed for a patent, using the data they had provided under the rubric of participation.

As the 'participatory turn' broadens in scope and scale through digital self-reporting, researchers need to consider the implications of these mutual and fragile trust relationships for research and knowledge production practices. How self-reported data, constructed by subjectivities, become rhetorically empowered through various discursive practices on the web influences how we understand the social life of digital data. Our critical analysis of *23andMe* and its practices in relation to the collection and use of self-reported data mediated by digital technologies therefore forced us to examine our own practices in using the data we found online, generated by people who, in contrast to *23andMe* customers, may have no interest in supporting or being part of research. Of course, another important difference is that we do not seek to gain financial advantage even if we would like academic rewards in the form of peer-reviewed publications and citations. Digital research practices raise important questions for both medical and social science research: what is good research, or reliable and valid research? Are digitally mediated self-reported data any different from other types of self-reported data? What do we mean by a representative sample? How does the internet frame trust in data sharing between researchers and research participants, and amongst researchers? What constitutes ethical research practice? As data become more social and more mobile, do researchers need

to re-consider how they interact with data and the people the data may or may not represent?

Acknowledgements

We are grateful to the reviewers and guest editors of this special issue for their encouragement and advice. Harris, Wyatt and Kelly are also grateful to the Netherlands Organization for Scientific Research and the UK Economic and Social Research Council for funding received under their Bilateral Agreement Scheme (grant number 463-09-033, 'Selling genetic tests online').

Notes

1. http://quantifiedself.com/about/
2. https://www.23andme.com/legal/privacy/
3. These are the more explicit forms of participation. Consumers also 'participate' in research and development activities less visibly by being part of aggregated sets for internal validation experiments and to develop new features and products; through providing comments on fora and blogs; and in their web activity collected using log files, cookies and other technologies.
4. http://mediacenter.23andme.com
5. http://www.genomeweb.com/23andme-acquires-community-based-health-site-curetogether
6. A highly contested aspect of genetic research more broadly.
7. http://www.genomicslawreport.com/index.php/2012/06/01/patenting-and-personal-genomics-23andme-receives-its-first-patent-and-plenty-of-questions/
8. The company also relies on so-called traditional science articles in order to justify their choice of genetic markers for analysis.
9. The research did not technically require ethics approval because there was no 'interpersonal contact between investigator and participant (that is, data and samples are provided without participants meeting any investigator)' (Gibson and Copenhaver, 2010).

References

Abbate, J. (1999) *Inventing the Internet*. Cambridge, MA: The MIT Press.

Adams, S. (2010) 'Sourcing the Crowd for Health Experiences: Letting the People Speak or Obliging Voice through Choice?', pp. 178–193 in R. Harris, N. Wathen and S. Wyatt (eds) *Configuring Health Consumers: Health Work and the Imperative of Personal Responsibility*. Basingstoke: Palgrave Macmillan.

Agar, J. (2006) 'What Difference Did Computers Make to Science?', *Social Studies of Science* 36: 869–907.

Akrich, M. (2010) 'From Communities of Practice to Epistemic Communities: Health Mobilizations on the Internet', *Sociological Research Online* 15(2): 10. URL: http://www.socresonline.org.uk/15/2/10.html.

Allison, M. (2009) 'Can Web 2.0 Reboot Clinical Trials?', *Nature Biotechnology* 27(10): 895–902.

Arnquist, S. (2009) 'Research Trove: Patients' Online Data', *The New York Times*, 24 August.
Bakardjieva, M. and Feenberg, A. (2000) 'Involving the Virtual Subject', *Ethics and Information Technology* 2: 233–240.
Beaulieu, A. and Estalella, A. (2011) 'Rethinking Research Ethics for Mediated Settings', *Information, Communication & Society* 15(1): 23–42.
Bishop, L. (2009) 'Ethical Sharing and Reuse of Qualitative Data', *Australian Journal of Social Issues* 44(3): 255–272.
Bowker, G.C. and Star, S.L. (1999) *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: The MIT Press.
boyd, d. and Crawford, K. (2012) 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon', *Information, Communication & Society* 15(5): 662–679.
Brown, N. (2003) 'Hope Against Hype: Accountability in Biopasts, Presents and Futures', *Science Studies* 16(2): 3–21.
Bruyninckx, J. (2013) *Sound Science: Recording and Listening in the Biology of Bird Song, 1880–1980*, PhD dissertation, Maastricht University.
Callon, M. and Rabeharisoa, V. (2003) 'Research "in the Wild" and the Shaping of New Social Identities', *Technology in Society* 25(2): 193–204.
Callon, M. and Rabeharisoa, V. (2008) 'The Growing Engagement of Emergent Concerned Groups in Political and Economic Life', *Science, Technology & Human Values* 33(2): 230–261.
Carusi, A. (2008) 'Data as Representation: Beyond Anonymity in e-Research Ethics', *International Journal of Internet Research Ethics* 1(1): 37–65.
Collins, H.M. and Evans, R. (2002) 'The Third Wave of Science Studies', *Social Studies of Science* 32(2): 235–296.
Do, C.B., Tung, J.Y., Dorfman, E., Kiefer, A.K., Drabant, E.M., Francke, U., Mountain, J.L., Goldman, S.M., Tanner, C.M., Langston, J.W., Wojcicki, A. and Eriksson, N. (2011) 'Web-based Genome-wide Association Study Identifies Two Novel Loci and a Substantial Genetic Component for Parkinson's Disease', *PLoS Genetics* 7(6): e1002141.
Epstein, S. (2008) 'Patient Groups and Health Movements', pp. 499–539 in E.J. Hackett, O. Amsterdamska, M. Lynch and J. Wajcman (eds) *The Handbook of Science and Technology Studies*. Cambridge, MA: The MIT Press.
Eriksson, N., Macpherson, J.M., Tung, J.Y., Hon, S.L., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I. and Mountain, J. (2010) 'Web-based, Participant-Driven Studies Yield Novel Genetic Associations for Common Traits', *PLoS Genetics* 6(6): 1–20.
Ess, C. (2009) *Digital Media Ethics*. Cambridge: Polity Press.
Ess, C. and AoIR Ethics Working Committee (2002) 'Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee'. URL:http://aoir.org/reports/ethics.pdf.
Fearnley, L. (2006) 'Beyond the Public's Health: Constructing National Syndromic Surveillance'. Working paper, Anthropology of the Contemporary Research Collaboratory.
Foster, V. and Young, A. (2012) 'The Use of Routinely Collected Patient Data for Research: A Critical Review', *Health* 16(4): 448–463.
Foucault, M. (1963) *Naissance de la Clinique: Une Archéologie du Regard Medical*. Paris: Presses Universitaires de France (*The Birth of the Clinic: An*

Archaeology of Medical Perception, trans. A. Sheridan. London: Routledge, 2003).
Genes, N. (2006) 'Diabetes Patient Offers Goldmine of Information and Support', *Medscape Med Students* 8(2). URL: http://www.medscape.com/viewarticle/544166.
Gibson, G. and Copenhaver, G. (2010) 'Consent and Internet-Enabled Human Genomics', *PLoS Genetics* 6(6): e1000965.
Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009) 'Detecting Influenza Epidemics Using Search Engine Query Data', *Nature* 457(19): 1012–1014.
Goetz, T. (2010) 'Sergey Brin's Search for a Parkinson's Cure', *WIRED Magazine*, 22 June.
Gordon, N., Hiatt, R. and Lampert, D. (1993) 'Concordance of Self-reported Data and Medical Record Audit for Six Cancer Screening Procedures', *Journal of the National Cancer Institute* 85(7): 566–570.
Hall, W., Mathews, R. and Morley, K. (2009) 'Being More Realistic about the Public Health Impact of Genomic Medicine', *PLoS Medicine* 7(10): e1000347.
Harris, A., Wyatt, S. and Kelly, S. (2013) 'The Gift of Spit (and the Obligation to Return it): How Consumers of Online Genetic Testing Services Participate in Research', *Information, Communication & Society* 16(2): 236–257.
Jutel, A. (2010) 'Sociology of Diagnosis: A Preliminary Review', *Sociology of Health & Illness* 31(2): 278–299.
Kotz, J. (2012) 'Bringing Patient Data into the Open', *Science Business Exchange* 5(25): doi: 10.1038/scibx.2012.644.
Langstrup, H. (2011) 'Interpellating Patients as Users: Patient Associations and the Project-ness of Stem Cell Research', *Science, Technology & Human Values* 36(4): 573–594.
Lawrence, A. (2006) '"No Personal Motive?": Volunteers, Biodiversity, and the False Dichotomies of Participation', *Ethics, Place & Environment* 9(3): 279–298.
Levallois, C., Steinmetz, S. and Wouters, P. (2013) 'Sloppy Data Floods or Precise Social Science Methodologies?', pp. 151–182 in P. Wouters, A. Beaulieu, A. Scharnhorst and S. Wyatt (eds) *Virtual Knowledge: Experimenting in the Humanities and the Social Sciences*. Cambridge, MA: The MIT Press.
Lipworth, W., Forsyth, R. and Kerridge, I. (2011) 'Tissue Donation to Biobanks: A Review of Sociological Studies', *Sociology of Health & Illness* 33(5): 792–811.
Lupton, D. (2012) 'The Quantified Self Movement: Some Sociological Perspectives', *This Sociological Life*. URL (consulted November 2012): http://simplysociology.wordpress.com/2012/11/04/the-quantitative-self-movement-some-sociological-perspectives.
McCray, W.P. (2006) 'Amateur Scientists, the International Geophysical Year, and the Ambitions of Fred Whipple', *Isis* 97: 634–658.
Mol, A. (2002) *The Body Multiple: Ontology in Medical Practice*. Durham: Duke University Press.
Nissenbaum, H. (2010) *Privacy in Context: Policy and the Integrity of Social Life*. Palo Alto: Stanford University Press.

Oudshoorn, N. (2011) *Telecare Technologies and the Transformation of Healthcare*. Basingstoke: Palgrave Macmillan.
Panofsky, A. (2011) 'Generating Sociability to Drive Science: Patient Advocacy Organizations and Genetics Research', *Social Studies of Science* 41(1): 31–57.
Pentzold, C. (2011) 'Imagining the Wikipedia Community: What do Wikipedia Authors Mean When they Write about their "Community"?', *New Media & Society* 13(5): 704–721.
Piras, E.M. and Zanutto, A. (2010) 'Prescriptions, X-rays and Grocery Lists: Designing a Personal Health Record to Support (the Invisible Work of) Health Information Management in the Household', *Computer Supported Cooperative Work* 19: 585–613.
Pols, J. (2012) *Care at a Distance: On the Closeness of Technology*. Amsterdam: Amsterdam University Press.
Prainsack, B. (2011) 'Voting with their Mice: Personal Genome Testing and the "Participatory Turn" in Disease Research', *Accountability in Research* 18(3): 132–147.
Savage, M. and Burrows, R. (2007) 'The Coming Crisis of Empirical Sociology', *Sociology* 41(5): 885–899.
Shapin, S. and Schaffer, S. (1985) *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton: Princeton University Press.
Smith, B., Chu, L.K., Smith, T.C., Amoroso, P.J., Boyko, E.J., Hooper, T.I., Gackstetter, G.D., Ryan, M.A. and the Millennium Cohort Study Team (2008) 'Challenges of Self-reported Medical Conditions and Electronic Medical Records among Members of a Large Military Cohort', *BMC Medical Research Methodology* 8(37).
Star, S.L. and Griesemer, J.R. (1989) 'Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939', *Social Studies of Science* 19(3): 387–420.
Sterckx, S., Cockbain, J., Howard, H., Huys, I. and Borry, P. (2012) '"Trust is not Something You Can Reclaim Easily": Patenting in the Field of Direct-to-Consumer Genetic Testing', *Genetics in Medicine*, doi: 10.1038/gim.2012.143.
Swan, M., Hathaway, K., Hogg, C., McCauley, R. and Vollrath, A. (2010) 'Citizen Science Genomics as a Model for Crowdsourced Preventive Medicine Research', *Journal of Participatory Medicine* 2: e20.
Thomas, G. and Wyatt, S. (1999) 'Shaping Cyberspace: Interpreting and Transforming the Internet', *Research Policy* 28: 681–698.
Tung, J.Y., Do, C.B., Hinds, D.A., Kiefe, A., Macpherson, J.M., Chowdry, A.B., Francke, U., Naughton, B., Mountain, J., Wojcicki, A. and Eriksson, N. (2011) 'Efficient Replication of over 180 Genetic Associations with Self-reported Medical Data', *PLos ONE* 6(8): e23473.
Tutton, R. and Prainsack, B. (2011) 'Enterprising or Altruistic Selves? Making up Research Subjects in Genetics Research', *Sociology of Health & Illness* 33(7): 1081–1095.
Waitzkin, H. (1991) *The Politics of Medical Encounters: How Patients and Doctors Deal with Social Problems*. New Haven: Yale University Press.
West, C. (1984) 'When the Doctor is a "Lady": Power, Status and Gender in Physician-Patient Encounters', *Symbolic Interaction* 7(1): 87–106.
White, M. (2002) 'Representations or People?', *Ethics and Information Technology* 4(3): 249–266.

Wicks, P., Vaughan, T.E., Massagli, M.P. and Heywood, J. (2011) 'Accelerated Clinical Discovery Using Self-reported Patient Data Collected Online and a Patient-matching Algorithm', *Nature Biotechnology* 29(5): 411–414.
Wyatt, S. (2012) 'Ethics of e-Research in Social Sciences and Humanities', pp. 5–20 in D. Heider and A. Massanari (eds) *Digital Ethics: Research and Practice*. New York: Peter Lang.
Zimmer, M. (2008) 'The Externalities of Search 2.0: The Emerging Privacy Threats When the Drive for the Perfect Search Engine Meets Web 2.0', *First Monday* 13(3). URL: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2136/1944.
Zimmer, M. (2010) '"But the Data is Already Public": On the Ethics of Research in Facebook', *Ethics and Information Technology* 12(4): 313–325.

Author Biographies

Sally Wyatt is Professor of Digital Cultures in Development at Maastricht University and Programme Leader of the e-Humanities Group of the Royal Netherlands Academy of Arts and Sciences. Her research focuses on the use of digital technologies in healthcare, and on what digital technologies mean for social science and humanities research practices.

Anna Harris is a postdoctoral researcher in the Department of Technology and Society Studies, Maastricht University. She is working on an ethnographic project concerning how doctors learn to listen to sounds which is part of a larger project entitled 'Sonic Skills: Sound and Listening in the Development of Science, Technology and Medicine (1920–now)'.

Samantha Adams is Assistant Professor of Patient-centred e-Health in the Department of Healthcare Governance at the Institute of Health Policy and Management (iBMG) of the Erasmus University Rotterdam, the Netherlands. Her research focuses on personalized web-based health information and new media in healthcare.

Susan E Kelly is a Senior Research Fellow and Director of the Health, Technology & Society research group at the ESRC Centre for Genomics in Society (Egenis) at the University of Exeter. She is also Senior Lecturer in Medical Sociology in the Department of Sociology, Philosophy and Anthropology. Her research focuses on clinical implementation and societal implications of genetic and genomic technologies.