BIG DATA AND HEALTH

REVOLUTIONIZING MEDICINE AND PUBLIC HEALTH

Report of the Big Data and Health Working Group 2013

Professor Alex Pentland,
Dr Todd G Reid, and
Dr Tracy Heibeck

WISH
مؤتمر القمة العالمي للابتكار في الرعاية الصحية
World Innovation Summit for Health
DEC 10-11 DOHA 13

an initiative of Qatar Foundation

BIG DATA AND HEALTH

REVOLUTIONIZING MEDICINE AND PUBLIC HEALTH

Report of the Big Data and Health Working Group 2013

Professor Alex Pentland,
Dr Todd G Reid, and
Dr Tracy Heibeck

CONTENTS

Professor The Lord Darzi

Professor Alex Pentland

FOREWORD

Since the beginning of time, most people have been isolated, without information about or access to the best health practices. But in just the last decade, this situation has changed completely: through the spread of cell phone networks, the vast majority of humanity now has a two-way digital connection that can send voice, text, and most recently, images and digital sensor data. Healthcare is suddenly something that is potentially available to everyone; all across the world, we are beginning to see healthcare workers collecting health information and delivering telemedicine consultations in even the most remote areas.

This new digital nervous system is also driving a more subtle and potentially even more profound change known popularly as "Big Data." The proliferation of wireless devices such as cell phones provides an enormous stream of data about human life and behavior. When these new capabilities are combined with existing health data, they create new opportunities to detect and monitor disease and disease vectors, and to provide non-traditional interventions that increase access to – and reduce the costs of – healthcare. Examples include the ability to track and control flu propagation, to identify sources of malaria and food poisoning, and to co-ordinate disaster recovery. Big Data is giving us the ability to know about health status everywhere and in near real-time. Historically we have always been blind to health conditions outside central cities; diseases could spread to pandemic proportions before the news would make it to the ears of central health authorities. We are now beginning to be able to see the health conditions for all of humanity with unprecedented clarity.

This new view of human health and behavior is also beginning to enable tremendous strides in medical science. By combining fine-grained, ubiquitous monitoring of human behavior with standard medical data and standard genomic data, we are taking the first steps towards generating a new, holistic understanding of disease and disease processes. A comprehensive, Big Data understanding of phenotypic, genetic, and treatment variables promises to revolutionize medicine and medical treatment.

The use of Big Data in health is a new and exciting field, full of promising case examples, but it is not as fully tested as most health and medical systems. As a consequence, while there is enormous promise, there are also practical problems to be worked out, such as data privacy and ownership issues. There are also dangers to be avoided, such as the risks of misuse of personal data and new types of medical error. This report aims to give a view into the future of Big Data in health, and to map out concrete steps that will help ensure that we can realize its full potential.

Professor The Lord Darzi PC, KBE, FRS
Executive Chair of WISH, Qatar Foundation
Director of Institute of Global Health Innovation,
Imperial College London

Professor Alex Pentland
Director, Human Dynamics Laboratory
Massachusetts Institute of Technology

EXECUTIVE SUMMARY

THE ROLE OF BIG DATA IN HEALTH

Everyday devices such as cell phones now provide us with an enormous stream of data about human life and behavior. Combined with existing health data, the behavioral data obtainable from these devices may greatly enhance opportunities to predict long-term health conditions and identify non-traditional intervention points, as well as to design better diagnostics tools, prevent diseases, and increase access to – and reduce the costs of – healthcare. Significant application areas include: chronic and infectious diseases, mental health, environmental health, nutrition, healthcare cost and quality, accidents and injury, and social health. While there is enormous promise, there are also dangers to be avoided, including the following: data privacy and ownership issues, risks of misuse of personal data, and new scientific risks.

MANAGING BIG DATA

Both regulation and technology must continue to evolve in order to provide us with the potential benefits while not exposing citizens to the dangers of exploitative companies or unreasonable government oversight. A taxonomy framed in terms of data control is discussed, including:

- **Open Data Commons:** Our data are worth more when openly shared, because they can inform improvements in healthcare and public health systems.
- **Personal and Proprietary Data:** New regulations in the EU, the US, and elsewhere require the use of both new computer security algorithms and new contractual agreements to specify and audit how data may be used and shared. Currently, trust networks provide the best practice for data sharing of personal and proprietary data.
- **Government Data:** To ensure adequate security and oversight, restricted government data should be both physically and logically distributed, and should have heterogeneous computer and encryption systems.

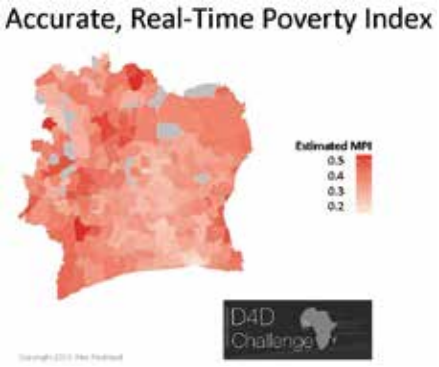
RECOMMENDATIONS

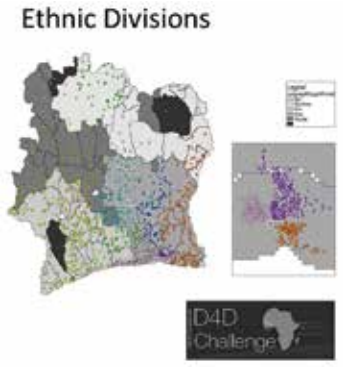
A medical and health science based on the use of Big Data is emerging. How can we best support the development of Big Data health systems?

- **Encourage public-private partnerships:** This kind of collaboration can serve to underwrite costs and accelerate deployment; special sector banks are a useful method.
- **Ensure data access:** Update privacy and data ownership policies to ensure that data are accessible to patients and their healthcare providers, and require trust network technology in order to provide safe data sharing.

- **Allow for open data:** Pool unrestricted government data and non-proprietary private data in an open data commons, in order to promote development of a "Big Data" health ecosystem. We suggest that there needs to be an international Charter for Open Data Sharing, which specifies best practice, and commits nations to sharing health data for their mutual benefit.
- **Promote Big Data health science:** Create centers of excellence to train Big Data behavioral and health scientists in the use of open-source tools for data analysis.
- **Accelerate Big Data health practices:** Support partnerships between physicians and Big Data behavioral scientists to create "living laboratories" that develop new Big Data health solutions.

BIG DATA IN THE DEVELOPING WORLD

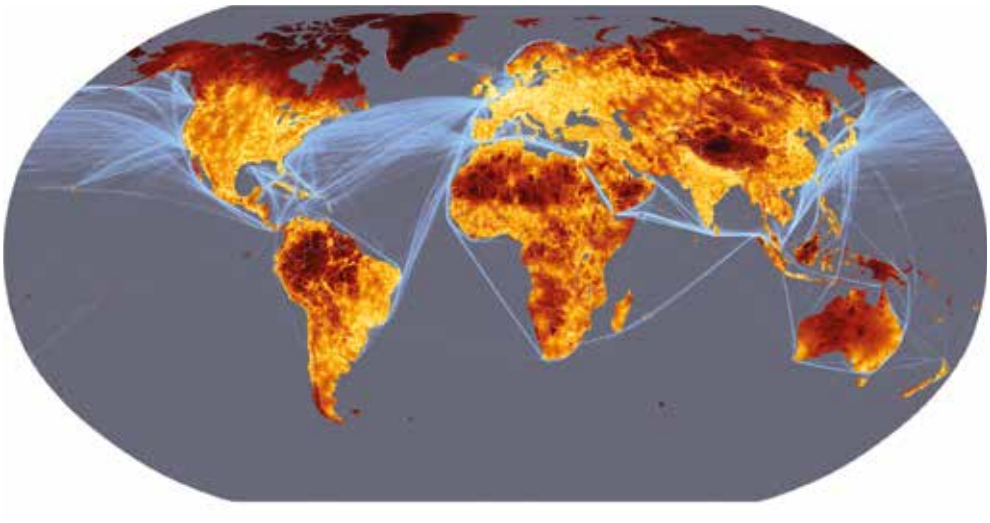




Big Data is not only for developed nations; indeed, important applications of Big Data are taking place in some of the world's poorest countries. As an example, consider the Data for Development (D4D) initiative.[1] In this collaborative effort, 90 research organizations from around the world reported hundreds of results from their analyses of cell phone data. They described the mobility and call patterns of the citizens of the entire African country of Ivory Coast, a country struggling with poverty and the aftermath of a recent civil war.

In several projects, the D4D data were utilized to understand and promote the operational efficiency of health systems. For instance, analysis of human mobility patterns showed that small changes in the health system could potentially cut the spread of flu by 20 percent as well as substantially reduce the spread of HIV/AIDS and malaria.

The D4D data were also used to address social health issues. An example: the development of a method for mapping poverty from the diversity of cell phone usage. As people have more disposable income, their patterns of movement and patterns of phone calls become increasingly diverse. Another example of using the D4D data for social health: the development of a method for mapping of ethnic boundaries. This method relies on the fact that ethnic and language groups communicate far more within their own group than they communicate with others. Mapping social boundaries is important because, while we know that ethnic violence often erupts along such boundaries, the government and aid agencies are usually uncertain about the geography of these social fault lines.



Worldwide distribution of real-time behavior-sensing systems (aka cell phones)

INTRODUCTION

Around the globe, we are increasingly living our lives in digital networks. We wake up in the morning, check our messages, make a phone call, commute to work, and buy lunch. All of these activities leave behind digital breadcrumbs – tiny, detailed records of our daily experiences that together comprise what we call "Big Data." These digital traces are accumulating in both wealthy and developing countries; cell phones are already commonplace in almost every part of the world, and "digital cash" is spreading quickly through many of the poorest nations.

These digital crumbs offer an unprecedented, ubiquitous and continuous view of our individual lives and behavior: where we live and work, our activity level, travel patterns, shopping habits, what we eat and drink, and which people we interact with. By linking these petabytes of raw information to health records, demographic data and genetic information, we secure novel opportunities to uncover population health patterns, predict long-term conditions and identify non-traditional intervention points. Improved disease prevention is now achievable, as are better diagnostic tools, as well as increased access to – and reduced cost of – healthcare.[2]

SCOPE OF THE PAPER

Big Data analysis has now reached every sector in the global economy. Awash with data, the traditional stakeholders from the healthcare industry have struggled to turn these data into information that guides care decisions more effectively. If, for example, the US healthcare industry were to use Big Data creatively and effectively to drive efficiency and quality, it is estimated that the sector could create savings of more than US$ 200 billion every year, reducing US healthcare expenditure by about 8 percent[3] (see the costs case studies in the Appendix).

The potential uses of Big Data extend far beyond enabling more efficient healthcare systems. So far, these extended uses have mostly remained theoretical possibilities, owing to a number of barriers, including privacy and data ownership issues. There are emerging best practice examples, however, which demonstrate the potential of these largely untapped health and behavioral data. Therefore, the purpose of the WISH Big Data and Health Forum is to provide a framework that will inform discussion of the role of Big Data, summarize existing best practices, highlight the remaining barriers, and develop policy recommendations to overcome the barriers in the intersection of Big Data, health, and medicine.

Within this paper, we will direct our attention to areas outside of the complex hospital systems of developed nations. Our reasons for doing so are these:

- The transformative potential for Big Data seems greatest where there is currently the least data;
- The vast majority of humans do not have access to advanced hospital systems; and
- Many of the challenges of using Big Data within hospitals concern entrenched financial interests and legacy legal barriers, and therefore require detailed, specific discussion outside the scope of this paper.

While use of Big Data holds out enormous promise for improving health systems, there are also dangers that must be avoided. There is scientific risk: the unfamiliar, correlational nature of Big Data raises the possibility of misinterpretation that can cause serious harm. Consequently, we must devise new procedures for developing health systems that incorporate Big Data. There is also risk of misuse: stemming from the danger of putting so much personal data in the hands of either companies or governments. So we will discuss how new approaches to regulation and technology have been developed – approaches that can help protect personal privacy from exploitation, and can also mitigate the problem of government overreach.

ANALYSIS: THE ROLE OF BIG DATA IN HEALTH

A myriad of everyday devices already provide us with an enormous stream of data about human behavior. Sensors in cell phones, security cameras, "smart card" readers, digital wallets, loyalty cards, smart electricity meters – these and large-scale e-commerce all enable the measurement of human physical and social activity. In addition, there is all the data generated from online social networks, internet documents, digital video, and digital photography – such data expand our ability to understand our collective mental and cultural life. All of these billions of digital traces provide scientists with a new lens to examine society in fine-grained detail. This new method of observing human behavior, sometimes called "reality mining," was identified by Technology Review as one of "ten emerging technologies that could change the world," helping us measure, document, and ultimately better understand the dynamics of human life.[4]

A familiar current example of using Big Data for health purposes is Google Flu Trends. This resource predicts outbreaks of flu by counting the number of flu-related internet searches using the word "flu" that occur in each state or region of the US. Regions with a strong increase in the number of online "flu" searches are likely to be experiencing increases in the number of flu cases. Reality-mining techniques similar to this – but based on patterns of pharmacy purchases, commuting traffic, and records of school and work attendance – have long been used by the US Centers for Disease Control and Prevention (CDC). They allow us to detect new strains of flu, to predict the amount of medicine that will be required, and to help hospitals, cities, and companies anticipate the number of sick patients, residents, and employees that they may have.

New sorts of continuous behavior-sensing technologies, such as cell phones and digital payments, are now becoming operational across the entire human population. As with the earlier behavior-sensing systems used by the CDC, these are beginning to find their first application in health systems as a sort of "extended nervous system," detecting the very first signs of disease and making "just-in-time" medicine a potential reality.[5] Such focused early interventions could, of course, provide dramatic improvements in health outcomes as well as cost savings (see the Punjab case study in the Appendix).

Within the domain of medical care, technological advances such as on-line health forums and digital data commons[6] are also opening up new opportunities, including: understanding both chronic and infectious disease on a population level, new approaches to diagnosis and patient- and treatment-monitoring, improved surveillance of disease and risk factors, and improved health investigation and disease control.

These data about human behavior and belief, together with electronic medical records (EMRs) and genomics information, can potentially provide us with a far more complete picture of human health. They can also provide us with new opportunities and methods for encouraging healthy behavior, and new capabilities for medical intervention.

To achieve these goals requires new sorts of experimental techniques; the scientific method as currently practiced in the health sciences is becoming increasingly inefficient, and threatens to collapse in an era of Big Data. We need controlled experimentation in order to develop Big Data health systems. But owing to the novel and unfamiliar nature of inferences made using these new data sources, it is difficult to squeeze such experimentation into a traditional framework of treatment-control experiments.

In consequence, we need to construct living laboratories in order to test and confirm our ideas for building data-driven health systems. What is a living lab? Imagine being able to place an imaging chamber around an entire community, and to record, analyze, and display every facet and dimension of individual behavior, genetic background, and every medical measurement within the member population. And imagine doing that for several years, while the members of the community go about their everyday

lives. Living labs are the Big Data equivalent of real-time fMRI scanning, giving the most complete picture possible of the entire health ecosystem (see the CATCH case study in the Appendix).

Restaurants, stores, and entertainment venues color-coded by chronic-disease risk of their patrons.

This map of San Francisco was obtained by analysis of city-wide cell phone mobility data and interviews with local citizens.

CHRONIC DISEASES

The overwhelming majority of chronic diseases in humans arise from a complex web of causes that act over years and often decades before the disease is manifested. The mission and challenge of epidemiologic research is to unravel these causes, as a prerequisite for prevention. The Achilles heel of such research is to accommodate the complexity of the data and adequately ascertain information about causes. In this regard, the new technologies that help collect, analyze, and correlate large volumes of personal health data might offer entirely new opportunities. Indeed, interactive prevention via such accessible technologies could be used to reach segments of the population that cannot access medical care otherwise.

Much of the information that we have now about preventing chronic, non-communicable diseases (NCDs) has come from traditional longitudinal studies. For instance, studies based in North America and Europe have heavily influenced preventive health policy such as dietary standards for schools and restaurants, laws regulating smoking, and air-quality standards.[7] In some ways, these studies could be considered the progenitors of Big Data, in that they typically enroll thousands of people and assess them periodically, usually by questionnaire. These questionnaires can be web-based, mailed, or conducted in-person or by telephone interview. For the most intensive of such studies, the frequency of assessment is typically every 6-24 months. The data from these research instruments, however, are limited to what individuals have said (or think) that they have done. With reality-mining innovation, we get to measure what the individual has actually done, and that leads to important new insights.[8]

Rapidly developing countries face the challenge of replicating such research in ways suitable for their context, given the obstacles to disseminating questionnaires (such as limited mail and land-line telephone service) and the limited internet access for their populations. Big Data and reality-mining techniques give scientists in these countries the opportunity to "leapfrog" over these obstacles, collect far more and far better information than before, and dramatically improve health data at the same time (see the PaCT case study in the Appendix).

Behavioral data also offer intriguing possibilities in other areas of chronic disease. There is research suggesting that some chronic health-related conditions and behaviors are "contagious," in the sense that individual-level outcomes are linked to other individuals with whom one shares social connections. For example, both smoking behavior and obesity have been shown to spread within social networks. And the same is likely to be true for other health-related behaviors as well, such as diet, exercise, general hygiene, sexual habits, and so on. As such, reality mining might yield specific points of leverage for effective health interventions. That is, if certain behaviors are indeed contagious, then targeting individuals in key parts of the social network could generate more powerful approaches to intervention and more effective ways to promote behavior modification. Of course, privacy issues are paramount here (see the discussion on Management of Big Data).

Map of the MIT campus showing risk of contracting infectious disease at the current time, obtained from analysis of cell phone mobility patterns and short cell phone health surveys of selected individuals.

INFECTIOUS DISEASES

As the world becomes increasingly interconnected through the movement of people and goods, the potential also increases for global pandemics of infectious disease. In recent years, following outbreaks of SARS and other serious infectious diseases in widely separated but socially linked communities, the need has clearly arisen for fundamental research on disease transmission and effective prevention and control strategies. In developed countries, health officials typically investigate cases of serious infectious disease (such as tuberculosis, SARS, and malaria) to identify other cases and the source of infection, and to prevent further transmission. The investigations are typically difficult and time-consuming, and while they are underway, transmission continues unabated. Moreover, informants often forget all the locations they have visited, even for recent periods. Similarly, they might not know many of the people to whom they were exposed or might have exposed themselves. Given such difficulties, disease control would potentially benefit hugely from any systematic analysis of location data and social-behavioral data, both readily obtained from cell phones. Logs of location-tracking data from patients' cell phones can be examined to identify places where the patients might have acquired or transmitted infection, thereby facilitating the investigation. Recently, this approach has demonstrated its effectiveness in elucidating the transmission of malaria[9] and food poisoning.[10]

Reality-mining tools too could assist in the detection of disease outbreaks. For instance, acute illnesses such as influenza – illnesses causing sufferers to reduce their physical activity and mobility patterns (even confining them to bed) or to change their communication behavior – are identifiable in several types of reality-monitoring data streams.[11] At the population level, fluctuations in digital traces of these behaviors may indicate outbreaks of such diseases. At the individual level, the emergency-room or clinical-intake process would include an examination of data about an individual's exposure summary and indicate, for example, if the patient had eaten or spent much time near known outbreak areas; information that might not have been captured through self-report alone. In the future, such tools could offer a formidable defense against pandemics: a recent pilot study has demonstrated the potential for real-time tracking of flu propagation from individual to individual, using only behavioral data collected from smartphones.[12]

APPLICATION AREAS

DIAGNOSIS, TREATMENT, AND FOLLOW-UP OF HUMAN DISEASE

Big Data obtained from the continuous monitoring of motor activity, metabolism, and so on can be extremely effective in tailoring medications/treatments for individuals. Once a course of treatment (behavioral, pharmaceutical, or otherwise) has been chosen, it is important for a clinician to monitor the patient's response. For that purpose, the clinician can use the same types of Big Data as used for diagnosis. The patient's compliance, response, and side effects to treatment then become clearer, especially when the patient's pre-diagnosis data are available and can serve

as a baseline for comparison. Even when these data streams are not relevant for diagnosis, they can be useful in assessing side effects of treatment, such as reduced mobility, activity, and communicative behavior. Because these data can be collected in real-time, a clinician would be able to adjust treatment according to the patient's response, perhaps leading to more effective treatment and preventing more costly office visits.

Currently, doctors prescribe medications on the basis of population averages rather than individual characteristics. They assess patients for the appropriateness of the medication levels only occasionally – and expensively. With such a data-poor system, it is not surprising that medication doses are often over- or underestimated, and that unforeseen drug interactions result. Such adverse effects account for a sizable proportion of hospitalizations, notably among the elderly. Many or most of those could be avoided once data are optimized. The trick would be to correlate a continuous, rich source of behavioral data to prescription medication use for millions of people. This could make drug therapies more effective, and help medical professionals detect new drug interactions more quickly[13] (see the mPedigree case study in the Appendix).

MENTAL HEALTH

Mental diseases rank among the top health problems worldwide in their cost to society. Major depression, for instance, is the leading cause of disability in established market economies. Diagnoses of psychiatric disorders are overwhelmingly based on reporting by the patient, or by a teacher, family member, or neighbor. In fact, many symptoms of psychiatric disorders concern patterns of physical movement, activity, and communication – all things that can be measured by cell phone data. Accelerometers can reveal fidgeting, pacing, and abrupt or frenetic motions. Location tracking can reveal changes in places visited and routes taken, as well as the overall extent of physical mobility. The frequency and pattern of individuals' communications with others, and the content and manner of their speech, can also serve as key signs of several psychiatric disorders.[2, 14]

Now that we have the ability to use inexpensive, pervasive computational platforms such as cell phones to monitor these sensitive indicators of psychological state, we can dramatically improve the early detection of disorders such as depression, attention deficit hyperactivity disorder, bipolar disorder, and agoraphobia.[15] In addition, because the data streams provide direct, continuous, and long-term assessment of patterns and behaviors, we can develop new avenues of monitoring and assessing treatment in mental health.

ENVIRONMENTAL HEALTH

In making epidemiologic investigations of the links between various health conditions and individuals' exposures to airborne pollutants (such as particulate matter, carbon monoxide, nitric oxide), we have in the past relied on a variety of methods for measuring exposure. To date, most studies compared aggregates of persons (residents of particular neighborhoods or cities, or students at specific schools) with exposure measurements applying to all individuals in a given group. Air pollution

levels, however, can vary dramatically over short distances and time scales in urban and other environments. Thus environmental health experts have called for more precise and dynamic measures of time-activity patterns in relation to exposures. Such measures could well be provided by location-tracking data generated by cell phones, when coupled with measurements of ambient air pollutions at numerous places in a community (gathered from existing air-quality monitoring stations and/or inferred from vehicle traffic patterns and locations of industrial facilities).[16]

NUTRITION

Innovations driven by Big Data could also help to revolutionize nutrition epidemiology. Dietary record-keeping has long been a major challenge in health research: it is often riddled with bias from the inaccurate recall of what people have eaten over the course of a week, month, or year. As it is now possible to track dietary intake at almost every point of consumption, previous inaccuracies in recording can be markedly minimized. For example, GPS-enabled cell phone applications could track whether individuals frequented fast-food restaurants vs farmer's markets, or even the produce aisle vs the snack-food aisle in a community's grocery store. Detailed consumer-purchase streams will not only serve as enormous data-rich sources for dietary record keeping, but will also offer unprecedented opportunities to track and analyze important correlated behavioral data associated with nutrition health outcomes. In addition, food security and availability can be improved, to ensure more steady and adequate supplies of nutrition on a population level.

SOCIAL HEALTH

Despite compelling evidence, most efforts to encourage healthy behavior and medical compliance are still organized around conscious, individual decision-making only, and the social dimension is almost entirely neglected. By using Big Data to gain a better understanding of social conditions, we could achieve more in terms of behavioral modification and hence health advancements. For example, Big Data can be used to provide the social pressure needed to establish new, healthy behavior norms. Online "friend networks" have successfully been used to promote higher levels of physical activity and to increase pro-social behaviors such as voting and energy conservation.[17,18]

The D4D initiative highlighted on page 3 provides another good example of the way that Big Data can help to address social health issues. In these studies, cell phone call and mobility data, together with more traditional sources of information, were used to provide ubiquitous, up-to-the-minute mapping of poverty and ethnic boundaries throughout Ivory Coast.

PERSONAL HEALTH

In the last few years, the number of "quantified self" applications on cell phones has surged. These applications were originally designed mainly for self-improvement rather than health maintenance or healthcare. Recently, however, many companies, hospitals, and healthcare systems have begun testing mobile applications that deliver personalized feedback about health issues, as well as giving healthcare

providers a more complete picture of patient or employee health. The success of these applications has been variable at best, and the health feedback they deliver to the individual has been fragmented.[19] More recently, a new generation of health apps – mostly driven by passive sensor measurement rather than by human answering of questions, and therefore less intrusive – have been more successful at integrating personal health measurement and feedback into the overall healthcare system.[20]

IMPROVING HEALTHCARE QUALITY WHILE REDUCING COST

Big Data will help to increase quality of care, and also to reduce costs, by making it possible to identify best practices and reveal inconsistencies in healthcare delivery. It is not uncommon for nearly identical health services in different places to have drastically different costs. Much to the frustration of the patient, these service costs are often not very transparent, have little to do with quality of service, and are typically driven by the relative bargaining power of the providers. For example, with Big Data analytics, patients will be able to access information on the doctors who generate the highest costs for specific procedures; doctors can also work to lower costs by better understanding which tests are unnecessary. As a case in point, it is estimated that using Big Data to recommend specific actions to combat coronary disease – such as taking aspirin, undergoing cholesterol screenings, and quitting smoking – could reduce US healthcare costs by US$30 billion annually.[3] To achieve these quality improvements and cost reductions, however, requires that the underlying cost and outcomes data be available for analysis. Consequently, creation of an open data commons, containing both cost and outcomes data, is a high priority for any effort to improve health through the power of Big Data (see the Ghana case study in the Appendix).

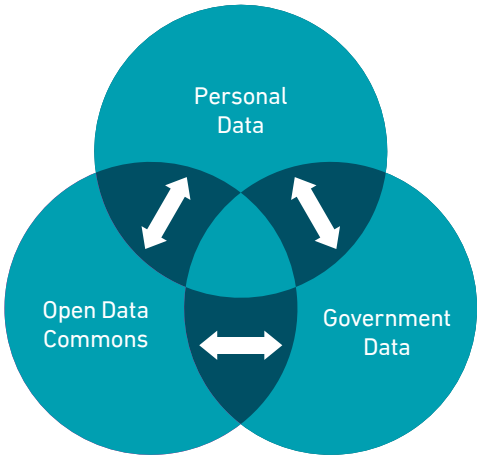
ACCIDENTS AND SAFETY

Accidents constitute a major source of death and long-term disability, so data about where, when, and how accidents occur must form a large element of any Big Data health strategy. In developed countries, effective accident databases exist because healthcare providers are required to file digital accident reports containing exact location, time, and other contextual data. However, in rural and poor regions, such reporting is usually not possible. But it is possible to report digital health information collected by rural healthcare workers, usually on smartphones or similar devices. In some of the best rural health systems, nurse midwives visit each family once or twice a year, and collect a wide variety of digital health information. These data can then be combined with information such as traffic patterns or environmental conditions, and thereby help to improve policy and set priorities (see the Dimagi case study in the Appendix).

MANAGING BIG DATA

Reality mining of Big Data for behavioral information is still in its infancy. In the near future, though, it may be common for smartphones and other ubiquitous devices to continuously monitor a person's motor activity, social interactions, sleep patterns, and other health indicators. These data can be used to build a personalized profile of an individual's physical performance and nervous system activation throughout the entire day. If these rich data streams were combined with personal health records, including medical tests taken and medicines prescribed, that creates the potential for dramatic improvements in healthcare.

These new tools, with their view of life in all its complexity, could well be the future of medical science and public health policy. There is risk in deploying this sort of data-driven health system, however, because of the danger of putting so much personal data in the hands of either companies or governments. Fortunately, new approaches to regulation and technology have been developed that can help protect personal privacy from exploitation, and can mitigate the problem of government overreach as well. Both regulation and technology must continue to evolve in order to provide more scientific, real-time public policy without exposing citizens to the dangers of exploitative companies or governments. This section of the report will outline the current best practices in this area.



Big Data taxonomy: Overlapping regions require a trust network for data sharing.

Open data commons: Includes geo-located and time-stamped statistics about aggregate costs, health outcomes and behaviors (eg, mobility, smoking, drinking, crime, accidents).

Personal data: Includes digital breadcrumbs (mobility, call, and purchasing patterns, etc), personal annotations about eating, subjective variables, and standard health data (temperature, glucose, genomic, etc). Collection of these data often requires participation by service providers.

Government data: Includes detailed healthcare system costs, individual performance ratings, fine-grained outcomes data.

A BIG DATA TAXONOMY

It is probably hopeless to provide a detailed taxonomy of data types and uses, because the technology is progressing so quickly. But it is possible to provide a broad taxonomy framed in terms of control. The three main divisions within the spectrum of data control are:

- Data commons, which are available to all, with at most minor limitations on use;
- Personal or proprietary data, typically controlled by individuals or companies, for which there needs to be legal and technology infrastructure that provides strict control and auditing of use; and
- The secret data of governments, which typically has less direct public oversight and more stringent controls.

We will look at each of these three types in turn.

Open Data Commons. A key insight is that our data are worth more when shared, because they can inform improvements in systems such as health, transportation, and government. Using a "digital data commons" could give us unprecedented instrumentation for assessing the way our policies are performing, so we can know when and how to take effective action to address the situation.

We already have many data commons available: maps, census data, and financial indices. With the advent of Big Data, we can potentially develop many more types of data commons. These commons can be both real-time and unprecedentedly detailed, because they depend mostly on data that are already produced as a side effect of ongoing daily life (digital transaction records, cell phone location fixes, road toll records, and so on). That is, they can be produced automatically by computers without human intervention.

One major concern with such a data commons is that it could endanger personal privacy. Another concern involves the tension between personal and commercial interests: these proprietary interests might reduce the richness of such a commons, and diminish its ability to deliver public goods.

To explore the viability of a Big Data commons, the D4D initiative hosted what is perhaps the world's first true Big Data commons, which included data describing the mobility and call patterns of the citizens of Ivory Coast as well as more traditional data sources.[1]

The work of the 90 research groups involved in D4D suggests that many of the privacy fears associated with the release of data about human behavior may be generally misunderstood. In this data commons, the data were processed by advanced computer algorithms (for example, sophisticated sampling and the use of aggregated indicators), so it was unlikely that any individual could be re-identified. In fact, no path to re-identification was discovered, even by several of the research groups that studied this specific question.

In addition, while the data were freely available for any legitimate research that a group was interested in, the legal contract specified that the data could be used only for the purpose proposed and only by the specific people making the proposal. A similar technology-legal framework is used in trust networks, as described in the next section. This use of both contract law and advanced computer algorithms to specify and audit how personal data may be used and shared is the goal of new privacy regulations in the EU, the US, and elsewhere.

Personal and Proprietary Data. This second kind of data is typically controlled by individuals or companies, and requires legal and technology infrastructure that can strictly control and audit use of the data. The current best practice is a system of data sharing called trust networks.[21] A trust network is a combination of a computer network – to keep track of user permissions for each piece of personal data – with a legal contract that specifies the permissible uses of the data and the penalties if the permissions are violated. This is the model of personal data management that is most frequently proposed within the World Economic Forum Personal Data Initiative.[22]

In such a system, all personal data have attached labels specifying what the data can and cannot be used for. These labels are exactly matched by terms in a legal contract between all the participants, stating penalties for not obeying the permission labels and giving the right to audit the use of the data. Once all the permissions and the data provenances are on the system, data use can be automatically audited, and individuals can change their permissions and withdraw data.

Today, there are longstanding versions of trust networks that have proved to be both secure and robust. The best-known example is the SWIFT network, reliably handling trillions of dollars per day for inter-bank money transfer. Its most distinguishing feature is that it has never been hacked. Until recently, such systems were only for the "big guys." To give individuals a similarly safe method of managing personal data, researchers have built open-source software systems such as openmhealth and openPDS (open Personal Data Store), and are now testing these systems with a variety of industry and government partners.[23]

Government Secret Data. This third category typically includes tax data, detailed census data, detailed expenditures, and social health factors. The advent of Big Data health systems may dramatically expand the depth and breadth of these secret government data to include all types of individual behavior data.

A major risk of deploying data-driven policies and regulations stems from the danger of putting so much personal data in the hands of governments. But why might governments choose to limit the data they keep? The answer is that governments themselves, not just the citizenry, can suffer when data about citizens' behavior is inappropriately accessed. Consider the NSA's response to the recent Edward Snowden leaks in the US:

"This failure originated from two practices that we need to reverse," Ashton B. Carter, the deputy secretary of defense, said recently. "There was an enormous amount of information concentrated in one place," he said. "That's a mistake." And second, no individual should be given the kind of access Mr Snowden had, Mr Carter said.
www.nytimes.com/2013/08/04/sunday-review/a-washington-riddle-what-is-top-secret.html?_r=0

Therefore the government must organize Big Data resources in a distributed way, with each different type of data separated and dispersed among many locations, using many different types of computer systems and encryption. Similarly, human resources should be organized into cells of access and permission that are localized both spatially and by data type. To prevent overly powerful central actors, computer and human resources need to be redundant and distributed.

The logic is this: when databases are physically and logically distributed and also have heterogeneous computer and encryption systems, they are hard to attack, both physically and through cyber means. That is because any single raid would likely gain access to only a limited part of the whole database. Similarly with organizations having a heterogeneous cell-like human and permissions structure: that is how intelligence and terrorist organizations maintain their resilience.[24]

The computer architecture for this type of system is similar to the trust networks described in the previous section: distributed data stores with permissions, provenance, and auditing for sharing between data stores. The architecture is very similar to the citizen-centric personal data stores envisioned by most advocates of electronic medical records (EMRs), so adopting this architecture enables easier and safer sharing of data between citizens' EMRs and government. For this reason, several US states and EU countries are beginning to test this architecture for both internal and external data analysis services.

RECOMMENDATIONS

A medical and health science based on the use of Big Data is emerging. This new science leverages the capacity to collect and analyze data with a breadth and depth that was previously inconceivable. The goal of the following recommendations is to accelerate the emergence of operational Big Data health systems.

1. **Encourage public-private partnerships.** This kind of collaboration can serve to underwrite costs and accelerate deployment.

   Big Data health systems require some investment in data handling infrastructure, but are not as intrinsically expensive as many civil systems are. On the other hand, they require a continuous partnership between the healthcare system and private companies, private individuals, and healthcare professionals, since all of those are required to obtain the necessary data. So the question arises: how might we conceive a "public-private partnership" investment framework for Big Data in healthcare? In the past, a standard approach has been to assign "basic systems" investment to the public sector and "applied systems" investment to the private sector. Financial entities such as special-purpose banks – to underwrite the required capital investment at low interest rates – should prove especially useful in promoting a Big Data health ecosystem.

2. **Ensure data access.** Update privacy and data-ownership policies to ensure that data are accessible by patients and their healthcare providers, and that trust networks provide safe data sharing.

Some of the thorniest challenges posed by new digital capabilities on Big Data revolve around data access and sharing. Robust models of collaboration and data sharing – between government, industry, and academia – need to be developed, but it is vital to safeguard both the privacy of consumers and the legitimate competitive interests of corporations. Combined computer and legal systems, such as trust networks, could be the answer here: they can allow safe, controlled, and auditable sharing between hospitals, company proprietary data stores and individuals' personal data stores (see www.idcubed.org). Recently, some vertically integrated health systems, such as the UK's National Health Service (NHS) and the US Veterans Administration, have begun experimenting with such systems with very promising results (see www.va.gov/bluebutton/).

3. **Allow for open data.** Pool unrestricted government data and non-proprietary private data in an open data commons in order to promote development of a Big Data health ecosystem.

   Today, data in healthcare – particularly data stemming from pharmaceutical and medical R&D, clinical settings, patient behavior, and payer activity are highly fragmented and not generally accessible by the health researchers or even patients themselves. The creation of broad, open data commons that support research is critical. Efforts such as "Patients Like Me" or www.openmhealth.org are successful examples of creating data commons by combining EMRs with contributions of personal behavior and measurement data. Unfortunately, such examples are the exception, and are not connected to most formal healthcare data systems. We recommend an international Charter for Open Data Sharing, to specify best practice and to commit nations to sharing health data for their mutual benefit.

   As part of an open data initiative, there must be discussion and public engagement about the idea of Big Data systems. For instance, the UK's NHS is about to launch a large information campaign to explain to patients the importance of using their primary care data for the advancement of science. Encouraging public and patient engagement is very important; it can be facilitated by cutting-edge visualizations of the Big Data analyses. The goal is to communicate effectively with the public about the issues of privacy and trust, and about the trade-offs between potential benefit and cost.

4. **Promote Big Data health science.** Create centers of excellence to train Big Data behavioral and health scientists in the use of open-source tools for data analysis.

   Centers of excellence could serve to train human analysts in the testing of interventions that use Big Data. To that end, the academic community needs to train more computational social scientists and develop Big Data experimental methodologies, such as living laboratories and rich open data repositories. In addition, if more efficient collaboration could be encouraged by the use of sophisticated visualization techniques and collaborative data analysis infrastructure, that would enable more individuals to come together, generate more comprehensive insights, and thereby solve complex interdisciplinary problems.

The availability of easy-to-use tools would greatly accelerate a Big Data health science. Just as mass-market computer-assisted design software revolutionized the engineering world decades ago, common analysis tools and data-sharing protocols could now bring about important advances. An open architecture would allow the ecosystem to evolve faster, more efficiently, and in a way that is more responsive to clinical and human need. We suggest development of a "best practice kit" that lowers the barrier to entry for interested countries.

5. **Accelerate Big Data health practices.** Support partnerships between physicians and Big Data behavioral scientists to create "living laboratories" that develop new Big Data health solutions.

   By incorporating medically-related datasets into the analyses of other computational and behavioral scientists, we can magnify the potential impact of each dataset. Significant potential health solutions are likely to be those developed in partnership with more inquisitive front-line physicians, who best understand the problems to be solved, who are interested in applying new approaches and piloting promising ideas, and who, most importantly, are committed to the iterative development of new solutions. After all, most technological success arises through a rapid, iterative process with motivated early adopters.

   It is very important to foster this combination of pioneering physicians and behavioral scientists with experimental platforms such as living labs that support rapid innovation; that is precisely the way to rapidly develop successful Big Data health practices. Medical schools should integrate Big Data analysis into their curricula, to enable the next generation of healthcare providers to integrate these advances into their medical careers. Such efforts, led by teams of computational social scientists and medical faculty, could also foster much-needed collaboration among the academic disciplines.

SUPPORTING RECOMMENDATIONS FOR NGOs

1. Create "Big Data for health" loan facilities.
2. Support a "best practice" kit for data access and sharing.
3. Support a Charter for International Data Sharing.
4. Support research in Big Data medical science.
5. Create a "best practice" kit to bootstrap national centers of excellence.

ACKNOWLEDGMENTS

BIG DATA AND HEALTH WORKING GROUP MEMBERS

- **Hans-Olov Adami**, Adjunct Professor and Former Chair of Epidemiology, Harvard School of Public Health; Professor Emeritus, Karolinska Institutet, Stockholm, Sweden
- **Dennis Ausiello**, Jackson Professor of Clinical Medicine, Harvard Medical School; Chief of Medicine, Massachusetts General Hospital
- **Francis Bajunirwe**, Lecturer in Epidemiology and Biostatistics, Mbarara University of Science and Technology, Uganda
- **Caroline Buckee**, Associate Professor, Harvard School of Public Health
- **Stephen Brobst**, Chief Technology Officer, Teradata
- **Jorge Chavarro**, Assistant Professor of Nutrition and Epidemiology, Harvard School of Public Health
- **Kenneth Cukier**, Data Editor, The Economist
- **Shona Dalal**, Epidemiologist, Centre for Disease Control and Prevention, Uganda
- **Ahmed Elmagarmid**, Executive Director of Qatar Computing Research Institute
- **Deborah Estrin**, Professor, Computer Science, Cornell Tech, NYC
- **Axel Heitmueller**, Director of Strategy and Commerce, Imperial College Health Partners
- **Michelle Holmes**, Associate Professor of Medicine, Harvard Schools of Public Health and Medicine
- **Tim Kelsey**, National Director for Patients and Information, NHS England
- **Vikram Kumar**, Founder, Dimagi
- **Jeremy Nicholson**, Head of Department of Surgery and Cancer, Imperial College London
- **Marina Njekelekam**, Executive Director, Muhimbili National Hospital, Tanzania
- **Beth Noveck**, Founder, The Governance Lab
- **Alex (Sandy) Pentland**, Director, Human Dynamics Laboratory, Massachusetts Institute of Technology
- **Todd G Reid**, Research Scientist, Harvard School of Public Health
- **Bright Simons**, Founder, Mpedigree
- **Guang-Zhong Yang**, Director and Co-founder for the Hamlyn Centre for Robotic Surgery, Imperial College London
- **Thomas Zeltner**, President, Science et Cité, Switzerland, former Secretary of Health, Switzerland
- **Jakob Zinsstag**, Deputy Director, Swiss Tropical and Public Health Institute

Our special thanks to Will Warburton and Sarah Henderson for their help and guidance in preparing this report.

APPENDIX

This Appendix includes a sampling of current uses of Big Data and examples of its promise in a variety of health outcomes. Projects similar to these could be used by governments, agencies, and organizations to help set health policies for their respective countries. The case studies are arranged by topic as follows:

I. Living Labs for Health
   CATCH Living Lab: Personal Health and Continuous Health Assessment
   PaCT Living Lab: Chronic Disease Surveillance
   mHealth Living Labs

II. Big Data and Longitudinal Follow-up
   Epidemiologic Cohort Studies

III. Population Health
   Punjab: Preventing the Spread of Infectious Disease

IV. Data Ownership
   Personal Health Data Ownership

V. Health Systems
   Dimagi: Health Worker Data Recording
   mPedigree: Medication Purchases and Diagnostic Trends
   Ghana: Medical Claims
   Qatar: Electronic Health Records
   NHS England: Care.data
   Costs: Big Data and Healthcare Costs

I. LIVING LABS FOR HEALTH

CATCH LIVING LAB: PERSONAL HEALTH AND CONTINUOUS HEALTH ASSESSMENT

For hundreds of years, medicine has been episodic and symptomatic. When a patient has a headache or backache, for example, a narrow snapshot of that condition is captured, some action is taken, and the patient is then thrown out into a "black hole." Even with the sickest of patients, the monitoring systems and assessments are relatively trivial. CATCH (Center for Assessment Technology of Continuous Health) aims to move medicine into pre-symptomatic, continuous assessment that is both minimally invasive and minimally intrusive.

CATCH projects combine passive and active analytics with very sophisticated molecular and genetics assessment. One current CATCH pilot study aims to combine passive behavioral and physiologic sensors to improve patient-centered management of type 2 diabetes (T2D). Several aspects of T2D may benefit from the analysis of non-traditional data. For instance, behavior modification (notably, diet and exercise) has been shown to play an important role in T2D. Co-morbid conditions (such as depression, sleep disturbances, and cardiovascular disease)

affect individual's quality of life as well as T2D outcomes. Another study will integrate measurements, from scientific "omic" measurements to more holistic measures of individual health and behavior. Individuals will contribute periodic blood samples for analysis of genotype, and gene expression analysis of immune cell activity. The stool microbiome will be analyzed, using next-generation sequencing, to establish which species of bacteria are prevalent and which metabolic pathways are active across the bacterial community. Patient-level assessments will include symptom diaries, SMS questionnaires to assess mood, and passive behavioral and activity measurements obtained through a smartphone app. These seemingly disparate data types will be analyzed in concert.

From the perspective of basic science, the patient-level assessments will better identify the earliest onset of symptoms consistent with a disease flare, to focus analysis of the "omic" data. From the perspective of the patients (and their physician), the patient-level measurements can quantify their daily symptoms in detail, and provide a novel type of data that can help them better manage their own health. These studies utilize several classes of potentially sensitive data in addition to traditional patient health information, including questionnaires on symptoms, GPS location, and metadata on smartphone communications. Indeed, the techniques used in CATCH for continuous assessments of health have the potential to provide rich and powerful data tools that help shift some of the management performed in hospitals and clinics to the hands of patients for more personalized medicine.

An academic partnership between Massachusetts General Hospital (Boston, US) and the Massachusetts Institute of Technology, CATCH includes several private partnerships, including Siemens, Pfizer, Merck, and a variety of Venture Capital firms as well as sensor and device manufacturers.

For more information, please see: www.catch-health.org

PACT LIVING LAB: CHRONIC DISEASE SURVEILLANCE

Africa/Harvard School of Public Health Partnership for Cohort Research and Training (Africa/HSPH PaCT)

Longitudinal studies of chronic disease follow people over time and collect information about their lifestyle, diets, physical activity, medication usage, smoking habits, reproductive health, and so on. Well-known examples of this type of study include The Framingham Heart Study and the Nurses' Health Study. Typically, the data are correlated with disease outcomes; such epidemiologic studies have been among the primary tools (at least in North America and Europe) for setting policy on preventive health. For example, in Boston, New York, and many other parts of the US, certain types of fat have been banned in restaurants, and labels are required on food packaging containing these fats, because study results show that they increase risk for heart disease.

PaCT, currently in the pilot phase, is based in five sites in four African countries (Nigeria, South Africa, Tanzania, and Uganda), and eventually plans to enroll 500,000 participants. Africa is the last place on earth to have such studies, but it greatly

needs them as it is now experiencing an epidemic of chronic disease that will overtake infectious disease. In collaboration with MIT's Media Lab, PaCT accordingly plans to collect much data via cell phones – eminently feasible in Africa, which is the world's fastest-growing cell phone market. This approach would allow for a continuous flow of data, in contrast to the annual data influxes that are the current standard. The data available through PaCT's research will help ministries of health in African nations set their own policies on chronic disease.

For more information, see: www.pactafrica.org

mHEALTH LIVING LABS

Open mHealth

For mHealth to most effectively improve health for each of us, integrated solutions are necessary. Unfortunately, integration is difficult, because most mHealth solutions are being built in silos, without the technical means or a market ecosystem for data sharing. Open mHealth is building an open architecture to break down barriers to integration, by enabling the following features:

1. **Adaptive and integrated solutions.** "Building blocks" of software can be flexibly selected and combined through a shared set of open APIs to build more effective applications. Through the same APIs, fully built applications and devices can also "talk" to each other, and can be integrated to suit a particular individual or health need, and thus achieve greater commercial success.
2. **Meaningful insights.** Data-processing and visualization modules make it possible for separate data streams to be integrated and exchanged, allowing the relationship between them to be more accurately understood, and enabling patients and clinicians alike to gain meaningful insights.
3. **Evaluation.** Research modules embedded directly in an application provide an effective way to test scientifically the impact of a particular treatment, or of the application itself. Clinicians and developers can use "application analytics" to find out what is working for whom, and to rigorously demonstrate and continuously improve health outcomes.

Every mHealth solution – whether proprietary or open-source, public or private – can benefit by adopting the Open mHealth architecture, which will enable integration with the wider health IT ecosystem, thus resulting in more connected solutions. Open approaches like this are particularly indicated for solutions aimed at under-served communities, where sharing, re-use, and joint learning are vital for achieving sustained impact.

Open mHealth is funded by the Robert Wood Johnson Foundation, and proud to be partners with organizations including the VA, Ginger.io, Kaiser Permanente, Intel, Microsoft, Runkeeper and more.

For more information, see: www.openmhealth.org

mHealth Greenhouse

The mHealth Greenhouse, part of Cornell Tech's activities in Healthier Life research and innovation, is being launched in New York. The mission is to catalyze development and use of clinically informed and patient-centric mobile technologies to improve personal health and clinical treatment.

Although standalone applications are available to help users monitor exercise, diet, symptoms and side-effects, there are few clinically-informed, integrative applications for patients to self-manage and inform their care. The mHealth Greenhouse will help clinicians to create novel and effective personal treatment tools for patients using smartphones, mobile apps, and cloud-based data analytics. The mHealth Greenhouse will bring clinical expertise into the smartphone information loop.

The mHealth Greenhouse is working with clinicians and patients to conduct rapid, iterative, small-scale pilots focused on managing prevalent, chronic conditions, such as depression, inflammatory diseases, and chronic pain.

The development team is collaborating with clinicians to design, build, test, and perfect mobile health apps through rapid, small-scale pre-clinical pilots. The project will germinate new applications, processes, data-capture techniques, visualization, and analytics. The long-term outcome will be faster mHealth innovation, thanks to an enhanced toolkit of reusable modules; and robust, innovative proof-of-concept solutions that can be used by others to obtain funding for large-scale clinical trials and/or further commercial development.

For more information, see: http://smalldata.tech.cornell.edu and www.tedmed.com/talks/show?id=17762&videoId=224255&ref=about-this-talk

II. BIG DATA AND LONGITUDINAL FOLLOW-UP

EPIDEMIOLOGIC COHORT STUDIES

Nurses' Health Study

The Nurses' Health Studies (NHS1, NHS2, and NHS3) are three studies aimed at identifying nutritional, biological, and lifestyle risk factors for chronic diseases among women. The original study, NHS1, started following 121,700 Registered Nurses (RNs) in 1976. NHS2 began following 116,430 RNs in 1989. In these two studies, participants are followed every two years with mailed questionnaires (and more recently with web-based questionnaires in NHS2), where women update information on a wide variety of lifestyle factors, including detailed assessments of their diet every four years. They also report any new diseases they have suffered. In addition, participants have contributed a very large number of biological specimens that can be related to the questionnaire-collected data, including disease outcomes. Specifically, 68,213 women provided toenail clippings in 1982, 25,264 women provided blood and urine samples in 1989, and 18,649 provided a second set of

blood and urine samples in 1999. In addition, 33,744 women who had not provided blood samples provided cheek swabs in 2002-2004. In NHS2, 29,269 women provided blood and urine samples, and 29,859 provided cheek swabs. In addition, paraffin-embedded tumor blocks are recovered from participants developing certain cancers (such as breast cancer and colon cancer). This bio-repository has been used, among other things, to perform multiple genome-wide association studies.

Building on the success of both NHS1 and NHS2, and with the goal of obtaining a better characterization of how adolescence and early adulthood may contribute to chronic diseases, NHS3 was launched in 2010, with the aim of enrolling 100,000 women. By the fall of 2013, 37,000 had been enrolled. Unlike its predecessors, NHS3 is completely web-based. Participants receive questionnaires every six months. Web-based questionnaires allow more detailed characterization of specific areas among subgroups of women, without increasing the burden on other participants. Pregnancies and planned pregnancy attempts are closely followed, allowing the assessment of critical exposures at specific times before conception or during pregnancy. NHS3 is currently piloting bio-specimen collections, and using smartphone-enabled questionnaires to facilitate the collection of GPS, accelerometer and other smartphone-generated data.

For more information, see: www.channing.harvard.edu/nhs and www.nhs3.org

The Danish and Norwegian National Birth Cohorts

The Danish National Birth Cohort is a study aimed at identifying the risk factors for pregnancy complications, adverse peri-natal outcomes, and the developmental origins of chronic disease. The study enrolled 101,042 pregnancies from 91,827 women between 1996 and 2002; approximately one-third of all the pregnancies in Denmark during that period. Women were informed about the study by primary care providers at the first pre-natal visit, and were subsequently followed during pregnancy. The study collected blood samples at gestation week (GW) 6-12 from about 98,000 women, a second blood sample at GW 24 from about 77,000 women, a detailed assessment of diet from about 70,000 women at GW 25, and cord blood samples at birth. Children are actively followed after birth. The age 7 years follow-up had an active take-up of 52 percent, and the age 11-13-years active follow-up is currently underway. Multiple sub-studies focusing on particular populations are also being conducted. For example, women who experienced gestational diabetes and their offspring are currently participating in a sub-study where detailed metabolic data is collected from both mother and offspring.

The Norwegian Mother and Child Cohort (MoBa Study) adopted a similar design, and enrolled about 110,000 pregnancies from about 90,000 women between 1999 and 2008.

A major advantage of these cohort studies over similar studies is that both studies can rely on passive follow-up of participants via national registries that capture health and other data. So even if all participants were lost to active follow-up, the cohorts could still identify risk factors for virtually any disease for many decades to come. In addition, by combining data from both cohorts, the researchers will be able to identify risk factors for rare diseases in a level of detail that has never been possible before in this type of epidemiologic cohort study.

For more information, see: www.niehs.nih.gov/research/atniehs/labs/epi/studies/moba

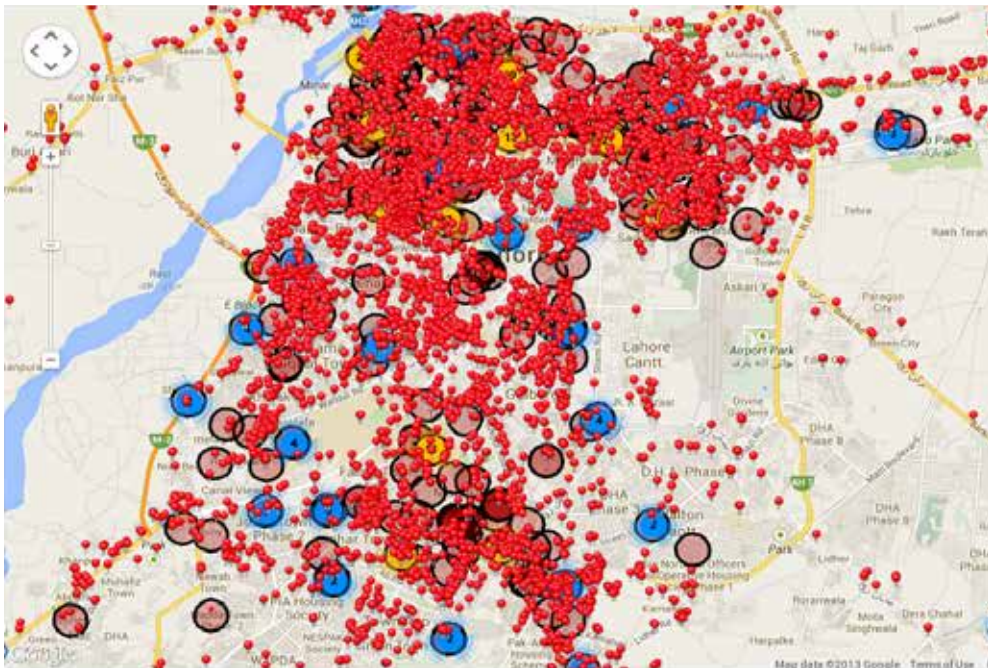
III. POPULATION HEALTH

PUNJAB: PREVENTING THE SPREAD OF INFECTIOUS DISEASE

Use of Big Data to Prevent Dengue Fever in Pakistan

In recent years, dengue fever has been a serious health risk for the citizens of Lahore, Pakistan. Despite efforts to curb the disease, infectious mosquitoes had become a huge problem. In 2011, the city experienced its worst outbreak of dengue fever in history. But 2012 and 2103 have been very different, thanks to the use of Big Data analytics and smartphone technology. In a collaborative project between the government of the Punjab province of Pakistan and the IT University of Punjab, a solution was developed to help curb future epidemics of the disease, by using a modification of software originally created and posted into an open-source repository by the CDC in the US. The software for early detection of dengue fever epidemics in Pakistan was based on software designed to detect outbreaks of flu epidemics in the US.

Enhancements to the application enabled investigators to identify the residential locations associated with infected patients. By analyzing the location of the dengue fever cases, it was possible to identify high-risk areas for infection, and then to aggressively eliminate breeding grounds for the mosquito larvae. Government employees were given smartphones to take photos and geo-code the location of standing water where mosquitoes breed. The breeding grounds were eliminated by a combination of two techniques: removal of standing water when possible, and the use of tilapia fish, which eat the mosquito larvae in larger bodies of water. The smartphone photos, before and after treatment of the breeding grounds, along with geo-coding, allow tracking of progress in the battle against the mosquito population. The application is easy to use even for people who are not technically savvy, and provides much-needed data in the fight against dengue fever.



Disease surveillance. This map of Lahore shows dengue patients (circles) and dengue-carrying larvae (red pins).

For more information about the program, please see:

www.technologyreview.com/news/506276/pakistan-uses-smartphone-data-to-head-off-dengue-outbreak/ and

www.npr.org/blogs/health/2013/09/16/223051694/how-smartphones-became-vital-tool-against-dengue-in-pakistan

IV. DATA OWNERSHIP

PERSONAL HEALTH DATA OWNERSHIP

Linked National Cooperatives as Citizen-Owned and Citizen-Controlled Health Data Repositories

The Challenge: A dysfunctional and unsustainable data model.

Health data are stored and controlled by physicians, clinics, hospitals, labs, pharmacies, insurance companies, and government agencies in innumerable, incompatible data silos. And although citizens (individuals with medical needs and healthy individuals) legally own their health data, they lack access to and control over such data. This data model – dysfunctional and unsustainable – substantially increases the cost and reduces the quality and effectiveness of healthcare globally. Furthermore, the move towards personalized healthcare requires large complex datasets from millions of people. These cannot be obtained without the active participation of citizens across the world.

The Solution: Cooperatives that empower citizens by giving them control over their data

The cooperative is an old and successful form of corporation that is entirely owned by citizens. The DNA of the cooperative is: we do it ourselves, on our own terms; we are self-supporting rather than dependent on government or pure capital investors. In Switzerland and in other countries, successful retail stores (MIGROS, COOP) and banks (Raiffeisen) are cooperatives. Since all data in health are personal, and since each citizen (whether from Switzerland, parts of Africa, or elsewhere) has about the same amount of potential health data (ie, the same genome size), the cooperative's traditional one-member-one-vote principle is particularly suitable for a health data repository. Members who join the cooperative will safely store and manage all their health data (medical, mHealth, genome, and so on) in their account. In this way, they will have access to all their data from anywhere in the world, and they can share subsets of the data, or all the data, with doctors, friends, or biomedical research. Since citizens are the owners of their data, their informed decision to share their own data for research is not subjected to the same data protection regulations as when third parties request to access personal health data. Aggregated personal health data have a high economic value. Pharmaceutical companies and other research institutions will pay significant sums to study the data that users have consented to share. The revenues will stay in the cooperative, and members will decide whether to invest the gains into research, information, platforms or continued education. By being in control of all their data and deciding how capital gains will be used for the common good, the members of the cooperative (the citizens) will enjoy true citizen empowerment, and the large potential of aggregated health data will be unleashed for personalized prevention and treatment.

Great benefits will come from setting up a federation of national health data cooperatives that share the same IT infrastructure and central data storage as Switzerland: the result will be a true democratization of the global health data space, and a welfare boost for the national communities of citizens.

For more information, see www.datenundgesundheit.ch

V. HEALTH SYSTEMS

DIMAGI: HEALTH WORKER DATA RECORDING

Over the past decade, Dimagi has been working on systems for cell data collection by frontline health workers (FLWs) in rural parts of the world. Its core platform, CommCare, is being used by 50 organizations in over 30 countries. More than 2.5 million patient forms have been submitted through its system. Dimagi uses the depth of the data to make some important inferences.

One such inference is whether FLWs are submitting "real" or "fake" data. Dimagi's cofounder, Dr Vikram Kumar, along with Dr Neal Lesh and his collaborators at the University of Washington, conducted a study using CommCare in Tanzania and Uganda to investigate this question. At one of the study sites, the FLWs were supervised in the

data collection of real patients, and were later asked to enter data on "fake" patients at a "fake data party." Unlabeled data from fieldwork of these FLWs, and from FLWs at the second site, were obtained. The team built classifiers to automatically detect "fake" from "real" data. The depth of the CommCare dataset facilitated the detection of the fake data with a high sensitivity and specificity.

What gave away the fakers? When an FLW entered fake data, he or she spent on average 148 seconds, compared to the 240 seconds on validated real data entry. Moreover, it turns out that the fakers over-estimated the incidence of diarrhea in the households they visited. In the fake data, 35 percent of households reported someone suffering with diarrhea, while in the real data, only 5 percent of the households contained someone who had diarrhea. Data systems like this could also serve as an important assessment tool in determining the value of FLW views of patients' medical needs, since physicians often have too little time to record information.

For more information, see: www.dimagi.com

mPEDIGREE: MEDICATION PURCHASES AND DIAGNOSTIC TRENDS

mPedigree is a system commercially deployed or piloted in eight African countries. It enables consumers to validate the certification and authentication status of a pack of medicine through a simple text message. mPedigree has been working on two auxiliary platforms: Acodion and Ovasight, both of which use the text messages to discern broader patterns, such as changes in doctor prescription preferences. These patterns can be of great help to epidemiological programs that track the changing face of disease in Africa. The data can also be helpful in evaluating the shifts in the economic burden of disease, by looking at variations in drug-purchasing behavior over time. And thanks to the two-way communication opened up between the patient and the datacenter, it now is possible to harness even more complex information, which can be used for modeling diagnostic trends in specific geographies.

For more information, see: www.mpedigree.net

GHANA: MEDICAL CLAIMS

Ghana Health Insurance System

Ghana has a national health insurance program managed by the government. All citizens are entitled to access, although premiums differ across broad categories determined by income. The government pays healthcare providers in both the public and private sectors for the services and procedures that they carry out on behalf of subscribers. The growth of the program has, however, been significantly hampered by claims fraud, in which some providers submit reimbursement claims that are overpriced or that relate to services and procedures that were never actually carried out. To limit such fraudulent claims, the government of Ghana has embarked on a number of data-analysis projects to track suspicious-looking transactions and claims patterns. To design better and broader responses, however, a more integrated approach will be needed. It should combine data from multiple sources –

dispensaries, consulting rooms, medical depots, clinical records, diagnostic records, and so on – to data mine for insights and devolve use of processing capacity to the frontlines.

For more information, see: www.nhis.gov.gh

QATAR: ELECTRONIC HEALTH RECORDS

Qatari Hospitals

All the major hospitals in Qatar have adopted the same Electronic Health Record (EHR) solution, which enables easy consolidation of health data for a large part of the population. Along with the patients' EHRs, other health data such as hospital readmission records, medication prescriptions and online social data are also available. Such rich health data can be used to enhance the healthcare system in Qatar in several ways: by providing personalized health risk profiles and care plans, by improving communication with patients in order to facilitate behavioral changes, by enabling hospitals and healthcare providers to deliver better services, and by assisting government in planning for further improvements in healthcare provision in Qatar.

Personalized health risk profiles and care plans

One objective of using Big Data in healthcare is to make actionable data available to healthcare providers, who can then decide the optimal care option for the patient. That means developing individual risk profiles (determining the probability of various diseases) by collating the patient's EHR with the data collected from other sources –social media or smartphone applications, for example – and filling the gaps through data-mining rules and analysis of doctor's notes. What emerges is a personalized health plan, including suggested preventive measures.

Improved communication with patients in order to facilitate behavioral changes

Once the personalized health plan has been created, alerts can be sent to patients to ensure adherence; that is, to remind them about medications or hospital visits, especially in the case of chronic diseases. The alerts would typically be made by cell phone, and could involve an animated character or avatar explaining health conditions, disease, test reports, discharge summaries, or precautions to be taken, and so on – all in the patient's own language. Thus no privacy or language issues should arise, and the patient can engage with the app freely, asking simple questions or repeating the question without inhibitions.

Improved service from hospitals and other providers

Healthcare providers can use Big Data to develop a better understanding of the triggers of hospital readmission. Such an understanding would enable them to intervene proactively and thereby reduce readmissions. It would also help them to analyze the clinical effectiveness of various treatments, and to identify hospital-acquired conditions. Big Data will also enhance the management of facilities and inventory, such as drugs and medical equipment.

Optimized government planning for healthcare provision in Qatar
Heat maps could be created for city planners (within the various municipalities) or Public Health authorities on actual Geographic Information System (GIS) maps or Google Maps. The mapping would record factors such as environmental pollution, disease occurrence data, demographic data and availability of medical services in the vicinity, and would thereby enable the authorities to take more informed decisions about public health.

For more information, see: www.cerner.com/about_cerner/newsroom/hamad_medical_corporation_signs_agreement

NHS ENGLAND: CARE.DATA

Care.data is a new initiative aimed at dramatically increasing the availability and value of clinical data across all health and care services in England. The information, extracted securely from all clinical systems, includes details of patient demographics, symptoms and investigations, diagnoses and treatments. It is linked together at an individual-level to give a complete picture of a person's care experiences. It is made available in a range of formats, including aggregated data (published online), pseudonymized data (made available to approved analysts under strict regulation), and data from the NHS (made available to patients to be used as they wish).

This is the first time that clinicians and researchers have had access to data across the pathway of care for all public services. The initiative will improve the measurement of comparative clinical outcomes, transform data resources for the life sciences, and enable NHS England to allocate its resources more effectively. In addition, the program is designed to support innovation in the development of digital services for patients and the public.

COSTS: BIG DATA AND HEALTHCARE COSTS

The following four case studies demonstrate how Big Data is being used to help control healthcare costs.

UK
Mastodon C, a Big Data analytics company affiliated with the Open Data Institute in London, has analyzed drug-prescribing patterns in the UK in a way that could significantly reduce national healthcare costs. The company has specialized in using cloud computing to analyze Big Data more efficiently and with a lower carbon footprint than other, similar companies. Using Big Data from the UK's NHS, Mastodon C has conducted a government-funded analysis of variations in prescribing patterns across the UK. The analysis shows that in some areas of the country, expensive drugs are being prescribed for medical problems where generics would work just as well. The variation appears to be due to regional differences in doctors' prescribing patterns, rather than to medical necessity. This analysis suggests that the NHS could save more than £1 billion per year (US$ 1.6 billion) by changing the way that some doctors prescribe medication.

US

The nonprofit company FAIRHealth was founded in 2009 with funding from the New York State Attorney General's Office, following a settlement with the insurance industry over the way companies set reimbursement policies. FAIRHealth collects data from hundreds of private payers and others involved in handling health claims, with all identifiable patient information removed. In return for contributing data, payers get discounts on license fees. As of 2012, FAIRHealth had data on 13 billion different cases from across the US, all coded by zip code. The result is a website that gives consumers unique tools to predict their out-of-pocket costs, choose among health plans, negotiate with providers, and decide whether or not to seek services outside the healthcare network. The database has also been a valuable resource for researchers and industry analysts trying to reduce national healthcare expenditures in the US.

For more information, see: www.fairhealth.org

MedStar Washington Hospital Center in Washington, D.C.

Working with Microsoft Research and using Microsoft's Amalga software, this hospital center analyzed several years of its anonymized medical records – patient demographics, tests, diagnoses, treatments, and more – for ways to reduce readmission rates and infections, which are among the costliest parts of healthcare. The technique uncovered some surprises. Consider the list it created of all conditions that increased the chances that a discharged patient would return within a month. Although some of these correlations are well-known and have no easy solution, the system spotted an unexpected top predictor: the patient's mental state. The probability of readmission increased markedly if the initial complaint contained words that suggested mental distress, such as "depression."

The Health Care Cost Institute

Created in 2012 by a handful of America's largest health insurers, this nonprofit institute has combined data amounting to five billion (anonymized) claims, involving 33 million people. By sharing their records, firms are able to spot trends that might have been indiscernible in their smaller individual datasets. Among the first findings was that US medical costs had increased three times faster than inflation in 2009-10, though with pronounced differences at a granular level: emergency-room prices grew by 11 percent, while nursing facilities' prices actually declined. Health insurers would never have handed over their prized data to anything but a non-profit intermediary: a non-profit's motives are less suspect, and the organization can be designed with transparency and accountability in mind.

REFERENCES

1. The D4D data were donated by the mobile carrier Orange, and the research initiative was organized with help from the University of Louvain (Belgium) and the MIT Human Dynamics Lab (US), along with collaboration from Bouake University (Ivory Coast), the UN's Global Pulse, the World Economic Forum, and the GSMA (which is the mobile carriers' international trade association). The D4D program was led by Nicolas De Cordes (Orange), Vincent Blondel (Louvain), Alex Pentland (MIT), Robert Kirkpatrick (UN Global Pulse), and Bill Hoffman (World Economic Forum). See www.d4d.orange.com/home for more details.
2. Pentland A, Lazer D, Brewer D, Heibeck T. Using reality mining to improve public health and medicine. Studies in Health Technology and Informatics. 2009; 149: 93-102.
3. www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health _care
4. www2.technologyreview.com/article/409598/tr10-reality-mining/
5. See QuantifiedSelf, CogitoHealth.com, Ginger.io, GoodlifeMe.com, CommCare, and others.
6. PatientsLikeMe, HealthVault, UN Global Pulse, and others.
7. en.wikipedia.org/wiki/Nurses'_Health_Study
8. Little M, Wicks P, Vaughan T, Pentland A. Quantifying short-term dynamics of Parkinson's disease using self-reported symptom data from an internet social network. *Journal of Medical Internet Research*. 2013; 15(1): e20.
9. Wesolowski A, Eagle N, Tatem A, et al. Quantifying the impact of human mobility on malaria. *Science*. 12 October 2012; 338(6104): 267-270.
10. Sadelik A, Brennan S, Kautz H, Silenzio V. nEmesis: which restaurants should you avoid today? *AAAI*. 2013. See www.cs.rochester.edu/~sadilek/publications/Sadilek-Brennan-Kautz-Silenzio_nEmesis_HCOMP-13.pdf
11. Madan A, Cebrian M, Moturu S, et al. Sensing the "health state" of a community. *Pervasive Computing*. 2012;11(4): 36-45.
12. Dong W, Heller KA, Pentland A. Modeling infection with multi-agent dynamics. In *Proceedings of Social Computing, Behavior-Cultural Modeling, and Prediction (SBP)*. 2012; 172-179.
13. Duan N, Kravitz RL, Schmid CH. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of Clinical Epidemiology*. 2013;66(8 Suppl):S21-8. doi: 10.1016/j.jclinepi.2013.04.006.
14. Pentland A. Honest signals. Cambridge, MA: MIT Press; 2008.
15. CogitoHealth.com
16. www.wilsoncenter.org/sites/default/files/participatory_sensing.pdf
17. Aharony N, Pan W, Ip C, et al. Social fMRI: investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*. 2011;7(6): 643-659.
18. Mani A, Rahwan I, Pentland A. Inducing peer pressure to promote cooperation. *Scientific Reports*. 2013; 3(1735). doi:10.1038/srep01735
19. See www.openhealth.org and http://smalldata.tech.cornell.edu
20. Ginger.io
21. Pentland A. Reality mining of mobile communications: towards a new deal on data. *In The Global Information Technology Report 2008-2009: Mobility in a Networked World*, eds. S. Dutta and I. Mia. Geneva: World Economic Forum. 75-80. See www.insead.edu/v1/gitr/wef/main/fullreport/files/Chap1/1.6.pdf

22. WEF 2011, Personal data: The emergence of a new asset class. See www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf

23. www.idcubed.org and www.openmhealth.org

24. This architecture can help prevent the use of Big Data from trampling on individual freedoms. The key insight is that for these types of data systems, each type of data analysis operation has a characteristic pattern of communication between different databases and human operators. In consequence, it is possible to monitor the functioning of the *data analysis process without access to, or endangerment of, the analysis content. In short, one can use "metadata about Big Data" in order to monitor the use of Big Data, and can with reasonable confidence ensure that only "normal" analysis operations are being conducted, without reference to specific content. Governments that structure their data resources in this way can more easily monitor attacks and misuse of all kinds.

NOTES

WISH ACADEMIC PARTNERS