Big Data for Personalized Healthcare

Liseth Siemons, Floor Sieverink, Annemarie Braakman-Jansen, Lisette van Gemert-Pijnen
Centre for eHealth and Well-being Research
Department of Psychology, Health, and Technology
University of Twente
Enschede, the Netherlands
(l.siemons, f.sieverink, l.m.a.braakman-jansen, j.vangemert-pijnen)@utwente.nl

Wouter Vollenbroek
Department of Media, Communication & Organisation
University of Twente
Enschede, the Netherlands
w.vollenbroek@utwente.nl

Lidwien van de Wijngaert
Department of Communication and Information Studies
Radboud University
Nijmegen, the Netherlands
l.vandewijngaert@let.ru.nl

***Abstract* - Big Data, often defined according to the 5V model (volume, velocity, variety, veracity and value), is seen as the key towards personalized healthcare. However, it also confronts us with new technological and ethical challenges that require more sophisticated data management tools and data analysis techniques. This vision paper aims to better understand the technological and ethical challenges we face when using and managing Big Data in healthcare as well as the way in which it impacts our way of working, our health, and our wellbeing. A mixed-methods approach (including a focus group, interviews, and an analysis of social media) was used to gain a broader picture about the pros and cons of using Big Data for personalized healthcare from three different perspectives: Big Data experts, healthcare workers, and the online public. All groups acknowledge the positive aspects of applying Big Data in healthcare, touching upon a wide array of issues, both scientifically and socially. By sharing health data, value can be created that goes beyond the individual patient. The Big Data revolution in healthcare is seen as a promising and innovative development. Yet potential facilitators and barriers need to be faced first to reach its full potential. Concerns were raised about privacy, trust, reliability, safety, purpose limitation, liability, profiling, data ownership, and loss of autonomy. Also, the importance of adding the people-centered view to the rather data-centered 5V model is stressed, in order to get a grip on the opportunities for using Big Data in personalized healthcare. People should be aware that the development of Big Data advancements is not self-evident.**

Keywords - Big Data; personalized healthcare; eHealth.

I. INTRODUCTION

The "Big Data" revolution is a promising development that can significantly advance our healthcare system, promoting personalized healthcare [1]. Imagine a system that analyzes large amounts of real-time data from premature babies to detect minimal changes in the condition of these babies that might point to a starting infection. Science fiction? No, IBM and the Institute of Technology of the University of Ontario developed a system that enables physicians to respond much sooner to a changing condition of the baby, saving lives, and leading to a significantly improved quality of care for premature babies [2].

We are standing at the beginning of the "Big Data" revolution. Many different definitions exist for "Big Data". Where Mayer-Schönberger and Cukier [2] focus on the new insights and economic value that can be obtained from Big Data in contrast to traditional smeller settings, Wang & Krishnan [3] refer to Big Data as complex and large data sets that can no longer be processed using the traditional processing tools and methods. Yet another definition comes from Laney [4], who defines Big Data according to 3 assets (often referred to as the 3V-model) that require new, cost-effective forms of information processing to promote insight and decision making, including: 1) high-volume (i.e., the quantity of data), 2) high-velocity (i.e., the speed of data generation and processing), and 3) high-variety (i.e., the amount of different data types). Marr [5] expanded this 3V model to the 5V model by adding 2 additional Vs: veracity (i.e., the accuracy or trustworthiness of the data) and maybe the most important asset: value (i.e., the ability to turn the data into value).

Though this is just a grasp out of all the definitions available, there is one thing they have in common: The use of Big Data for analysis and decision making requires a change of thought from knowing "why" to knowing "what". Where we focused on small, exact datasets and causal connections in the past (i.e., knowing "why"), we now focus on gathering or linking large amounts of (noisy) data, with which we can demonstrate the presence of (unexpected) correlational connections (i.e., knowing "what") [2]. As a result, we will obtain (and apply) new insights that we did not have before. Insights that can not only be lifesaving, as demonstrated by the example of IBM and the University of Ontario, but that also opens the door towards more personalized medicine [6-8]; i.e., where medical decisions, medications, and/or products are tailored to the individual's personal profile instead of to the whole patient group. For example, when genetic biomarkers in pharmacogenetics are used to determine the best medical treatment for a patient [6] or when data from thousands of patients that have been

treated in the past is being analyzed to determine what treatment best fits the individual patient that is under treatment now (e.g., in terms of expected treatment effects and the risk for severe side-effects given the patient's personal characteristics like age, gender, genetic features, etc.).

This shift towards more personalized healthcare is reflected in the change of focus within healthcare from a disease-centered approach towards a patient-centered approach, empowering patients to take an active role in the decisions about their own health [8]. As a result, an increasing number of technologies (e.g., Personal Health Records) are being launched by companies to support chronically ill people in the development of self-management skills [9].

The past decades have also shown a rapid growth in the amount of (personal) data that is digitally collected by individuals via wearable technologies that may or may not be stored on online platforms for remote control [2, 6-8, 10], or shared via other online sources like social media. Social media have become socially accepted and used by a growing group of people [11]. They use it, for example, to share data collected by activity, mood, nutrition and sleep trackers on a variety of online platforms (such as Facebook, Twitter, blogs or forums). These data provide new opportunities for healthcare to personalize and improve care even further [12-14]. Furthermore, the data and messages shared via these tools provide insight in vast amounts of valuable information for scientific purposes. For example, [14] used the data from Twitter to predict flu trends and [15] used social media as a measurement tool for the identification of depression. The information gleaned from social media has the potential to complement traditional survey techniques in its ability to provide a more fine-grained measurement over time while radically expanding population sample sizes [15].

By combining clinical data with personal data on, for instance, eating and sleeping patterns, life style, or physical activity level, treatment and coaching purposes can be tailored to the needs of patients even better than before and are, therefore, seen as the key towards a future with optimal medical help [6]. However, it also confronts us with new technological and ethical challenges that require more sophisticated data management tools and data analysis techniques. This vision paper aims to better understand the technological and ethical challenges we face when using and managing Big Data in healthcare as well as the way in which it impacts our way of working, our health, and our wellbeing.

This paper builds on first insights obtained from Big Data experts as already described in [1] and adds the perspectives of healthcare workers (HCWs) and the online public. Section I describes the background of Big Data in literature. Section II describes the procedure of the meetings with experts (focus group; individual meetings) and HCWs (interviews), and describes how the online public's associations with Big Data in a health context were assessed. Section III presents the results, which are discussed more into depth in Section IV. Finally, Section V concludes this paper, describing a number of implications for research using Big Data in healthcare and addressing some future work.

II. METHODS

The impact and challenges of Big Data will be examined from three different perspectives: 1) from the perspective of Big Data experts [1], 2) from the perspective of HCWs, and 3) from the perspective of the online public. Different methods were used to gather information from each group. Where a focus group was planned with the Big Data experts, this turned out to unfeasible with HCWs because of their busy schedule. That is why individual interviews were scheduled with them. Finally, to evaluate the perspective of the online public, social media posts were scraped and analyzed.

A. Focus group with experts

Many potential issues regarding the use of Big Data have already been mentioned in the literature, newspapers, social media, or debates, and panel discussion websites. However, many of these media sources do not specifically address the healthcare setting and only focus on a limited set of issues at a time (e.g., the privacy and security issues).

To gain more in depth insights into the pros and cons of using Big Data in personalized healthcare, a focus group was organized [16]. The aim was to gain a variety of opinions regarding the scientific and societal issues that play a role in using and managing Big Data to support the growing needs for personalized (and cost-effective) healthcare.

Purposeful sampling was used in the formation of the focus group, meaning that the selection of participants was based on the purpose of the study [16]; i.e., to map the experts' variety and range of attitudes and beliefs on the use of Big Data for (personalized) healthcare purposes. To gather a broad perspective of viewpoints, multiple disciplines were invited to join the expert meeting, resulting in a panel of 6 experts in Big Data research and quantified self-monitoring from different scientific disciplines: psychology, philosophy, computer science, business administration, law, and data science. Participants were recruited at the University of Twente (the Netherlands), based on their societal impact, expertise, and experiences with conducting Big Data research. Individual face to face meetings were conducted to validate the focus group results.

The focus group took 2 hours in total and was facilitated by LS and FS (authors). All participants signed an informed consent for audiotaping the focus group and for the anonymous usage of the results in publications. LVGP and ABJ took additional notes during the discussion. Group discussion was encouraged and participants were repeatedly asked to share their concerns and thoughts.

In preparation of the focus group discussion, literature and multiple sources of (social) media were searched for information on potential Big Data issues that might play a role. During the discussion itself, experts were asked to write down as many issues as they could think of that might become relevant using Big Data for healthcare. Flip-overs were used to express the issues and experts had to categorize these issues into overall concepts that covered the issues. They named these overall concepts themselves by thinking aloud. These concepts are presented in this vision paper. The focus group was audio taped and transcripts were made by

authors of this paper. Loose comments without any further specification were excluded to ensure the results are not a representation of the authors' interpretation. Ethical approval for the scientific expert meeting and the consent procedure was obtained by the ethics committee of the University of Twente.

B. Interviews with healthcare workers

Based on the results of the focus group, an interview scheme was constructed to assess how HCWs perceive and experience the issues that were identified. Questions were formulated open-ended to encourage HCWs to elaborate on their perceptions of and experiences with Big Data (or eHealth applications) in healthcare. For each question, HCWs were asked to think aloud and to elaborate on their thoughts. Interviews were transcribed verbatim afterwards and a coding scheme was developed.

A total of 6 physicians with experience in Big Data were interviewed. Participants received a first description of the aim of the interviews by e-mail and, in addition, each interview started with a 1.5 minute long movie presenting the interview's subject: Big Data eHealth applications in healthcare. All participants were interviewed in their work setting. Interviews were semi-structured and continued until the interviewer felt that all questions were answered and no new information could be expected. This took about 60 minutes on average. Participants gave informed consent for audiotaping the interviews and for the anonymous usage of the results in publications. The study was approved by the ethics committee of the University of Twente. No additional ethical approval was necessary from the medical ethical committee.

C. Online public's associations with Big Data in a health context

Though Big Data receives a lot of attention nowadays, little is known about the publics' associations with the term Big Data in health contexts. With the digitalization of society, the online public that uses social media channels encloses a large proportion of the potential users of Big Data-driven technological applications in healthcare. Furthermore, the content within the social media provides new opportunities to identify the associations made by the online public in relation to Big Data in a healthcare setting. These associations provide a better understanding of the concerns, opportunities, and considerations that the health sector must take into account.

As such, Coosto, a social media monitoring tool (www.coosto.com), was used as a first explorative analysis to analyze these associations among social media users, using multiple data sources (social networks, microblogs, blogosphere, forums) in both Dutch and English.

The identification of the online public's perceptions regarding the terms they use when discussing about Big Data in relation to healthcare was completed by following three phases. In the first phase, social media posts were scraped, based on 6 search queries in the Dutch social media monitoring tool Coosto (Table I). To avoid issues caused by word variations (for example: healthcare, health-care, health care) and synonyms, an extensive list of different spellings was used in each search query. The terms selected for the search query were derived from a systematic analysis of synonyms in academic literature, popular literature, and websites and Google search results. To treat (longer) blog posts and (shorter) tweets equally, each sentence of all posts was analyzed separately in the second phase. More citations and shares means that more online users are interested in that particular topic.

TABLE I. SEARCH QUERIES

Search query
1. "big data" "e-health" OR "ehealth" OR "e health"
2. "big data" "healthcare" OR "health care" OR "health-care"
3. "big data" "care"
4. "big data" "sensors" OR "health" OR "e-health" OR "e health" OR "ehealth" OR "care" OR "healthcare" OR "health care" OR "health-care" OR "wellness" OR "wellbeing" OR "well-being"
5. "big data" "wearables" OR "health" OR "e-health" OR "e health" OR "ehealth" OR OR "care" OR "healthcare" OR "health care" OR "health-care" OR OR "wellness" OR "wellbeing" OR "well-being"
6. "big data" "domotica"

The second phase was aimed at extracting the most commonly used (combinations of) terms in the collected social media posts and measuring the proximity (the relative distance (similarity)) of these terms. The more frequently two terms are mentioned simultaneously in the whole dataset the higher the proximity between these two (combinations of) terms. Based on a codebook (Appendix 1) consisting of terms that are related to and associated with Big Data and/or healthcare, the most frequently mentioned terms in the social media posts were identified. The codebook terms were selected based on a systematic analysis of scientific and popular literature (including references) and websites (including links to other websites), news articles, social media posts (e.g., Twitter), and Google search results. In our design, the sentences of analysis were considered as the cases, and the terms in these messages – after properly filtering for example the stop words, hyperlinks and @mentions – as the variables. The next step in our analysis was to find the terms (e.g., privacy) or phrases (e.g., Internet of Things) from the codebook in the sentences (cases). The sentences without any of the terms were omitted from further analysis. Thus, a matrix was operated that contained terms as the variables in the columns and sentences as cases in the rows. The cells in the matrix consisted of binary data (whether or not a particular term occurs in the sentences). A proximity measurement [17] indicates what combinations of terms are most prevalent.

In the third phase, the main objective was to determine what terms are mostly associated with Big Data in the context of healthcare within the social media. To do so, the open-source network analysis and visualization software package Gephi (http://gephi.github.io/) was used to visualize the interrelationships between (groups of) terms in a semantic network [18]. The binary matrix formed the basis for the semantic network graph. Due to the reasonably large dataset and the minimum agreement between the terms, the correlations have been relatively low. Therefore, all correlations higher than 0.02 were included in the actual

analysis. The terms which served as the basis for the search queries were then removed from the semantic network analysis, since the preservation of these terms in the search results would produce biased results because they will occur significantly more than in reality may be assumed.

III. RESULTS

Results are presented separately for each group: 1) Big Data experts, 2) HCWs, and 3) the online public.

A. Focus group meeting with experts

The results can be subdivided in 3 categories: 1) empowerment, 2) trust, and 3) data wisdom.

1) Empowerment

What does it mean when you monitor your activities, food intake, or stress 24 hours a day using technologies like smart wearables? What drives people to use these 24 hour monitoring devices and what do they need to understand the data generated by these systems? Do they understand the algorithms that are used to capture our behaviors and moods in pictures and graphs? Who owns the data and how to control the maintenance of that data? How to avoid a filtered scope on our lives ignoring others that are out of our affinity groups? The concept of empowerment captures topics as autonomy, freedom, and having control.

Big data evokes a discussion about freedom and autonomy. Autonomy concerns our critical view on how to use technology, while freedom is more about our way of living and thinking. It might, therefore, be more important to focus on freedom instead of autonomy: understanding how you are being influenced and taking a stance against that instead of trying to keep everything away. The focus group made a distinction between positive freedom and negative freedom; two common concepts within the field of philosophy. Positive freedom is the freedom to do something yourself (e.g., to decide for yourself that you want to share your data), whilst negative freedom is the freedom to keep things away, protecting yourself (e.g., when you do not give permission to companies to link your data with other sources). Not losing control, being able to use, share and understand your data is one of the topics when discussing freedom, self-efficacy using self-monitoring technologies.

Empowerment forces us to think about having control, who has the power through the use of Big Data? There might be just a small elite that understands the algorithms and with the increasing complexity, this elite will become even smaller in the future. This can create a division between people who can access and understand the algorithms and people who do not.

Empowering by personalization is one of the aims of the participatory society. Big data can be a leverage to realize this by creating a personal profile, providing the right information, at right moments to enable just in time coaching. Though it can be useful to put people in a profile, the danger of profiling is that you can never leave the assigned group again; once assigned to a group means always assigned to that group. Profiling might be suffocating to people because it creates uncertainty about what people know about you, what data are being collected, and for what purposes. Also, it is often unclear how to determine the norm to which people are compared when assigning them to a group (i.e., standardization, losing freedom). Furthermore, being assigned to a profile might lead to discrimination and certain prejudices/biases. Questions that arise are: How can profiling be used in a sensible/sound way? And who is responsible when mistakes are being made based on a certain profile?

2) Trust

Trust will become a key concept in a data driven society. This concept captures more than privacy and security issues. Trust refers to topics as how to create faith in data management and data maintenance, and how to make sense of these data for humans.

Privacy issues become particularly relevant when the linkage of anonymous datasets leads to re-identification. Encryption of the data might prevent identification of individuals, but transparency is not always possible (e.g., when analyzing query logs with search terms). In the end it is all about creating trust to overcome uncertainty or anxiety for a digital world.

People often give consent to institutions to use their data for certain purposes in return for the (free) use of the product or service. However, data can be (re)used for other purposes as well or can be sold to other interested parties, even though that is not always allowed. This leads to great concerns: e.g., healthcare insurance companies who use treatment data for other purposes on a more personal level (for instance, for determining a personalized health insurance premium based on your personal data about your health and lifestyle). It is not that people do not want to share data, they already do this using Facebook or Google services, but they want to understand what happens with the data, in particularly when it concerns the health domain.

Self-monitoring technologies, with no doctors or nurses involved in the caring process, are provided more and more by institutions. Smart algorithms can be applied to personalize data in such a way so you can manage your health and wellbeing yourself. However, these algorithms decide what information you get to see, based on information about you as a user (e.g., search history, Facebook friends, location). This will influence trust in the healthcare system, using data from your device compared to personal advices given by your doctor or nurse.

3) Data Wisdom

There is a rapid growth of self-monitoring technologies, but little is known about the reliability and validity of these systems. The lack of evidence for causality can lead to unreliability as well. Furthermore, how can you tell what you are actually measuring? How can the correlations that are found be validated? Does it really say what we think it says or are it just assumptions?

Data wisdom is the concept that captures scientific and societal topics. Scientific refers to how to create data wisdom, in several ways. Those who generate data are not the ones that have the knowledge to analyze, those who analyze lack domain insight (technologies, behaviors). Different kinds of expertise will be needed in the future to

deal with Big Data. For instance, expertise to analyze Big Data, expertise to develop and understand the working of algorithms, or expertise in data interpretation and visualizations. The use of data to personalize healthcare demands for new knowledge to support critical and creative thinking to understand data driven decisions and to watch the impact on science, health and society. We all know the disaster with google flu trends, but we have to learn from these failures to set the agenda for future research in using several sources of data (geospatial data, medical data, technology device data) to develop predictive models about health and wellbeing. We have to search for new models, methods to deal with huge datasets, search for patterns rather than testing hypotheses based on small data. Results are not causal-driven but correlational-driven. This requires a change in thought. The golden rule for Randomized Clinical Trials will no longer be the ultimate format for health sciences. New methods are needed to get a grip on "big", how many data (critical mass) is needed and how rich and mature should data be to make meaningful decisions? How to add qualitative experiences and expertise to Big data? Numbers do not tell the whole story, and a clinical eye is important to interpret data in the context of individual health and wellbeing.

Societal refers to the implications for healthcare, addressing topics as ethics, values for a meaningful life. How to avoid a division between people who can access and understand the data and analytics that rule the decisions about treatments and lifestyle advices, and people that cannot? Knowledge and skills are needed to empower people and people should participate in debates about the values of data for self-regulations on the level of individuals, communities and society. Transparency and trust are the key-topics in that debate. Digging into data starts with a scientific and societal debate on the vales of data for a smart and healthy society.

B. Interviews with healthcare workers

Again, the results can be subdivided in the 3 categories: 1) empowerment, 2) trust, and 3) data wisdom.

1) Empowerment

Physicians recognize the advantages of data sharing. For instance, it provides them easy insight in treatment outcomes, which is an important instrument for quality of care. Yet though the large majority of people probably do not have any problems with sharing their data, patients cannot be forced to share their data and should give informed consent first. Nevertheless, patients often do not understand where they give permission for. What if they change their minds, is it possible to undo their data sharing? What if the data is already shared with different disciplines, will they all be refused access after withdrawing the data sharing approval? Possibly, an independent supervisor should be appointed the task to safeguard the proper handling of patient data (at least till data encryption).

There was no consensus among the physicians about data ownership. Some argue the data is primarily of the patient, but the hospital or healthcare practitioner should be able to gain access to it as well when they have to give account for their actions. On the other hand, it is the physician who writes most of the medical data down in the patient's personal health record, so it can be argued that he owns (that part of) the data (as well).

Next, the physicians argued that profiling as a concept is nothing new. Current practice is already to gather as much information as possible from a patient and to "go through a checklist" of characteristics, symptoms, or complaints before deciding which treatment might be best. Yet profiling based on Big Data might make this process more accurate, improving treatment outcomes. Especially in complex cases, profiling based on Big Data might be of significant value. It promotes personalized healthcare. Concerns about profiling mainly involve drawing conclusions with far-reaching consequences based on incomplete/imprecise data and the unauthorized misuse of data by third parties, like insurance companies. Also professionals might lose professional skills when they do not have to think for themselves anymore.

When using Big Data for predictive modeling proposes (e.g., to predict the chance you might get lung cancer in the next 5 years), people have the "right not to know". Within a certain boundary that is. If national health is in danger, personal rights do not weigh up to national security. Patients should be informed at an early stage about their rights and about who is liable when something goes wrong. In case of treatment decisions, the physician is most of the time liable if something goes wrong. In the Netherlands, physicians are guilty until proven innocent in case of an accusation. When using Big Data algorithms in treatment decisions, physicians should still think for themselves whether the provided advice by a system appears to be reasonable, because they make the final judgment about the best possible treatment for their patient. That is not different from the current process in which a physician also has to deal with information from, for instance, radiology, long-term research, laboratory results, etc. It does not matter whether choices are based on Big Data or not, the physician needs to keep thinking as a doctor. Yet liability is not always clear. For instance, what if a patient wears a smart watch that registers his blood pressure, but the device has a defect. Who is liable? The manufacturer or supplier of the device? But how to deal with liability when the patient uses the device in an ignorant way? Who is liable then? Who has to prove what and how is it regulated? It will be a long juridical procedure because rules are not clear and straightforward in cases like this.

2) Trust

Physicians definitely recognize the importance of secure data management. Data management might be outsourced to a third party, on condition that the liability in case of data leaks is properly arranged. Yet physicians might also play an important role in data management, since they are needed for data interpretation. Furthermore, there should be restricted access to sensitive patient data. For instance, when implementing a patient portal, it can be decided to only grant the patient and his own physician access to the portal. In addition, it is of utmost importance that the portal's communication and data transport mechanisms between the

patient and the physician are thoroughly thought through, to minimize the risk on data leaks and cyber-attacks.

Opinions on the need to understand the underlying algorithms of the system were divided. Yet most physicians do want to know the basic reasoning behind the algorithms. They fear to lose their professional clinical skills when blindly following the system, without thinking for themselves anymore. They also argue that the standards and guidelines in the medical field exist because the reasoning behind them are clear. They would never trust a computer without at least some insight into the underlying principles, the reliability of the results, and the parameters that were taken into account. Others, however, express no need to understand the algorithms, but they do argue that the algorithms should be checked by a group of experts.

3) Data Wisdom

When asking about their (future) profession, physicians acknowledge the added value of Big Data systems. The medical field encompasses an enormous amount of data that has been collected over the years. It would be helpful if the data from all the existing different platforms could be combined into one overall system to make it surveyable for both clinical and scientific purposes. When the Big Data revolution offers new opportunities for combining this data using innovative data management tools, many new medical and scientific research questions can be explored. When the right algorithms are being developed, Big Data systems might be more accurate than the patient and physician together could ever be, improving quality of care. For instance, when monitoring a patient's condition at home, the data is send to the physician via an interactive application right away (i.e., real-time, 24/7), changes can be detected at an early stage, and (remote) treatment can be provided timely. Also, certain risk factors can be detected and controlled at an early stage. Such applications can improve treatment outcomes and significantly reduce healthcare costs.

A certain expertise is needed to translate the enormous amount of data into clinically relevant pieces of information. These Big Data specialists might be appointed from outside the hospital, although the hospital might not have the financial means to do this. Nevertheless, when implemented, Big Data applications can serve as medical decision aids that help the physicians to combine and process all the available data real quick or it could be used for e-consults, where the patient can login onto the system to see lab results, to make appointments, or to ask questions. As a result, patients need to visit the physician less often. An important note that was made, is the need for guidance to the patients. For instance, if a patient suddenly notices a drop in his blood pressure from 120 to 90, he might think: "oh, I am dying". However, it is still within the norm and this should be explained to the patient to reassure him.

The physicians also recognize the challenges that come with the application of Big Data systems in healthcare. First of all, not all patients will be able to use such systems. For instance, elderly or cognitively impaired patients might not understand the working mechanisms and do not know how to interpret the results. Furthermore, the physicians already always base their decisions on data and it does not really matter where that data comes from (and, as such, whether it is called "Big Data"). What really matters is that the data is reliable. When you start digging into the data without any clinical expertise, chances are that you will find a significant result. However, it might have no clinical meaning and relevance at all. Data collection and data interpretation cost time (and money) as do the development and testing of the underlying algorithms. People (and insurance companies) should be aware of the imperfections of the system and should not blindly follow it as if it is the golden rule. The system does not replace the physician. The physician still makes the decisions and they will only use a Big Data system if it is proven to be more effective than current practice.

C. Online public's associations with Big Data in a health context

For the social media analysis, a total of 5.852 social media posts (blogs, forums, microblogs, and social networks) were crawled and scraped, with a total of 59.281 sentences from a time period of five years, with a focus on Big Data in the context of healthcare. In Appendix 1, the frequencies of emerging words are given. Fig. 1 shows the semantic graph consisting of 49 nodes (interrelationships) and 174 edges (collections of terms). The larger the node, the higher the frequency with which the term is mentioned in relation to Big Data in the context of healthcare. The weight of the edges is determined by the proximity between the terms.

With the modularity algorithm [19], ten clusters of terms were established that are often used together. The modularity metric is a well-known exploration concept to identify a network that is more densely connected internally than with the rest of the network [20]. Five of the major ten clusters are (Fig. 1): concerns (red), opportunities (violet), personalized healthcare (green), infrastructure (yellow), and applications (blue). These clusters cover 92 percent of the associations. The remaining clusters (8 percent) are solitary nodes or dyads and have a lack of power.

When evaluated per cluster, the cluster concerns shows the most frequent associations with the terms: 1) privacy, 2) regulations, 3) reliability, 4) algorithms, 5) transparency, and 6) legislations. The terms that are mentioned the most in the cluster involving the opportunities of Big Data in the context of healthcare are 1) innovation, 2) future, 3) development, 4) technology, 5) challenges, 6) start-up and 7) revolution; whereas the personalized healthcare cluster shows the most frequent associations with the terms 1) quantified self, 2) medicine, and 3) personalization. In the infrastructure cluster, the most associated terms with Big Data and healthcare are 1) cloud, 2) service, 3) platform, and 4) software. And finally, the majority of social media posts related to applications focus on 1) S-Health app, 2) Healthkit, and 3) Healthtap.

On average, the vast majority of terms are related to more than one other term (average degree: 7.102). The average degree is a numerical measure of the size for the neighborhood of an individual node [21]. The terms most associated with Big Data in the context of healthcare that

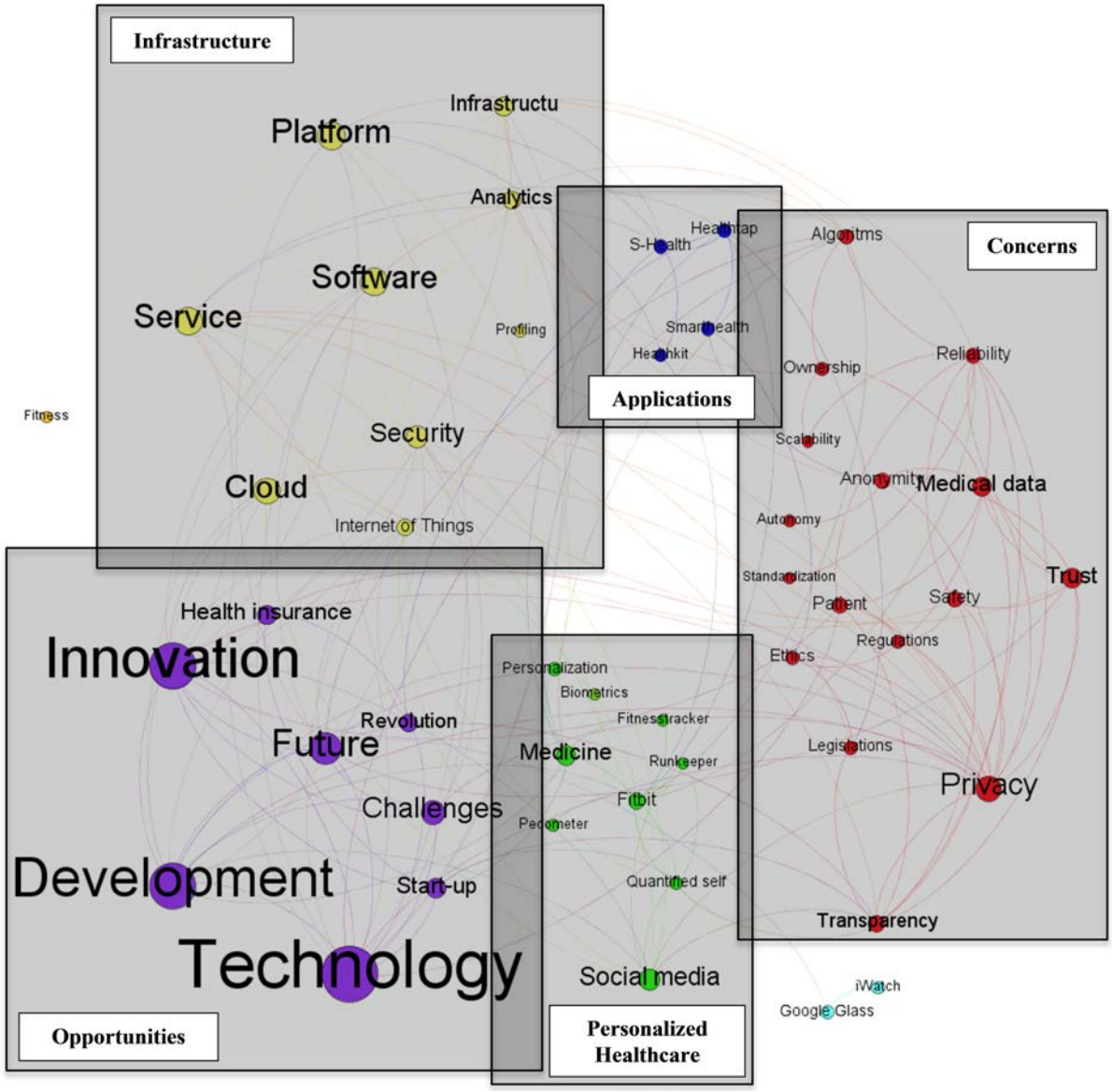

Figure 1. Total semantic network graph (49 nodes – 174 edges).

have a reciprocal degree with other terms in this study are: technology (11), cloud (11), privacy (8), innovation (8), software (8), service (7), development (7), and platform (6). The most frequently mentioned terms in the social media are: technology (1.726), innovation (1.361), development (1.337), future (794), service (638), platform (618), software (610), cloud (581), and privacy (545).

IV. DISCUSSION

With this study we aimed to set a first step in understanding how Big Data impacts healthcare and which critical factors need to be taken into account when using Big Data to personalize healthcare. This was examined from three different perspectives: 1) scientific Big Data experts, 2) HCWs, and 3) the online public.

Results show that Big Data touches upon a wide array of issues, both scientifically and socially. In general, experts and HCWs discussed the future of Big Data on a meta-level, from the perspective of their expertise and their discipline, while the online public considered Big Data more from a consumer-perspective, as end-users of wearables and other technologies. The experts and HCWs make a distinction between promises and concerns depicted as crucial for successfully using and managing Big Data to support the growing needs for personalized healthcare and two rather identical clusters (concerns and opportunities) were found in the social media analyses as well. Concerns are mainly about trust, reliability, safety, purpose limitation, liability, profiling, data ownership (which is unclear), and autonomy, which is consistent with literature [2, 6, 7]. Perhaps the most

well-known concern bears upon our privacy [6, 7]. For a great deal, these privacy concerns are associated with potential misuse of data by, for instance, insurance companies [6, 10]. If these privacy concerns are not dealt with appropriately, the public's trust in technological applications might diminish severely [10]. According to the HCWs, patients are often unaware of what is being collected, who is able to view it, and what decisions are being made based on that information. Transparency is needed, people should know what they give informed consent for when they decide to share their data and what happens if they change their mind after some time.

Both experts and HCWs acknowledge the need for a new sort of expertise to be able to understand the algorithms (or at least the basic reasoning behind them) and to interpret the data that is being generated by the technology. At the end, the technology does not replace the physician but they supplement each other. Quality of care can be improved and personalized healthcare becomes the future. Personalized healthcare received a lot of attention in the expert group and among the HCWs. Telemonitoring is seen as a promising development, enabling the physician to react quickly on changes in the clinical status of a patient 24/7, improving the patient's prognosis. In social media personalized healthcare showed clear associations with the Quantified Self movement, medicine, and personalization. Though the semantic network graph (Fig. 1) only visualized the interrelationships between the (groups of) terms without providing it with an interpretation, this result does demonstrate that the emergence of personalized healthcare and the Quantified Self movement receives a lot of attention in science as well as in society. Yet the main themes discussed by the online public did not include personalized healthcare that much, but rather focused on the technological innovation brought by Big Data, the infrastructure that is needed to make this happen, and privacy issues. The need for a good infrastructure was something the experts also stressed, whereas HCWs focused less on this technological aspect of Big Data.

Some other differences between the groups could be identified as well. An aspect specifically addressed by the HCWs is a concern about a potential loss of their autonomy, control, and professional skills if they " blindly" follow an algorithm and do not have to think for themselves anymore. The technology has to respect and keep into account their medical autonomy. Furthermore, experts were rather concerned about the misuse of profiling, whereas HCWs stated that profiling in itself is nothing new. According to the experts, the danger of profiling is that you can never leave the assigned group again. Also, profiling might be suffocating to people because it creates uncertainty about what people know about you, what data is being collected, and for what purpose. Profiling might lead to discrimination and certain prejudices/biases and people might experience the feeling that they lose control. On the other hand, HCWs claim that "profiling is something we have always done, otherwise you cannot start any treatment". HCWs believe that Big Data has the potential to increase its accuracy even further. One concern they do have, in correspondence with the experts, is about the potential misuse of profiling by third parties like insurance companies.

Though these results provide a broad overview of promises as well as barriers that need to be taken into account when using Big Data in healthcare, a few important limitations should be taken into consideration when interpreting the results. At first, we only performed one focus group. This provided us with diverse insights, but we are not able to determine if saturation has been reached [16]. Still, we do expect that we covered a rather broad area, given the multidisciplinary composition of the group and the large variety of expertise they brought into the discussion. To ensure the accuracy of the results and to prevent that the results represent the interpretation of the researchers, clarification and follow-up questions were asked in case of ambiguity to ensure the validity of the results. As such, we believe that the findings provide an accurate exploration of issues that play a role when using Big Data for personalization purposes in healthcare, from a scientific perspective as well as from a societal perspective.

Secondly, only 6 physicians were interviewed, potentially providing a rather limited view on how HCWs in general think about Big Data. However, that was also not the intent of this study. The aim was to gain a better understanding of technological and ethical challenges that need to be faced when using and managing Big Data in healthcare, as well as to gain insight into its impact on our way of working, our health, and our wellbeing. The interviews with the physicians provide some important first insights for this that can be studied further. All physicians had knowledge of and/or experience with Big Data in some way, to ensure they were able to discuss the topics that were addressed. The interview scheme was constructed based on the input from experts to make sure the same themes were addressed, allowing us to compare the results. At the same time the interview scheme was semi-structured and questions were formulated open-ended to allow the physicians to raise other thoughts as well, enriching the data.

Another limitation is that the results from the perspective of the online public might be colored, as the data are restricted to those who use social media. Therefore, a completely reliable reflection of how the general (online) public thinks or speaks about Big Data in the context of healthcare cannot be given at this moment.

Finally, none of the experts or healthcare workers turned out to be a strong adversary of Big Data in healthcare, even though they did provide some critical comments. As such, it would be interesting to extent the results with the opinions of strong adversaries. After all, for sake of implementation, it is important to take their concerns into consideration as well.

V. Conclusion and Future work

Big Data is seen as the key towards personalized healthcare. However, it also confronts us with new technological and ethical challenges that require more sophisticated data management tools and data analysis techniques. This vision paper aimed to better understand the technological and ethical challenges we face when using and managing Big Data in healthcare as well as the way in which

it impacts our way of working, our health, and our wellbeing. A mixed-methods approach (including a focus group, interviews, and an analysis of social media) was used to gain a broader picture about the pros and cons of using Big Data for personalized healthcare from three different perspectives: Big Data experts, HCWs, and the online public. All groups acknowledge the positive aspects of applying Big Data in healthcare, touching upon a wide array of issues, both scientifically and socially. By sharing health data, value can be created that goes beyond the individual patient. The Big Data revolution in healthcare is seen as a promising and innovative development.

Yet the development of these advancements is not self-evident and potential facilitators and barriers need to be addressed first. Concerns were raised, mainly about privacy, trust, reliability, safety, purpose limitation, liability, profiling, data ownership, and loss of autonomy. Also, trust in the technological applications is essential to overcome uncertainty or anxiety for a digital world. To achieve this, a first condition is that privacy and security issues are dealt with appropriately. People should be able to decide for themselves whether or not to share their data and with whom. Also, algorithms should be transparent (at least to a certain degree) to the users (e.g., physicians) to make them meaningful. Reliability should be assured and different kinds of expertise need to evolve. Expertise to analyze Big Data, to develop and understand the working of algorithms, and to interpret and visualize the data in a meaningful way. Moreover, technology should be embedded in our way of working and living. As such, technology should supplement the work of physicians, not replace it, respecting the medical autonomy. The digitalization of society is an ongoing process and the “Big Data” revolution is already changing science, healthcare, and society.

In general, Big Data is described according to the 5V model (Volume, Velocity, Variety, Veracity and Value) [5]. Yet this paper stresses the importance of adding the people-centered view to this rather data-centered 5V model, in order to get a grip on the opportunities for using Big Data in personalized healthcare. Following this view, this vision paper aimed to discuss Big Data topics for personalized healthcare that need to be investigated further to 1) develop new methods and models to better measure, aggregate, and make sense of previously hard-to-obtain or non-existent behavioral, psychosocial, and biometric data, and 2) to develop an agenda for Big Data research to transform and improve healthcare. Topics include:

- Health analytics: Advanced methods (machine learning) and models to analyze Big Data.
- Predictive modelling: To set up smart models to predict behaviors, to prevent diseases, and to personalize healthcare.
- Visualization of data: How to present data meaningful (to the patient as well as the HCW) to support decision making?
- Integration of (mobile) technology with data-platforms to enable automated services and to tailor feedback.
- Disruptive models (new actors, role-players in data driven systems).

ACKNOWLEDGMENT

The authors thank the participants of the focus group as well as the healthcare workers that were interviewed for their valuable input in this study.

REFERENCES

[1] J. E. W. C. van Gemert-Pijnen, F. Sieverink, L. Siemons, and L. M. A. Braakman-Jansen, "Big data for personalized and persuasive coaching via self-monitoring technology," The Eigth International Conference on eHealth, Telemedicine, and Social Medicine (eTELEMED 2016), IARIA, 2016, pp. 127-130, ISBN: 978-1-61208-470-1.

[2] V. Mayer-Schonberger and K. Cukier, Big data: a revolution that will transform how we live, work, and think. New York, NY: Houghton Mifflin Harcourt, 2013.

[3] W. Wang and E. Krishnan, "Big data and clinicians: a review on the state of the science," JMIR Med Inform, vol. 2, pp. e1, 2014, doi:10.2196/medinform.2913.

[4] D. Laney, 3D Data management: Controlling data volume, velocity, and variety. Stamford, CT: META Group Inc., 2001.

[5] B. Marr, Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance. West Sussex, United Kingdom: John Wiley & Sons, 2015.

[6] S. Klous and N. Wielaard, We are big data. The future of the information society [Wij zijn big data. De toekomst van de informatiesamenleving]. Amsterdam: Business Contact, 2014.

[7] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," JAMA, vol. 309, pp. 1351-1352, 2013, doi:10.1001/jama.2013.393.

[8] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: a patient-centered framework," J Gen Intern Med, vol. 28, pp. S660-665, 2013, doi:10.1007/s11606-013-2455-8.

[9] F. Sieverink, L. M. A. Braakman-Jansen, Y. Roelofsen, S. H. Hendriks, R. Sanderman, H. J. G. Bilo, et al., "The diffusion of a personal health record for patients with type 2 diabetes mellitus in primary care," International Journal on Advances in Life Sciences, Vol. 6, pp. 177-183, 2014.

[10] P. Kamakshi, "Survey on big data and related privacy issues," International Journal of Research in Engineering and Technology, vol. 3, pp. 68-70, 2014.

[11] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," Business Horizons, vol. 53, pp. 59-68, 2010.

[12] R. Batool, W. A. Khan, M. Hussain, J. Maqbool, M. Afzal, and S. Lee, "Towards personalized health profiling in social network," Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in (IEEE, 2013), 2012 pp. 760-765, ISBN: 978-89-94364-20-9.

[13] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson, "Harnessing the cloud of patient experience: using social media to detect poor quality healthcare," BMJ quality & safety, vol. 22, pp. 251-255, 2013, doi: 10.1136/bmjqs-2012-001527.

[14] R. Nagar, Q. Yuan, C. C. Freifeld, M. Santillana, A. Nojima, R. Chunara, et al., "A case study of the New York city 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives," J Med Internet Res, vol. 16, pp. e236, 2014, doi: 10.2196/jmir.3416.

[15] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations,"

Proceedings of the 5th Annual ACM Web Science Conference (ACM 2013), 2013, pp. 47-56, ISBN: 978-1-4503-1889-1.

[16] R. A. Krueger and M. A. Casey, Focus groups - a practical guide for applied research. Thousand Oaks, CA: Sage, 2015.

[17] D. Wishart, CLUSTAN user manual, 3rd ed., Edinburgh: Program Library Unit, Edinburgh University, Inter-university research councils series, 1978.

[18] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," Proceedings Third International ICWSM Conference, 2009, pp. 361-362.

[19] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of Statistical Mechanics: Theory and Experiment, vol. 10, pp. P10008, 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[20] M. E. Newman, "Modularity and community structure in networks," Proceedings of the National Academy of Sciences, 2006, vol. 103, pp. 8577-8582, doi: 10.1073/pnas.0601602103.

[21] J. Scott, Social network analysis. London: Sage, 2012.

SUPPORTING INFORMATION - APPENDIX 1

WORD FREQUENCY TERMS "BIG DATA" – "HEALTHCARE.

Term*	Word frequency	Term	Word frequency
Care	**7469**	Safety	167
Health	**4196**	Anonymity	163
Technology	1726	Patients	153
Innovation	1361	Algorithms	142
Development	1337	Reliability	139
Healthcare	**1320**	Legislations	91
Future	794	Ethics	90
eHealth	**668**	Healthtap	90
Service	638	S Health	89
Platform	618	Ownership	83
Software	610	Smarthealth	81
Cloud	581	Personalization	81
Privacy	545	**Wellbeing**	**80**
Challenges	484	Regulations	74
Security	400	Google Glass	69
Wearable	**374**	**Domotica**	**55**
Social media	374	Quantified self	53
Big Data	**337**	iWatch	44
Start-up	324	Healthkit	42
Medical data	311	Autonomy	37
Health insurance	310	Profiling	36
Trust	310	Pedometer	34
Medicine	309	Wellness	31
Infrastructure	291	Biometrics	19
Analytics	249	Scalability	13
Revolution	242	Runkeeper	12
Sensors	**229**	Fitnesstracker	11
Transparency	196	**Health condition**	**4**
Fitbit	185	Fitness	2
Internet of Things	185	Standardization	1

*** Bold**: Not included in the data-analysis, since they were also present in the search query. Not bold: Included in the data-analysis.