11-2016

Big Data and Public Health: Challenges and Opportunities for Low and Middle Income Countries

Sundeeep Sahay

University of Oslo, sundeeps@ifi.uio.no

Follow this and additional works at: <http://aisel.aisnet.org/cais>

Recommended Citation

Sahay, Sundeeep (2016) "Big Data and Public Health: Challenges and Opportunities for Low and Middle Income Countries," *Communications of the Association for Information Systems*: Vol. 39 , Article 20.

DOI: 10.17705/1CAIS.03920

Available at: <http://aisel.aisnet.org/cais/vol39/iss1/20>

This material is brought to you by the Journals at AIS Electronic Library (AISeL). It has been accepted for inclusion in Communications of the Association for Information Systems by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Big Data and Public Health: Challenges and Opportunities for Low and Middle Income Countries

Sundeep Sahay

Department of Informatics, Faculty of Mathematics and Natural Sciences

University of Oslo, Norway.

sundeeps@ifi.uio.no

Abstract:

Big data is creating the potential for a revolution in many domains, including public health. However, big data represents only one of the ingredients to materialize this revolution, and we need to surmount many other socio-technical elements required to enable it and other various challenges. From the perspective of ministries of health in low and middle income countries (LMICs), I examine some of the drivers of this data revolution and some of the challenges and opportunities big data provides for strengthening health systems. Based on this analysis, I draw some key implications for LMICs governments on how they could seek to harness the potential of big data to address public health concerns they are currently engaged with or expect to do so in the future.

Keywords: Big Data, Data Revolution, Global Health, Low and Middle Income Countries.

This manuscript underwent editorial review. It was received 03/17/2016 and was with the authors for 1 month for 2 revisions. Devinder Thapa served as Associate Editor.

1 Introduction: The Problem of “Big Data” for Public Health

A data revolution in health signifies a radical and exciting shift in data generation, collection, and use practices from what currently exist. A key driver of this revolution is arguably “big data”, which represents large volumes of machine-readable data and tools for visualizing and interpreting it (Pentland, Reid, & Heibeck, 2013). Further, big data is data whose volume, velocity, and variety makes it difficult for an organization to manage, analyze and extract value from it using conventional methods and systems (Grossmand & Vogel, 2014). Some have described big data as the “new oil” to fuel innovation in the new economy (Beardsley et al., 2014). This data revolution has not fully arrived or static but is continuously in the making and evolving. Zuboff (2014) makes a “digital declaration”:

When it comes to “big data” and the digital future, we are at the very beginning. Despite the rapid pace of connection and the oceans of data it generates, our societies have yet to determine how all this will be used, to what purpose, and who decides.

Zuboff (2014) highlights the point that big data is in its infancy and that our agency has an important role in shaping the promised revolution. Most discussions on big data focus on the “supply side” and emphasize the (promised) availability of large volumes of data in different domains, such as retail, logistics, credit ratings, banking, social media, health, and various others (Chen, Chiang, & Storey, 2012). We have seen more limited discussions on skills, capacities, structures that represent the “demand side, to leverage this promise, although some authors have raised privacy and security concerns (Guarda, 2009). In health, most discussions on data, driven by fiscal compulsions, concern the context of clinical settings in the West, and applications for public health, particularly in low- and middle-income countries (LMICs), remain limited. Researchers have largely presented only anecdotal and speculative evidence for and against the potential of big data.

As public health challenges in LMICs arising from epidemics (e.g., Ebola, Zika), natural disasters (e.g., earthquakes, storms), and humanitarian crises (e.g., migration, conflict, security) continue to exponentially rise, LMICs are grappling with new and innovative ways of identifying and dealing with them. Information and supporting ICTs will be a key resource, and big data undoubtedly comes with huge potential to aid in this quest. But we know little about big data’s nature (especially when contextualized from the perspective of LMICs), about its challenges and opportunities, and about what kind of specific applications big data can be useful. Thus, we urgently need to go beyond the hype of big data and concretely examine how, why, and when one can use it for strengthening public health systems in LMICs. I examine some of these questions in this paper.

This paper proceeds as follows. I discuss the nature of big data as it is generally understood in Section 2 and from the context of a LMIC perspective in Section 3. I conclude Section 3 with a framework to understand the nature of big data initiatives. In Section 4, I provide some concrete examples of some state- and non-state-driven big data applications in public health, and, in Section 5, discuss some other applications that are more futuristic in nature. In Section 6, I discuss some of the challenges and opportunities that come with big data, especially for public health in LMICs. I also discuss some specific implications for ministries of health in LMICs.

2 Understanding Big Data’s Nature

Simply speaking, big data refers to large volumes of computer-readable data. Popularly, it has the following characteristics (known as the six Vs) (Normandeau, 2013):

1. Volume: the data comes in large amounts
2. Variety: the data (structured and unstructured) have different sources
3. Velocity: the data has a real-time and continuous nature
4. Veracity: the data can be triangulated from multiple sources
5. Validity: the data reflects primary sources of collection, and
6. Volatility: the data is available over time.

An important, seventh V concerns data’s value (Global Pulse, 2012). Data’s primary beneficiary is the global IT industry: a McKinsey report (Manyika, Michael, Richard, Peter, & Alex, 2013) has predicted that big data, fueled by “disruptive” technologies (e.g., mobile Internet, Internet of things, the cloud, advanced robotics, genomics, and social media, will significantly increase by 2025. Big data’s value comes in two

types. Big data's "use" value represents how it helps to address some local problems and use conditions. Big data's "exchange" value represents its reusable intellectual capital and how one can apply data in a similar way in multiple contexts to generate value. Large corporations (those dealing with health insurance, clinical trials, and data brokerage) have typically leveraged big data's exchange value, which may increase in the future. As such, many LMICs face the challenge of learning how they can start to generate and leverage big data's exchange value.

Each of big data's characteristics provides challenges for how one gain value from big data. Big data's enhanced volume and variety correspond with increased biases, noise, and abnormalities, which raise questions about the data's validity. One needs to eliminate inherently "dirty data" to gain value from it. While one can access and store historical data for future periods, one also faces questions about how long the data remains valid, who owns it, and what rights citizens have to the data. Epidemiologists conducting research in LMICs may find little use for big data because of the limited nature of available patient based data, which reflects how big data has challenges and opportunities for different user groups (Matjaz, 2014).

To understand this question of what value does big data provide and how one can create and leverage it, we need to expand our perspective on big data to go beyond the data itself to include the whole network of people, institutions, and things that go into its production, exploitation, and use. More than two decades of information systems research has explained that "airplanes don't fly, airlines do", which implies technology or data on its own can do little, and, to be effective, one needs to align it with social-institutional heterogeneous networks in a historical context. Consequently, the notion of "information infrastructures" is useful to emphasize the inter-connected and evolving nature of relationships between technology, data, people, things, practices, and institutions (Braa & Sahay, 2012). One can find it useful to think of big data in terms of "data-enabled knowledge infrastructures" that emphasize that data on its own, small or big, can achieve little if not taken in context of its history, future possibilities, spatial attributes, underlying technologies, use, exchange value, and the technologies and expertise (human and institutional) required to interpret, visualize, analyze, and take action on it. Big data on its own does not represent a revolution but can serve as an important and necessary condition for it. For example, a LMIC's health system would need to leverage big data's potential and develop other ingredients (e.g., capacity, the use of cloud and other infrastructure, and knowledge) in conjunction to create a data-enabled knowledge infrastructure for public health to cultivate sufficient conditions to make improvements in health outcomes—a non-trivial challenge.

3 Contextualizing Big Data from the Perspective of Public Health in LMICs

Public health deals with populations who live in villages and communities and for whom health services are provided beyond the confines of structured settings of hospitals and clinics (such as their homes and communities). Data on these services is, by definition, "big" due to numbers and scale but is relatively ambiguous and unstructured. Thus, public health data is fraught with contestations around data quality, with census data's being notorious to derive consensus on. LMICs have traditionally captured and processed such health services related data (reported typically as aggregate statistics) through national routine health management information systems (HMIS). Examples of such data are "number of babies born" or "numbers of BCG vaccinations given" or "number of maternal deaths". LMICs typically circumscribe these data in a period (typically a month or quarter) and provided for the catchment population a health facility is expected to serve. In addition to the routine HMIS, national governments use survey data coming from, for example, district-level household surveys, sample registration systems, household and demographic surveys, and various others. Governments often rely on survey data due to the lack of trust in the perceived quality of the routine HMIS even though many consider it as the most reliable source (Keen, Calinescu, Paige, & Rooksby, 2013). Survey data tends to be delayed and does not offer the level of granularity required for supporting local decision making.

HMIS typically deals with "big data". For example, more than 200,000 public health facilities in India each report on about 2000 data elements monthly, and the Tajikistan Ministry of Health collected about 30,000 data elements quarterly until recently. In today's parlance of "big data", one may not see this many elements as "big" (i.e., because the data only has volume and not big data's other characteristics). While HMIS data has a lot of variety (e.g., service, infrastructure, financial, and other data), it has traditionally been narrowly focused to reproductive- and child health-related services. The health system also includes

individual reporting systems of multiple vertical health programs such as HIV, Malaria, and TB—data from which are not easily visible in the national HMIS. The velocity of this data tends to be “slow” because it is typically a snapshot and retrospective rather than real-time data. The veracity of the data remains suspect because limited opportunities for both technical and institutional reasons exist to triangulate the “truth value” of this data with other sources. As a result, it also becomes difficult to establish the validity of this data. Given that many countries still rely on manual recording and reporting systems, they have a limited ability to deal with big data’s volatility that arises from the its historicity.

Many LMICs in the last decade or so have undertaken efforts to reform their national HMIS by computerizing them. Reform efforts have also been driven by the efforts of international donors who have tried to strengthen accountability of their aid money through improved information systems. With the wide and increasing proliferation of cloud infrastructure, Web-based technologies, mobile devices, and other technologies, LMICs are rapidly starting to create data that is becoming “big” in nature. Institutional processes of introducing universal health care and insurance systems are slowly making it mandatory for countries to be producing and using big data.

Having contextualized big data in LMICs, I describe a framework to help better understand the nature of big data in Section 3.1.

In summary, big data in LMICs will vary based on different dimensions, which includes the application area one uses it, data type and sources, technology platforms, and others. In Table 1, I present a diagnostic framework to help understand some of these differences and unpack the nature of big data (which has implications for its challenges and opportunities they provide).

Table 1. Various Examples of Crowd Sourcing in Health

Application domain	Type of data	Data sources	Time frame of data	Technology platform	Data integration required	Ownership/partnerships
Data repository/ specific analytics	Aggregate/ patient based	Primary/ secondary	Historical/ current	Custom made/public domain	Ongoing/batch mode	Public/private/ community owned
Single/multiple purpose	Quantitative/ qualitative	Real time/ periodic	Project/non- project based	Internet/non- Internet based	Generic/custom made tools	Single/multiple owners
Prior/post-hoc defined application	Structured/ unstructured	Single/multiple sources	Short/long/very long term	Open source /proprietary Single vendor/open sourced/crowd- sourced	Integration with legacy data	

4 Some Examples of Big Data Initiatives in Health

A wide ranging nature of state and non-state based efforts are ongoing in the domain of big data for health. These efforts range in scale (national to small projects), types (population based surveillance to person focused records), technologies used (mobiles, sensors, social media, genomics), geographies covered (country specific or a regional set of countries), and various other dimensions. I now present some examples in Sections 4.1 to 4.2.

4.1 Some State-based Initiatives

I discuss two national level initiatives from contrasting contexts. The first initiative is the U.S. Government’s efforts towards building a national infrastructure of electronic health records (EHRs) and health information exchange (HIE). The second initiative is the Rwandan Ministry of Health’s initiative to set up a national scale Rwanda health information exchange (RHIE). I discuss the example from the US even though it isn’t a developing country for two key reasons: 1) many LMICs such as Rwanda draw inspiration, design, technical, and financial support from the US to guide their own efforts and 2) the U.S. example emphasizes the challenges that one faces even in well-resourced contexts with a strong

regulatory environment and political commitment. LMICs without adequate resources and/or a poor regulatory environment and/or political commitment will likely face even greater challenges and require different approaches as the Rwanda case emphasizes.

4.1.1 U.S. Efforts towards a National Health Information Network (NHIN)

President George Bush signed the Executive Order 13335 to create the Office of the National Coordinator for Health Information Technology (ONC) and the NHIN. Further, he called for the nation to use interoperable EHRs. The ONC came up with a report in 2005 that advocated establishing a decentralized architecture built on interoperable standards using open source software and guided by a proactive policy framework that provided incentives for use. The ONC also advocated for stronger mechanisms for data privacy and security. Furthermore, the ONC required decentralized networks of networks and all federal health information organizations (HIOs) to adhere to the national interoperability standards through the Executive Order 13410 in August 2006. In 2005, the ONC contracted agencies to develop prototypes of the NHIN to integrate and interface with EHRs, personal health information, health information organizations, and other entities engaged with public health management and research. In 2008, the ONC announced the connecting of 20 federal agencies in the NHIN, where individual participants exchanged live data. In 2009, the Federal Health Architecture Group (FHA) released code for open source software for a NHIN gateway (called CONNECT) to promote and incentivize the use of a NHIN as a common resource to be used by all FHAs.

On January 17, 2014, President Obama in a speech at the Justice Department reiterated the importance of big data technologies in influencing various socio-economic, political, and institutional developments. The Obama Administration initiated efforts such as My Data, Open Data, and Data.gov where people could access public data in which personal data was secured. The administration modified existing laws (such as the Health Insurance Portability and Accountability Act (HIPAA) of 1996 and Safe Harbor Act) to accommodate the changing needs of security and privacy enabled through this new type of data and infrastructure.

The Department of Health and Human Services (DHHS) (healthit.hhs.gov) (Blumenthal, 2010) made two legal announcements in 2010. The first announcement described how hospitals, physicians, and other healthcare professionals can qualify for billions of dollars of extra Medicare and Medicaid payments by “meaningfully using” EHRs. The second announcement described standards and certification criteria of EHRs for users to collect payments. The Health Information Technology for Economic and Clinical Health (HITECH) Act, part of the American Recovery and Reinvestment Act of 2009 (also known as the Stimulus Bill), authorized these actions. The Obama Administration structured the HITECH Act to include four sets of interventions: define meaningful use, encourage and support attainment of meaningful use through incentives and grant programs, bolster public trust in EHRs by ensuring their privacy and security, and foster continued health information technology innovation. It defines these interventions over three stages: stage 1 focuses on collecting critical data elements in electronic form, sharing key information with other providers and patients, and reporting quality measures to government; stages 2 and 3 emphasize rewarding providers for using EHRs to improve care processes and outcomes. Gale, Hartley, and Croll (2014) summarize what entails “meaningful use” across the three stages (see Table 2).

The HITECH Act provides various measures to address obstacles to this evolution in meaningful use across the three stages. For example, eligible healthcare professionals could earn up to US\$44,000 in extra payments between 2011 and 2015 if they became meaningful EHR users. Meaningful hospitals users could collect an initial bonus and extra payments per discharge of a Medicare patient. Further, the U.S. Government committed US\$650 million under the HITECH Act to create a network of up to 70 regional health information technology extension centers to provide advice to healthcare providers for procuring EHR systems and becoming meaningful EHR users. The NHIN, which enabled jurisdictions to share data among themselves, further attracted US\$560 million under the HITECH Act to different state governments. Furthermore, the HITECH Act strengthened existing privacy protections under the HIPAA.

While the US has largely focused on establishing a NHIN to promote the generation and use of big data from the top down, it is relevant to understand the practical benefits achieved. A survey conducted in 2014 (Gale et al., 2014) in about 225 rural health clinics (RHCs) on the adoption of EHRs concluded that most clinics were unlikely to adopt an EHR or would struggle with their implementation. Approximately 25 percent of responding clinics had not adopted an EHR and about 17 percent had no plans to do so or were unsure of their plans. RHCs experienced significant challenges in accessing technical support from the extension centers set up under the HITECH Act provisions. The challenges extended beyond

information and technical support and included issues of acquisition and maintenance costs, lack of capital for small healthcare providers, and concerns about loss of productivity and income during the inception phase. Small clinics experienced fear of loss on their economic performance and their ability to access technical assistance. While the implementation process was at the end of stage 2, most RHCs were primarily in stage 1.

Table 2. Meaningful Use of Data as the HITECH Act Defines It

Stage 1: 2011-2012 Data capture and sharing criteria	Stage 2: 2014 Advanced clinical processes criteria	Stage 3: 2016 Improved health outcomes criteria
Electronically capturing health information in standardized format	More rigorous health information exchange (HIE)	Improved quality, safety, and efficiency leading to improved health outcomes
Using that information to track key clinical conditions	Increased requirements for e-prescribing and incorporating lab results	Decision support for national high-priority conditions
Communicating that information for care coordination processes	Electronic transmission of patient care summaries across multiple settings	Patient access to self-management tools
Initiating the reporting of clinical quality measures and public health information	More patient-controlled data	Access to comprehensive patient data through patient-centered HIE
Using information to engage patients and their families in care	Improved population health	

The HITECH Act contributed to the development of a single-vendor and monoculture with respect to the supporting software (Koppel & Lehman, 2014). Many healthcare systems and practices in US have shifted to Epic's EHR (EpicCentre Inpatient Clinical Systems) based on other software first developed at the Massachusetts General Hospital in 1968. Today, EPIC captures more than 50 percent of the United States' new large hospital contracts, which cover 51 percent of the U.S. population. This monopolistic situation establishes and proliferates standards around formats, data, and user interfaces, which improves data interoperability, promotes clinicians' geographical mobility, reduces training and maintenance efforts, and promotes standardization of patient care and meaningful use. However, it comes with high costs (ranging from US\$250 million to US\$1.1 billion to initially set up the application and about 50% more—in relation to the original set up cost—to upgrade new versions) and dependence on fixed-design software with little scope for customization. Additionally, some of the system adopters lost historical data because of migration problems and experienced magnified lock-in costs. The system was based primarily on market needs rather than requirements of clinicians and patients, and it stifled innovation because other competing products were not allowed to thrive. With this monopoly, EPIC influenced not just the IT systems but also the healthcare logic including decision support rules, order sets, quality measures, and care coordination with implications on patient safety and care.

In summary, the US experience reflects the advantages and disadvantages of a top-down and nationally driven system for EHRs and HIE expected to drive the generation and use of big data. While national ownership has sought to create a common infrastructure and standards along with a coherent policy, its implementation is fraught with various challenges about system uptake, limited use of the system, and monopolies.

4.1.2 The Rwanda Health Information Exchange (RHIE)

The RHIE is a Rwandan Government effort to promote a large-scale infrastructure that can potentially drive the generation and use of big data for health. The RHIE seeks to collect and aggregate health data consistently and promote patients, providers, and organizational decision makers to reuse it. The RHIE represents one of the world's first effort towards a HIE in a LMIC context supported by global entities such as PEPFAR, CDC, and the Regienstrief Institute, USA. The RHIE project started in March 2012 in one district (covering 14 health centers and one referral hospital) and focused only on the Maternal Health program, a priority area for Rwanda's Ministry of Health.

An evaluation of the RHIE carried out in 2014 first examined the efficacy of the public-private partnership that comprised multiple globally dispersed stakeholders and covered skills in IT, medicine, global health,

and project management. Governance challenges were seen to have been underestimated, and while various providers had bilateral agreements between them, the initiative lacked an overall vision and coordination mechanisms. Further, various other factors hampered this coordination, such as partners' being located in different time zones and challenges of Internet connectivity and reliable power supply in the clinics. Discussions often took place along disciplinary lines with the medicine or IT people talking to each other, which limited the building of an overall vision. The initiative did not factor in adequate time to address these governance challenges.

Discussions during project design, the project planners analyzed procuring "off-the-shelf" software compared to developing from scratch, which also involved addressing ideological differences between open source versus proprietary software. The project plan developed emphasized an open architecture design, which helped pave the way to select an open source platform to develop the architecture from scratch. Although the RHIE focused on providing access to national-level data, users had to access networked desktop computers running OpenMRS software. There were concerns of activities' being uncoordinated and unplanned as local network infrastructure and computers took a long time to establish even as the software was ready. It took nearly a year to get the network cables in place in the clinics, and some partners felt the mobile and 3G networks could have been more effectively exploited. The infrastructure adopted was "complex but not elegant" and impeded scaling. The agile approach of starting small and building incrementally ran contrary to the governance contracts established, which specified payment on finished work.

In terms of actual use, while most respondents of the evaluation were satisfied with the system, they felt they spent lesser time than before with their patients. Infrastructure bottlenecks (Internet and electricity) impeded them from effectively using the system. While IT staff, not surprisingly, liked the software because it helped them to enhance their computer skills, they were largely unaware of the information-exchange capabilities the system provided (Catalani, 2014). Further, the evaluation showed that, in the data about at-risk pregnant women between the HMIS and HIE, more referrals were reported through the HMIS, which was also the case for registrations of ANCs. Overall, the study concluded that no significant health impact emerged from the initiative.

In summary, LMICs face significant challenges in establishing an appropriate infrastructure and governance mechanisms to support something as complex as a HIE. Clinicians still preferred to use paper even though IT staff saw value in using the system. The RHIE has had limited to no impact on health outcomes. . While scalability was a desired end of the RHIE, after about three years and significant resources invested, the architecture has not expanded beyond the 14 clinics and one health program it started with. On the contrary, as of 2016, only eight clinics were operational with the system.

The RHIE experience highlights valuable lessons in the challenges of generating and harnessing big data for health in LMICs. Challenges include building appropriate infrastructure and adopting a frugal and incremental design. There are significant challenges in enabling use of data, and in managing the transition from a paper based to a complex electronic one. Clinicians still prefer to use paper, and are not so impressed by the potential the infrastructure provides to share data through the information exchange.

4.1.3 Some Other State-based Initiatives

India has developed a national health portal to integrate all health data into a national health repository. A high-level expert group (HLEG) created to design initiatives to scale universal health coverage made recommendations for developing a national health-information technology network based on uniform standards to ensure interoperability. To overcome interoperability and standardization issues, a committee on meta-data and data standards (MDDS) was established to propose data definitions in the public health system. The Indian Government has also poised to establish a "center of excellence" in big data under the Department of Science and Technology (2014). Similarly, premiere research institutes such as the Indian Institute of Sciences have started big data initiatives. The Ministry of Electronics and Information Technology has focused on devising a policy on IT use that will incorporate e-commerce, enterprise architecture, and the Internet of things. The Indian Government also sees the creation of a national knowledge network and national innovation foundation as forward-looking initiatives toward developing a big data culture. The Centre of Scientific and Industrial Research (CSIR) has created a traditional knowledge digital library (TKDL) from which one can access, analyze, and interpret digital data. Many of these efforts still represent largely policy statements, are fragmented, and focus primarily on establishing of a technological infrastructure than on strengthening IT use.

Kenya was the first country in East Africa to launch (with World Bank funding) an open data portal featuring information from government census and economic, social, health, and education data. Users can now potentially access information about healthcare facilities, doctors, and other medical information on their mobile by using the MedAfrica application. Scientists use these systems to track hotspots for malaria transmission in the country and design a strategy for control of the disease. A study by Global Alliance for Vaccine Initiative (GAVI) reported that the creation of digitized databases in Kenya has resulted in the better tracking of patients, improved health service delivery, and better monitoring of hospital admissions. The same study noted the database's potential to provide useful evidence-based decision support for administering vaccines and monitoring pneumococcal diseases (Germann, Jabry, Njogu, & Osumba, 2012). An ongoing challenge concerns establishing institutional responsibilities for continually updating the data and enabling its use. This challenge is not just limited to Kenya and remains widespread across LMICs.

Bangladesh, in recent years have made significant improvements in their national HMIS, which innovative uses of ICTs have powered. For example, recently, Bangladesh has provided tablets to thousands of community health workers to collect name-based and longitudinal data on pregnant women and children for immunization from villages. The government is synchronizing with the national-level database at the central Ministry of Health. The data being collected is indeed "big" and lays down the potential for its use for strengthening health care. These processes of strengthening use are currently underway in the country.

In analyzing these various state-driven big data initiatives, the U.S. one exemplifies establishing adequate IT infrastructure supported by significant policy and fiscal measures to support a national-level architecture to support the generation and use of big data (The White House, 2014). However, not all healthcare providers have adopted EMR systems and the practices for enabling meaningful use. However, ongoing monopolistic tendencies around the software can potentially endanger patient safety and care. Historically, LMICs have dealt with aggregate data systems, and some are now attempting (as of 2016) to transition to EHRs and HIEs, which involves collecting name- and event-based data. This transition in the kind of data being collected will potentially contribute to this data revolution in terms of data's volume and velocity. However, strengthening data use is a far more complex endeavor than just the data collection as even the experience of the resource-rich US illustrates. Achieving meaningful use of data in LMICs will be far more complex, which the RHIE case illustrates despite significant investments and political support. Countries with fewer resources and less political support face far more issues in generating and using big data.

4.2 Some Non-state Project-based Initiatives

Various civil society organizations, corporations, NGOs, university research departments, and others drive various non-state project-based big data initiatives target specific problems.

For example, Google Flu Trends, a Web service operated by Google since 2008 that provides estimates of influenza activity for more than 25 countries, represents an archetypical example for using big data in health. By aggregating Google search queries, it attempts to make accurate predictions about flu activity. The application monitors millions of users' health tracking behaviors online and analyzes the search queries to reveal if there is the presence of flu-like illness in a population (identified by IP address of each search). Google Flu Trends compares these findings to a historic baseline level of influenza activity (based on 50 million queries entered weekly in the USA between 2003-2008) for its corresponding region and then reports the activity level as either minimal, low, moderate, high, or intense. Some researchers have reported these estimates to be generally consistent (both nationally and regionally) with conventional surveillance data collected by health agencies (Ginsberg et al., 2008). The World Health Organization uses a similar text-mining methodology to analyze and predict disease outbreaks based on the media reports with its Global Public Health Information Network. Arguably, such analyses can help national governments in identifying early warning signals and disease outbreaks.

The Indianapolis Emergency Medical services (IEMS) has used big data analytics tool to track influenza-like patterns for controlling and managing the Ebola outbreak in real time by visually displaying influenza-like symptoms (Indianapolis EMS News, 2014). The application uses mobile mapping to track Ebola trends in Western Africa where other reliable data sources are lacking (Wall, 2014). Based on historic data collected by mobiles, one could track people's movement and allocate resources and treatment centers at places that needed it most.

The Public Health Foundation of India (PHFI) has developed an initiative called KGB (Kooda, Gandgi, & Badboo) to collect, collate, and interpret pictures of garbage, dirt, and bad smells posted in social media in real time in to track the location and forecast the occurrence of seasonal diseases and to report this information to concerned departments so they can take the necessary action. International organizations such as Global Pulse and Paris 21 under the aegis of the United Nations promote the use of big data analytics in developing countries by providing technical assistance, building capacity, and disseminating knowledge through conferences and workshops. IBM has announced plans to create the world's largest and most comprehensive data sets and data repository for cancer in Sub-Saharan Africa (where there is less than 1% cancer registration done currently) together with the Union for International Cancer Control (UICC) to help track and predict future disease patterns.

In Table 3, I summarize various health-related big data crowdsourcing initiatives.

Table 3. Various Examples of Crowdsourcing in Health

Article one Partners: patent engines for medical instruments, drugs.
PatientsLikeMe: enabling platform that exchanges health and disease information with 300000 members that provide real time health experiences.
Open Research Exchange: platform that allows researchers to develop, test, and conduct surveys within a community of 200000 plus members.
HealthMap: from Boston Children's Hospital for disease outbreak monitoring and surveillance of emerging public health threats.
Outbreaks Near Me: a mobile application for real-time data on infectious diseases
Prize4Life: from Harvard Business School to accelerate Amyotrophic Lateral Sclerosis (ALS) research in partnership with InnoCentive.
Project Data sphere: platform to share historical cancer clinical trial datasets. Cancer Cell Identification: supported by Cancer Research, U.K.
SNOMED CT: a fuzzy subset for semantic tagging of medical documents.
Indiegogo: A crowdsourcing initiative to support various entrepreneurial campaigns, including in health.

These examples highlight the multi-faceted examples of big data-based applications in health being pioneered by various crowdsourcing efforts typically driven by different kinds of organizations and focused on addressing a particular problem. These efforts focus on integrating and standardizing various disparate datasets, including repositories of patient medical records, clinical trials, and other information across various organizations for predicting trends, conducting research, and providing useful information. Brown, Chui, and Manyika (2011) identify three contributory factors to the rise of such big data initiatives: the growth of sensors and electronic gadgets being interconnected to computing resources, the availability of data in the public domain (especially social media), and the growth of suitable technologies (e.g., smart phones, mobile technology, the Web) and methods for data mining (e.g., data-analytics tools such as Hadoop and SAS). The success of these efforts will also result in increasing pressure on governments and businesses to provide more efficient, economic, and effective real-time data mining-based solutions to, for example, build public awareness on changing patterns of disease outbreaks. Big data analytics provides the basis for integrating knowledge, using evidence-based medicine, strengthening epidemiological foundations, and supporting translational research. In the resource-constrained environments of LMICs, low-cost and widely accessible technologies such as the mobile phone can potentially drive big data-based innovations.

However, these efforts come also with key challenges. The first concerns scale. Projects, by definition, have finite scope, timeframe, and budget. The proliferation and death of multiple mobile pilot projects, aptly termed the phenomenon of "pilotitis", highlights the challenges in reaching scale (Germann et al., 2012). A second challenge concerns integrating these project-based efforts with national health systems, which is often essential to scale and sustain them. However, due to institutional incompatibilities, issues of ownership, differing agendas, and variations in technological platforms, these project-based initiatives tend to not scale, which limits their broader relevance to the population. Governments need to develop institutional mechanisms in order to fully realize the benefits of these various non-state efforts.

5 Selected Applications for Big Data in Health

Big data has various drivers, including those related to technology (such as the growth of cloud hosting services), institutions (such as the growth of insurance-based care systems), and emerging application areas. In this section, I discuss some of the drivers that come from new application areas.

As policy makers, private sector, national governments, and donors become actively involved in pursuing big data initiatives, new areas of applications in public health also open up. Here, I discuss four such future areas: genome sequencing techniques, personal-based health records, universal health coverage, and global surveillance of diseases.

5.1 Application of Genome Sequencing Techniques

New-generation sequencing techniques in health have potential applications in the diagnosis of rare diseases, pharmacogenetics/treatment guidance, cancer diagnostics and treatment guidance, microbiology, and predictive and risk assessment. The transition from traditional methods (partial/targeted sequencing of particular genes) to full/whole-genome sequencing (FGS/WGS)—also called high throughput sequencing (HTS)—is driving several changes related to data, including big data's generation. Further, falling prices have made replacing several targeted tests with one “big” test financially feasible (e.g., today, each newborn undergoes between 30 and 40 targeted tests). Doing one big (full-genome) test then drives the rationale of storing individuals' full genome because, when it has been sequenced once, it makes sense to keep it and go back and query it for whatever reason later (since one's genome doesn't change throughout life). New technologies (such as HTS) are driving a “data turn” and comes with an associated need for computational storage using high-performance computing solutions and highly parallelized processing. Further, the availability of full genome data makes other data-analysis approaches possible and also a shift from hypothesis-driven targeted testing for health interventions (e.g., interactions between gene mutations (testing specific relations)) to data-driven studies, which look for correlations between a variant and a disease in a large dataset as a part of a big-data strategy (such as genome wide association studies (GWAS)).

Henry T. Greely, Director for the Centre for Law and Biosciences at Stanford University, notes that FGS has moved from being a sport for the rich and famous to being applied for those who really need it; from where it becomes possible to sequence all rather than three to five genes to where it becomes a must in certain areas such as the screening of the new born and an integral part of the individual's electronic patient record. The value of data for research and knowledge building increases enormously if one couples genomic data with other data (clinical and/or biomedical data). However, the potential of applying these sequencing techniques raises several questions such as “Do we really want to know everything?” and “How do we know what to do?”, issues of data security and privacy, the legality of DNA registries, and what role insurance companies have.

Google X's moonshot project called “Baseline Study”, which Google intends to use to help one better understand the human body and its functioning (Amadeo, 2014) seeks to create a picture of a healthy human based on genomic sequencing techniques. The project collects data on bodily fluids such as urine, blood, saliva, semen, and tears and information on how well the subject metabolizes food, nutrients, and drugs. The project also performs FGS to monitor subjects' heart rate, family genetic history, and how chemical reactions change behavior of the genes. The project hopes replace some of the follow-up tests with wearable devices that continually collect data such as subjects' heart rates and oxygen levels. To alleviate privacy concerns, a third-party will collect the data, anonymize it, and then give it to Google. Once Google has a good baseline of what a healthy human looks like, it will compare that data to data from other individuals to discover potential problems before symptoms become obvious. With the Baseline Study, Google intends to build a database of its own and use its big data prowess to search for “biomarkers” or specific molecules that indicate something is amiss. Such analytics should help to identify, for example, a biomarker associated with the efficient breakdown of fatty foods. People identified who are lacking this biomarker could eventually suffer problems with fatty build ups in their arteries and, thus, treated early to could help them modify their behavior and avoid a potential health problem.

The value of genomic sequencing techniques for public health, especially for LMICs, remains questionable because it focuses primarily on individuals, especially from rich countries, and has limited use for larger populations in resource-constrained environments. The Google X quest of creating a healthy human currently seems utopian given the immense amount of variations that exist between

individuals. While genome sequencing techniques will indeed create big data, it will arguably be limited to the richer countries.

5.2 Personal-based Health Records

A personal-based health record (PBHR) is one where patients maintain health data and information related to their care. Many corporations have adopted initiatives to develop PBHRs, such as Dossia, Microsoft, and Google.

Dossia, a PBHR provider that began as a nonprofit consortium of large employers, introduced the Dossia Health Manager, a system that is linked to employer-sponsored health plans. Dossia promoted the slogan: “Everyone has a unique health story, Dossia empowers you to author yours”. It started its PBHR initiative in 2006 through efforts of a consortium that AT&T, Intel, Walmart, and some other large enterprises. It provided a voluntary open-source based platform (called Indivo) to store records of about two million employees and five million dependents/retirees from these corporations. Its long-term goal was to “provide a portable and secure lifelong record that will be available regardless of employer, insurance provider or plan” (Steinbrook, 2008, p. 1654).

Microsoft began its Microsoft Health Vault in 2007 to help people in the US “take control of [their] own health”. It provides additional features for search, accessing emergency health services, and a connection center to enable people to directly upload heart rate, blood pressure, and blood glucose information via compatible devices to the PBHR. Google Health was another high-profile PBHR initiative: it started in 2008 but shut down soon after. When it closed, the Google Health website told people who had entered their personal data that their data was lost forever.

The story of these different initiatives, starting about the same time, tell the same story of PBHRs’ unrealized potential and lack of collective impact. Various concerns around PBHRs such as of confidentiality, security, and technological constraints have greatly constrained their uptake (Pushpangadan, S. & Seckman 2015). Today, most people use emails to communicate with their doctors, schedule appointments, request prescriptions, and share test results. Typically, doctors do not like being told by patients how they should treat them. Further, PBHR data is not subject to HIPAA standards, which creates uncertainty on how clinicians can use this data. The value of such systems may potentially be more beneficial for patients with chronic diseases with episodic needs for care and treatment rather than the general population. Similar to GSAs, PBHRs focus on the individual with limited utility to population-based health interventions, especially in LMICs with weak regulations for such data. While many LMICs (e.g., India) have or are now considering PBHR initiatives, policy makers and designers will find it useful to understand the experiences in rich countries such as the US in this regard to ensure they do not repeat others’ mistakes.

5.3 Universal Health Coverage and its Needs for Data Support

When implemented, universal health coverage (UHC), positioned as a key global health reform agenda, will require information support that is significantly different and “big” as contrasted with traditionally existing HMIS in most LMICs. UHC needs to expand the basket of health services to also include common NCDs such as mental health, hypertension, obesity, diabetes, and other lifestyle-related diseases. UHC involves proactively caring for populations based on measuring population rates for diseases, costs of care, service use, and outcomes. UHC seeks to strengthen continuity of care, which requires a seamless referral and mutual follow-up system between primary and secondary/tertiary care facilities. Given its emphasis on providing financial risk protection to individuals, UHC needs to promote a budget allocation based on facilities’ needs and caseloads and to ensure care with minimum out-of-pocket expenditure.

For UHC to be effective and meet its lofty aims, it will require robust health information support, which is radically different from what exists today in most LMICs. Firstly, while traditional HMIS focus primarily on reproductive and child health services, a UHC health information system would need to maintain patient or family-based health records of the entire population and be able to track for each treatment encounter the details of the interventions, outcomes, and cost of care. The system would need to enable the state to measure the population rates for diseases, service use, service provision, quality, and health outcomes attributable to healthcare services. To ensure individuals receive continuous care, the information system would need to include secure information flows of patient-based referral information from primary to secondary care facilities and follow-up action as a part of the reverse flows. The UHC health information

system would need to enable the state to generate indicators that respond to needs-based case loads and the actual package of care provided for all facilities.

Taken together, the HIS to support UHC would need to firstly cater to collect and process data based on names of patients and their individual encounters with the health system (including costs of treatment outcomes). Further, the system should have the ability to aggregate to the population from individual levels, cater to information needs of multiple state and non-state financial risk coverage providers, and measure progress towards reducing poverty. By definition, UHC data would be “big” in terms of volume, variety, (clinical, financial, biomarkers), and continuity. The UHC would also combine a variety of datasets, such as census, household surveys, civil registration, and various others. As LMICs start to implement UHC in earnest, they will need to harness various big data techniques (such as data analytics for modelling longitudinally financial inclusion progress) to provide effective UHC.

5.4 Big Data to Support Global Disease Surveillance

I recently participated in the development of a proposal to globally monitor why children under five years' old die. This proposed 20-year project to be funded by the Gates Foundation seeks to establish first 10 and subsequently 50-60 sentinel sites in Sub-Saharan Africa and South East Asia. Each site will cover a population of 200,000 and record details of all under-5 deaths and its causes in that area on a continuous basis. Data on deaths will flow from multiple sources including routine HIS, hospital medical records, verbal and full autopsies, reference and local laboratories, civil registration systems, demographic and household surveys, and various others. This big data initiative will involve large volumes and varieties of data, including clinical, demographic and administrative. It will be real time and involve detailed data analytics for estimating and predicting causes of death using historical trends from comparable sites. Data would be stored and shared globally using the cloud infrastructure and managed by educational institutions where such expertise resides. This initiative is expected to contribute with expertise on designing interventions, such as for the discovery of new vaccines to fight against vaccine preventable deaths. This initiative will involve a multiplicity of institutional partnerships and cross-disciplinary expertise in epidemiology, data analytics, open software, cloud infrastructure, coupled with in-country institutions with expertise in research, policy, and public health issues.

In summary, these four application areas represent potential future drivers of big data, where the UHC and global surveillance cases arguably have much greater relevance and significance for LMICs as compared to genomic technologies and PBHRs. These drivers will undoubtedly be supported further with the development of new technologies, continued evolution of the internet and Internet of Things, significant enhancements of data analytic techniques, improved methodologies for development of software applications based on crowdsourcing and open source communities and various others. Growth of business opportunities will attract more firms in the domain providing further impetus to the growth of big data. But are LMICs ready for this, and what kinds of institutional, policy, legal and capacity related measures they need to take for effectively leveraging on the potential provided by big data? These are important questions for debate and action. We discuss some of these issues in Sections 5.5 and 5.6.

5.5 Challenges Emerging from Big Data

5.5.1 Privacy and Security Challenges

Various authors (e.g., Appari & Johnson, 2008) have identified privacy and security issues as the foremost challenges related to big data, particularly how to reconcile potential benefits to public with risks to individual privacy. Privacy and security represents multi-faceted challenges in which technical developments continue to outpace legal advances and require interventions at multiple levels. At the individual level, we need to determine how to take informed consent of people to enable data use for agreed-on areas. Given the scale of big data collection, it is logistically difficult to get consent of all people, prior to data collection. Further, a fundamental ambiguity surrounds personal data's definition and its associated rights. Some have expressed concerns that technical solutions to anonymize personal data, such as for de-identification, are inadequate and vulnerable to hacker break-ins. Big data's potential to link across databases also creates the vulnerability of an individual's being identified even with anonymized data. Once electronic storage stores personal data (potentially forever), individuals lose adequate protection of their right to be forgotten and deleted from the cloud. At the organization level, there are many unknowns around contracts made, for example, by ministries of health and cloud service providers. Many ministries are bound by regulations that health data should not leave their national boundaries and

are not adequately informed on how to deal with the technical and institutional issues concerning the cloud. Technically, data held in servers in U.S. jurisdiction are subject to U.S. laws, which implies that other countries' data held in the US may be vulnerable to scrutiny by U.S. authorities.

Many argue that these concerns of privacy and security have no ground in LMICs because they deal primarily with retrospective and aggregate data. Further, due to the historical lack of regulation and issues of illiteracy and poor awareness, these concerns do not feature prominently in public debates in LMICs. However, LMICs should not disregard these concerns. For example, the Indian Government collects name-based data on pregnant mothers and children for immunization (without those individuals' consent). The government stores this individual-level data in a national server and makes it available to private call centers for follow-up and data-verification purposes. If not properly regulated, such data could end up in the hands of commercial vendors (e.g., those dealing with baby products) who would derive commercial value from marketing to these individuals. Governments need to address these privacy and security problems while making plans for collecting individual-level data.

5.5.2 Ownership-related Challenges

Big data's ownership-related challenges are significant for various reasons, which includes the actors involved in big data and the various regions that collect, store, use, and reuse the data. Agencies often collect data for a particular purpose and send it to others, often for commercial purposes. For instance, the example above (about data on mothers and children for immunization) has serious ownership questions at stake. Firstly, does the ministry own the personal data to whom they provide health services? Further, does the ministry have the ownership right to give this data to a third party outsourcing agency? David Cameron, ex-British Prime Minister, has noted similar ownership concerns about health data by saying that: "Every NHS patient would henceforth be a research patient whose medical record could be opened up for research by private healthcare research firms" (Tene & Polenesky, 2013). Such a situation raises concerns of who pays and does the research and who benefits from its results. There are power asymmetries inherent between about who the data covers and who collects data and who derives value from it.

The above questions are especially relevant from the perspective of LMICs who are often the subjects of research and data collection. They would need to engage with questions similar to those experienced by countries such as India in the early days of software outsourcing. These questions related to whether or not the countries providing the labor had a fair share in the intellectual capital they were contributing to generate (Sahay, Nicholson, & Krishna, 2003). But there is a difference, of course, when we talk of public health data as contrasted with commercial software. A key challenge here would be to see how one could position big data and its supporting technologies and techniques as a "public good" to benefit the population, which effective policy and regulation would need to back.

5.5.3 Epistemological Dilemmas

These challenges relate to who and how knowledge from and about big data is accessed. In public health, traditional models for generating knowledge involve a hypothetic-deductive logic followed by epidemiologists that involves developing a priori hypothesis about a phenomenon (e.g., testing the correlation between smoking and cancer), establishing experiments or clinical trials based on scientifically defined criteria (of sample sizes, control groups, and statistical tests for hypothesis testing), and executing these experiments to develop a "truth" about the phenomenon and make statistical generalizations with accompanying confidence levels. Big data analytics in contrast emphasizes identifying correlations between different types of data often collected post priori for varying purposes. Without a guiding conceptual framework, finding such correlations using sophisticated data analytic techniques backed with strong computing power may sometimes be akin to trying to find a needle in the haystack or, even worse, a needle not in the haystack. Does this trend imply the "death" of theory-driven knowledge generation in favor of correlation-driven post hoc computer-supported analysis? An important concern focuses on who drives this knowledge generation: will epidemiologists (who have the domain-specific knowledge) or will the IT and big data specialists and corporations (such as Google) who have the IT and big data analytics powers drive it? We may also wonder what role insights and creativity in developing knowledge will have as compared to the more computational-based methods used in big data applications. There is also the real danger of "good but small data" being overpowered by "bad but big data" in this game of big numbers.

Epidemiologists base their science on patient-based data; in its absence (as is the case in most LMICs), the diseases of the poor may remain increasingly marginalized in favor of diseases of the rich on whom

one can more easily collect personal data. The Economist (2013) recently attributed the poor levels of medical knowledge being generated to scientists' focusing on trying to develop "new" knowledge as contrasted with doing replication and verification studies to validate the results of earlier studies. Replication studies are notoriously hard to publish, which goes against the incentive structure of current academia which rewards the "new". Big data initiatives run the danger of further reinforcing the trend towards attempting new discoveries as it becomes easier to fish for new types of correlations in datasets at the disposal of researchers.

5.5.4 Infrastructure-related Challenges

These concerns, especially how one designs, develops, and maintains health information infrastructures are significant in big data initiatives. The infrastructure required for implementing big data initiatives is necessarily complex because it involves numerous stakeholders, technological platforms, and data types. A key concern for designing such infrastructures relates to enabling effective participation mechanisms from patients and clinicians given the rising domination of IT and data specialists enabled through a market-based logic and the increasing reliance on market forces. Developing these infrastructures is resource intensive and involves global expertise and money, which creates vendor lock-ins that can endanger scalability and sustainability of systems, especially in LMICs. As the Rwanda case demonstrates, some countries face extreme governance challenges in forging sustainable partnerships that naturally involve geographical and disciplinary boundaries. Further, infrastructures need to be flexible to be able to evolve over time. However, flexibility and evolution also pose challenges of rapid obsolescence of technology and sunk costs, which LMICs cannot afford.

Some may say that big data and the cloud are inextricably intertwined. However, to understand the cloud, we need to unpack both its underlying business model and also its technical hosting model. Cloud service models vary from offering only the infrastructure, infrastructure, and software to those that offer infrastructure, software, and end-to-end service. While end-to-end service models may seem attractive to capacity-constrained health ministries in LMICs, they are extremely complex to institutionally manage.

5.5.5 Capacity Challenges

Capacity challenges have historically plagued healthcare organizations in LMICs. Most health ministries in LMICs have a limited technical capacity and an insufficient number of specialists to help reform their routine HMIS. Further, public sector organizations typically place restrictions on new recruits, which creates dependencies on external expertise and donor funds. Big data demands more capacity (both in type and scale) routine HMIS reform efforts do and requires a new breed of professionals who are IT savvy; understand data modelling, visualization, and analytics; and possess a sound understanding of the public health domain. LMICs currently lack such professionals. Further, educational institutions in LMICs have not been agile in incorporating many necessary skills (e.g., related to data analytics, visualization, public health domain knowledge, the cloud, etc.) in their curriculum. Technological developments far outpace the speed and capacities of universities to transform, which contributes to an ever-widening gap of demand and supply of skills.

In summary, effectively leverage big data's potential is fraught with major challenges that relate to technology, data, institutions and the law and that span the individual, institution, and the nation levels. Addressing these challenges is a non-trivial task, which will take years and decades rather than months to implement. Furthermore, building appropriate capacity is a moving target, as skill sets required will constantly be changing. However, addressing these challenges are necessary for LMICs to start engaging with now, at both policy and strategy levels, and also operationally.

After discussing challenges, some opportunities from big data for public health in LMICs are discussed.

5.6 Opportunities that Big Data Provides

5.6.1 Supporting Transitions in Models of Healthcare Delivery

Big data initiatives can provide the health system to provide more patient-focused and proactive care. In many LMICs, such as India, even though health is a state subject, care has been typically organized around vertical, national-level top-down disease-based models. While much success has been achieved against Polio, Malaria, TB, and HIV, there have been associated challenges of fragmentation and weak sharing of data across the health programs. UHC requires patient and disease specific data collected over

the lifetime of the patient. This data needs to be continual and in real time, and it needs to pertain to aspects in addition to healthcare, such as individuals' financial, social, and lifestyle characteristics. For example, a diabetic patient can potentially receive better care if his/her blood sugar levels are continuously monitored, integrated with the clinical record, made available to the medical practitioner, and combined with data on diet, weight, BMI, exercise regime and other parameters. By definition, big data enables the collection and processing of such different types of data on a continuous and long term basis from multiple sources, and provides for analytics of historical trends. Similar big data applications can be found for diseases such as TB, mental health and HIV/AIDS. Further, as there are improvements in software and techniques of data aggregation from patient based to population levels, public health interventions can potentially be based on more accurate baseline data.

Traditionally, state systems notwithstanding outreach care services, have largely catered to the health needs of those who come to the state facility, implying a reactive model of care. In such a system, many people and diseases, typically the most marginalized and those who need care most, remain invisible and, thus, do not receive the necessary care. For example, much of the population remains oblivious to their diabetic condition or tendency and reach the facility often in diseases' advanced state. Big data can enable a more proactive model of care that can lead to early diagnosis and provision of more cost effective treatment.

5.6.2 Addressing the Challenge of Scale

Scale refers to a system's ability to expand from the position it began from. This expansion can be across various dimensions including relating to geography, numbers, functionalities, maturities of use, vertically (across administrative levels) and also horizontally (across business units at a particular level, such as different health programs at the district and state). Big data potentially enables health information systems to reach larger scale and also can provide support for health system interventions to achieve scale. The failure of HISs to reach scale has been an endemic problem, contributing to premature closure of many systems. The cloud based infrastructure, Internet and social media, provides the technical possibility of collecting data from widely distributed sources, on an ongoing and real time basis, and of different granularities. Similarly, ensuing analysis can also be disseminated to a larger audience and in real time, and can also help in the provision of technical support. This can help surmounting to a certain extent capacity related limitations also at facility levels and support scaling. Technical enablers of scale will naturally need to be complemented with institutional scaling measures, and further supported through the building of regional and global partnerships.

Ability to access low cost and community based software solutions. Growth of open source communities and crowd sourcing efforts helps build big data initiatives implies that national governments can also access using cost-effective models. An example of such an initiative is the University of Oslo coordinated network around the open source application DHIS2 (District Health Information System) now in the process of being deployed (at different levels of maturity) in about 50 countries. The increasing maturing of DHIS2 means more countries can now access this product without having to reinvent the wheel, and have financial investments be directed primarily towards implementation and capacity strengthening rather than on building software from scratch. Given that health systems needs are continuously evolving requiring the supporting information systems to also forever grow, the ability to draw upon community resources can accelerate response time to these changes. The Oslo effort would not have been successful without the active support from global agencies such as WHO and Global Fund emphasizing the key role of cultivating global partnerships. The DHIS2 case provides useful pointers and implications on how big data initiatives may be similarly harnessed, evolved and maintained over time.

5.6.3 Enabling More Effective Global Surveillance of Diseases

The WHO and many national governments currently seek to strengthen systems for reporting mortality and morbidity. Successful networks such as INDEPTH and ALPHA have helped to strengthen such reporting systems by bringing together researchers from different countries to collect, analyze, and disseminate data on a longitudinal basis and conduct regional capacity-building programs. The ALPHA network started in 2005, links existing African HIV cohort studies, and runs training workshops to help researchers undertake comparative studies and meta-analyses on comparable data sets and impose common formats on data collection, storage, and analysis. Similarly, the INDEPTH network spans 20 countries in Africa, Asia, and Oceania and connects researchers through a global network of health and demographic surveillance system (HDSS). The network has contributed to the development of data

repositories, the dissemination of analyzed statistics, and the conduct of various capacity-building programs. While these networks already take advantage of big data-like technologies and techniques, current advancements can help to further strengthen these networks and also create new ones.

6 Conclusions: Some Recommendations for Policy in LMICs

In this section, I develop some recommendations for policy from the perspective of LMICs.

National ministries of health in LMICs serve as the nodal agency for monitoring diseases, dealing with health crises and emergencies, serving as planning advisory to states and districts, and monitoring issues ranging from clinical, public health, and biomedical research and how to strengthen data systems and fill data gaps through enabling innovations. I discuss some policy-related implications to help support this enabling function of big data.

The WHO has articulated a national health systems strengthening framework, which includes six building blocks, including health information systems. This framework could be expanded to include “big data technologies and analytics” as an additional block, which includes specific technology enablers, principles, standards, frameworks, and relevant capacities. As a policy statement, such an expansion would help express a government’s seriousness towards big data.

Ministries of health should seek to enable new forms of partnerships for strengthening big data initiatives. Various similar initiatives have occurred both in and external to the U.N. system and the post-2015 development agenda framework. Initiatives could include creating an inventory of big data efforts and the different actors involved. While public-private partnerships would be beneficial, ministries would need to build these efforts in an open and public framework to ensure larger benefits. Ministries would also need to identify priority areas that they can best address and how best to link data with practice such as forging relationships with physicians and data scientists.

6.1 Building Mutual Synergies between National HIS and Big Data Initiatives

Ministries could aim to develop strategies and guidelines to strengthen linkages between national HIS and big data initiatives. National HIS systems have large gaps, such as poor cause-specific mortality statistics, which big data initiatives can meaningfully help to strengthen. National policy makers need to identify and prioritize specific problem areas to ground big data efforts in concrete problem contexts, which would also help allay fears that big data seeks to strengthen and not replace national HIS.

6.2 Promote the Establishment of Health Data Science Expertise in National and Regional Centers

Ministries should seek to establish expertise in a framework of regional centers in big data science to help develop their countries’ capacity. This expertise will involve hybrid skills in ICTs, data analytics, systems integration, the interpretation and use of data, and regulation issues. Such centers of excellence can help develop and propagate state-of-the-art-knowledge through workshops, short courses, and partnerships with leading universities. Such centers can serve as focal points for the sharing of knowledge and best practices in the country.

6.3 Provide Guidelines for Regulatory Systems to Address Ethical and Legal Issues

Ministries need to provide leadership on establishing an enabling regulatory environment that will effectively address ethical, legal, and intellectual property-related issues at the individual, facility, and national levels. A guiding normative principle is the need to balance between data use for the larger good with individual privacy concerns. These frameworks need to address issues concerning consent, personal data’s definition, data anonymization, the right to be forgotten, and the law (e.g., jurisdiction and liability). Such a comprehensive regulatory environment will help to alleviate mistrust that people may have about big data and also protect their rights in situations of conflict and disagreements, and it would also require ministries to establish appropriate governance mechanisms to manage inter-sector and inter-stakeholder coordination issues, including building guidelines for public-private partnerships. This governance structure needs to oversee the intersections of three domains of data analytics, IT, and public health.

6.3.1 Promote Practices of “Frugal Innovation”

Establishing national-level IT infrastructure to promote the use of big data for improving a population's health is a resource-intensive and long-term endeavor. Considerations of cost and its optimal utilization is an important consideration in these efforts. “Frugal” innovation, which encourages doing more with less, should be a guiding design principle in formulating big data efforts. Furthermore, for these innovations to be successful, they need to consider not only technology but also institutional and social innovation, and these different domains intersect.

Acknowledgments

I thank Dr. Ties Boerma, WHO Geneva, for his valuable and insightful guidance on the development of this paper.

References

- Amadeo, R. (2014) Google X's "baseline study" applies big data techniques to healthcare. *Ars Technica*. Retrieved from <http://arstechnica.com/business/2014/07/google-xs-baseline-study-applies-big-data-techniques-to-healthcare>
- Appari, A., & Johnson, M.E. (2008). *Information security and privacy in healthcare: Current state of research*. Retrieved from <http://www.ists.dartmouth.edu/library/416.pdf>
- Beardsley, S., Enriquez, L., Grijpink, F., Sandoval, S., Spittaels, S., & Strandell-Jansson, M. (2014). *Business driven technology*. New York, NY: McGraw Hill.
- Blumenthal, D. (2010). Perspective. Launching HITECH. *New England Journal of Medicine*, 362, 382-285.
- Braa, J., & Sahay, S. (2012). *Integrated health information architecture—power to the users—design, development and use*. New Delhi: Matrix Publishers.
- Brown B., Chui, M., & Manyika, J. (2011). Are you ready for the era of big data? *McKinsey Global Institute*. Retrieved from <http://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/are-you-ready-for-the-era-of-big-data>
- Catalani, C. (2014). 5 Key Insights about users experience with RHIE: An analysis of qualitative evaluation data. Retrieved from <http://emr.moh.gov.rw/wiki/display/RHEA/Five+Key+Insights+into+RHIE+Users%27+Experiences>
- Chen, H., Chiang, R. H. L., & Storey, V.C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165-1188.
- Department of Science and Technology. (2014). *Call for proposals under big data initiatives (BDI)*. Retrieved from http://dst.gov.in/whats_new/whats_new14/bdi-adv-t-webpage.pdf
- Gale, J. A., Hartley, D., & Croll, Z. (2014). *Meaningful use of electronic health records in rural health clinics* (working paper 52). Maine, USA: Maine Rural Health Research Center.
- Germann, S., Jabry, A., Njogu, J., & Osumba R. (2012). The illness of "Pilotitis" in mHealth—early lessons from the KimMNChip partnerships in Kenya. *Global Health Forum*. Retrieved from <http://ghf.g2hp.net/2011/12/07/2154/#.VJARStKUeS8>
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature*.
- Global Pulse. (2012). *Big data for development: A primer*. Retrieved from http://www.unglobalpulse.org/sites/default/files/Primer%202013_FINAL%20FOR%20PRINT.pdf
- Grossman, R. E., & Vogel, K. P. (2014). Organizational models for big data and analytics. *Journal of Organizational Design*, 3(1), 20-25.
- Guarda, P. (2009). *Electronic health records: Privacy and security issues in a comparative perspective*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1528461
- Indianapolis EMS. (2014). Indianapolis EMS using big data to track potential Ebola upsurge in real time. Retrieved from <http://www.emsworld.com/news/12015836/indianapolis-ems-using-big-data-to-track-potential-ebola-upsurge-in-real-time>
- Keen, J., Calinescu, R., Paige, R., & Rooksby, J. (2013). Big data + politics = open data. *Policy and Internet*, 5(2), 228-243.
- Koppel, R., & Lehmann, C.U. (2014). Implications of an emerging EHR monoculture for hospitals and healthcare systems. *Journal of American Medical Information Association*, 22, 465–471.
- Manyika, J., Michael, C., Richard, D., Peter, B., & Alex, M. (2013). Disruptive technologies: Advances that will transform life, business and the global economy. *McKinsey & Company*. Retrieved from http://www.mckinsey.com/insights/business_technology/disruptive_technologies
- Matjaz, J. (2014). *Information revolution, from data to policy action in low-income countries: How can innovation help?* Retrieved from <http://www.paris21.org/sites/default/files/PARIS21-DiscussionPaper3-Innovation.pdf>

- Normandeau K. (2013). Beyond volume, variety and velocity is the issue of big data veracity. Retrieved from <http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/>
- Pentland A., Reid, T. G., & Heibeck, T. (2013). Big data and health: Revolutionizing medicine and public health. *WISH*. Retrieved from <http://www.wish-qatar.org/app/media/382>
- Pushpangadan, S., & Seckman, C. (2015). Consumer perspective on personal health records: A review of the literature. *Journal of Nursing Informatics*, 19(1).
- Sahay, S., Nicholson, B., & Krishna, S. (2003). *Global IT outsourcing: Software development across borders*. Cambridge, UK: Cambridge University Press.
- Steinbrook, R. (2008). Personally controlled online health data—the next big thing in medical care. *The New England Journal of Medicine*, 358(16), 1653-1656.
- Tene, O., & Polenetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property*, 11(5), 240-272.
- The Economist. (2013). *Trouble at the lab—how science goes wrong*. Retrieved from <http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble>
- The White House. (2014). *Big data: Seizing opportunities—preserving values*. Retrieved from http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf
- Wall, M. (2014). Ebola: Can big data analytics help contain its spread? Retrieved from <http://www.bbc.com/news/business-29617831>
- Zuboff, S. (2014). A digital declaration. Retrieved from <http://www.faz.net/aktuell/feuilleton/debatten/the-digital-debate/shoshan-zuboff-on-big-data-as-surveillance-capitalism-13152525.html>

About the Author

Sundeep Sahay is a Professor of Informatics at the Department of Informatics, University of Oslo, Norway, since 1999. His research interest is in the domain of design, development and implementation of open source-based health information systems in developing countries. He has worked in more than 20 countries in Asia and Africa and supervised about 25 doctoral studies.

Copyright © 2016 by the Association for Information Systems. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than the Association for Information Systems must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or fee. Request permission to publish from: AIS Administrative Office, P.O. Box 2712 Atlanta, GA, 30301-2712 Attn: Reprints or via e-mail from publications@aisnet.org.