May 2014

An Epidemiology of Big Data

John Mark Young
Syracuse University

ABSTRACT

Federal legislation designed to transform the U.S. healthcare system and the emergence of mobile technology are among the common drivers that have contributed to a data explosion, with industry analysts and stakeholders proclaiming this decade the big data decade in healthcare (Horowitz, 2012). But a precise definition of big data is hazy (Dumbill, 2013). Instead, the healthcare industry mainly relies on metaphors, buzzwords, and slogans that fail to provide information about big data's content, value, or purposes for existence (Burns, 2011). Bollier and Firestone (2010) even suggests "big data does not really exist in healthcare" (p. 29). While federal policymakers and other healthcare stakeholders struggle with the adoption of Meaningful Use Standards, International Classification of Diseases-10 (ICD-10), and electronic health record interoperability standards, big data in healthcare remains a widely misunderstood phenomenon. Borgman (2012) found by "studying how data are created, handled, and managed in multi-disciplinary collaborations, we can inform science policy and practice" (p. 12).

Through the narratives of nine leaders representing three key stakeholder classes in the healthcare ecosystem: government, providers and consumers, this phenomenological research study explored a fundamental question: *Within and across the narratives of three key healthcare stakeholder classes, what are the important categories of meaning or current themes about big data in healthcare?* This research is significant because it: (1) produces new thematic insights about the meaning of big data in healthcare through narrative inquiry; (2) offers an agile framework of big data that can be deployed across all industries; and, (3) makes a unique contribution to scholarly qualitative literature about the phenomena of big data in healthcare for future research on topics including the diffusion and spread of health information across networks, mixed methods studies about big data, standards development, and health policy.

AN EPIDEMIOLOGY OF BIG DATA

by

John Mark Young

B.S., University of Maryland, 2001
E.M.L., Georgetown University, 2007

Dissertation
Submitted in Partial Fulfilment of the requirements for the degree of
Doctor of Professional Studies in Information Management

Syracuse University
May 2014

ACKNOWLEDGEMENTS

First, I want to acknowledge my doctoral committee, Dr. Art Thomas, Dr. Michelle Kaarst-Brown, and Dr. Jian Qin, my advisor who guided me through this journey with an informed and fair disposition. I also want to acknowledge Dr. Andrew London and Dr. Derrick Cogburn for serving in instrumental roles on the dissertation committee. I also want to recognize my classmates in each of the professional doctoral cohorts, particularly those in cohorts three and four for unwavering encouragement and support – these are lifelong friendships.

Though they remain anonymous, this dissertation thesis would not have been possible without the nine study participants who were very generous with their time in spite of busy schedules and a federal government shutdown. Their insights and perspectives were inspirational.

I would like to acknowledge Rita Charon, M.D. (Columbia University) the foremost authority in Narrative Medicine who encouraged me to think deeply about the distinction between Narrative Medicine verses Healthcare Narratives and Dr. Rebecca Gardner (Upstate Medical Center), who generously provided me with timely scholarly works in medical humanities and who introduced me to other scholars in this important field of medicine.

I would like to acknowledge the leadership of the ECRI Institute (Washington, D.C.) and Brent James, M.D. (Intermountain Health) who liberally offered unparalleled insights into big data in healthcare; and, Dr. Paul Cleary (Yale University) whose insight on the importance of healthcare consumer information needs was timely.

Finally, I am blessed to have had parents, Odell and Susie Young and my siblings who early in my life planted the seeds of this journey. And to my wife, Erie Sampson: thank you for your extraordinary emotional support and patience while I pursued my educational dream.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I. INTRODUCTION

Big data is a phenomenon of data usage closely linked to the "Information Age" (Heudecker, 2013). The term is common to many industries, in which 15 of the U.S. economy's 17 sectors, companies with more than one thousand employees, store on average more data than is contained in the U.S. Library of Congress (Brown, Chui, & Manyika, 2011). With the advent of health information technology (HIT), namely electronic health records (EHRs), big data in healthcare has emerged as a "natural resource" that could potentially revolutionize how we deliver personalized medicine and improve the health of populations. Consider the following vignette which describes a vision of the future of health and healthcare, fueled by big data:

> *At the level of the healthcare consumer, "big data" facilitated health improvement by applying massive computational utilities and the profound knowledge of systems biology to rich data clouds around each person. The billions of bits in each cloud came from inexpensive microfluidic devices enabling nearly continuous testing of blood for circulating proteins with bio-monitoring devices that could interface with personal simulations to predict future wellbeing. By collecting a person's genetic code, zip code and everything in between, these systems offered the capacity to predict when people were likely to get a major disease and to die. Personal avatars (digital health coaches) helped people recognize and leverage the extent to which their health was shaped by social, psychological, and behavioral factors. Most cancers were effectively preempted and managed by 2030. Former Type I and II diabetics now faced happier and longer lives due to the ability to grow and transplant pancreatic islet cells from pluripotent stem cells. Healthier communities, more effective personal healthcare and more sophisticated self-care decreased the demand for physician services and hospital care. In the eyes of many, the revolutionary transformation in both health and healthcare in the decades leading to 2032 was inevitable given the rapid diffusion of knowledge to an engaged population with a deeply held aspiration to be healthy.*[1]

[1] Institute for Alternative Futures. Health and Health Care in 2032: Report from the RWJF Futures Symposium, June 20-21. Alexandria, VA: Institute for Alternative Futures; 2012. http://www.altfutures.org/pubs/RWJF/IAF-HealthandHealthCare2032.pdf

The aforementioned vignette is not merely a pipedream – it is a likely reality. But a major roadblock persists: the definition of big data is hazy (Dumbill, 2013) and remains a buzzword (Davenport, Barth, & Bean, 2012). While big data in healthcare fervently grows weekly by some unknown order of magnitude, the difficulties and realities of sharing, linking, visualizing, and using big data in healthcare are magnified.

Research Question

This study addressed an important research question:

Q1: *Within and across the narratives of three key healthcare stakeholder classes, what are the important categories of meaning or current themes about big data in healthcare?*

While not usually a focal source of data for public policymaking, my intuition lead me to believe that 'stories,' or narratives, from the perspective of those who live the experience would yield rich, in-depth descriptions of the big data phenomenon in healthcare. In large part, this study was inspired by the science of epidemiology which studies the origin, patterns, and spread of an epidemic. Eysenbach (2002) coined the research discipline, infodemiology, which "identifies areas where there is a knowledge translation gap between best evidence (what some experts know) and practice (what most people do or believe)" (p. 763) about the distribution of information and misinformation on the internet. In "An Epidemiology of Big Data," this study aimed to determine the practical meaning about big data and fill the translation gap between what some experts know about big data offered through the wealth of 'grey literature' and what healthcare leaders believe through their cohesive 'lived experiences' of the big data phenomena.

Healthcare at a Glance

Recent estimates released from the Office of the Actuary at the Centers for Medicare and Medicaid Services (CMS) project that aggregate healthcare spending in the United States will grow at an average annual rate of 5.8 percent for 2012–22, or 1.0 percentage point faster than the expected growth in the gross domestic product (GDP). The healthcare share of GDP by 2022 is projected to rise to 19.9 percent from its 2011 level of 17.9 percent (CMS, 2012). Not to be confused with the life sciences, translational bioinformatics (Butte & Shah, 2011) or biomedical sciences, which produced the groundbreaking Human Genome Project that propelled the life sciences to the forefront of big data by generating approximately one terabase (trillion bases) of sequence data per month (Hey, Tansley, & Tolle, 2009), healthcare (or *heath care*) differs from other commodities because it is typically provided in a series of separate but related delivery episodes (Hornbrook, Hurtado, & Johnson, 1985; Lameire, Joffe, & Wiedemann, 1999) and can be thought of as a bundle of attributes (e.g., diagnosis, treatment, prevention of disease, illness, injury, appointments, technology, insurance) that vary in cost as well as importance to the buyer (Weisbrod, 1991). The bottom line in healthcare is cost savings, which have been extremely difficult to achieve in the absence of a major health system transformation.

The healthcare system possesses a large and growing elderly population that threatens to push the pace of upward spiraling healthcare price increases even higher than their already faster-than-inflation rates. Expensive medical treatments, end-of-life care, health inequities, new technologies, fraud and waste are just some of the intended and unintended expenditures that wreak havoc on healthcare delivery system budgets. Unchecked healthcare inflation creates ever-larger federal budget deficits, and pushes up the embarrassingly large number of Americans without adequate health insurance. Brown (2011) estimates potential “savings from big data in the sector could be upwards of $450 billion annually” (p. 2). This unprecedented potential for cost reductions within the healthcare system has captured the government’s imagination and attention, as over $200 million in new federal commitments were announced in an effort to improve the nation’s ability to manage, understand, and act on big data (Re, Nter, & Mill, 2012). Big data’s role in healthcare cost reduction is vital. To understand big data in healthcare, big data in a general context must first be understood.

Big Data in a General Context

Big data is not a new concept or idea; however, there is no clear definition for big data (Zaslavsky, Perera, & Georgakopoulos, 2013). The term "big data" originated as a tag for a class of technology with roots in high-performance computing, as pioneered by Google in the early 2000s (Hopkins & Evelson, 2012). A representative search of the term big data from Google yields a multitude of references : Big data BANKS (Kates, 1969); Big data BASES (Boehm, 1975); Big data FILTER (Ernst, 1976); Big data POOL (Porth, Badke, & Mieth, 1982); Big data SETS (Kinnstaetter, Lohmann, Schwider, & Streibl, 1988). One of the earliest references to big data was found in a dissertation that used the term big data as a subject key. The dissertation topic considered the problem of the optimal hardware architecture for advanced data management systems (Neches, 1983).

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data (Davenport & Jarvenpaa, 2008). Big data is routinely referenced in many industries including banking, defense, and oceanography, whose information technology and computational methods are mature and robust. Consumer retail has been a commonly cited industry that has taken advantage of big data's benefits. Large retailers like Target and Wal-Mart have used big data to develop business intelligence on consumer shopping patterns and behavior. By assigning a unique identifier to each customer that uses a credit card, fills out a survey, or provides their phone number, retailers are able to employ sophisticated statistical models to create targeted marketing campaigns.

The volume of stored information in the world is growing so fast that scientists have had to create orders of magnitude of data, including zettabyte and yottabyte, to describe the flood of data (Kuner, Cate, Millard, & Svantesson, 2012). The digital world is expected to hold a collective 2.7 zettabytes of data by year-end, an amount roughly equivalent to 700 billion DVDs (Hardy, 2012). As hardware and software advance, the capacities of large computational resources provide us with the only practical and reliable sense of what "big" means. This is particularly characteristic in an emerging digital information economy, where clickstream data give precisely targeted and real-time insights into consumer behavior. Our purchases, searches, and online activities are being tracked to improve everything from websites to social movements intended to democratize entire countries.

Earlier mainstream notions of big data were limited to a few organizations such as Google, Yahoo, and Microsoft, which did not produce scholarly communications but did produce reputable, credible marketing whitepapers found in the grey literature. Big data as a marketing and services tool has emerged as a profitable growth opportunity for many firms across industries. But there is a dearth of scholarly articles on big data, particularly in healthcare, as it has not been widely studied in academic circles; hence, many of the attempts to define big data are found in grey literature, including conference proceedings, briefing documents and sophisticated marketing materials that target buyers of services and goods. The following big data definitions sampled from mostly grey and some scholarly literature show just how wide-ranging and troublesome it is to adopt a definition of the term "big data:"

- "Big Data" is a science of fielding algorithms that enable machines to recognize complex patterns in data. It fuses machine learning with a very deep understanding of computer science and algorithms and that, of course, is key to being able to take machine learning and deploy it in a very scalable way (Paredes, 2012).
- "Big Data" exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it (Dumbill, 2013).
- "Big Data" is the ability to mine and integrate data, extracting new knowledge from it to inform and change the way providers, even patients, think about healthcare (Roney, 2012).
- "Big Data" is not a precise term; rather, it's a characterization of the never-ending accumulation of all kinds of data, most of it unstructured. It describes data sets that are growing exponentially and that are too large, too raw, or too unstructured for analysis using

relational database techniques. Whether terabytes or petabytes, the precise amount is less the issue than where the data ends up and how it is used (EMC2, 2012).

- “Big Data” is the ability to collect, process, and interpret massive amounts of information. One of the biggest potential areas of application for society is healthcare (Rooney, 2012).
- “Big Data” is a bubble just filled with hot air – at least for now. Everyone is talking about it but when you dig a bit deep with a pointed question, very quickly you discover that it has nothing much to do with the Big Data (Shah, 2013).
- “Big Data” are datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze (Manyika et al., 2011).
- “Big Data” is techniques and technologies that make handling data at extreme scale affordable (Hopkins & Evelson, 2012).
- “Big Data” is more data than our current systems and resources can handle (Fogarty 2012).
- “Big data” is an explosion of available information, a byproduct of the digital revolution (I. Thomas, 2013).
- “Big data” does not really exist in healthcare settings (Bollier & Firestone, 2010).
- “Big Data” n: the belief that any sufficiently large pile of s--- contains a pony (Arbesman, 2013).

Recent trends suggest big data is a philosophy: an organizational culture that embraces the complexities of integrating, analyzing and transforming vast amounts of data into a valued organizational asset. Young (2012) suggests “big data is only applicable to life and biomedical sciences research and not capable of adding value to the bedside delivery of healthcare, where patient encounters are counted - not petabytes” (p. 8).

Big Data in Healthcare

It is an impossible task to accurately count the number of patient encounters and transactions because of the current fragmentation of the care delivery system and the abundance of information technology platforms that do not interact. Big data in healthcare is slowly changing with the advent of system development approaches, wireless grids, and semantic web technologies that are highly compatible with widely distributed systems. The expansion of digital technology is capable of synthesizing data sources from other industries including housing, transportation, and social services to create an explosion of data in every aspect of an individual’s personal health profile.

Big data will enable the notion of personalized medicine, which provides physicians with a comprehensive understanding of a person's health and genomic makeup, rather than relying on a superficial understanding of other patients' histories (Horowitz, 2012). Underlying the data’s sheer volume are valuable relationships among datasets and social networks, implying that data integration can expose new information that was not discoverable in the past.

What has changed dramatically in the last twenty years is that computers have become more mobile, creating a robust mobile health (mHealth) industry where it is commonplace, if not necessary, for clinicians to carry handheld devices into exam rooms. Millions of smartphones, tablets, and other portable devices are generating and consuming data of increasing variety. Clinicians continue in 2013 to adopt mobile computing devices at a rapid rate, with nearly ninety percent expected to use smartphones in 2014 and almost as many using tablets.[2] Microsoft's Google Glass is gaining a reputation as a potential disruptive innovation. The wearable device is now deployed during certain surgical procedures and outpatient visits and is not as impersonal and distracting as a handheld device. The masses of small, mobile devices represent enormous computational capacity; albeit each individual physician typically generates or consumes a modest amount of data.

Big data is a challenge for industries such as defense, transportation, and banking. For healthcare it is even more formidable largely because patient data records cannot be so easily collected and freely shared; there are all sorts of technical, ethical, and public policy barriers to making such liquid data – liquid (Bollier & Firestone, 2010). Healthcare data remain in silos, fragmented and distributed across thousands of physician offices, hospitals, and clinically-integrated delivery systems that themselves are composed of autonomous units (L. R. Burns et al., 2002). The real revolution is not in the machines that calculate the data but in the data itself and how we use it (Mayer-Schönberger & Cukier, 2013).

Healthcare's Big Data Drivers

Bringing intelligent healthcare informatics to bear on the national problems of improving healthcare (Robertson, Dehart, Tolle, & Heckerman, 2009), reducing healthcare costs, and improving quality and health outcomes relies on an ability to take raw data and transform it into information that becomes knowledge for decision making. This is what fundamentally drives big data in healthcare. The next section provides a brief, but important acknowledgement of three

[2] Data taken from Epocrates' Mobile Trends Report based on a survey of 1,063 clinicians in May 2013. Internet Source: http://www.healthdatamanagement.com.

significant drivers beyond cost reduction pressures of big data in healthcare: health information technology, federal healthcare legislation, and healthcare consumers.

Health Information Technology (HIT)

The advent of health information technology (HIT) is expected to improve the management, analysis, and deployment of a tremendous amount of granular-level, patient-centered data. For instance, a possible transition to International Classification of Disease – Version 10 (ICD-10) will require physicians across all clinical specialties to transition from 20,000 codes under ICD-9 to 155,000 under ICD-10 – an almost eight-fold expansion.[3] In an information-rich healthcare industry, basic HIT interoperability is still a daunting problem. Even with HITECH legislation that encourages widespread adoption of HIT across all healthcare settings including physician practices, hospitals, and laboratories, there are different *scales* of data, both structured and unstructured, that do not have the ability to connect on a single platform. Much of medical knowledge and information remains in paper form. And even where data is digitized, it often resides in disparate datasets and repositories in diverse formats.

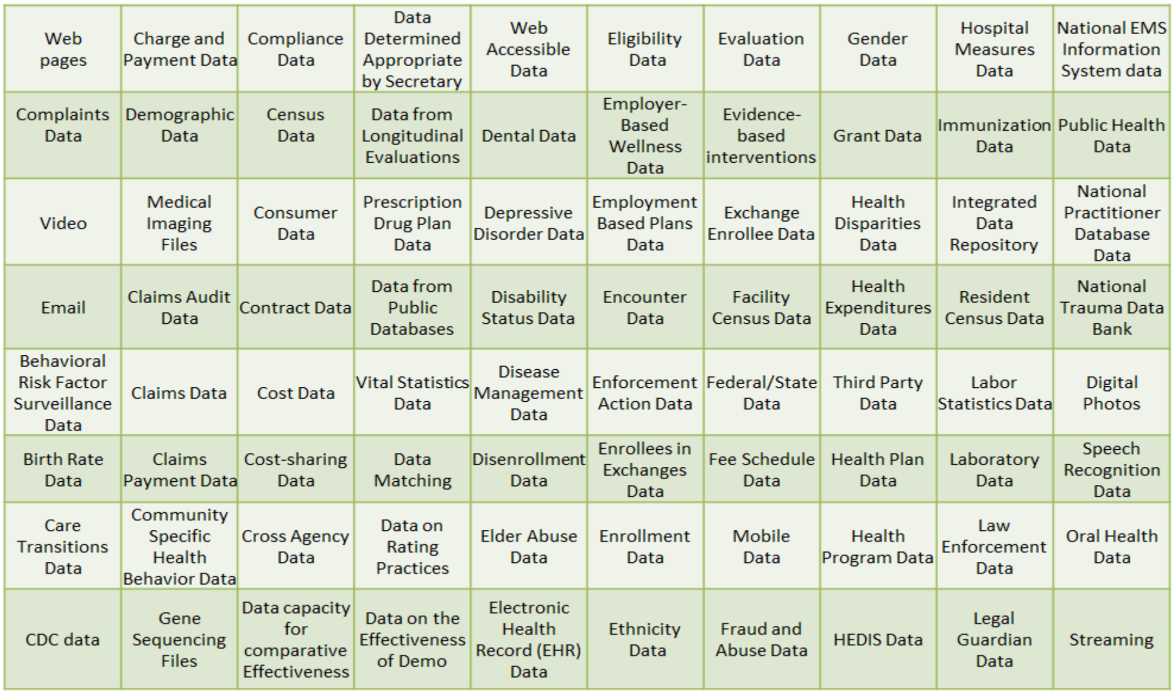
It is expected that through the adoption of HIT an extraordinary amount of structured and unstructured data will be generated, requiring a new level of computational strength and synthesis. As such, this data can be used as information to create knowledge to inform healthcare providers, consumers, and policymakers alike about topics ranging from highly complex questions at the point of care to pandemic forecasting.

[3] McKesson. Source: http://sites.mckesson.com/achievehit/files/ICD-10_FAQs_McKesson.pdf

With EHRs slowly filtering into physician practices, it is the data - and lots of it - that has healthcare thought leaders and visionaries anticipating the threshold moment of “Healthcare Singularity” (Buchan, 2009), when healthcare knowledge becomes instantaneous (remember the 2032 vignette?). When data was once considered tedious to manage and costly to store big data is now considered an asset to both individuals and organizations. Although the potential of new laws that promote information technology interoperability across stakeholder classes and consumer demand for “liberated” data on the health of communities are exciting, the spread and diffusion of medical knowledge is slow (Porter & Teisberg, 2006).

Federal Healthcare Legislation

Healthcare is a highly regulated field, with various laws guiding how healthcare data is used and reported (Sullivan, 2011). Healthcare legislation designed to reform an inefficient healthcare “system of systems” has been at the forefront of presidential political agendas for decades. Over the past ten years, several major bodies of healthcare legislation have been enacted to provide Medicare beneficiaries with Part D drug plans, which closes the metaphorical “donut-hole” prescription coverage gap that describes the variance between initial drug coverage limits and catastrophic drug coverage thresholds; the Health Insurance Portability and Accountability Act (HIPAA) has standardized the exchange of essentially all healthcare transactions between physicians, hospitals and their business partners while also providing guidance on patient privacy and systems security; and, the Patient Protection & Affordable Care Act (Affordable Care Act), which is the most sweeping body of legislation since Medicare was introduced over 45 years ago, will introduce, among many patient protections, innovative payment and coordinated care models designed to provide high quality healthcare at lower costs, rules against insurers dropping patients because of pre-existing conditions, and eliminates lifetime limits on medical expenses. The Affordable Care Act identified a host of old (administrative) and new (streaming) datasets (Figure 1) that must be collected, managed, and reported by healthcare stakeholders.



Web pages	Charge and Payment Data	Compliance Data	Data Determined Appropriate by Secretary	Web Accessible Data	Eligibility Data	Evaluation Data	Gender Data	Hospital Measures Data	National EMS Information System data
Complaints Data	Demographic Data	Census Data	Data from Longitudinal Evaluations	Dental Data	Employer-Based Wellness Data	Evidence-based interventions	Grant Data	Immunization Data	Public Health Data
Video	Medical Imaging Files	Consumer Data	Prescription Drug Plan Data	Depressive Disorder Data	Employment Based Plans Data	Exchange Enrollee Data	Health Disparities Data	Integrated Data Repository	National Practitioner Database Data
Email	Claims Audit Data	Contract Data	Data from Public Databases	Disability Status Data	Encounter Data	Facility Census Data	Health Expenditures Data	Resident Census Data	National Trauma Data Bank
Behavioral Risk Factor Surveillance Data	Claims Data	Cost Data	Vital Statistics Data	Disease Management Data	Enforcement Action Data	Federal/State Data	Third Party Data	Labor Statistics Data	Digital Photos
Birth Rate Data	Claims Payment Data	Cost-sharing Data	Data Matching	Disenrollment Data	Enrollees in Exchanges Data	Fee Schedule Data	Health Plan Data	Laboratory Data	Speech Recognition Data
Care Transitions Data	Community Specific Health Behavior Data	Cross Agency Data	Data on Rating Practices	Elder Abuse Data	Enrollment Data	Mobile Data	Health Program Data	Law Enforcement Data	Oral Health Data
CDC data	Gene Sequencing Files	Data capacity for comparative Effectiveness	Data on the Effectiveness of Demo	Electronic Health Record (EHR) Data	Ethnicity Data	Fraud and Abuse Data	HEDIS Data	Legal Guardian Data	Streaming

Figure 1.A sample of structured and unstructured datasets collected under healthcare reform. Source: Patient Protection & Affordable Care Act 2010

Though the new healthcare landscape promises to provide high quality, cost effective care to millions of new beneficiaries through federally-mandated Health Insurance Marketplaces and to people with preexisting health conditions, the deluge of data will certainly test the system's ability to collect, store, and analyze big data. Still, there is skepticism that federal policies thus far have blunted big data's potential in the public sector (Konkel, 2013).

Although the government has a long history of making biomedical science data available to the public, the Obama Administration's Open Government Initiative has motivated federal agencies to make a wider variety of data available to "citizen scientists" at www.data.gov. This website has the potential to create a secondary market for visionaries, researchers, and entrepreneurs to create new tools and knowledge for many stakeholders including healthcare consumers who lately have been inclined to provide open access to their personal health records.

Consumers of Health and Healthcare

A new healthcare information economy has materialized. Healthcare consumers now demand a new scale of data liquidity enabled by EHRs, laboratory information systems, medication-management systems which are interoperable with their personally controlled health records (PCHR) where they independently decide (Mandl & Kohane, 2008) when and with whom they share their individually identified health information.[4] Healthcare consumers must now become researchers, or "citizen scientists." However, beyond initiatives like Blue Button® Connector, which provides a limited number of Medicare beneficiaries access to historical claims data, access to the tools and information on par to the sophistication and rigor of that afforded to policymakers and providers allowing, them to better manage their own healthcare in the new health information economy is at best, scant. While some healthcare consumers so happen to be highly skilled data scientists, the masses do not have the necessary technical skills

[4] The Health Insurance Portability & Accountability Act (HIPAA) which includes provisions for protection of individually identifiable health information (formerly protected health information (PHI)) does not apply to patients who wish to share their own health information.

or even the requisite health literacy (and e-health literacy) skills to harness big data for basic healthcare decision making.

Healthcare is growing rapidly in terms of the quantity and quality of data that is collected on a daily basis. The problem is that this data is growing faster than the consumers can use it. As the world’s population increases, the health and healthcare data problem will be exacerbated. Healthcare consumers are facing the challenge of not only selecting the best care for themselves and their families, but doing it in a cost effective manner based on the best available healthcare information and clinical evidence-base.

The once skeptical healthcare patient engagement movement is slowly gaining momentum with the advent of technological innovations such as wireless grids, semantic web applications, and social networking approaches that revolutionize the way healthcare consumers collaborate, identify potential collaborators or friends, communicate with each other, and identify information that is relevant for them (Eysenbach, 2008). These tools will produce better ways for consumers to take charge engaging with physicians, government, and other healthcare stakeholders to reduce wasteful spending and improve population health.

Big data also enables personalized medicine, which provides physicians with a comprehensive understanding of an individual's health, environmental, and genomic makeup, rather than relying on a superficial understanding of other patients' histories (Horowitz, 2012). In order for healthcare consumers to be effective participants in a reformed healthcare landscape, they require information from trusted, third-party sources. The Health 2.0 movement makes a uniform attempt to provide collaborative approaches to engaging healthcare consumers through credible information. For instance, Dr. Gunther Eysenbach coined the term “apomediation,” which encompasses a socio-technological information seeking strategy where people rely less on a traditional intermediary, such as a pharmacist giving relevant information to a patient. The difference between an intermediary and an apomediary is that an intermediary stands “in between” the consumer and information. In contrast, apomediation means that there are agents (e.g., people, tools) that “stand by” to guide a consumer to high quality information and services without being a prerequisite to obtain that information or service in the first place. While these distinctions are not absolute (in practice, there may be a mix of both, with people moving back and forth between apomediation and intermediation models), it has been hypothesized that they influence how people judge credibility (Eysenbach, 2008).

Who are the key healthcare stakeholders?

From Congress who drafts healthcare legislation to patients who require evidence-based information to inform their treatment decisions, there are many stakeholders with an interest in the delivery of h. The Agency for Healthcare Quality and Research (2014) defines healthcare stakeholders, “as persons or groups that have a vested interest in a clinical decision and the evidence that supports that decision. Healthcare stakeholders include: patients, caregivers, clinicians, researchers, advocacy groups, professional societies, employers, and policymakers” (p. 11).

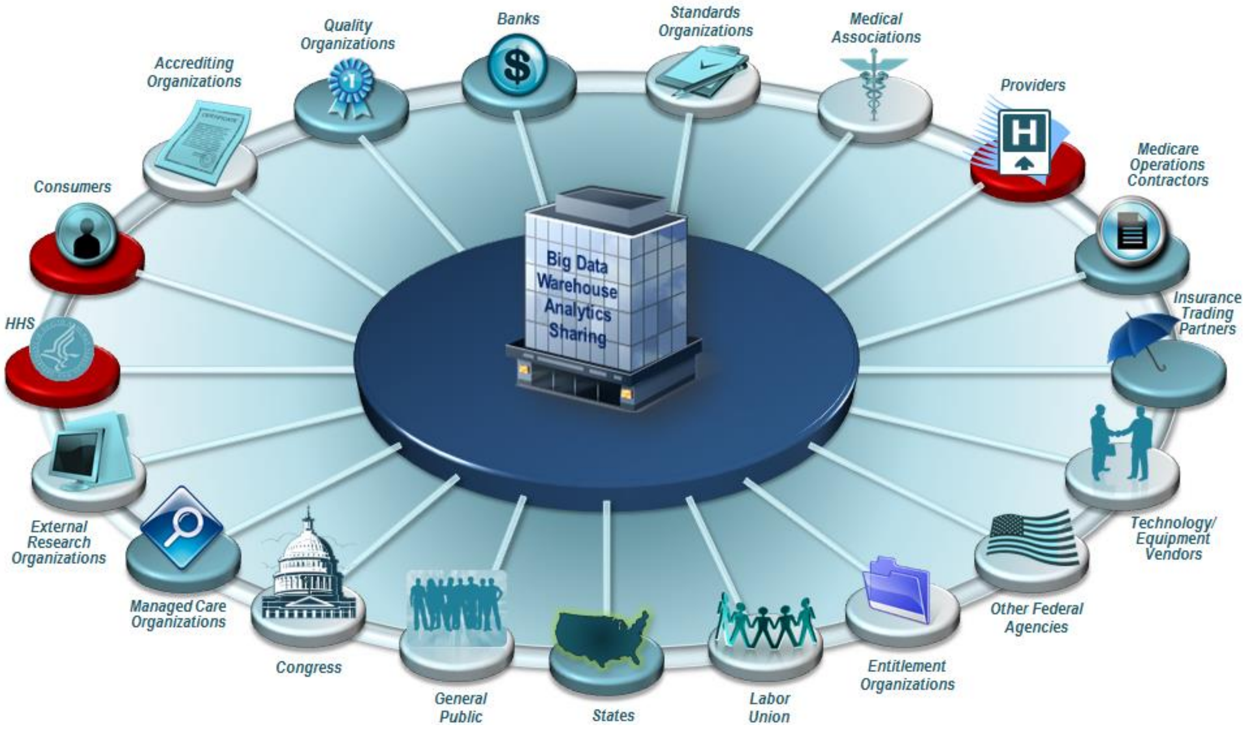

Big Data
Warehouse
Analytics
Sharing
Quality Organizations
Banks
Standards Organizations
Medical Associations
Accrediting Organizations
Providers
Medicare Operations Contractors
Consumers
Insurance Trading Partners
HHS
Technology/ Equipment Vendors
External Research Organizations
Managed Care Organizations
Other Federal Agencies
Congress
Entitlement Organizations
General Public
States
Labor Union

Figure 2. The many classes of healthcare stakeholders of the reformed healthcare ecosystem

It is a highly complex task to understand the interrelationship between many healthcare stakeholders of the ecosystem (Figure2). At a very basic level, an episode of care is initiated when a patient (stakeholder) initiates and follows through on a scheduled appointment to interact with a provider (stakeholder) for clinical consultation and treatment of an ailment or illness. This simple scenario does not even take into account whether the patient has employer-based insurance or is a beneficiary of a public healthcare entitlement program, such as Medicaid or Medicare. The scope of events that precede and succeed a single patient encounter entails synchronization of care coordination, data collection and analysis, information generation and exchange, and knowledge in the form of policies, procedures, evidence-based medicine, and provider report cards – underpinned by HIT. The point being, a single patient episode, regardless of its level of complexity, requires collaboration and exchange of information from as few as two to a multitude of additional healthcare stakeholders.

Data Sharing

Data sharing is complex and inconsistent within and across the many classes of healthcare stakeholders and are frequently hampered by the lack of foolproof de-identification for patient privacy, as data reside in many discrete data systems. The lines in Figure 2 depicts information technology interoperability where all stakeholders share their big data in a common data repository, creating massive amounts of data for healthcare decision making, shared knowledge for learning systems, and consumer choices. While such data repositories may exist locally or regionally, no such national data warehouse exists.

This issue alone impedes opportunities for data mining and analysis that would enable precise predictive and preventive medicine (Robertson et al., 2009). The use of EHRs is producing more data-in-depth healthcare environments in which substantially more data are captured and transferred digitally, flooding stakeholders with data, generating an urgent need for new techniques and tools that can intelligently and automatically assist in transforming (Fayyad et al, 1996) big data into better information for decision making.

An analysis of healthcare stakeholder classifications typically included federal, state, and local policymakers who create rules and regulations, consumers who demand healthcare services, and providers who supply healthcare services either at a cost or through charitable care. These three key healthcare stakeholders are central to achieving the industry adopted Triple Aim of improving the experience of care, improving the health of populations, and reducing per capita costs of health (Berwick, Nolan, & Whittington, 2008). Grossmann (2010), in an Institute of Medicine series wrote, “by providing greater insight to patients, providers and policymaker[s] … data hold the potential to help transform the U.S. healthcare system” (p. 69).

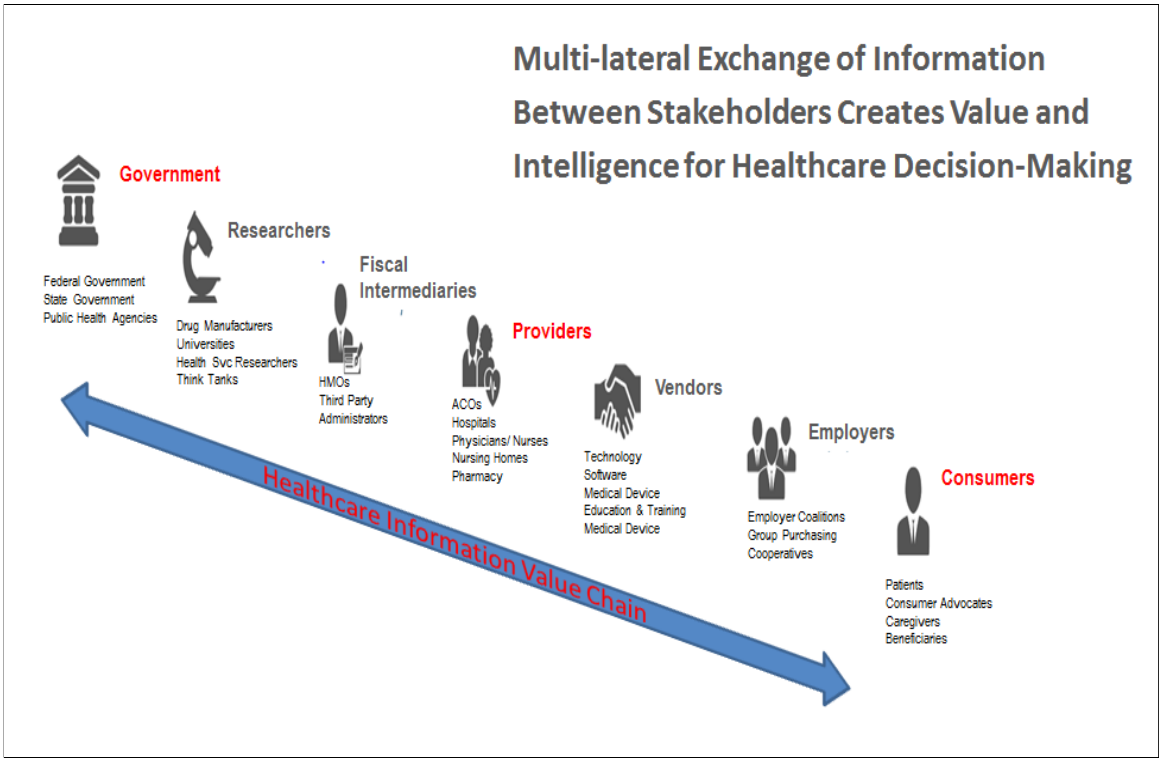


Figure 3. *Information flow along the healthcare information value chain*

These core health system classes are situated at the center and both ends of the healthcare value chain: government (producers), providers (deliverers) and consumers (users) (Figure 3).

The implementation of EHRs has contributed to a data rich healthcare environment in which substantially more data are now captured and transferred digitally, generating an urgent need for new analytical techniques and information management tools that can intelligently and automatically assist in transforming (Fayyad et al., 1996) big data into better information for decision making. Yet, the three stakeholder classes that are the focus of this study have different goals and hopes for big data (Feldman, Martin, & Skotnes, 2012). An assessment of the readiness of the three key stakeholder classes was explored:

- Government: As the largest producer of open source data for public use, government is a key contributor to the generation of information needed to achieve cost efficiencies in healthcare. Through government supported data initiatives like Healthdata.gov, providers, consumers and other healthcare stakeholders can have reasonable access to raw data for making choices about treatments (Clancy, 2006). Yet, government leaders **struggle with the sheer volume of data** they seek to manage.[5] They lack a systematic approach to classifying and sharing quality, cost, and outcome data with other interested participants of the delivery of healthcare. Also, what is the proper and practical role for government in the face of a deluge of digital data (Kuner et al., 2012)?
- Providers: They most frequently use data for healthcare delivery, value-based purchasing, and EHR reporting incentives. However, they often **lack sufficient data aggregation and analysis tools** to capture data and turn it into usable knowledge. The general perception is physicians are not prepared to use big data at the point-of-care for decision-making.
- Consumers: Consumers produce the bulk of big data. There is often an abundance of information available, but much of it is irrelevant to the decision-making process. **Little is actually known about what kinds of data and information consumers need** to

[5] Tech America Foundation Report (2012) Demystifying Big Data: A practical guide to transforming the business of government. http://www-304.ibm.com/industries/publicsector/fileserve?contentid=239170

make decisions (Edgman-Levitan & Cleary, 1996). Currently, consumers have more mobility, live longer lives and, healthcare is more shared than ever before. Consumers have little control within the contour of big data in an undefined, unregulated data environment. Ethical issues such as privacy, trust, and informed consent loom as major big data barriers.

Collectively, triangulating the perceptions of these three “key” healthcare stakeholder classes represent an optimal starting point to understand the phenomenology of big data in healthcare. This research study is about discovering the important categories of “meaning about” big data in healthcare verses the “meaning that” which many theoretical frameworks, including Grounded Theory, Information Diffusion Theory, or Dewey’s Theory of Experiential Learning seek to ground or test research data. However, a short discussion of information sharing provides the necessary breath to understand big data in the context of healthcare. Value chain analysis in healthcare provides an intriguing framework that encompasses the vertical and horizontal integration of the strategic relationships and information sharing among healthcare stakeholders. In the next section, I introduce an aspirational value chain framework: An epidemiology of big data.

An Epidemiology of Big Data

Value chain analysis originally sought to examine the operations of a manufacturing enterprise by looking at the value or cost of inputs in terms of the value or price of outputs. In a typical value chain, money, products, services, information, or other goods are multilaterally exchanged between two or more participants. L. R. Burns et al. (2002) describes the value chain as “a virtual network designed to help move a produce (information) from the producer (government) through an intermediary purchaser (provider), and eventually down to the consumer. However, in healthcare, a value chain framework “represents more aspiration than reality” (p. 11) because of its many “broken links.” Similarly, in epidemiology the “chain of infection” posits that for an infection to develop, each link of the chain must be connected. Breaking any chain of the link can stop the transmission of the infection. Analogous to the epidemiological chain of infection, in healthcare, information generated by big data might typically spread among healthcare stakeholders, at least in theory. When a link in the value chain that characterizes big data and information sharing in healthcare is broken, evidence-based medicine is unachievable.

To express the origin, incidence, spread and control of information derived from big data shared between healthcare stakeholders, a notional and aspirational value chain framework, “an epidemiology of big data,” is potentially an important aspect of the big data “contagion” in the healthcare ecosystem. In the context of big data analyzed into information for knowledge, such a notional framework suggests that big data in healthcare evolves into information that is multilaterally spread among healthcare stakeholders, creating commodity value each time big data is exchanged and is “kinetically energized” by the “invisible hand” of efficient organization which is embodied in metadata (Zeng & Qin, 2008).

An epidemiology of big data is not a construct of an IT system. Information derived from organized structured and unstructured data (big data) whose value is presumably increased (or decreased) through standardized multilateral knowledge and information exchange among and between all healthcare stakeholders, creating value add and ultimately healthcare intelligence for policymaking, decision-making, and care delivery (Table 1). In short, data’s value needs to be considered in terms of all the possible ways it can be “spread” by members along the healthcare information value chain, not simply how it is used for its initial use (Mayer-Schönberger & Cukier, 2013; Porter & Teisberg, 2006).

STAKEHOLDERS	GOALS	CONCERNS
Consumers	• Understandable Clinical Information • Improved Data mobility • Improved decision making Better care coordination	• Access to care • Affordable care • Security and privacy of personal data • Trustworthiness
Providers	• Performance based payments • Reduced administrative paperwork • Improved care coordination • Business Intelligence for ACOs	• Additional regulations and paperwork requirements • Increased uncompensated care • Data Quality • Malpractice
Government	• Program Integrity • Quality Measures • Better health outcomes • Lower healthcare costs	• Budget for infrastructure change • Prioritizing resources • Value-Based Purchasing

Table 1. Information Goals and Concerns of Key Healthcare Stakeholders

Collectively, little is known about how much key healthcare stakeholders really know about the magnitude of big data challenges and whether consumers are even aware of big data, much less how to leverage it for their own benefit. To support this claim, I immersed myself in an extended review of the literature which provided contextual background and supported identification and refinement of the research question and research problem (Ridley, 2009).

CHAPTER II. LITERATURE REVIEW

This chapter describes the processes I used to conduct the literature review and identifies results and emergent themes derived from scholarly and grey literature. This information served as a backdrop to the data collected from the in-depth interviews. The literature review is foundational for an phenomenological research study; the literature does not guide and direct the study but serves as an aid once patterns or categories have been identified (Creswell, 2009). The preliminary literature review that began January 2011 underwent several revisions through March 2013. A modified systematic literature review (Frehywot et al., 2013; Mays, Pope, & Popay, 2005) was used to provide structure. In this research study, *An Epidemiology of Big Data,* an extensive reference list of scholarly (87) and grey (1,380) literature was reviewed and assessed for validity and usability.

The questions, context, and content of healthcare management and policy are generally broader and more diffuse than those of the clinical world (J. L. Bellamy, Bledsoe, & Traube, 2006), requiring the use of 'grey literature' in this study. Web of Science/MEDLINE alone cannot be used to effectively gather data about social science and humanities citations (Hutton, 2009). The broad function of the literature review for policy relevant research is to help decision makers see and conceptualize the breadth of issues and broad models that can inform decision making about a policy problem. Reviews can involve a policy problem that has remained unchanged for years or it can involve a policy problem that is likely to emerge in the future. Increasingly, health policy decision makers and professionals are turning to research-based evidence to support decisions about policy and practice (Bell 2006).

Lomas (2005) concluded that historically, “policymakers are less commonly seen as a target audience for systematic reviews” (S1:36). Such rigorous literature reviews are relegated to clinical research. This approach is changing; the discipline of “systematic-type” reviews has filtered into policymaking, though not in the format or approach found in comprehensive systematic reviews.

Mays (2005) found “there is no single agreed upon approach” (p. 1) to policy-related systematic reviews. But in answering policy questions, policymakers and managers will often need to draw on diverse sources of evidence – not only quantitative and qualitative research, but also other evidence such as expert opinion and explicit value judgments (Mays et al., 2005).

A Dearth of Scholarly Literature

While the grey literature on big data has exploded with vendors adding the term “big data” to marketing materials just to drive hype (Hopkins, 2011), there is a shortage of scholarly works, and therefore, we are no closer to defining the term for stakeholders to make sense of its true potential and application. In a Google search (09 Sept 2013), the term “big data” generated over 9.1 million hits. Most of the literature addressed big data collected and synthesized for providers while touching on big data in government including its policy implications (Konkel, 2013) and its funding prowess (Leinweber, 2011; Re et al., 2012). Consumer-related big data research is almost nonexistent, as little is actually known about what kinds of data and information consumers need to make decisions (Edgman-Levitan & Cleary, 1996).

Modified Systematic Literature Review Approach

The modified systematic literature review for gathering, summarizing, and synthesizing published and unpublished research is narrower than state-of-the-evidence reviews, but broader than traditional systematic reviews and may include not only published and unpublished research, but also published and unpublished non-research literature (Benzies et al, 2006). A systematic review essentially summarizes the best available research on a specific question by using transparent procedures to find, evaluate and synthesize the results of relevant study questions. The methodological approach to modified systematic reviews found in Mays et al (2005, p.9) study was adapted for this proposed course of research (Table 2).

COMPONENT	RESULT
Explicit research question	*Within and across the narratives of three key healthcare stakeholder classes, what are the important categories of meaning or current themes about big data in healthcare?*
Explicit search strategy	Search Web of Science/MEDLINE and Scopus on the search string: "big data"[All Fields] AND "healthcare"[All Fields]. Limitations are animal science related articles and the availability of free articles and citations.
Explicit statement about what types of research evidence were included and excluded	Continue to refine selection criteria that contain “big data” and “healthcare” in peer-reviewed journal articles, systematic reviews, government supported research, and meta-analysis. Also the discovery of new themes and keywords are the objective.
Critical examination of the quality of the studies included in the review	Examine relevancy to research question as reviewed in journals and authors frequently appearing in searches; examine relevancy of citations.
Critical and transparent process of interpretation of the findings of the review:	Assess applicability to the “delivery of healthcare" and identify a proven method of qualitative content analysis.

Table 2. Five components of the modified systematic literature review

Where peer review is a key part of the systematic literature review process (Thyer & Myers, 2011), in this modified approach, I chose to forgo this important activity due to time constraints. I instead relied mainly on the credibility of the journal's peer review process.

Peer-reviewed scholarly literature that met the aforementioned criteria was identified by electronically searching the following resource databases: Web of Science/MEDLINE (Syracuse University Library) and Scopus.[6] Google Scholar was used to identify additional sources of scholarly and grey literature when Web of Science/MEDLINE or Scopus did not produce links to full text articles. Hutton (2006) found that "considering Web of Science, Google Scholar, Google and Web link information, through a varied approach to gather citations produces unique, relevant instances of the use of grey literature" (p. 12). Target literature included books (electronic and print) and scholarly articles on the primary search string and Boolean operator: "big data"[All Fields] AND "healthcare"[All Fields]. This approach was used to restrict the search to potential articles of interest and covered all possible combinations. Other key indicators were added as the literature review was refined. The term "large data sets" was often found in the literature but was not used in this study so to maintain consistency with the study term, "big data."

Investigator-led systematic reviews appear to be a clear method of progressively focusing and refining analyses so that policymakers (Lavis et al., 2005; Tranfield, Denyer, & Smart, 2003), providers and consumers will find the resulting information both persuasive and usable.

[6] At the time, PubMed was searched; however, no requisite citations were found.

Systematic Review Challenges

As an emerging scholar-practitioner in healthcare, there are material challenges of employing a “less than summative systematic review” as found in Cochrane Collaboration Studies in epidemiological and economic research on clinical care. The first challenge is to minimize bias (Benzies et al., 2006). Systematic reviews provide specific methodological requirements, explicitness, and transparency in regard to the specific research question (Lomas, 2005) that helps to mitigate researcher bias. Another major challenge that persists with systematic reviews is to gain credibility (Lomas, 2005) among academic researchers, who firmly embrace the rigidity of gold standard scholarly methodological approaches. As with grey literature, scholar-practitioners must weigh whether the advantages outweigh the challenges of employing such methods. The intent with this research is to mitigate all bias by adopting an approach that is replicable and proven to researchers and policymakers in the discipline.

Grey Literature Approach

In credible, scholarly research, the use of grey literature should only be used in two contexts. First, grey literature could be used to supplement and triangulate information from empirical scholarly literature that meets the gold standard for evidence. A second way of treating grey literature is to trace the experience of a community and its policymakers with a particular policy problem (Bell, 2006). This research study utilized both approaches where applicable, with the goal of supplementing the scholarly literature found in the modified systematic literature review. The prevalence of the term big data in conference proceedings, corporate marketing materials, newspapers, and blogs provided needed breadth and depth to frame and understand the definition, growth, and uses of big data in the absence of scholarly citations in peer-reviewed journals related to the subject. Grey literature was searched using Google and Google Scholar databases and was limited to consultancy reports, government briefings, conference proceedings, web articles, white papers, dissertations, newspapers and blogs from major corporations that are known to produce high quality industry sector documents.

While not customarily the target of Cochrane Collaboration-style research, increasingly, health policy decision makers and other allied health professionals are turning to research-based evidence to support decisions about policy and practice. Decisions about whether to include grey literature in a state-of-the-evidence review are complex (Bell, 2006). To reduce the complexities of using grey literature, the following criteria were used to evaluate the grey literature cited in this study:

- Source of the Report: Grey literature was from reputable consulting firms that conduct extensive industry studies in big data, from IBM, McKinsey, Forrester, Deloitte, SAS, Becker's Hospital Review and Microsoft will be included.
- Transparency of Methods: Data and other types of information about where the report came from, how it was analyzed, and how the final report was compiled were accessible.
- Currency: Consultancy reports, government briefings, conference proceedings, web articles, white papers, dissertations, newspapers and blogs were sourced between January 2010 and April 2013.[7]

[7] Source: http://guides.library.upenn.edu/content.php?pid=286667&sid=2454523

Article Selection Criteria

A two-step process to select literature was used. First, an independent screen of titles, keywords and abstracts (when available) of search results was carried out to ascertain if a document met the general inclusion criteria. Subsequently, an independent assessment was conducted of the full text file of each source based on the predetermined inclusion and exclusion criteria. This review was limited to published references that directly described (1) big data in a broad context to capture wide variations in its definition and application across industries and (2) big data with a specific magnitude, to capture the specificity of themes in the healthcare-related literature. Additional constraints included removing citation only references, and veterinary-related (e.g., animal science) research.

Aggregate results from the systematic review and the grey literature searches were entered into a Thomson Reuters Endnote x6 Reference Manager ® bibliographic management database and sorted by themes and important categories described in the following section.

Preliminary Literature Search

A preliminary literature search was conducted through Web of Science/MEDLINE and was used exclusively to initiate the modified systematic literature review approach to capture the scholarly literature. A secondary search was conducted in Scopus to find reputable articles from additional peer-reviewed journals in which the full-text of the article was available. An analysis was conducted of duplicate documents and relevance. Where no full text or abstract was available, I searched Google Scholar and found many of the PDF and HTML files used in this dissertation thesis. Google was searched to find select grey literature based on the inclusion criteria (consultancy reports, government briefings, conference proceedings, web articles, white papers, dissertations, newspapers, and blogs from major corporations that are known to produce high quality industry sector documents). Initial searches on the three databases led to an inclusion of documents procured from U.S. Federal Government administered websites.

Results from the Literature Review

The literature search began with use of the following search terms and Boolean operator: "big data"[All Fields] AND "healthcare"[All Fields]. Through the Web of Science/MEDLINE database, 87 documents were found in peer-reviewed healthcare management related journals (69). Government research support, reviews, letters, and editorials (18) constituted the balance of the documents found. Prevalent research areas were computer science (25), medical informatics (19), and healthcare science services (17). However, the most unanticipated research area that tied for second (19) was information science/library science. Journal articles specifically focused on research or life science disciplines including bioinformatics, genetics, biology, and engineering, or non-health related disciplines, including computer science and information science. Other areas of inquiry on big data are found in the energy and aerospace industries.

Because of the paucity of results, a second search was performed with the key indicator of "big data"[All Fields] only, using the Web of Science/MEDLINE database. The return was significantly larger, yielding 562 articles in various journals, including *Sensors, National Academy of Science* and *Journal of Animal Science*. The journals on computer science had a wealth of information on big data. Also conference proceedings were rich in usable information. During the literature review, the term “big data” was still trending in healthcare. A review of my results from the Web of Science/MEDLINE search found no author who emerged as a thought leader on the big data phenomena. The following argument table (Table 3) justified the fundamental reasoning for conducting this study.

ARGUMENT STEPS	RELEVANT REFERENCES
Big Data is exploding in healthcare	(Cukier, 2010; Dumbill, 2013; Feldman et al., 2012; Lomas, 2005; Villars, Olofson, & Eastwood, 2011)
Big Data has been slow to adapt in healthcare	(Porter & Teisberg, 2006; Young, 2012)
Big data in healthcare requires a clear definition and subsequent taxonomy	(Brown et al., 2011; Brynjolfsson & McAfee, 2011)
Stakeholders are central to the healthcare information value chain	(L. R. Burns et al., 2002; Gorman, 1995)
Dialogue between policymaker social scientist, and consumer (healthcare information value chain) must grow	(L. R. Burns et al., 2002; Dumbill, 2013; Leinweber, 2011; Lomas, 2005; Porter & Teisberg, 2006; Roper, Winkenwerder, Hackbarth, & Krakauer, 1988)
Metadata is fundamental to big data, interoperability, and information exchange in healthcare	(Burns, 2011; Gantz & Reinsel, 2011; Parsons et al., 2011; Pavolotsky, 2012)
Drawing together published literature, ‘grey’ literature, decision maker’s experience, and researcher’s knowledge and experience make the best practice and policy decisions	(Lavis et al., 2005)
Data scientist and trusted apomediation are necessary; data scientist profession consists of many titles, some of which have existed for years in healthcare	(Brown et al., 2011; Chen, Chiang, & Storey, 2012; Davenport & Patil, 2012; Eysenbach, 2008)

Table 3. Argument chart to conduct phenomenological study

The initial search was organized into four themes: “Big Data,” “Drivers,” “Methods_LitReview,” and “Methods_Qualitative.” As I conducted a deeper analysis of the literature through citation analysis to discern themes, additional categories emerged, including “Stakeholders,” “Data Scientist,” “Privacy,” “Ethics,” ”Narrative Medicine” and “Metadata.” These categories shaped the refinement and development of a credible research question.
The process of arriving to a very clear and concise research question was an iterative process that took skill and time. The literature presented a compelling case to conduct this research.

Most articles included in this study mentioned big data in the context of healthcare delivery.[8] In some cases, the general application of big data across industries where the term has matured was used for definitional purposes. Additionally, bibliographies of all documents retained (peer-reviewed and grey literature) were reviewed as part of a “snowballing” technique to find further relevant resources, including other documents and applicable websites. In all, over 200 documents are included in the review.

Analyzing the Evidence

There is a strong correlation between the categories of “Big Data,” “Information Sharing,” and “Stakeholders.” This seems like a logical relationship, but patients within the consumer stakeholder class were often left out of the information value chain; I believe there is great potential for further study on this topic. The notion that “modern medicine is an information science” (Hood & Friend, 2011; Litvin, Cavanaugh, Callanan, & Tenner, 2008) is

[8] The term “healthcare” was often used as part of a reference list of industries where big data is or could be used. The context of the article was not directly related to healthcare. These articles were eliminated.

intriguing and is viewed or articulated in different ways: “personalized medicine, information-based medicine.”

The master codes that synthesize across categories are:

- Big data (large data sets)
- Information Sharing
- Metadata
- Stakeholders (focusing on “patient” as the stakeholder)

The initial categories/columns helped me to organize the main points of each article and provide a map which I used to look back to either further study the work of the authors cited or find literature where I found potential gaps. As I scanned the literature a second and third time, I found some additional codes:

- Computation & Analytics
- Data quality
- Knowledge management
- Privacy (HIPAA)
- Data Scientists

Observations from the Literature Review and Emergent Themes

The modified systematic review of the literature on “big data” and “healthcare” produced the following initial cohesive observations:

- There is a dearth of scholarly research on big data in healthcare. This is significant because of the exponential growth in healthcare data types, the volume of data, and the speed at which data flows. Further inquiry requires rigorous study.
- The gulf between “life sciences” and “healthcare” is closing – fast. Big data is entrenched in life sciences research, including genetics, biomedical research, computational biology, and nanomaterial science. However, these advances are quickly making their way into point-of-care decisions (e.g., shared physician and consumer decisions about treatment plans).
- There is no consensus on what big data means in healthcare. Depending on the stakeholder, big data has different meaning, even within stakeholder classes. This makes achieving an interoperable platform almost impossible. Of the many big data definitions in both scholarly and grey literature, only one article was found that attempted to define “big data in healthcare” big data refers in the healthcare context to longitudinal medical claims data for millions of patients linked to their EHRs (Begley, 2011).
- Consumers do not have enough trustworthy, credible information to understand the scope and depth of big data and its impact on their health and healthcare.
- Patient informed consent and privacy regarding the use of an individual’s big data are as challenging to overcome as interoperability of HIT.
- Data Scientist is a generic term that requires no unique skill beyond that of a statistician. In fact, depending on one’s need for big data, a basic level of education will suffice (e.g., citizen scientist).
- Industry and marketing firms have dominated the proliferation of big data through conference proceedings, marketing materials, white papers, and blogs.

Based on the grey literature, there were six dimensions to big data that conveniently began with the letter "V". Gartner, the information technology and advisory firm, captured the industry's attention by introducing the popular "3 V's" of big data - Volume, Variety, and Velocity. The table below (Table 4) provides an inclusive overview and characteristics of the six dimensions of big data that are noted in various documents in both scholarly and grey literature.

CHARACTERISTIC	DEFINITION
High Volume (G)	Enterprises are awash with ever-growing data of all types, easily amassing terabytes—even petabytes—of information. • Turn 12 terabytes of Tweets created each day into improved product sentiment analysis. • Convert 350 billion annual meter readings to better predict power consumption.
High Variety (G)	Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together. • Monitor 100's of live video feeds from surveillance cameras to target points of interest. • Exploit the 80% data growth in images, video and documents to improve customer satisfaction.
High Velocity (G)	Sometimes two minutes is too late. For time-sensitive processes such as catching healthcare fraud, big data must be used as it streams into an enterprise in order to maximize its value: • Scrutinize 5 million trade events created each day to identify potential fraud; • Analyze 500 million daily call detail records in real-time to predict customer churn faster
Veracity	One in 3 business leaders don't trust the information they use to make decisions. Establishing trust in big data presents a huge challenge as the variety and number of sources grows.
Value	Value in healthcare is the health outcome per dollar of cost expended (Porter & Teisberg, 2006).
Variability	A variety of formats as opposed to just one relationally structured data set (Hopkins & Evelson, 2012).

Table 4. Six Characteristics of Big Data, Including Gartner's 3 V's

The review of the literature demonstrated that scholarly journals in Web of Science/MEDLINE provided relevant scholarly works about big data in healthcare compared to select grey literature. A possible reason for the lack of literature includes publication lags: grey literature found in industry cycles through peer review much faster than scholarly journals. Companies providing solutions in information technology, engineering, and other science-based firms have a mission to drive revenue and can quickly publish marketing research and other materials (e.g., white papers, conference proceedings, blogs, etc.). Companies have sought to capitalize on a subject few outside of their disciplines understand. The literature review was continuously revisited and refined throughout the course of this study for accuracy and relevancy and to ensure adherence to required elements of the modified systematic review standards.

The next two subsections of this chapter are important themes that emerged from the literature review: metadata and data scientist. A third theme, privacy, also stood out, but requires a full research paper to do justice on this very important topic. The intent is to provide a brief overview and discussion of these important themes. While I did not expect the key healthcare stakeholder narratives to capture the full essence of these two themes, each topic serves as important background information to the interpreted ‘story’ that this research study produced.

Metadata

At the very core of HIT interoperability is metadata. The Office of the National Coordinator for Health Information Technology (ONCHIT) issued an advanced notice of proposed rulemaking (ANPRM), which solicits public comments on the metadata standards. Metadata standards provide guidelines regarding structure, values, and content (Zeng & Qin, 2008). The metadata standards under consideration relate to:

- Patient Identity Metadata – These metadata relate to patient identity and include: a patient's name; date of birth; address; zip code; and relevant patient identifier(s).
- Provenance Metadata – These metadata would be used to provide information on the "who, what, where, and when." Provenance metadata would include: a tagged data element (TDE) identifier; a time stamp; the actor; and the actor's affiliation.
- Privacy Metadata – Privacy metadata would include a policy pointer and content elements descriptions such as data type (e.g., consultation note) and sensitivity (AMIA, 2011).

Metadata is foundational to healthcare data trustworthiness. Various sources of big data are generated by all key healthcare stakeholders who have the potential to create unimaginable amounts of data from structured and unstructured sources of data. Where administrative claims data (e.g., financial, procedure codes, place of service, demographics, etc.) were once the primary source of data for healthcare decision making, the underuse of unstructured sources of data puts organizations at a severe competitive disadvantage. Data quality and origination loom large in the reformed healthcare market. With competition for healthcare consumers and limited financial resources, healthcare organizations, including hospitals, Accountable Care Organizations, and technology vendors must share data and knowledge to remain viable. Systems integration, or interoperability, of fragmented information systems is the conduit to information sharing among stakeholders. While it is believed that the Volume, Velocity and Variety of big data are unmanageable, data about data, or metadata, is growing twice as fast as the digital universe as a whole (Burns, 2011).

Fundamentally, metadata helps interpret and transform data into information (Gudea, 2005). One kind of metadata is provenance (also referred to as lineage and pedigree), which tracks the steps by which the data was derived and can provide significant value addition in such data in-depth e-science projects (Simmhan, Plale, & Gannon, 2005). Metadata in the form of provenance information records the *how, where, what, when, why, which, and by whom* of data generated in a scientific experiment (Sahoo, Sheth, & Henson, 2008). Metadata provenance is broadly referred to as a description of the origins of a piece of data and the process by which it arrived in a database. Most implementers and curators of scientific and healthcare databases would like to record provenance metadata, but current database technology does not provide much help in this process. Databases are typically rigid structures and do not allow the kinds of ad hoc annotations that are often needed for recording provenance in an EHR and personal health record environment (Acar et al., 2010). Better understanding of how to create, harvest, and exploit metadata is a very near-term problem to be addressed by today's information management professionals. New capture, search, discovery, and analysis tools can help organizations gain insights from their unstructured data, which accounts for more than 90% of the digital universe (Burns, 2011).

Data Scientist

The term data scientist is a generic term that includes business analyst, data architect, engineer, and research analyst. Indeed, with the rapid increase in the Volume and Variety of health information, clinicians that interact with information systems departments are in high demand and the chief medical informatics officer (CMIO) and chief nursing informatics officer (CNIO) are recent additions to the ranks of data scientists. Even with these developments, demand for data scientists has raced ahead of supply. The shortage of data scientists is becoming a serious constraint in some sectors.

Roper (1988) suggested in his seminal article on data and health information that “the science of healthcare evaluation, still in its formative stages, requires certain resources: money, data, and people trained in the evaluative sciences, such as statistics, mathematical modeling, and epidemiology” (p. 3). The data scientist has received an excessive amount of attention with the emergence of big data. The definition of data scientist has many connotations. The National Science Foundation (2006) identifies the following capabilities as core to the role of the data scientist:

- conduct creative inquiry and analysis;
- enhance through consultation, collaboration and coordination the ability of others to conduct research and education using digital data collections;
- be at the forefront in developing innovative concepts in database technology and information sciences, including methods for data visualization and information discovery;
- implement best practices and technology;
- serve as a mentor to beginning or transitioning investigators, students, and others interested in pursuing data science; and,
- design and implement education and outreach programs that make the benefits of data collection and digital information science available to the broadest possible range of researchers, educators, students, and the general public.

Harvard Business Review touted the data scientist as the sexist job of the 21st Century (Davenport & Patil, 2012). The U.S. alone will need 140,000 to 190,000 people with deep analytical skills by 2018 just to keep up with the pace of innovation (Brown et al., 2011) and the explosion of big data. As big data emerges as a driver of value (Porter & Teisberg, 2006) for public and private sector companies across every industry, analytics is a competency required for essentially every position.

Pryor and Donnelly (2009) identified four data analytic roles: data creator, data scientist, data manager, and data librarian. They acknowledge that “in practice, there is not yet an exact use of such terms in the data community, and the demarcation between roles may be blurred” (p. 160). In their definition of these four roles the crucial words “training” and “formal qualification” are for the most part absent. Data creators are described typically as researchers who have acquired a high level of expertise in handling and manipulating data; data scientists appear to be working closely with data creators and may be involved in creative inquiry and analysis; and, data managers tend to be computer scientists, information technologists, or information scientists who have taken responsibility for the facilities necessary to store, access, and preserve data. Data scientists understand analytics, but they also are well versed in IT, often having advanced degrees in computer science, computational physics, biology, or network-oriented social sciences (e.g., social network analysis). Their advanced data management skill set — including programming, mathematical, and statistical skills, as well as business acumen and the ability to communicate effectively with decision makers — goes well beyond what was necessary for data analysts in the past. This combination of skills, valuable as it is, is in very short supply (Davenport et al., 2012).

CHAPTER III. METHODOLOGY

Study Design

This section describes the research design, data collection methods, and analysis approach used to conduct a phenomenological study using narrative (Clandinin, 2013; Amedeo Giorgi, 2009; M Van Manen, 1980), with the aim of discovering important categories of meaning about big data in healthcare through the insights and perspectives (Cyr & Reich, 1996) of three key healthcare stakeholder classes. To allow the study participant narratives to remain the focus of this study, a more detailed description of the research methodology can be found in Appendix A.

In exploratory qualitative research, social phenomena are investigated with minimal a priori, or presumptive, expectations in order to develop explanations of a phenomena (Guba & Lincoln, 1985). I contemplated grounded theory or another theoretical framework, but decided against doing so since exploratory qualitative research does not rely on the creation or adoption of a conceptual framework where the abstraction of the subject to be studied may alter or even not capture the most important characteristics to be analyzed (Guba & Lincoln, 1985). Rather than being constrained by a structured framework, I chose to stay true to the tenets of phenomenology, which allowed a cohesive ‘story’ to emerge from the frequent, dominant, or significant themes inherent in the raw data (D. R. Thomas, 2006). As such, this method required me to be able to thoughtfully, and unbiasedly, interact with the participants of the study and to better understand their individual and collective experiences (Creswell, 2009).

A purposive sampling strategy, which allowed me to exercise my expert judgment with inclusion and exclusion of study participants, was used to identify the best healthcare stakeholders to provide “thick descriptions”(Creswell, 2009; Geertz, 1973; Guba & Lincoln, 1985) about the big data phenomena in healthcare. The literature review complemented the discussion, description, and interpretation of the participant’s stories.

Study participant narratives were analyzed using a general inductive approach for qualitative data analysis (D. Thomas, 2003). This study produced three important contributions to the understanding of big data in healthcare: (1) thematized experiential knowledge about the meaning of big data in healthcare; (2) produced an agile definition of big data that can be deployed across all industries; and, (3) added to the dearth of scholarly qualitative literature about the phenomena of big data in healthcare for future research on topics including the diffusion and spread of health information across networks, quantitative studies, standards development, healthcare value chain analysis, and health policy.

As a rising scholar-practitioner who has been deeply immersed in many traditional and innovative practices of generating evidenced-based methods in healthcare including integration of patient preferences (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996), randomized clinical trials, and quasi-experimental studies, my current interests in big data in healthcare are exploring phenomena through multidisciplinary narratives (e.g., government, providers, and consumers) and subsequent scientific analysis of the collective data to ultimately inform further health policy. The results of this study confirm a natural collaboration and research agenda between the disciplines of information science and health policy, as medicine adopts the discipline of information science (Hood & Friend, 2011; Lester, Zai, Grant, & Chueh, 2008).

Such an approach resonated with my curiosity about the social and lived experiences of individuals who use and rely on big data as information and knowledge to meet their professional and organizational objectives. Borgman (2012) suggests “that by studying how data are created, conceived, handled, managed, and curated in multi-disciplinary collaborations, we can inform science policy and practice. Data are the ‘glue’ of collaboration, hence one lens through which to study the effectiveness of such collaborations is to assess how they produce and use data” (p. 7). This study was designed to answer the following research question:

Research Question

Q1: *Within and across the narratives of three key healthcare stakeholder classes, what are the important categories of meaning or current themes about big data in healthcare?*

Study Influencers and Rationale

In addition to the construct of epidemiology, my approach to this research study was initially influenced by a study design used in Cyr and Reich (1996), *Scaling the Ivory Tower: Stories from Women in Business School Faculties*, which provided powerful detailed narratives about “women’s personal choices, trade-offs, risks and chances that unfolded as they built their careers in competitive academic organizations” (p. 1). Independently, each story chronicled women in various stages of their academic career: early-career, mid-career, and leaders in academia. Most compelling to me is that *aggregately*, their stories were the impetus for action, policy change and influence for other women in academia and other fields facing the same trials of overcoming personal and professional challenges and the satisfaction of fulfilling dreams. Summaries of each story followed their narratives and a brief snapshot of each contributor, recurrent themes, interesting issues or challenges, and the lessons learned from their collective experiences was provided in a final summary chapter. This important study provided a “methodological blueprint” to guide my study approach to answering the study’s research question. I am hopeful I executed their methodology with the same rigor and preciseness.

As I began a deeper dive into the practical application of big data in healthcare, I was also strongly influenced philosophically by an emerging social dimension to medicine: narrative medicine. Traditionally, healthcare organizations have used troves of quantitative data (e.g., laboratory values), qualitative data (e.g., text-based documents and demographics), and transactional data (e.g., a record of medication delivery) to understand a clinical phenomenon of interest. Narrative medicine “describes the practice of medicine supported and reinforced by the ability to listen to, absorb, and act on stories” (Charon, 2006, p.1). I contacted Dr. Rita Charon at Columbia University. Dr. Charon is considered the foremost authority on narrative medicine. I believe our conversation was mutually informative; her perspective influenced my ideology about *healthcare narratives* which is fundamentally different from *narrative medicine*, which Dr. Charon describes as “a private conversation between a patient and a skilled physician.” I posited that *healthcare narratives* have a theoretical orientation that applies narrative inquiry skills across and between all healthcare stakeholders involved. Narrative skills are those that enable one person to receive and understand another person’s story, including the skills needed to listen actively, to understand what another person’s story means, to attain a complex and accurate interpretation of the story, and to grasp the situation of the other person and their perspective, in all of its complexity (Roscoe, 2009).

The next chapter provides details on the data collection procedures and each study participant “lived experience” of the big data phenomena in healthcare.

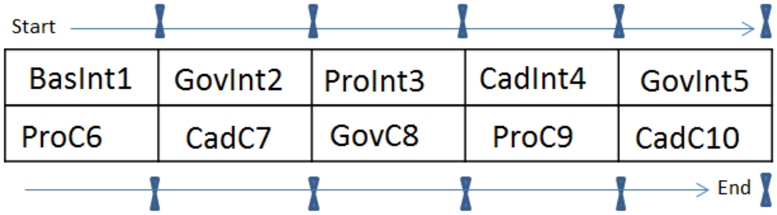
CHAPTER IV. DATA COLLECTION

This chapter contains a review of the data collection procedures and the data collected from the semi-structured interviews of the nine study participants. For an in depth description of the methodology, see Appendix A.

The unit of analysis is the narrative – narratives of individuals that have shared experience with the phenomena (Creswell, 2009) of big data in healthcare. Study participants were identified through a purposive sampling method. Semi-structured interviews were conducted with a total of ten key healthcare stakeholders: three policymakers; three providers; three consumers (advocates); and, one healthcare leader with a global perspective across the three healthcare classes (Figure 4). However, the global perspective interview (BasInt1) was omitted because it did not meet the established parameters described in the Interview Guide and eventually created a fourth stakeholder category that fell outside of the study design. Thus, nine interviews were used as part of the data explication, results, and discussion.

Sampling Frame

Boyd (2001) regards "two to ten study participants" (p. 93) as sufficient to reach saturation and recommends "long interviews" (p. 95).



BasInt1	GovInt2	ProInt3	CadInt4	GovInt5
ProC6	CadC7	GovC8	ProC9	CadC10

Figure 4. Interview sequence of selected healthcare stakeholders

Most studies of narratives are based on small samples, fewer than 50 cases (Bernard, 2006), simply because there is so much work involved. I chose cases on purpose – not randomly – dividing the sampling frame into three strata (e.g., government, providers, and consumer advocates). I selected three study participants within each stratum to capture their experiential narratives. This method allowed me to discover, describe, and interpret in detail themes, challenges, and categories of meaning that were similar and different across the subgroups (Teddlie & Yu, 2007). This sampling method is not to be confused with quota sampling, in which the researcher decides on the subpopulations of interest and on the proportions of those subpopulations in the final sample (Bernard, 2006). This was a small study that fit the purposive sampling approach.

Snowball Sampling as a Supplemental Strategy

When necessary to mitigate the risk of study participants falling out of the study, I relied on snowball sampling, which produced a sample of study participants through referrals made among people who shared or knew of others who possessed the same characteristics that are of interest to this research (Biernacki & Waldorf, 1981). Snowball sampling as a supplementary sampling strategy was invaluable as it allowed me to capture a geographically disperse study participant sampling frame but also required me to slightly modify the study design’s data collection method from exclusively face-to-face interviews to a mix of both telephone and Skype interviews. Such a modification was appropriate because the study did not require me to elicit emotions and body language through observation – only study participant narratives.

Recruitment

Key informants are people who know a lot about their culture and are, for reasons of their own, willing to share all their knowledge (Bernard, 2006). It is critical to be certain of the knowledge and skill of the informant when doing purposive sampling, as inappropriate informants will render the data meaningless and invalid (Tongco, 2007). During the recruitment phase, I made initial verbal inquiries through email, phone, and in-person, with sixteen potential study participants who met the following selection criteria (Table 5).

	POLICYMAKER	PROVIDER	CONSUMER
Title	Senior Executive Service (SES) (ES – Level I - Level V) Upper management	MD or DO Register Nurse Manager Hospital Executive	Director Executive Director Chief Executive CIO
Responsibility	Provide leadership in a federal or state healthcare agency that provides or supports the development of national healthcare policy	Provide senior executive leadership for a large integrated delivery system, accountable care organizations, or hospital	Provide executive leadership in a recognized patient/consumer entity; Advocate for healthcare issues or part of a multi-advocacy agenda
General Criteria across the three stakeholder classes	• Be of at least 18 years of age and be willing to participate in a qualitative research study; • Have at least ten (10) years of work experience in a healthcare related field; • Currently represent a federal government, provider, or consumer advocate organization, in the healthcare sector; • Possess a working to expert knowledge of “big data” and “healthcare” and possess in-depth insights into the current challenges and future opportunities for big data in healthcare; • Fully participate in both initial and follow up interviews; • Be willing to speak freely and engage in a conversational, two-way in-depth interview sharing rich, detailed narratives about professional “lived experiences” in big data and healthcare.		

Table 5. Selection criteria based on a purposive sampling strategy

CHAPTER V. KEY HEALTHCARE STAKHOLDER 'STORIES'

To maintain the confidentiality of each study participant's name and professional organization, I assigned a unique code for analysis and a pseudonym generated by an online tool to each study participant and provided a general description of the type of organization where each is employed. Each study participant's pseudonym is found under the title of their story. To further protect their identities, I deleted any references to their educational institutions, board appointments, research centers, and proper names of colleagues mentioned in their respective narratives. I also omitted references to geographic locations that appeared in the narratives. Before offering the key healthcare stakeholder narratives, below is a brief profile on each study participant categorized by their respective key stakeholder class.

Study Participant Profiles

Government Stakeholders

Mr. Peter Erazo is a director at a federal government agency. His role is to provide leadership, strategic vision, and execution around data, data analytics, and data dissemination. He has held a variety of healthcare roles.

Dr. Myles Renneker is a director at a federal agency. After completing medical school, he was assigned to work on projects dealing with quality, patient safety, and electronic health records. Beyond his medical education, he earned an M.B.A.

Dr. Matthew Blocher is a senior fellow at a government agency. His education is mostly in mathematics, physics, and chemistry. After graduation he began working in the life sciences industry and has been conducting research for more than two decades. As a healthcare thought leader he directs various scientific projects.

Provider Stakeholders

Dr. Nickolas Thompson is chief clinical information officer (CCIO) of a regional integrated delivery health system. His primary responsibilities are to sequence the health system's technology and optimize the data analytics of the organization.

Dr. John Boyken is an associate dean at a medical school. After medical school, he became very interested in informatics and computers and the role that information technology and information management would play in healthcare.

Dr. Barry Jensen is the chief quality officer at an integrated delivery system. He leads research that has an immediate impact on care delivery operations within the delivery system.

Consumer Stakeholders (Advocates)

Dr. Darwin Watkins is executive director of a patient-centered healthcare organization. He earned a medical degree and a master's in epidemiology. After epidemiology training, he became the director of a research department at a regional health maintenance organization.

Dr. Arnold Daniels is executive director of a non-profit patient advocacy organization that helps patients find money to pay for medical co-pays and premiums. He completed a doctor of pharmacy degree and has a master's degree.

Dr. Frances Milburn is medical director at a patient-centered quality association. His responsibilities include oversight of clinical informatics and quality improvement. With a public health background, his medicine background complimented work in chronic illness care from a population health perspective.

The following table (Table 6) is a brief summary of study participant’s profiles:

CLASS	TITLE	ORGANIZATION TYPE	EDUCATION
Government	Director	Federal Agency	Statistics
Government	Director	Federal Agency	Medicine/M.B.A.
Government	Senior Fellow	Federal Agency	Math/Physics
Provider	CCIO	Integrated Delivery System	Medicine
Provider	Associate Dean	Teaching Hospital	Medicine
Provider	Chief Quality Officer	Integrated Delivery System	Medicine/Physics/Biostatistics
Consumer	Executive Director	Patient Research	Medicine/Epidemiology
Consumer	Executive Director	Nonprofit	Pharmacy/Research Methods
Consumer	Medical Director	Quality Improvement	Medicine

Table 6. Profile of study participants occupation and education

Study Participant Narratives

The following study participant a priori narratives on big data in healthcare are presented in the study participants own words. The interview data was abridged without losing the essence of their stories. To reiterate, pseudonyms and generalizations of people, places and organizations were used to strictly protect the identity of each study participant. Narrative titles were chosen from the study participants own words that best demonstrated the spirit of each ‘story.’

Government Stakeholders

"A Whole Heap of 1's and 0's"

"Peter Erazo, M.S."

Professional and Academic Experiences

Mr. Peter Erazo is a director at a federal government agency. His role is to provide leadership, strategic vision and execution around data, data analytics, and data dissemination. He's held a variety of healthcare roles.

Meaning of Big Data

I think that frankly the expression of big data has become a little overused. I personally prefer the term "smart data," but if we are talking about big data it's traditionally defined by volume, variety and velocity. Again, I think for that breakdown I think you can have many, many important data driven activities that contain some but not all of these. I think obviously the rapidly emerging technologies in this area do allow people to crunch ever larger numbers of data in helping us bridge the gap between structured data analysis and unstructured data analysis, which I think is very important.

I think big data in healthcare can manifest itself in a number of ways. We can get the data to market quicker whether that's for internal analysis or distributing it to people externally. So, big data could mean getting researchers data that is weeks old instead of years old. Big data could mean routinely giving providers granular information of the beneficiaries they treat instead of shrugging your shoulders and not being able to do anything about it. Big data could be large scale hypothesis free data mining to maybe find an insight to correlations that weren't available.

Big data could mean the integration of administrative clinical and other patient generated data. So, it means lots of things in my mind.

I think the jury is still a little out, again to the extent that big data helps inform clinical files as far as effectiveness and real operational type medical decision making. Then, yes, I think it can help evidence-based medicine. As far as the attributes of big data that are different from traditional analysis, again, I think it's the ability to quickly secure in an agile manner to combine different datasets and have developed insights that we may not have from administrative data alone. So part of that is storage and part of it is new data matching techniques.

Medicine as an Information Science

I'm not sure I'm qualified as a non-clinician, but yes.

Healthcare Big Data Drivers

Practically one of the biggest drivers of big data in health care is the Affordable Care Act because what it does is places data and the ability to harness and leverage data at multiple points throughout the healthcare ecosystem at the center as opposed to at the trenches. Data used to be a byproduct of healthcare delivery. Now for successful healthcare delivery and healthcare transformation, data, it used to be you could almost argue it needs to be the center with providers and beneficiaries orbiting around it or at the very minimum it needs to be on the same level as what was previously considered the other core components in healthcare delivery, clinical knowledge, etc.

Sources of Data and Data Scientists

The sources of data that we use are pretty varied even though I'd say that administrative data is the foundational component. It actually meets the volume and the variety criteria we use records for multiple parts of the Medicare system, the Medicaid system, the enrollment data, hospital data, physician data, assessment data, laboratory data, Medicare data, and Medicaid data. I know we'd obviously be interested in adding other paired data to the mix. Then there's survey data and there is some pretty rudimentary Meaningful Use attestation data but we don't have any actual Meaningful Use data yet.

We're working hard to integrate quality data for the various cooperative reporting mechanisms and it's important we get a reliable clinical data stream we'd obviously be interested in incorporating that. So, that's what we work with. Again, everybody's conception of big data is different.

I have people who manage big data for me. I have a team of skilled data scientists who are part IT knowledge, part systems integrator, part subject matter experts, part analyst programmer; a data scientist isn't necessarily one person. You have a data scientist practice in which people specialize but talk to each other but you might have somebody doing the IT integration stuff and another separate subject matter expert and another programmer. So to put it in perspective, again, I know that some people consider big data not to be "big" until it's in the trillions, but we manage 400 billion discrete pieces of information that talks to each other pretty well and pretty efficiently, and it's growing by about four or five billion data records a year.

Organizational Challenges

I think the main challenges are cultural and leadership for this agency to truly transition to data-driven decision making. The impetus or the commitment is at the very top and its other people's sense that the commitment is not there. Ultimately, data driven decision making won't gain traction. Another challenge to data driven decision making is that occasionally government agencies are not necessarily in control of their own destiny and they may be subject to external political pressures that render data driven-decision making moves. These are the biggest challenges.

Unintended Consequences

I think one of the unintended consequences in the case that I have seen is that people think that big data is a panacea and again this gets back to the mix of human capital that you need to integrate big data successfully into your enterprise. I think there's a mistakenly held belief, not necessarily at my agency, but you know among other aficionados of big data that if you just install a minute stack that everything will magically be solved.

I think another unintended consequence is purely relying on machine learning without the application of subject matter expertise and also the application of a clearly defined set of goals can lead to an organization of big data actually distracting an organization from its core goals and outcomes.

Big Data's Future in Healthcare

I think everybody's hope is that in five years' time, there will be widespread integration of administrative, clinical and patient generated data that will be available through big data; it's assuring that the right person gets the right data at the right time and the right format for them. So it will be a big plus for analytic purposes directly to patients if that’s appropriate to providers, etc. while obviously obeying all privacy laws and regulations. So the one thing I know very little about is that people tend to get excited about biometric data. I’m not even sure I know enough about biometric data to get excited about it. But I know when people talk about big data they mention that a lot. I think also integrating device interoperability and the data that comes from medical devices is potentially very important.

Metaphors and Symbols

People like buzzwords but there’s no question that we’re dealing with great volumes and types of data than we ever have before, and we have the tools to deal with it. I think that the true challenge is you can have all the data in the world but until you translate it into actionable information, it’s really just a whole heap of 1’s and 0’s.

Closing Thoughts

I think big data is an area of incredible promise for healthcare that is also currently fraught with hype and over promising. So there will be hits, there will misses, and hopefully again in five years’ time, we’ll have a lot better idea of what exactly we should be doing with all this data.

“Mapping the Knowledge Base of Medicine”

“Myles Rennaker, M.D.”

Professional and Academic Experiences

I’m a physician by training. I had a mandatory service requirement and was assigned to work on projects dealing with quality, safety, and electronic health records. I became very interested in that and I went to business school instead of going back into residency thinking that there were many things that were going to change about healthcare, including increasingly information technology changing healthcare which was apparent even back in those days and the whole quality issue became fascinating to me – how you actually measured clinical performance. I’m interested in the issues of quality, safety, and how you can use IT to enhance the quality and safety of care including through electronic health records.

Meaning of Big Data

What big data means to me is just using information technology to analyze databases that have large units of whatever it is, whether it’s patients or accounts, or customers – just getting beyond small scale and having very large volumes of data to analyze. Nobody’s ever defined it for me. I’ve heard it used a lot and I guess that’s what I’m thinking it means. I would also say that I never thought about it until you asked me. I just assumed that I sort of knew what it was.

The advent of our increasing capacity to store things and the processing speed has allowed us to do things that were very hard to do even a fairly short time ago. I can remember working with computers and processing stuff where it would actually go overnight and at least in the realm I’m familiar with you don’t have to do that very much anymore, you can process so many transactions, so many records in such large databases and it’s so fast that I think that has ushered in the concept of big data. I do think that to some extent big data has become the latest buzzword, the latest fad, the latest craze, and to some extent I don’t know how much new there is in big data other than the fact everybody is getting excited about it. At many conferences they talk about big data as if suddenly somebody invented big data and then came along and it’s a new thing.

It really is an evolutionary thing and I think that it has a potential to perpetuate a myth that persists in IT generation after generation: That if somehow information technology can sell substance problems that people haven’t put their minds to, the computer just does what you tell it to do and if you haven’t solved the problem of structuring the analysis right, the computer is going to do it for you.

An example is the electronic health record where we’re very poor at structuring clinical information so we come along and we turn everything into electronic form and we somehow expect that electronic records to solve all our problems and it doesn’t do that unless you think through how you’re going to structure the data before it goes in and what everybody else is doing. You’re going to have big data and right now there are over 2,000 records that have been certified by CCHIT as meeting the Meaningful Use Stage One criteria and they’re all written in different languages, different interfaces, different databases and they can’t talk to each other. So it’s kind of a mess.

Medicine as an Information Science

It's clearly an information science. If you're going to treat a patient you're going to use symptoms of the patient, you're going to use physical findings from an exam, you're going to use laboratory values, you're going to use imaging, and those are all data. But at the same time they don't all get put into a computer and processed to get the answer. The computer is not a human brain and while we have computers that attempt to match many of what people do, much of what doctors do we don't have computers that can do all that doctors can do and that final step of processing, especially in complicated cases really needs to take place in the human mind, but it is processing of data for sure.

So I would say yes it's an information science, but it's one that has not been entirely encompassed by man-made; it's aided by man-made IT.

Healthcare Big Data Drivers

I don't think big data has had the kind of impact in healthcare that it's had in other industries and that doesn't mean I don't think it can down the line, but I think we haven't structured the information in healthcare to the extent necessary to allow big data to have the kind of impact it will potentially have on the future and it's not an indictment of the healthcare industry. So many people are critical of healthcare and say healthcare is in the 18th Century and healthcare is extraordinarily complex. I was giving an international speech in Europe. While I was talking about measuring quality and safety somebody got up and asked me, 'well why don't you do it just like a bank has an ATM,' and I didn't laugh but I felt like it. I just said because everything isn't an integer, it's not as simple as a balance sheet or income statement, or a checking account. It’s a whole different ball game and the relationship between processes and outcomes in healthcare is not totally defined. You can give the same drug to two patients in the same way, the same age, the same diagnosis and they’ll react differently, and it’s just we’re not making patients.

I mean in most other industries, the service is defined by the industry or is produced by the industry, we’re dealing with patients that are highly complex organisms which in some respects, many respects, black boxes. We know something about them, but we don’t know how they’re going to react to everything and they have many complicated problems and it’s all underneath the surface and we have to do diagnostic tests to get a little bit of it. So healthcare is enormously complex and so it’s just a whole different realm.

It’s not like big data allows the retail industry to behave differently just by the volume of processing because we’re still not processing things that are very elementary in other industries because we haven’t structured the knowledge to be able to go into the computer. For instance, I’ll give you a concrete example, let’s say we have three different electronic health records, three different offices and they get three patients in there with abdominal pain, an elevated temperature and elevated white count have tenderness in the upper right quadrant of the abdomen, well those are the classic signs of appendicitis. So the way that information first of all, most of that probably gets put in the lab IOB and the temperature will be in there, the patients symptoms will be free text, it won’t be probably won’t be in a defined field and there’s no program in there that says this is the definition of an acute abdomen or even with the probability of 95% or whatever, it’s the definition of an acute abdomen and therefore you should think about appendicitis. Those laboratory values will just sit in the lab area of the electronic record. The temperature will sit in the vital sign section, the narrative will sit in the narrative, nothing ties them together, there's no way to compare, there's no way to go into a database of thousands of patients and say how many of them had an acute abdomen. The data aren't structured that way. Could they be? Yes. So I think big data is not able to move things in healthcare the way it is in other industries.

Having said that, the drivers that are pushing IT, getting us more into big data that will invite us to try and answer the questions that will allow computers to be more helpful are certainly driving costs. The question, healthcare cost is making people say we've got to marshal information technology to make this whole power of data more cost effective and produce more for our providers.

Then the increase in technology, the improvements in technology for other purposes as well as in medicine are really, really good. The improving technology is making it easier to do the things that you need to do in healthcare to be of more assistance to the people providing care. So I think that's changing. What is not happening in an organized structure way is to try to analyze clinical medicine and represent it electronically in defined fields so that everybody can talk to each other and we could represent all the complexity in medicine. I can't ever foresee a time when you won't want to have the ability to collect narrative for at least some of the electronic record. But we need to get, right now probably the majority of most records, it's certainly true, the majority of most clinical information records is in narrative form and you can't use big data on. So we need much more of a structured knowledge base and that work isn't really going on in a very organized way right now.

I want to get people to use the same definition for enough time so that we can aggregate data, match it up against reality and then refine the definition so that the sensitivity and specificity of it, the accuracy, when the true positives and not a lot of false positives, not a lot of false negatives, so that all gets worked out by making the definition very precise and having people record it that way. It's the opposite end from natural language processing which says put it down in a precise or sloppy way we don't care about. We're going to go in and search for whatever we can find and we're going to hopefully be able to find things that are similar with a clear degree of accuracy. I want to go on the other end and say we're going to be very precise and then we're going to use that precision to refine the definition over time based on big data.

I've actually been engaged in such a process. The first thing you found out about is whether the standards worked or not, and so you could actually refine the measures by processing large amounts of data against those standards and validating it with the actual real life circumstances and that way the definitions could get more and more precise over time. We need to go through the whole knowledge base of medicine that way and map it. No one is even talking about doing that right now so we're a very long way from getting medicine to the point where we can do the kinds of things that they can do in other industries where the structure of data is simpler.

Sources of Data and Data Scientists

All of those skills of epidemiologists, biostatistician, physicians, all of those things are important skills. What I found in the quality area is there's a kind of unique skill of being able to think logically and distill the measurement process into binary form so that words like 'consistent with' or you know anything like that can't be measured. You have to find a way to triangulate what you're after and use binary thought processes to try and reduce highly complex situations to something that can actually be measured in concrete terms. So I don't know if that makes any sense but it’s sort of like you look at the beautiful color that you have on your module on different colors and it looks analog really, but at the end of the day you got a machine language that’s all 1’s and 0’s. When you measure quality, everything has to be in 1’s and 0’s at the end of the day. Then you have to realize that it’s as good as you can get with it. You always have to be humble about whether you’re right in an individual case or not, but the better you get with measuring so you can at least be right about trends and populations.

Organizational Challenges

I have found it difficult to find clinicians who have the ability to stop practicing medicine and to turn around and think about things in very objective binary ways. It’s not impossible but it’s hard. But one of the things when you’re looking at quality, you’re basically looking retrospectively. If you want to do big data, it could be populous in real time patients, but then you have problems with denominators and patients that are evolving. If you want to look at a population where it can be static and you can have denominators that allow you to draw conclusions, its material that’s going to have to be completed at some point in time and to get people to look retrospectively and think that way instead of thinking prospectively on the terms of uncertain conditions. That it might seem it would be easy to do, but apparently it isn’t so easy.

For instance, you can’t use pathology reports to find out whether a surgeon made the right decision to operate because you didn’t have them at that time or she didn’t have them. You have to use the presenting symptoms and lab values and so on and so forth, it’s a time the decision had to be made to operate or not, so that may sound simple but I’ve actually tried to set standards with people to say things like just go use the pathology report, you can find out whether the operation was needed or not.

Big Data's Future in Healthcare

My hopes for big data in healthcare would have to do with the fact that I hope we have common standards for how to represent the major clinical problems that patients have, both processes and outcomes of care, so that when electronic health vendors revise their programs, they write to those common standards and data get collected in defined fields in electronic records in a way that we can begin to compare apples to apples and we can begin to understand that what we're doing with treatments across the board because the results from one record can be compared with results from another record.

Also, that incidentally would make transferring information from one provider to another a lot easier. Right now, we have thousands of different health records and then a handful of other major vendors and then a whole bunch of do it yourself. Overall, there's just an enormous variety of electronic records out there that are not interoperable and can't produce information that can be benchmarked or compared or learned from really. So my idea would be that information could be moved more easily, could be benchmarked, compared, trended over time and I don't think that's unique to me – everybody has that vision. But I think it's going to take a little longer because I think the complexity of structuring the knowledge base of clinical medicine is a job that we haven't even defined how to do that job yet. Nobody has said much about doing it in a regular way.

Closing Thoughts

I have one last thought and that is in healthcare we tend to spend a lot of our time analyzing healthcare data that exists in electronic form knowing that it's incomplete and inaccurate to do the job, mainly billing and administrative information and sort of throwing up our hands and saying we know the analyses aren't very complete because the billing data doesn't have everything. But it's the only data we have so we're going to use that and we're going to base judgments on it. Since the billing data represents probably some tiny fraction of 1 or 2% of the clinical information about a patient in any setting, those data are not sufficient to make the kinds of judgments that one needs to make in terms of quality, safety, reimbursement, or policy.

So I think that when we get to the point where we define the data we need to then figure out how to get it in an efficient and effective way. We'll be far better off than saying okay what data do we have, how can we shoehorn that in, or try to stretch it to make what we need to do. So on defining what the objective of whatever endeavor we're in, whether it's quality or safety, or policy, defining the objective then defining the data that we need, the questions that we need to answer in order to drive that objective and then getting the data to answer the questions, doing so in that order instead of taking the data that we have is an essential step in moving this whole field forward.

We have been churning in terms of analyzing, re-analyzing, and making more and more powerful sophisticated programs to analyze administrative data for 30 years now and we haven't really moved along very well because the essential information you need isn't in electronic form.

"My Big Data – Your Big Data"

"Matthew Blocher, Ph.D."

Professional and Academic Experiences

My education is mostly in mathematics, physics, and chemistry. I graduated college and worked in the healthcare and life sciences industry and have been working in that field in one way or another ever since. It wasn't a trajectory that was straight into the medical field, but physics, mathematics, problem solving, handling data, and understanding analysis is one of those skills that you can apply to just about anything. It's one of those things full of interesting doors that opened and once I got into it and really understood what could be done, it was a lot of fun. I was really into the mathematics because it was much more rewarding.

I started out primarily as a drug discovery analyst, a person who was doing computer aided drug design in a lab and helping other researchers do their research. Basically, I did the computational part whether it was designing drugs or explaining how proteins interacted and doing simulations. I quickly understood that one of the biggest issues that I had interacting with people was trying to explain the amount of data that they had and how much I generated to them, so I started looking into visualization as well and got more into the graphics and visualization as I tried to communicate more and more information to the investigators.

Meaning of Big Data

I'm going to be like a lot of the folks that I'm reading on a lot of the blogs right now. Big data has become I think an over-bloated word. What I mean when I say big data is ingesting and integrating lots of data, lots of complex data that may be able to be used to answer questions either more rich questions, or answer questions more deeply and get down to the causes as opposed to just kind of scratching the surface. I'm literally interested from the etiology all the way down to the molecular and cellular levels. I have data coming in about your age, your background, your genetic background, hopefully the microbiome and all of the bacteria that live inside of you. How do those interact? When I start looking at not just the data, but all of the possible connections between all of the data, then I have a huge explosion of the data space that I need to explore to be able to find answers to the questions that I'm asking and trying to eliminate red herrings and false starts quickly. To me that's big data.

If I can answer those questions, it can then lead me to more relevant questions of causes and the etiology of the disease. Once I understand, if a particular gene is mutated in a way that isn't necessarily obvious, that it causes the problem but it leads to something two steps down in its pathway, I now can develop a drug against that and correct that disease. That's something that's important and right now we're not able to easily mine that. I'm searching for the holy grail of biomedical research, to be able to go and say I can find those hopefully, true associations and then we can ask the critical question that really is, if you will, the question to be able to address that disease.

Medicine as an Information Science

I would say that information science pre-dated medicine and allowed medicine to prosper. Until you were able to collect evidence objectively and to classify that in terms for differential diagnosis, the idea of classification of information and really the application of what your definition of information science is actually allowed, in my opinion, the development of medicine. So of course the answer is yes; it's absolutely an information science domain.

I'll take some of the examples that we've gotten recently and trying to get into. That kind of streaming data coming in, that kind of availability and the fact that it's the human view of the data and it's transferring, it's gaining knowledge out of that data, transferring it to the human so that the human can actually do something useful with it. At the same time there's a cultural shift going on where people are much more willing to share it.

You know it was taboo to talk about things like that let alone put it on a public space where the whole world can get to it. There's a real shift where people are much more willing to post their genomes online. I could just go to Amazon and pull it down and do an analysis, but that's clearly what a thousand genomes project and now you know the 10,000 genomes project and all the other projects are going. People are now making data available in the hope that somebody can come along and use it in a much more meaningful way. Further, we're now recognizing even more acutely that it isn't the professional scientists that will always find that link. There are other people out there that are citizen scientists and allowing them to have access to this data as well. They may come up with a solution that no one ever thought of.

Healthcare Big Data Drivers

I think there are a whole lot of drivers to this. The data complexity, the data volume is increasing and the richness of what is there is increasing to go out to ask questions we never could ask before. We're also collecting a lot of junk but clearly that's the big deal, right? You go gold mining and it's not all gold.

I think genomics and the ability to do sequencing is the first statement of it. I think that monitors that have come from telemedicine are really driving that as well. I no longer have to own all the computational facilities to be able to store my data and to analyze my data. It can be distributed around the globe and I think that that's driving the decision for people to both collect and store a lot of the data. I think that clearly the internet is a huge factor. We also are have an aging population that grew up relatively privileged and they're viewing mortality differently now.

Cancer and other diseases are big problems and I think too that the change in lifestyle we have where we're starting to see metabolic diseases are much more prevalent in the world. We're starting to see what were at one time typically western diseases or health issues becoming a global problem. A problem that was let's say antibiotic resistant, the disease that occurred in some small country in Africa that people in America didn't care about it and now all of a sudden within 12 hours that disease could be sitting here in New York LaGuardia Airport and spread across the United States just like SARS. I think that was a giant wake-up call for people. So we're realizing that focusing just on a small area is not going to solve this. The problems are global now and the data has grown globally.

Sources of Data and Data Scientists

I bristle just a little bit at the term 'data scientists' because every scientist whether they're a professional scientist or not is a data scientist because a scientist without data is a philosopher. So I understand what people are saying. But at the same time it kind of lets people off the hook that if they're doing science that means that they don't mean data. So going back to where does data come from? We generate a lot of data here and I’ve been talking about a data tsunami coming ever since about 1999-2000 as we were just getting ready to see the human genome project made publically available. We were getting lots and lots of data coming in from the genome sources from that point and we were ingesting it and trying to be able to analyze it at that point. You also have MEDLINE/PubMed, you have the National Library of Medicine has a ton of information that they store and they serve up free to the public. You’ve got a lot more sources coming in now, such as I said in the Human Genome Project, you’ve got the Human Microbiome Project, you’ve got European projects, even the Chinese now are starting to contribute and make their data available. So you’ve got a lot of information coming in from just the research world, but I think the healthcare world is starting to throw information out there as well and I think people making their health histories available through direct consumer marketing like at 23andMe.com where you send in your DNA and they start giving you information. People can argue whether or not that’s a good thing or a bad thing. Of course everything can be abused in one sense or another so you know there’s much more of that risk. And if you have children and there’s a genomic disease or whatever I think people are just much more aware and motivated to go in and try to explore. So I think that the source of the data is coming from everywhere.

I think that we’re starting to see the boundaries of the different sciences break down which is a good thing. You know the people used to be in either medicine and research and biology or another discipline. Now it’s crossing back and forth. Chemistry crosses back and forth; physics comes in and crosses. There’s a lot more information now for people coming in and bringing in physics information into what goes on in oncology and what goes on in various other fields. Cross fertilization is bringing new views into the questions that we ask and the answers that we can get.

Organizational Challenges

Organizations don't want to share and we're not necessarily incentivized to share. I think the monetization of big data causes some biases and perhaps maybe even causes certain things to be left out which might be critical, even in a large federal organization. Just remember a large federal organization is made up of people and you've got people who are trying to advance their career, they want to keep their job, they want to grow their lab, they believe their research is at least as important if not more important than everyone else's, so they want to drive that. That means having a competitive advantage over somebody else and in today's world that is information and sometimes that's data, and especially if I haven't mined all of the data; therefore, I want to hold onto that data forever because there may be another nickel I can get out of it. That's an unfortunate view the world that's short-sighted in my opinion but then that is human nature and I understand it.

All too often, the more data we have, the more fodder we have to beat it into submission and say what we wanted to from the start. Organizations, like people, suffer from such biases and challenges. Big data can also be used to open new areas to question and suggest new alternatives. People have shown that, in the case of ulcers, that there's a bacteria involved. We get caught in research bias and so we get rushed. I mean you can take big data and you can use it with your blinders on to prove a lot of different things, or you can take the blinders off and be surprised.

So we all too often have our blinders on. But I think it's mostly, even in a large organization like the federal government because it's made up of people.

Data Sharing

We do share, but in my opinion we don't share enough. No one has the answers by themselves and sharing data, sharing experiences, working together, working across in collaborations is a way that we really drive findings and unexpected findings where you didn't realize that 'A' and 'B' were connected because I've been studying 'A' for 25 years, you've been studying 'B' for 25 years and we never talked. But if we can make the data more available then it could be a person who's never done research in either 'A' or 'B' but mined the data set and came back and said did you guys know that there's a giant correlation here. But we don't incentivize that. We're just beginning to develop organizational programs to facilitate 'data science.'

As much as I appreciate privacy and I really like to be private as much as I can, we have to be able to share that data and we have to have as many eyeballs looking at it as possible. Are there going to be people that may do various things with it? Yeah, that's life. I think most of the barriers are cultural and legal as opposed to technical.

Unintended Consequences

So, let me tell you something what's going on, a transformation I've seen over the years. Very often in computational sciences some people are doing theory and they crunch on computers and do math, and there's those people over there that go in the lab and they do lab work and they generate data. Then it became people in the labs realized they needed these other folks and started to collaborate with them. That progressed to maybe we should incorporate them into the lab and started hiring some of them, but yet the problem was that data was still completely localized. So now there's a lot of folks who were bioinformatics professionals who were at universities who don't really have labs but they have now been able to say I can call up 'Company A' and get cell lines, I can have that company send it to 'Company B' and do the sequencing, I can have them send those cells to 'Company C' and have them look at proteomic analysis of it and I can have company 'B' and 'C' send me all the results so that I can do the data analysis. I never did an experiment and I never interacted with an experimentalist, but I have data and I'm now integrating it with all the other public data that I have and I'm getting some really interesting results.

We find the similar type of things here where often times we're asked to analyze one type of investigator's data and then we're asked to analyze a different investigator's data and we're going say maybe you guys should talk because we're finding commonality between them. And if that data was put together in a larger context then even other investigators that might be a little more inclusive to actually advance research and drug discovery and hopefully cancer therapeutics or even diagnosis at a much more rapid rate. I think that one of the commonly discussed things is when Google says we can start looking at searches and we can tell the CDC when there's about to be an outbreak. That's an authentic unanticipated finding by mining big data.

I think as we start doing more and more global sensors and people share their life we'll see even a lot more. We're in a really exciting time to see an even bigger explosion and understanding of the integration of data from bacteria and viruses in humans because my guess is we’re already starting to see such an interesting ecosystem. The ecosystem includes all of this and how we rely on bacteria and viruses, they rely on us and other animals that this is turning into a really complex scenario.

Big Data’s Future in Healthcare

I never want to see a parent have to lose a child over something as stupid as something that we can solve in a medical sense. There are a lot of things where we’re getting ideas integrated with genetics, we’re getting information about environment, and we’re getting information on health. Clearly this web of all of these interactions and the interaction between you and me affects our health.

We’re looking at human-beings holistically. The reductionist approach I think as useful as it has been. It needs to be augmented, I won’t say it needs to be replaced, but it needs to be augmented, a much more holistic approach. I can look at cells all day long but until they're organized into tissues, into organs and into systems and then into whole species and individuals it really sort of doesn’t matter. So it’s understanding human health, understanding how choices we make in shifting policy decisions so that we put investment where it matters the most as far as human condition and I think letting people realize their full potential as far as health and happiness as well. My vision is if it can lead to opening and the democratization of health, because clearly we’re having a fight over healthcare in this country.

There’s a huge disparity in terms of economics and wealth in this country, but you know, we have to somehow make it to the point where everybody has a fair chance to healthcare and education and emotional and mental well-being and I'm just hoping that maybe the data that we come across will help humanize everyone as opposed to just the few and fortunate.

Metaphors and Symbols

In my opinion, we don't have a language that's commonly accepted to discuss this issue, and when people first started talking about these various types of problems whether it's the Jim Gray's talking about the *Fourth Paradigm*, or whether you're talking about data tsunami or you're talking about big data, you're talking about whatever other metaphor people use it's the problem that I don't have a common language. I'm trying to communicate often times to funding agencies, policymakers, other thought leaders in the field that I need resources or trying to talk to other people in the field saying I'm trying to prepare for this or I'm trying to deal with that, or here's where other people are at and we don't have a common language to say this is what we're talking about and I think that the reason because it's relatively young. We come up with these terms to start building up some sort of language and we use a term like big data.

Well now you've got a lot of other people come into the field who think maybe this is something either interesting to them, something that they should know about, something they haven't dealt with yet, but maybe they think there's a problem or hearing somebody else talk about a problem that seems similar to what they're saying and they're using that term; therefore, that would be the same term as they have, so you've got a lot of people who don't understand what the original context was that maybe the first person or first few people used for that language and then repeat it. That's why I say if I look at it today and big data has kind of lost some of the meaning that it had at the early part and maybe even gained, and eventually will probably gain maybe a specific definition to make it useful again. I think it's the problem that we're all trying to communicate something that we're seeing .We're all trying to describe our view of big data and you know we have different experiences and we're all going to explain it a little bit differently and so big data to me is complexity and difficulty in analyzing and understanding it. Somebody else is just here for pure volume, you know, and other people it's the velocity of numbers and sensors coming in. It's all of that, and I think that's just right now it's a complex phenomenon that none of us fully understand; therefore, you get a lot of different colorful terms.

Closing Thoughts

Big data is a tool to solve problems and answer questions. Like any other tools, it can be used both for good and for bad and it's just the reality we have to live with. There needs to be a central policy of how we treat these large quantities of data and how we share the data and I think we could go back to I think to your central question of data sharing and acceptable use policy. I think it basically comes down to a sharing and acceptable use policy that is going to be very critical about how valuable big data can be to the population as a whole as we go forward in the future. Clearly, if this is all held by one secret government agency and used as to invade our lives that may not be a good thing. At the same time if it's trying to keep us safe from nefarious folks who are out to hurt us, then that's a good thing. The debate continues.

Provider Stakeholders

"Always Create Data for a Purpose"

"Boris Jensen, M.D., M.P.H."

Professional and Academic Experiences

I started a Ph.D. track for physics, heard about medicine, applied to one medical school and they accepted me. I joined the university faculty in biostatistics and was appointed professor of biomedical computing. I built the first computer network for a school of public health and established the computer network for the biostatistics department. My organization is an integrated delivery system of hospitals and employed physicians: about a 60% primary care, 40% of specialty mix. We are a charitable not-for-profit intended to be extremely patient-driven.

Meaning of Big Data

Turns out, there are many definitions for that term. Let me give you three. I'm going to start with some of the other ones that are commonly used in the marketplace and then finish up with mine. One definition of big data is that you have truly stunning amounts of data, but it's very well focused, it's not random data at all, just collected for specific purpose but just in truly massive quantities. The data that you collect you then analyze looking for rare events that was actually one of the original meanings of big data, right. Another related one, is if you're doing genetic sequence you know what you have are enzymes that will cut up DNA and you get them cut at particular points but in random lengths and then you can sequence the resulting lengths of DNA. What you get out of them you can analyze to figure out what the original genome was. You're dealing with truly massive amounts of data in doing this. So that's the first class.

A second class is data mining. The idea behind data mining is that you found a bunch of existing data of different types just anywhere you could find them. It tended to be again very large amounts of data. Then you applied automated statistical routines to them in the belief that this would give you some sort of insights that you would find associations. It's really a hypothesis generating exercise. You find things that became useful. This particular one, I've come to the opinion from having done research all my life that good answers come from good questions.

The idea that you just run statistical software and it's going to happen by useful association. You have to filter this with so many spurious associations finding anything that's useful that it's not a very productive use of time. That's called data mining. Ten years ago, it was massively counted, about leaving some of these computer programs that now run down through the databases and find these associations an almost magical learning from it. It never really materialized. It even felt like you'd think it would. At least a chunk of the current emphasis on big data is the reprise of that. Now, this is the cynical side of me talking. You see a consulting group selling this as some sort of a black box magical solution to a not very intelligent system leader.

The third class of big data for me is the kind I find useful. Dr. Deming, from who I learned quality theory, use to say that 'aim defines the system.' That's the fundamental truth. That is particularly true for data systems. You build data systems, they cost so much money to actually collect the data it's quite expensive. They're built for specific designated purposes and it's fairly important that you know what the purpose is before you start. Well, Deming also said that you should organize the thing around the processes, so quality improvement of course is the science of process management. While some years ago we went through and analyzed all the care we delivered and were able to identify a series of processes that make up the bulk of our work. We started with a little over 1,400 identifiable clinical processes. We used our existing hospital and outpatient data to prioritize them on the basis of number of patients affected and health risks to the patient which turns out they have a really tight correlation to cost of care.

We organized it through our enterprise data warehouse which contains roughly about two petabytes of storage. But what it is – is patient care data done over time organized along these processes of care and then you use those data to understand and systematically improve your care delivery. Now when I say you use it to organize and understand and systematically improve because of the way it's organized any patient who comes in to receive care us effectively was on a trial. But another way of thinking about it, for every patient who comes in we track what happens to them. We know what happens to them. By the way that we've structured that system as we care for patients I can measure what its actual outcomes are at least within our population the way that we delivered the care.

For example, I could track for medications and for complications that aren't recognized in their initial approval process. I can also track the actual outcomes of care associated with a particular treatment. So when we sit down to counsel a patient, its informed consent and here are your treatment options. I can tell them actually here's what you'll get with this treatment and these are the results you should expect.

Now it turns out those datasets are fairly extensive, they take the form of registries or data marts. They're effectively a registry but many times end up with millions and millions of records just because we're tracking all patients.

Medicine as an Information Science

I’ve been saying for 20 years that medicine is inherently an information science; the better data you have the better you can diagnose. The more effectively you can select treatment, the better you can actually see those treatments. It’s unquestionably an information science.

Healthcare Big Data Drivers

I think of it a little bit differently. I just did a carefully designed data system that is very purpose specific. They’re designed for a specific utility a specific purpose and they just happen to collect massive amounts of data but they’re always for a specific purpose. There are an infinite number of data points I could collect. There’s effectively no limit to them, so good answers come from good questions. Nearly always to answer that good question you have to have data that matched that question. They’re very purpose specific.

Now once you have the data it turns out that you tend to get really rich data because, explicitly because they are the right data for clinical management, clinical process management, and many times you can take those data and they’re more likely to be useful for other unanticipated applications, you darn well better have the ability to modify your data systems on the fly because more often than not that’s what you’re going to have to do, you’ll find that it will point you toward an interesting question. You’d really like to examine in detail but then as you start to examine that question you realize that you’re missing a few critical data elements without what you really can’t interpret the data, and so you’re going to have to go back and somehow add those data in order to properly answer the question. So you build that into the structure of your data system. Ask questions and then generate useful data on the fly.

Sources of Data and Data Scientists

I have master's level analysts whose time is assigned to a particular clinical area and they're the main analytic resource for the clinical teams for managing and improving care, testing changes in care, deploying best care, and tracking performance in the system. Now it's the funniest thing on this, most of my statisticians have some computer science background and regard themselves as fairly competent data architects. So as far as I can tell all of the data architects see themselves as analysts but when you're more than past familiarity with both fields you're different, and they're radically different. So you got to make sure that you have both of those areas available to you because it's specialty knowledge, really profound specialty knowledge on each side of the line and you have to get people working where they are most effective in that regard. So part of my job is to manage that and defend it.

Organizational Challenges

What I routinely get is an administrator who can understand the budget but they don't understand why I get so excited protecting that professional environment for my analysts. Now it's easy to show the performance that you get by protecting it. But on the other hand, somebody has to know and be able to manage them.

When I talk about having a rigorous method, we figure out what data I need to manage the specific process. So rather than it's called availability bias rather than just using the data that I happen to have available because I'm already collecting it for financial purposes. I understand that's big data where you're repurposing existing financial claims data and then trying to make it somehow work for these other things.

As you might imagine, you can completely justify some fairly significant outlays that you've invested for purposes of clinical management. Now the fact that it also becomes a full learning system that allows you to generate through knowledge at a paralleled rate, that's just a really nice side benefit. So what you're doing is a mining aim.

Data Sharing

We are discussing with some of our colleagues what might happen if they were collecting the same data fields for the same conditions and the other thing it means, imagine if somebody raises a question about best care. Effectively, my routine care becomes the control arm of the trial and so I can run 'X' therapies in amazingly short periods of time if we decide that it's worth the effort to do it. We had a fight that cropped up in the system about two medications that you can use for community acquired pneumonia which one is best for a patient. We eventually decided that it was worth the effort required to run a full trial on them, a full randomized controlled trial and we put about 5,000 patients in about three or four months. The routine treatment under that protocol was the control arm. So you kind of standardize treatment and so routine treatment was the control arm and then what you do is you just inform the patient, get informed consent in other words and then you randomize them and just have two therapies there so it becomes just part of routine care. That cost has dropped to a fraction of what it was before.

I think of it as sharing at two levels. The first level you share is existing data and that depends a little bit on the current capabilities of the systems collaborating together. You simply share existing data, whatever you happen to have. By far the most common data are financial data, whether technically claims data, it's not purely financial, but mostly financial. So you share claims data from the system. Now the next step which we're starting to do is rather than just straining existing claims data, you take a step forward and you start to generate specific clinical data that has a lot more meaning as you might expect has a whole lot more meaning.

Imagine that we're addressing a specific clinical topic like diabetes mellitus. The argument is that diabetes is kind of diabetes whether I'm in New Hampshire, or Minnesota, or Oregon. When you look at it, we ought to be collecting about the same data in about the same way as my process management system. You see the whole key is to be able to justify this thing on a financial basis as a care management system. That's how I get the justification for it, that's how I get money guys to put up a lot of money because it costs money. You've got to design them to that purpose so that will get better clinical performance. If I remember the goal, the best medical result at the lowest necessary cost. So the way I hope to sell it if I don't wait for my colleagues to come get me is I basically hold myself accountable.

Unintended Consequences

Well, first of all big data is never used for its intended consequence. So however you care to classify that in terms of being useful and actually managing care is so badly incomplete and there's a beautiful theory you can relate to it, it has to do with what's called decision layer in a process setting. An unintended consequence is it tends to create a group of clinical partners for a massively cynical ends to make change very difficult because it destroys trust. It's interesting because it's not just insurance companies, you can argue that most of the report card scoring systems of people are out there creating and using these datasets trying to repurpose existing data somehow and when you evaluate them technically they don't produce an actual result. By that I mean they, if you measure the confidence intervals or the scores that they produce the confidence intervals are so modestly useless. This is actually pretty well known.

Big Data's Future in Healthcare

There's this concept that effectively every patient goes on trial because of the way the data systems are structured. Now the jargon we used for that is a learning healthcare system where you build the learning, the knowledge management and it's an information science tool that comes out of this you quickly learn is it's perhaps the key capability in a system like this, it's knowledge management. How do you identify best practice knowledge, how do you systematically and routinely deploy it into routine use. What it means is I get much better clinical data in a real-time feed. Now the next piece beyond that is when you're using these tier process models you use the clinical processes to drive your care delivery, you can use it to integrate research. So I can justify this stuff purely on a financial basis see, that's the idea behind it. And then how do you use the resulting structure to rigorously learn from your experience. That's the learning system.

I want to get to the point where we will run at least 1,000 published papers in a single year. And by the way that's quite reasonable, that's not unreasonable. See for me that's big data. But it's interesting, it's not random data. It's big but it's not random. A lot of people seem to think it's random; no, it's not random. I've got colleagues in some of the other big integrated systems if we can start to collaborate together and as we work out the content of those data systems together so that we share the data back and forth it will accelerate the whole process. So the things that I could run a trial on that it would take me six months, I mean that's compared to ten years right now, the structure I might be able to get them done in six months. Under that structure we could do it in three weeks or at least that’s the idea.

Closing Thoughts

We probably did cover it, but here’s how I would say it: big data doesn’t mean unstructured data. You always create data for a purpose, right. That’s the human creation. It always has purpose, you have to understand the purpose if it’s going to be effective. And then everything else is just a tool.

"Big Data Means Greater Truth"

"John Boyken, M.D."

Professional and Academic Experiences

I became very interested in informatics and computers and the role that IT and information management would play in healthcare but I had no training or background in it. In my first two years of medical school, I had the opportunity to work in a lab that was focused on using technology to help transform the way we teach our medical students and I found it just fascinating.

I then started my clinical years of medical school and really became very interested in general internal medicine, mostly in-patient hospital medicine, and I did my residency in internal medicine and during that year became reconnected with this world of the power of health information technology to the point where I said this is going to be a big part of my life and I did a two year research fellowship in medical informatics. I began to realize how important big data was not only in our clinical and research missions but especially for me very important in our education mission and how we could use the same analytic approaches, we could use the same structured data collection, the same storage techniques, the same warehousing even the same software analytics tools to begin to transform the way we measure our students, our house staff and our faculty as we do our patients and our research subjects and our genes and proteins.

Meaning of Big Data

So that's a good question because it is a popular term that means a lot of different things to a lot of different people. I would say what it means to me and what it should mean to healthcare it means two things. First, it means turning data into knowledge and insight, that's not a database, that's not a data warehouse, it's the actual analytics. So that's the first part carrying data into knowledge and insight, but that only gets you halfway there. Second, I think the other part of big data is actually using that knowledge and insight to change practice, to change what you're doing into big decisions. A lot of people are heavily involved in producing knowledge and insight from massive data sets but that last of actually translating what you learn into agile dynamic operational changes and informing what you're going to do next. That's the part that I think has the least maturity in all of this. It's the most exciting and potentially powerful part.

I arrived at this definition through experience. It's experience of building systems, building dashboards, synthesizing very large amounts of educational data and seeing the power or the lack of power that those conclusion could have by whether or not people were embracing them and using them to make decisions and implement changes or just using them to make slides in a PowerPoint presentation.

Big data is different from data. The type of competencies of the person who potentiates the big data, your analytics people and the research people answering these questions, their competencies are fundamentally different from someone who's working with small data and it's more about the analytics than that, the algorithms and the causality sort of detection than it is about things like more straight forward regression analogies.

But when it comes to the volume of data, that's arbitrary and it's really a spectrum. It's big and it tends to involve from a very engineering perspective, it tends to involve database storage technology that is not your standard desktop or even your standard relational database, so that's important. I would say that often times people approach big data with the viewpoint that it contains answers to questions they don't yet know whereas they approach small data with how can we answer this existing question using the data in front of us. I think that both of those approaches have opportunities and pitfalls but I'd say that that's kind of one of the differentiating factors.

Medicine as an Information Science

Yes, absolutely, medicine has always been an information science. But whether or not that information has been at the individual patient level or at the group of patients a provider takes care of or at the population level has been the things that have changed. So when we see the big data revolution we're seeing that transformation of the maturity of information science in medicine go from that individual patient, the anecdote to the types of patients, the whole constellation of patients I've seen my career, to understanding the relationship of clinical and biochemical data across population which is truth, that is big data.

Healthcare Big Data Drivers

I think that the availability of big data is certainly something that's driving it and that definitely correlates with technology, whether it's clinical technology to measure biochemical signals from people or sequence genes or sequence proteins or sample the air or whatever. The availability of data is one thing that's driving it. I'd say that the willingness to base decisions and planning on truth and the desire to have more finely grained and precise measures of truth is another thing that's driving this thing.

People want to know especially in healthcare, how we're doing, what is quality, what is safety, how can we be more efficient both to make our patients healthier, but also to make the care we deliver less expensive and more efficient overall. I think those things are big drivers as well.

Sources of Data and Data Scientists

That's a complex question in our environment. So since we are an integrated academic medical center in our school, in our hospital our one entity we don't have many political barriers between data that is in our clinical world, in our research world and in our educational world. In fact, our leadership is extremely committed to transparency of those data and through as many people having appropriate access to them as possible so that we can make better decisions and we can be stronger because of them. If we don't have access to these things it's a missed opportunity. That being said we have safeguards in place with our IRB and we have other data access request review boards that say what people can and can't do with the data and who can and can't see things to protect our patients and to protect our students for the most part.

We have fairly robust resources of people who work on the data and infrastructure, so we have a large central data warehouse team an enterprise data warehouse team and then in my group for education we have a full time person whose job 100% is to run our education data warehouse and to create all of our reports and dashboards. Then we've also just created in my group a new division of education quality in analytics who are the scientists who work off of the data, who work off of creating the analytics and using the data and the knowledge and insights to translate them into decisions about how to improve our students, our faculty and our patients.

The tenant of what makes a psychomatrician or a data analyst or a data person are becoming much more about these competencies of managing large data sets of implementing analytics, of working with warehousing and non-relational databases, so these skills are key and they're not easy to find in people. We don't look so much for content expertise, in our world it's not like we're going to go out and find somebody who necessarily is an expert in health data but they can learn that here. The stuff that we really are looking for is for them to have the ability to use these tools to figure things out to invent new tools and event new knowledge and new techniques etc.

Organizational Challenges

So the organizational challenges are about, you know, are related really to agility, right, the ability to keep up with all of the conclusions and knowledge that you draw. There are often organizational challenges although we've been pretty lucky in respect to them about transparency and people willing to share the data or people worried about sharing data or fighting for silos or turf we haven't seen that much here. A big organizational challenge that is often overlooked is that you need to create value from the data for the people who are contributing the data. For example, here for our medical students and our faculty they conduct all these evaluations of each other and these evaluations are very important, they monitor the performance of people, they monitor the educational quality and if they're entering all these things and they don't see the value that aggregating all this data and analyzing it provides then they just view it as just an annoying server they have to keep fiddling out. So if all we do are create tools that show our deans and our vice presidents what these data mean and we don't give any feedback back to the people who are contributing it, it’s going to extinguish itself very quickly.

Data Sharing

Almost all of our sharing is where we’re giving data to somebody else has been internal. We have physically integrated our education data warehouse, our clinical data warehouse and our research data warehouse, we said let’s take this beyond this step of sharing, let’s just integrate these actual data and eliminate all of the technical silos and that has been amazingly powerful. Especially in healthcare it’s hard to share some of this data outside. The good news is that the government and the state government, federal government and state governments, are beginning to take the data that they’re paying us for with Medicaid and Medicare and many other things and put it out there for us to use, for researchers to use and so when it comes to some clinical data, performance data you can actually begin to download big datasets publicly online.

I’ll give you an example of something we just did in the last few days. The health department publishes every single hospital discharge of every patient per year online and you can download this massive dataset. It’s like a one gig CSV file that has every single discharge, what the diagnosis was, what the procedures that were performed, what zip code the patient had, what age they were, their gender and the license number of the doctor that took care of them so we know who the doctor was and that is this giant dataset that we can use that the state is facilitating by putting it out there, it’s terrific, it’s awesome.

Big Data's Future in Healthcare

My hope is that we get increasing transparency that data portability across the silos of organizations, of research settings, of educational settings is key because we need our big data to get bigger. We need to actually aggregate this stuff. So to do that what does the future need – it needs standards. It needs standards for clinical data, standards for research data, and standards for educational data. That's beginning to emerge but it's definitely not there yet. We need reasonable and rational policies around how to protect these data but also how we can flexibly use and release the data.

Often times the barrier to sharing is not political or financial -- it's regulatory. I'd say that we also need the ability of the consumer whether it's the patient or the student or the research subject to have access to their own data and be able to do more with their own data if they want to be able to move it around or integrate it with some other source, etc. But empowering the people whose data it is should be an important value for all of us as we go forward. I think that one of the things that we're not yet seeing and that we should is so big data, especially big clinical data has enabled things like hospital report cards and there's a hospital compare websites where you can go online and say is this hospital better at hip replacements than that hospital and make a choice based on it. So we haven't seen, we've seen a lot of big data being used to produce these things but we haven't seen the general public embrace those kinds of things to make their decisions. So we haven't seen people outside of these ivory towers, outside of these research topics where experimental pilots or you know clinical improvement that's real but it's happening top down as opposed to bottom up. We haven't seen that sort of grassroots use of big data, you know, there's lots of stuff that's happening in the consumer side with Twitter using big data for cure locations and detecting trends. It would be great to see the people say let’s use this data to help us make decisions. They already do on Amazon; they should do the same thing when it comes to picking where they’re going to have their hip replaced. Our hospital was just ranked number one for quality and safety, so I can say that with confidence that they should use those data to make their choice to come here.

Insurance companies and that’s how they’re going to run their practice, that’s how they’re going to negotiate with insurance companies, and that’s how they’re going to make sure they’re doing a good job and these things have not been extremely present in medical schools so we are really interested in changing that. The Affordable Care Act and the whole direction of the content of data and quality driving how we evaluate how we’re doing and how we make course corrections makes that even much more important. But these are the things that absolutely need to be very prominent in medical school, they are the critical skills of the future physicians, the present physicians, and they’re not taught nearly to the degree that they should be in medical schools in general.

Closing Thoughts

So there is new science that’s only potentiated by big data and that’s a whole other thing. But here like in the clinical world or the educational world, big data means greater truth. It means answering questions that were not answerable well before. But it also does mean potentially really empowering consumers and that’s one of the most important things.

The integration of genomic data and phenotypic data, which is clinical data in the electronic medical record, is something that every academic medical center is racing to do because the answers are not going to be in one or the other. The answers are going to be in the combination of both and so that is absolutely the future, a very reasonable approach.

I think that we’re going to see a lot of start-ups in this space, a lot of companies that are going to race to fill those voids inexpensively. I think that the federal government is also going to play a role in all of this and they’re going to provide some views of data from their perspective. So it’s uneven right now but I think it will rapidly become more uniformly used. And it will become cheaper.

"The Art of Applying Information and Evidence"

"Nickolas Thompson, M.D."

Professional and Academic Influences

I graduated medicine/pediatrics and took on an internal medicine position and at the same time was doing a pediatric hospital rotation at one of the local hospitals. I got pulled into the research informatics side of the equation. We had competed for clinical translational science award for a couple years and I wrote the informatics section and we got funded. Then I had a very unusual opportunity after having done some consulting work. While I was doing my work at the university, I had a chance to go to the Middle East and work at an ultramodern from the ground up pediatric and women's hospital.

As the chief clinical information officer (CCIO), my responsibilities are more around sequence in technology over time into the future and also working on kind of re-orchestrating the data analytics of the organization and other jobs not otherwise specified. When I got here, a lot of people were using beepers and pagers and so forth and so moved them all over to smartphones so we can leverage that platform. I had them use usernames and passwords across a bunch of applications so I'm moving them over to single sign-on tools so that they can just tap their ID and get into the systems if they need. I saw them using a fairly old version of EPIC so I accelerated the path to get to EPIC 2014 just to get to contemporary code.

Meaning of Big Data

Well, I like the definition that Gartner coined years ago where big data is a high volume, high velocity, high variety information asset that demands cost effective innovation, you know, basically something used to enhance insight in decision making. So the key I think is sort of insight discovery and hypothesis testing. So what does that mean for us? If I look at better velocity it means that if we look at current systems and how we copy in the traditional data warehouse model, so EPIC is a MUMPS-based programming system; it's not relational so every day from MUMPS to Clarity which is their relational data base. We need something to be able to get that to load faster and we need something that is going to be able to process that in a velocity fashion. Then for the health system to have better variety it means pulling more than just the data that we have, public data, other forces of data not typically used for healthcare really for more of the hypothesis generation.

Then for volume it means accommodating the ever increasing deluge of data that's coming from our own data sources. EPIC is the obvious one, but there are other things like location condition-based services, patient outcomes, all the biomedical device interfaces we have in the organization sort of like IV pumps and vents and physiologic monitors and so forth. And then I suppose you could add another of the Gartner's V's, Veracity, meaning that all the transactional systems work properly when people enter things perfectly the way it's supposed to that doesn't always happen. I think that can sometimes be an issue in terms of trusting the data or people finding the system to be too inefficient so creating a separate data warehouse that are cleansed within themselves but don't come back to the main data warehouse. Sort of some of the traditional data warehouse problems that we have.

Then for healthcare, I think we need to do our best to learn lessons from other industries because we're not the only regulated industry in the market area, banking, insurance others are regulated and still using big data more than we are.

Medicine as an Information Science

My direct answer is no. I would say medicine isn't just an information science; it's also about smartly moving information around clinicians. It's the application science of information as well so you know the art of translating patient's observations when they come in with signs and symptoms, their complaint is the history of the physical exam, the art of applying information and evidence in particular patients. That human therapeutic relationship between the patient and the team, between the patient and the doctor and so forth, so I'd call it that medicine includes information science, but amongst other arts and sciences that it has to dip into.

Healthcare Big Data Drivers

It's probably going to be a combination of some things that other industries are seeing and some things that are very specific to healthcare. So IBM will say that 80% of the data that is deemed collected is unstructured and therefore potentially untapped until we use big data tools. Also, I've seen several times that 90% of the data that is currently being produced ever has been produced over the past two years kind of suggesting that we're in sort of an accelerated exponential growth of the amount of data that's coming to us.

But from a healthcare perspective there's one very, very important part of the missing piece which is value based purchasing and moving away from the lack of accountability of fee—for—service. This whole 'sign and forget model' to get the patient and then send them off somewhere and if they come back is more money for me. So moving more towards the database means that I've got to start showing in, you know, connecting the dots of things that are outside of my line of sight. As I take care of a patient I need to really make sure they're actually doing better because otherwise somehow it's not going to work out for my practice for example. So I think value based purchasing is a pretty big driver to find out what other things can help fill the gaps for me in terms of understanding what's going on with a patient, could be behavioral economics, it could a number of other things. I think another one is the Office of the National Coordinator has been pushing these Meaningful Use Standards and that's resulted in an abundance of data, and there's more demand on doing analytics with the data and they'll be actually in Stage Three more expectations around producing outcomes and you can't really produce outcomes without data. So I think that's going to help as well.

I think there's a desire to maintain a competitive edge, you see all that, you're just doing the old data warehouse thing and just like anybody else would because the invented tool is beginning to mature at the warehouse level but for those of us who are kind of embraced in data science and data scientists and trying to push the envelope I think that we'll be able to keep that competitive edge.

I think another thing is the fact that hardware is getting faster and is available at low cost points. We compel them to use it as a result looking to solve some of the data problems by throwing more hardware at it to be able to have it crunch faster through new software applications including Hadoop, MapReduce, and NoSQL that Google has had for a while. I think healthcare organizations are starting to understand a little bit the fact that they're sitting on a mountain of data that they're not necessarily tapping into that's not really being acted upon. So I think they're trying to figure out if we have all these people that we're paying in healthcare to basically collect and digitize all the information from the patients and the EMR's are we really using that data that they're collecting to potentially affect the patient's health. I think other things like new data sources including genomics, senomics, and other '-omecs' that are out there. Metropolomics, for example, are generating huge amounts of data that need to be processed in ways different than we have in the past.

Sources of Data and Data Scientists

I want to caution one thing: I think our health system, is still reasonably early in the trajectory relative to big data. I think we're kind of proceeding kind of cautiously. I can give you some concrete examples. We do some work with natural language processing like most people do. We're finding some ability to go from unstructured text to structured text. Imagine out of the million radiology reports that were generated last year or this past year we were able to take what was being dictated and turn that into a structured text that's in a CDA mark-up and it kind of ends up in an XML format and you know the natural language processing is helping us do that. I can use that to be able to do correlations with other things, even the imaging data to understand health. If I understand that this report is normal and that the image that I have here is normal, I can send both of those to a machine loading tool and basically, over time, get the machine to help me figure out what's normal and what's not normal.

I think that there are some key things in terms of our desire to get closer to real time. I mean it's really not very useful for me to identify that a patient is in need of something 24 hours later after the opportunity has kind of come and gone. So our looking at our current system that's 24 hours behind is helping us in some ways but really not, it helps us maybe more in a population health side, but not so much on a prospective what am I doing with the patient right at the point of care side. We are still very SQL dependent and are slowly shifting over to other options to embrace NoSQL and MapReduce. We're just starting to look into social media and geospatial data from tweets. We're trying to get an idea of the behavior economics.

Organizational Challenges

I think one of the problems is trust in data quality and data fidelity. I think people don't really know yet if this is something worthwhile yet. We put everything in a little black box and it comes out the other end and it gives me a relationship. People are not so sure if that actually is true or not. So to the degree, at least initially people will be able to use it as hypothesis generation and maybe the hypothesis testing is actually occurring on the standard enterprise data warehouse tools.

I think there's still a very limited skill set out in the industry in terms of the people who know how to do this, so it's going to be hard to recruit a team of data scientists. There are some programs out there but there are not a whole lot of people that come out through them. They're going to get mopped up very, very quickly. I think a correlation to that is finding somebody who's got 10 years of experience in big data is going to be pretty impossible to find. So getting experienced people, there's going to be a lot of on the job training and that's going to be a challenge for people.

Then there's no proof points yet really that are real concrete in terms of projects that are out there especially in healthcare, but in terms of what the outcome is if it's going to be something that will be feasible from an economic or even a regulatory standpoint is still a little bit of an open ended question. I think that's still out in terms of being able to figure out if that's going to happen. Then you know, all this work may generate a lot of reports but I wonder to what end it will produce data but the question will be to what extent will the data be useful to actually produce an outcome and I have some examples of things that we are working but it’s right on the tip of the hype curve right now and I think it’s not going to be the kind of solution that’s going to solve all data problems.

Finally, it used to cost billions of dollars to sequence the DNA and now we’re down to like $1,000 and then it’s anticipated that in the next few years we’ll be down to a $200. It’s going to be pretty useful to create an account where you can go to Google and look up your genetic code and figure out what things are associated with that.

Data Sharing

First, it’s kind of important to talk about what our capabilities are in terms of our set up. We have a computer computational predictive modeling set up that basically we use for personalized predictive medicine. So we have some of our staff that are taking vast amounts of data from clinical and molecular radiographic economic data to create basically models that can help inform decision support the doctors make every day and we use high performance computer cluster that has the typical multicore and we have 400 core, 50 CPU’s at 2.2 terra bytes worth for computational ram that have some in memory database management systems which is kind of the newer way of doing it and plenty of dedicated storage. So the center basically is going to leverage this parallel cost of computing to be able to do some of the mathematical and computational modeling that’s necessary. With it, we’re part of a collaborative developed to identify what’s in your DNA and how the patient presents where there’s a relationship that can basically be put into the EMR itself. So that particular project has a couple of parties you know, one is to from the EMR get a precise phenotype and the other one is to basically return an actionable genomic result so that we do something different with the patient so that the use case is something I can't give medication for a patient like Warfarin but I usually start at 5 mm but this particular patient I may want to start only at 1 mm because if I were to start at 5 mm, they would have a brain bleed, so it helps me to understand where I'm starting certain medications based on a genetic code.

We also are involved in a collaborative project that established a virtual data warehouse of basically it simplified data sharing by having a very reasonable similar data model that's federated across all different organizations and it has demographic data, physical measures, personal medical history, management treatments, diagnosis, health claims and so forth and basically this data model retains control and stores data and stores kind of standardization across all sites and people can use this as a tool to do their research. It's an immense data depository as you can imagine. I think the third example is the Whole Genome Sequencing component. It's more of the genetics side of the equation over the patient's life span and helps predefine clinical context based on the genetic information. So I think that's hopefully going to help us with neurogenetic diagnosis decision support in the electronic medical record so some of the things you see in 23andMe.com by maybe more sophisticated in terms of patient genomic test reports and that kind of thing.

Big Data's Future in Healthcare

Again, I don't think it's going to necessarily replace what we currently do. I think you're still going to need people who are going to have to, you know, the big debate is will it be to the point where I don’t even need extraction tools and I don’t really even need semantic errors, and I don’t even need data marts and data warehouse anymore because I can basically just take information in an unstructured format and then just put it into this box and it’s going to tell me how the data is actually organized and what the correlations are and what’s the approximation size and so forth. I think that’s a little bit too nirvana. I’d love to get rid of the infrastructure and not to even think about it and have systems basically think about it for me.

But I think especially probably in healthcare there’s still always going to have to be somebody who’s going to be the data steward, who’s going to really make sure that people understand what something means. I want to recast our current system into a data warehouse model. I want to turn that into a logical data warehouse that has your standard component that has an ETL in tune data warehouse and then starts giving us different data marks, but also for certain data sources can tap into the big data needs and then for others that are more real time I use more of an operational data store as opposed to a data warehouse.

So something that hasn’t been fully mapped out into the analytical processing scheme that I want or something that’s more real-time feed that I can actually act upon much quicker. So there will be some components that are real time, some components that are like data warehouse and dashboard based and then some components that are for big data for large data sets and for better insight. So I can go to my big data to find the hypothesis. I can go to a data warehouse to test that hypothesis and I can use my real time data to basically put that hypothesis into action with decision support.

Metaphors and Symbols

You may be somebody who likes the whole quantitative self-movement. You may get on the scale and it gets WiFi'd and set up to your computer or you may track your sleep, you may track your mood, what you eat, a number of different things that could be tracked and you could do that every day. That information is going to become very helpful because it will help a bit of phenotype documentation, so when we're trying to match it up with a whole series of EMR derived data or decision support, you know, it kind of gives us a better idea of behavior economics of what's going on with a patient. There are other things that we were talking about in terms of pills and medication compliance and there are all kinds of tools that are now making themselves available that go beyond just the actual bottle having some sensor in it. In other words, you swallow this and like a potato chip that activates in your stomach and sends out a signal to a little sensor that's on your skin and tells me exactly when somebody has taken a medication versus not and it's actually been ingested and digested. So those kinds of things will be pretty helpful.

I think other things in terms of matching patients up with clinical trials will be helpful as well, getting a better idea of simulation platforms when you're trying to figure out how people respond to different medications. I think the promise of the big data is the fact that you can use all kinds of sources whether it's social media or even peer view literature like what IBM Watson is doing where they just consume all the literature so people don't have to read it. I'd much prefer this because I can't possibly get through the literature; yet, there's some useful stuff in there and if I can have a computer absorb it and then I can just ask it questions and it can tell me well based on the literature X, Y, Z then I think that could be beneficial. So a lot of the personalized medicine type of initiatives that match the patient to a treatment requires a lot of information and data to make that happen all in real time. I think the key thing is if you set the issue, if you generate data that's great, but we got to make sure we generate the causal relationship as well. So I think that's always a challenge.

Closing Thoughts

The battle for Accountable Care Organizations and the Affordable Care Act, you know, is really being fought here and we're able to demonstrate for example that we can make money on Medicaid patients and that we can make money on Medicare patients if we look at a population base level instead of this individual fee for service. I think that's the thing we differentiate ourselves with, we've invested in IT infrastructure, we've invested in bundles of quality care and we've invested in care coordination and we're now able to demonstrate as a result of having done that. We can get better mortality numbers and better outcomes for the patients in a way that's going to be compatible with where the legislation is going as opposed to being forced into it. So I think the fact that we're in the big data equation now is just testament to the fact that we like to stay on the leading edge and we want to be able to help solve the healthcare equation as much as possible and help share that information with everybody else so that we can just take better care of patients.

Consumer Stakeholders

"Learn to Talk to the Patient about Data"

"Darwin Watkins, M.D., M.P.H."

Professional and Academic Experiences

I was a family physician working at a neighborhood health center and later in my career I kind of got talked into going and getting a degree in public health/epidemiology and I was actually interested in doing that because computers were just coming to the neighborhood health centers in those days and I was very interested even then in what you could do with computerized data from healthcare delivery in terms of beginning to understand your patient population and what were the common problems and what worked and what didn't work. So as far back as 1983, I could see that that was a very good idea. After I took my epidemiology training, I wound up at a place which over the next ten years just moved hugely into computerized data. From 1984 through really 2000 we made huge investments; whereas, when I first got there, you had to do almost all research by abstracting paper medical records. By 2000, just about everything was in computerized databases and all you had to do was link them together. You had a population of three million people and you could build registries and you could do comparative effectiveness studies and other kinds of functions.

We were at the head of the curve then; others were too. We had a very large defined population and really good databases even at that time and they just kept getting better through the 20th Century and then they got a full-fledged electronic health record and that took a little getting used to because we were very used to the computerized data systems which kind of backed up this simpler electronic record. So we had all the lab results, all the prescriptions, all the diagnoses, all the visits and visit types and all the procedures. Subsequently, we switched to a real EHR so we had all the notes and stuff was in a very different structure. But certainly we continued to be able to do richer and richer analysis.

We think big data are important because we think the kind of studies we want to fund are really best done in real world settings and the best way to do some of those studies without completely disturbing the natural setting. In the process we want the whole enterprise to take advantage of the big data from electronic health records and other computerized databases that these systems have with the active involvement of the patients, and the active involvement of the clinicians, and the active involvement of the systems. We have a particular notion called patient engagement and we want the patients to be engaged but we also want to take advantage of the bigness of the data that are now accumulating and answer important comparative questions.

Meaning of Big Data

To me it simply means lots of data, lots more than you're used to and you know, the reason big data is important is because without it you wind up with studies that are almost always too small. Smaller than ideal, because it's just simply too costly to go out and collect all this data on the very large numbers of people that you need. So we're hopeful that the existence of these big sources of data allows us to do studies in a million people instead of 10,000.

And the reason that's important is because first of all everybody feels more confident generalizing from an unselected population of a million people than from a much smaller population where you had to really work hard to get these people to participate in your study and give you the data, a much more selected volunteer population. So advantage number one is you've got real world data now, unselected data.

Advantage number two is we're really interested in how treatments work for individuals. In the past, because studies couldn't be that large nobody could afford to fund a million person clinical trials study. You really always had to settle for the average affect, the average difference. You know, I had a randomized trial and I got 100 people in each arm and the average response rate was 70% in treatment A, and 60% in treatment B, the average difference, and that's about all you can do; with 200 people that's all you can do, and it wasn't statistically significant. You know, never mind that each arm had people of all ages and all levels of co-morbidity and certainly they differed genetically dramatically. So if you can increase that tenfold, then you can begin studying the same comparisons but you can subdivide them into males and females, over 75 and under 55 with a genetic marker versus without.

So big data number one is usually more representative and number two it's much more powerful and allows you to zero in and get much more refined answers and ultimately that comes back to being able to tell the individual patient this is what works better.

Medicine as an Information Science

Well medicine could become an information science I think if the clinicians and patients got actively involved in it. I think, I like to imagine that back in the 16th Century when somebody went to the doctor that doctor had maybe a few books, but he also he made mental notes, or perhaps he kept written notes of his patients and he learned from patient to patient and he passed on what he learned, he kept it on paper, he kept it in his head, did his best to learn everything he could from each experience and passed it on to future younger doctors and took it with him to see the next patient. I think that with the arrival of the computer we could see the same thing but on a much richer, more detailed, more accurate, precise scale. So I think if clinicians got into that frame of mind, they'd pay more attention to what they were writing in the electronic health record. If patients got into that frame of mind they'd answer patient reported outcomes measures, they'd participate in randomized trials at higher rates.

So I think that healthcare delivery, medicine as you call it, could become an information science. It could become clinical research if the patients and the clinicians become willing participants, and I think most clinicians in the long term if they had time and were incented properly would be happy to do that. Patients I think it's going to take a little bit more work just to get them to accept the fact that a lot of the things doctors are doing to them they're doing without good evidence you know. They're doing with uncertainty and so I think that we have work to do and elsewhere there's work to be done to convince patients, doctors, delivery systems that clinical care really ought to be research, you call it information science. Everybody participates in some kind of learning.

Healthcare Big Data Drivers

Well I think probably the main drivers are a desire to be able to bill accurately, okay, so that's a huge driver of electronic health records believe it or not and the second one is a, you know, this rapid rise of performance measurement. So you know, one of the things that I saw drive the deployment of computerized clinical data systems in EHR was when NCQA began asking for all these performance measures and Kaiser wanted to monitor its own performance and improve it. To do that they had to be able to measure that performance in an affordable way, you couldn't have millions of people sitting down with paper records trying to figure out what the blood pressure level was, so you needed it in the computer. So I think those are probably the two biggest drivers.

I think clinical efficiencies lagged way, way behind and in fact I think it isn't necessarily more clinically efficient. It might be higher quality care but it takes much more time. It's not you don't wind up going home faster at night because you have an electronic health record; in fact most people say the opposite. So I think billing, accurate billing with the increasing requirements of data related to billing and performance reporting are the two biggest drivers that occur to me. That's the reason we picked the electronic health record that we did pick at my previous job because it was the *leading* electronic health record for billing.

Sources of Data and Data Scientists

Others basically generate the data and we haven't made any significant effort yet to gain possession of copies of data I would say and that's maybe not even in our future, you know. We don't aim to become a big data processing shop. We are going to support this infrastructure and it will in fact ultimately become a data processing shop, but we won't be driving it, it won't be here.

We do require that everybody who's been funded to submit a final report and a version of that report gets put up on the website, so we do publish reports from our studies, but we also strongly encourage grantees to publish in the scientific literature and we use other means when the findings are really important and need to disseminate the findings more broadly.

I think the bigger the data sets the more complicated I think the platforms that are used for storing it and for analyzing it. So you know, all of a sudden you're moving to Oracle and beyond to places I don't even know. You no longer are just keeping little SAS data sets sitting around and so I think that's one thing that takes a lot of programmer expertise at a high level and then there are statistical, analytic kinds of questions that come into play and so you need the kind of expertise that asks the question.

Organizational Challenges

First of all it costs a certain amount of money to extract the data and analyze it. So organizations spend billions building up these systems but they have a hard time justifying spending a million to analyze all that data, so it just is crazy but that is seen as a challenge. Trusting the data and the methods that were used to analyze it can be a second. Changing practice based on what one sees in the data is a third because sometimes even though you see it still the incentives aren't necessarily aligned to make it easy to change. Let's say you have a big system and you've got a bunch of cardiac surgeons and you've got a bunch of cardiologists and you do an outcome study yourself and you find that either the surgery or the stents that the cardiologist placed are not doing as well as the alternative. You want to move in one direction. Well you know you're going to have a certain amount of opposition there and from the people who are being told to do less and so incentives, the incentives to act on the data. I'd say spending the money to analyze the data, trusting the results, and knowing that the results are really reliable and should be acted on and then rearranging the incentives in the organization so that you can actually make the move that the data suggests you should make.

When you’re using real world data you always have to ask yourself whether the fact that one treatment looks like it leads to better outcomes than another than a competing treatment. One explanation is that one treatment is better than the other. That’s what you’d like to think but first you have to resolve the possibility that it might be because the patients are different and that there’s confounding selection bias that patients who get one treatment are just different in ways that effect outcomes from the patients who get the other treatment. So I think another huge question is what do you do about missing data? So a lot of clinical data has lots of ‘missingness’ in it and how do you sort of account for that ‘missingness’?

Data Sharing

What do you do if five systems each have part of the data and they don’t actually want to send their data anywhere, they don’t want to share it? So this notion of distributed data collection, distributed queries, and distributed analysis is a big methodological issue that people are working on. Let’s say I’m the CEO of a health plan and some of the researchers in my organization are in part of a network and they’re in along with people from United Healthcare and Blue Cross Blue Shield, as a CEO I might be concerned that the data that we shared might be used for some purpose other than the stated research questions. So you know, you feel better saying couldn’t we accumulate the data here and be ready to look at it whenever you ask, but we’ll just send you the aggregated results on the questions you asked then you can figure out how to put them together with those from United Healthcare and Blue Cross Blue Shield. So I think there’s a lot of interest in the idea of distributed analyses.

Unintended Consequences

Well I think the critical thing is whoever provided the data and that includes the patients as well as the systems, need to be kept in the governance. If you get to the point where this data is getting repurposed, pretty soon you're going to have somebody that's very angry and decides to withdraw.

Big Data's Future in Healthcare

I think that our vision is that delivery systems, whether they are big HMO's, whether they are neighborhood health center networks, whether they are Accountable Care Organizations which are starting to come together all around the country and turning communities into systems of a certain type, they will begin to see it in their interest to capture the data, to ask and answer the questions, to share the findings broadly, and to drive quality up and cost down as part of what we call a learning healthcare system. So you know, you've got to get familiar with data and convinced that the data can actually lead you to decisions and then you've got to overcome those other barriers which are spending the money, trusting the findings, and changing the incentives. I think that that's got to happen, it will probably eventually happen but not as soon as it should.

Closing Thoughts

Well, just a couple last things, three things. Number one, something we didn't talk about today but is going to be part of big data pretty soon is genetic information. I think it's just a matter of time before doctors are ordering genomic screens, the whole genome, and somebody is going to have to store that information somewhere and when it's stored then somebody is going to want to use it for research, so that’s number one and that will take a lot of space, that will be really big data.

Number two, there’s a lot of work to be done on patient privacy and the oversight of this research. Yes everything has to be done to keep this data secured and protect people’s privacy. On the other hand, when these studies are not posing any physical harm to patients because we’re just looking at data, you do not need to require that a patient sign a 10 page or 20 page consent form. Even in certain randomized trials, yes, you need a consent form but it doesn’t need to be 20 pages long if it’s a very low risk question. So I think figuring out these issues about now that we’ve got big data, how do we work with IRB’s and human subjects oversight to rationalize how we use it and how we talk to patients about use. I think a subpart of this is we have got to learn to talk to the patients about how these data, how and why these data are being used and why it’s a good thing and certainly leave room for people to opt not to participate, but mainly beat that drum that, you know, we’re practicing with uncertainty. We don’t know what we’re doing and we can learn from the data and you could contribute.

The third thing is just the extreme costs. We have to look for ways to make this more efficient cost-wise and I think part of it is deciding that there’s enough value in getting the answers to the questions that can be asked and that you give up on the notion that your data is one of the ways you make money. I think some of the big HMO’s and others have seen their data as a commodity that they can capitalize on and that really gets in the way of data sharing and being in the learning healthcare system.

As for patients having the ability to make data driven decisions, part of the research we fund is research on how do you help patients make these kind of decision. So it’s one thing to do the studies and get the data, it's another thing to present the data in ways that patients can appreciate, even the clinicians can appreciate. When you get to genetics, most doctors wouldn't have a clue what to do with the results of a genomic screen, so you really need to figure out ways to take the data and take what's known and put it into a format that people can use it.

“A Sentinel that Signals Problems Ahead”

“Arnold Daniels, Ph.D., Pharm.D., M.A.”

Professional and Academic Experiences

Towards the middle to the end of the 90’s the drug costs were going up dramatically, the cost of the benefit. Some drug companies would ask us to do really unethical things to cut their costs without making hard decisions. I was always in a position where I could stump it and I was caught off guard by this so I sought some formal training and some help from some different ethicists that could help me figure this out and I wound up working with a couple very prominent philosopher ethicist types. From there, I wound up doing a doctorate in medical humanities because once you get into the ethics and you start to understand the place of illness and the human condition really want to understand that, it’s the humanities that renders it much more clearly than any science does. So I pursued the ethological end, that’s where my mind usually is and my greatest interest is where illness and the arts kind of work together.

The research we did was often using that huge administrative claims database and we had combined it with other things. So it could be prescription records, it could be prescription plus medical, but we had tens of millions of people that we had data on that we would use to do various research, look for various trends, help for our planning. We used it to affect drug selection, drug utilization by sending messages to docs, sending messages to pharmacists, messages to patients. We could send docs information about a certain patient’s patterns of drug utilization, and did, so they could take care of their patients better. We worked with public health officials from time to time to show them certain trends. Sometimes, we were just being silly and we would look at things like the effects of drug utilization after a certain TV show and focus on it.

My patient advocacy organization has been around for quite some time now. We are responsible for I think at the highest level beating back the utilitarian impulse we all have which is to do the most good for the most people. We also have patient assistant programs that can help patients find money to pay for co-pays, premiums, etc.

Meaning of Big Data

Big data has gotten to be almost like a parody. ‘Big pharma’ was a way of distinguishing the big institutional pharmaceutical giants that had almost unlimited resources and influence whereas those who were not ‘big pharma’ didn’t. So that’s how I understood the first use of big data in that fashion. I don’t think of big data as any particular company. I think of it as data sets that have huge numbers of elements and are organized in a way that can be mined to discover things, but also can be used to alter the course of events and improve performance and outcomes. I think of it mostly in a predictive way. We helped physicians and pharmacists make decisions about drug use for individual patients at the moment their deciding based on what is in the files, how we can access it, the sophisticated systems and the software allows us to access it, analyze it and produce something that’s usable in split seconds. So big data to me is not just a big data set; it’s also how it can be used to alter a course of events or approve some sort of outcome from parking to healthcare.

I arrived at this description by seeing it, being involved in it. I was part of the group that would write rules that would affect how certain prescriptions that came in through our system, how they would be adjudicated, you know, when would we send an alert out to a pharmacist that says look below. We tell a patient that more information is needed before we can adjudicate the claim or not based on what was in their file or not. I saw the power of big data or what was available from a lot of the big database sites we were in. I'm paying attention to what's going on out there like what IBM does with some of these cities: manage traffic, manage water, etc. I mean every time you turn around it's a big data conference; it's the big thing.

Medicine as an Information Science

I'm going to say no. Medicine can use it, it applies it but it's not it. Medicine still requires listening to the stories, it's touching and hearing and smelling, and all that. So information is part of it, like I said information is part of the decision support systems but that's all.

Healthcare Big Data Drivers

Some of it is the science. Now that we've sequenced the genome, now big data involves the computational biology, you need big data to just be able to make sense of all that genetic information. Some of it is just trying to make sense of all the information that we have now so it's an organizing approach. How do we make sense of all the data that we're getting from the science? There's a lot of pressure on clinical performance and there's a lot of pressure on the economics and with good reason. There's thinking that you can use a lot of information to create these decision support systems needed to come from big data. Again, it's kind of the predictive analysis. That helps people understand what they should do or help them make decisions. I would say that those things that are driving it in healthcare.

Sources of Data and Data Scientists

I think where big data could come in real handy is the diagnostic odyssey. The average time for people with a disease to get diagnosed is eight years. It ranges from eight minutes to 80 years, but it’s a long time and you just hear these stories constantly of people who have gone from doctor to doctor to doctor, test to test to test and it just takes a long time and while if you spent some time in healthcare like in the clinics and in the hospitals and while you see that the docs will always list the disease and their differentials to show their brilliance and their chops, it’s rare they’ll go after it because the mantra in medicine is common things happen commonly so don’t waste your time on the esoteric.

What big data could do is to help bring some precision to when a patient is presenting whether or not there’s a strong likelihood of a particular illness. This is what I suggested to my Watson friend there at IBM was you could add into this, they take a lot of the clinical data, scientific data, but I think you should add in the patient experience. That trajectory is meaningful. What docs do they go to first? Which ones sent them on their way? What was the sequence of docs they saw? What sequence of drugs they might have been given, what was the sequence of tests, are there certain things that you could make out of that that are consistent or at a high predictive value for a given disease that you could interrupt that odyssey early. That to me would be an important application of big data in healthcare.

Unintended Consequences

I’ll tell you, in my professional focus, time really matters, and there’s so much damage done by that duration that the time it takes to get diagnosed, so much damage done in that period of unnecessary tests, etc., that it lowers that risk. But that risk is there. All it can do is predict and so there's certain probabilities of being right but it comes with the probabilities of being wrong so it could be a problem.

Repurposing of data – it's happening actually. I don't know if we'd say data is being repurposed, there's a bit out there, there are a lot of drugs that are sitting in laboratories or were on the market and taken off for a variety of reasons that are being repurposed for rare disease. I suppose what comes with those drugs is information on them and they wind up, there's several drugs that have just either fallen by the wayside and get brought out because additional investigations find out that they have a role there. So it's a combination of bringing out the old drug and using the information available. But that's big and actually causes a lot of problems because it could be effective therapy where there wasn't any before. But sometimes they pull out these drugs that cost pennies and they get repurposed and then charge hundreds of thousands of dollars a year. But with the drug comes data. I suppose you could say it's been repurposed.

Big Data's Future in Healthcare

Well the hope is that its use is predictive in a few ways. One is to be able to use it for surveillance, to be able to find in clinical trials. And so those one in a million events that can kill people, one in 100,000 are myths and so big data should be able to conduct a surveillance that serves as a sentinel to find the signals or problems ahead. So part of it surveillance for problems … picking up signals that could not be picked up in the pre-clinical testing phase.

The other is to be able act as decision support for patients and for whoever is taking care of people to know that with a certain set of attributes and certain environments what's going to happen to these patients with a range of options. That’s the future state that I’d like to see. Now it’s threading the needle, I don’t know if you have, but I have, where there’s way too much of a reliance on technology and information I think on the part of healthcare professionals. Even just these clunky electronic health records, the doc doesn’t even look at you anymore. They’ve got their head in the computer, let alone take the stethoscope out and listen or just take a look at you or listen to your story. So the future state also somehow doesn’t create this automaton of a healthcare professional, nurses, you name it, who forgets that there’s a human-being setting right there and just produces these weird robotic type of interactions where they don’t think about the situation.

Metaphors and Symbols

I’ve used it but I count on experts to help me with it when I need it. So to me big data is very non-descript. I can’t come up with a metaphor because I really don’t know. It’s not descriptive enough for me. I don’t know if people in big data, I wouldn’t know what they’re talking about or when they talk about it, do they mean just how it’s set up, are they just talking about volume, are they talking about a certain type, are they talking about certain capabilities with it, or is big data just like lots and lots of information in a particular area or does that also encompass the things you can do with it, I don’t know. So to me it’s too vague to even come up with a metaphor.

Closing Thoughts

I think the cause of big data would be better served by characterizing it more clearly for a lot of us and their own constituencies if they want to be able to move forward. People can be afraid of it, so between not really understanding what’s meant by it and by being afraid of any big thing, it needs to be clarified and demystified because I’m not sure what the hell they’re talking about.

"An Unnecessary Euphoria"

"Frances Milburne, M.D."

Professional and Academic Experiences

I spent about 13 years at a managed care plan were my involvement with data and analytics really was about primarily learning how to think about populations in healthcare because as you probably are aware most clinicians really only think about the patient in front of them or a handful of patients at a time and aren't really used to thinking in terms of populations. Certainly my public health background and then the fairly exciting work that was going on at there at that time in the early 1990's on population management and how clinical medicine needed to be thinking about chronic illness care from a population perspective. That was work that was done by Ed Wagner and his colleagues. I found that very intriguing.

I left the managed care plan in the late 90's because I'd actually been sharing my practice with a clinician who had gotten informatics training and was also an internist and I became aware though just conversations with him about how important the electronic health records were going to be. This was long before there were any electronic health records but I then had the opportunity in the late 90's to help start up a clinical network for a university. So that was really where I began thinking about data from a clinical perspective on a more organized fashion. I do not have training in informatics; it was really on the job. So I moved after several years into a role of being the medical director for clinical informatics. While I was in that role what I basically did was start setting up reporting out the backend of the EHR's which nobody was doing at that time. In fact, even though we were using what is now probably the pre-eminent

EHR in the country, it did not have any way to report out of the backend and so I actually wrote a grant to one of the pharmaceutical companies to get a beta version of their relational database that's now just standard operating equipment for the EPIC installations but we tested it out and started generating clinical reporting out of the backend for things like diabetes, heart disease, hypertension and stuff like that. So that was really my experience, it was on a very, very practical application.

Another activity that I was involved in that sort of overlapped with that somewhat or at least from a privacy and security perspective was essentially set up at the request of several governors at the time try to come up with some national standards for health care information exchange simply because each state had its own rules and regulations and policies and procedures and it was becoming very, very challenging to exchange health information across state lines. So that was some insight into some of the issues around handling large data.

Meaning of Big Data

What big data means to me is that you assemble information from multiple sources that then get assembled in a large dataset and who knows where that actually resides in various servers around the country or even around the world I suppose which is sort of euphemistically referred to as the cloud and assuming that privacy and security constraints are being considerations for being adhered to which is I think a big question mark. Then you know there are certainly ways that large datasets can be used to recognize patterns which are otherwise hard to spot.

I don’t even know where it’s all coming from but I think there’s a sort of euphoria being built around big data that is not necessarily, I think it’s, and you get this sense there’s this sort of train leaving the station and everybody is supposed to get on it and yet as I said I think where information has the greatest potential to improve health and raise cost is in the way it’s used in individual provider’s practices and in delivery systems. But, I think there are some really risky things about big data.

Medicine as an Information Science

Clearly medicine is a very information rich endeavor. I mean it’s one of the most informative, it’s orders of magnitude more data rich than finance let’s say. A lot of the models that we come up with for certain thinking about how to handle big data are based on finance, but it’s just orders of magnitude more complicated. So I think there’s no question that it’s an information science. It’s more than that though because it doesn’t lose its human side. What I’m referring to is the fact that so much of the information we use in medicine is imprecise or irrelevant or just background noise. Computers are really good are really good at setting things up so that they flawless. They will do the same task with the same information over and over again so they’re very good for example at prompting humans to remember to do things that have to be done over long time intervals that are really easy for us to forget like screening and they’re very good at presenting information in patterns that we’ve programmed the computer to present information in. The other thing computers are really good at is when we have to actually slow down and think something through and figure something out then computers can organize information in ways that make it easy for us to solve difficult problems. So that’s really the use of computers. But we have to remember all the time that computers aren't anywhere near as smart as the person that's using them, so what the human brain is really good at is recognizing in patterns where we didn't know there was a pattern, computers can't do that at all.

But I think it's unrealistic to expect providers and patients to really adhere to strict privacy and security standards which for the past 10 years we've taken very seriously and then just find out that we just sort of shrug our shoulders. I think it really is a crisis that has to be addressed.

Healthcare Big Data Drivers

I think industry really sees big data as an opportunity to selectively market to the American people based on their individual healthcare needs. Now this may get sort of promoted as if we can figure out which patients need certain drugs and we can get those drugs to them we can improve their health. Unfortunately, the background reality is that Americans take way too many medications; only a fraction of them really provide benefit. Certainly blood pressure is a good example and some of the new biopharmaceuticals that are very specific for cancers or certain immunologic disease is they have some targeting potential on that. So I suspect that a lot of the push for big data is from industry, it's not all the pharmaceutical industry, a lot of it is medical devices.

I think the use of information in healthcare again as I said is really local and there's a cycle, I mean there's actually a pattern there that I find extremely interesting and it's basically this: Information gets entered into the computer and it comes in from multiple sources. It may come in from lab results and a lot of it comes in from just interactions between people and gets documented hopefully as structured data but also as text. Then that information now can be acted on by tools in the computer that have to be set up. So the way that it's of greatest importance in the delivery of healthcare is as in its use in two ways. One, its ease in being shared with other people including the patient so that more than just the doctor that has the chart can have access to that information, but other specialists, hospitals, emergency departments, and most importantly the patient can have that information. So that's one thing that the flow of information in the practice does that's revolutionary.

The other thing I would say is that the public health community particularly is quite interested in getting involved in the treatment or the management let's say of chronic illnesses and they really haven't been able to do that up until now. Public health is primarily involved with disease outbreak and particularly reportable diseases, infectious diseases, and then disaster response. But to start to get involved in chronic conditions, obesity, diabetes, and heart disease what public health is really looking for is information where they can identify places that public policy can be driven by patterns of diseases that are right now hard to see: smoking habits, nutrition, and places in the community where there are high rates of narcotic use.

Another one is just for public policy decisions. For example, where do we put our resources and what are the biggest health threats to the population? So I think for the potential to do that is from big data sets is really great. From the perspective of a clinician or a healthcare policy person, that's really driving improvements in quality.

Sources of Data and Data Scientists

It depends on how it’s structured. Providers really don’t interact with big data on that level. The closest thing that I could think of to that is a provider would have a hosted EHR so they don’t have the servers in their own organization as is the case for small groups and rural groups, and so that that information then is hosted externally. One example is the local Regional Extension Center. They’ve been heavily involved in helping providers install their EHR’s and have formed a collaboration with a data analytics company and so they then have access to these streams of information on providers who they have no relationship with and they can basically go to those providers and say do you know how you’re doing in managing your heart disease patients and they say no we have no way to get that and they say well look let’s show you and here is not only here is your population of patients with heart disease but here’s how you’re doing and actually you know here’s the gap. Here’s where you want to be and here’s how you’re actually doing. Let us help you fix your processes to improve. In fact, here’s a free iPhone app you can have if you join our system and you can actually go in and look for individual patients; you can see how they’re doing. You can also look at how they’re doing in aggregate and they get their data from the EHR’s and through the laboratories and then they get billing data and they can do this. So that is a way that I’m starting to see a developing and I think that sort of counts in the big data, it’s not, certainly not de-identified it’s in a service provider realm that’s developing.

Organizational Challenges

I think there's, it's really hard to make a case of big data if we're going to make it's where the money is and it's been my observation that if we start getting sort of distracted by these things that big data can do in terms of disaster response that that really leaves out where most of the work that happens that is involved with healthcare quality takes place, and that is in the provider's office. I think it's pretty hard to make the argument that it is going to provide a lot of benefit for providers and patients on a one-to-one basis. I'm pretty skeptical about that.

I think most of the action with information in the EHR has to do with getting it into the EHR so that it's accurate which is a major challenge because there's a lot of inaccurate information in most of the EHR's. Also, learning how to report out of it so that practices see how they're doing taking care of their populations on a very local level. So what percentage of my patients, who are my patients first of all, that's the big one, but after you get that, what percentage of them have been immunized properly, have been screened for cancers properly, if they have chronic diseases are being taken care of properly, that's not a big data issue that's an in the practice use of data so I think that is where the real action is as far as use of the EHR's.

Unintended Consequences

The downside I think there are two, and unfortunately I think this is where big data is largely being used and what's unfortunate is it doesn't get talked about very much. The advocates of even in public health tend to just sort of quietly move to another subject if you bring these of negative uses of big data up because they're so enumerative the potential. One of them is marketing.

The other place that I think big data entails a huge risk is simply in abuse from national security. I think it's very, very clear that the NSA is all over big datasets and there's no reason to think that healthcare is any different. Very clearly there's a need for government to identify individuals who may oppose risk to the rest of the population. But that's a very, very different proposition than gathering and doing large scale population surveillance and simply sucking up everything. I think we've tended to shrug off the revelations that have come out over the past six months or so about how large datasets are being used for security agency surveillance which is a major departure from what we've assumed was the case in the past.

I think that by setting up large datasets in healthcare in the cloud, I mean we can say that they're secure but those are now just words and so I think we may very well be coming to a point where, well there's something changing and it's not clear to me what's going to happen. I mean the Europeans are certainly starting to sort of disconnect themselves from interactions that are easily surveyed although they have the exact same issues there and certainly some of their security agencies have been part of the whole thing, but I think we're either moving into a field where we just kind of give up on privacy and say well that's kind of over or we have to say no look we actually do take privacy and security seriously.

Big Data's Future in Healthcare

Everything that I have talked about that I see as a benefit for big data can be accomplished if the data are completely de-identified. So public health surveillance, disaster preparedness, situational awareness, disease patterns, public policy, every single one of the beneficial activities or purposes can be accomplished if there is nothing in there that allows identification of individual patients. One can argue that if you put in some identifiers it allows some additional benefit. For example, if you really wanted more information out of how people, what percentage of people with asthma admitted to the emergency rooms are ending up in the ICU and then a week later bouncing back then you have to put patient identifiers on them. You can get that same information out of delivery system data on the local level if you find a place where you need to investigate further.

So my vision would be that big data sets first of all have strict purposes rather than let's create big data and then figure out what we might do with it. I think there needs to be a what are we addressing, what information do we need, and how are we going to use that information. I think that first of all patients need to be informed of where their information is going so that they know and I think they should have an opportunity to opt out if they don't want it. Then I think that any information that's gathered and compiled and aggregated and then looked at for public policy or for any of the public health purposes we've talked about should be strictly de-identified. That would be a vision there for what it might look like. Then I think you would avoid the two pitfalls which are marketing and a lot of the national security abuse I think would be avoided in that case.

Metaphors and Symbols

I spend a lot of time thinking about information, but as I said most of it is on a very local level. For example Accountable Care Organizations will really only work if everybody understands the metrics against which they're being compared and everybody is using the same metrics, every insurance company is using exactly the same metrics, so one is the saying well we're measuring the percentage of patients who have hypertension who have a blood pressure average on the last three or whatever under 140 over 90 let's say, and the next insurance company says yeah we're doing that too only it's actually greater to or equal to, less than or equal to 140 over 90. Those are two totally different targets because people round to whole numbers. So and blood pressure is imprecise, or one might say 130 over 80.

An ideal system would be one in which the providers actually have an internal dashboard that's identical to what the payers are seeing and that the providers are allowed to manage their outcomes to where they get to a point where they say we've got them where we want them, now we send them to the payer and they can come and audit us and make sure that our process is correct. So I think we're going to have to have a lot of transparency. So how do we set this up so that we can actually manage outcomes and costs and agree that we're both measuring the same thing?

Closing Thoughts

I'd like to point out that what providers need more than anything is analytics. They need to be able to use the information in their systems to tell how they're doing and to figure out what to do and to set priorities and that's what's lacking. It's not clear to me that big data, the way it's set up in cloud base will allow that to happen but it may. One way that big data could be used is to identify emergency high utilizers.

These are patients who are completely overwhelmed by their disease or their medical conditions and their social situation, so they end up going back to the emergency room over and over again and running costs up at an extravagant rate. It's very destructive, and big data could be used to identify places and people where that’s happening. Big data can desperately identify social and mental health services needs and that’s the type of thing that probably should be done inside a practice.

So there’s lots of ways that this could be brought in to help in specific situations. But I once again want to caution the way we’ve gone about this is let’s build huge datasets and then I’m sure some great social benefits will accrue. I think what we’ve done is we’re raising the risk of what I consider to be misuses of data in ways that are not necessarily in the public interest. Yes, there clearly is a role for big data in public health policy, public health intervention, and high utilizers and I’m sure other uses will show up. So having big data capacity is I think very, very useful. Again, I don’t see that it requires identifiable information but you know that may or may not hold up.

CHAPTER VI. RESULTS

Overview

This chapter provides the research study's findings assembled through a general inductive approach to qualitative research which is commonly used in healthcare (D. Thomas, 2003). Data were collected through nine semi-structured interviews with key healthcare stakeholders from three classes: government, providers, and consumers (advocates). The results describe (Amedeo Giorgi, 2009) and interpret, as accurately as possible, a first-hand account about the phenomenon of big data in healthcare across and within the three key healthcare stakeholder classes. Triangulation of the three stakeholder groups was an important strategy that facilitated any inclination towards researcher bias. The analysis was anchored by the following research question:

> Q1: *Within and across the narratives of three key healthcare stakeholder classes, what are the important categories of meaning or current themes about big data in healthcare?*

Further in-depth analysis produced "units of meaning" used to reconstruct key stakeholder narratives into a cohesive yet agile statement about the meaning of big data in healthcare – without losing the essence of narratives in their entirety. The words of Van den Berg, translated by Van Manen (1997, p. 41) profoundly captures the essence of phenomenology as both a philosophy and a research method:

> *[Phenomena] have something to say to us — this is common knowledge among poets and painters. Therefore, poets and painters are born phenomenologists. Or rather, we are all born phenomenologists; the poets and painters among us, however, understand very well their task of sharing, by means of word and image, their insights with others — an artfulness that is also laboriously practised by the professional phenomenologist (M. Van Manen, 1997).*

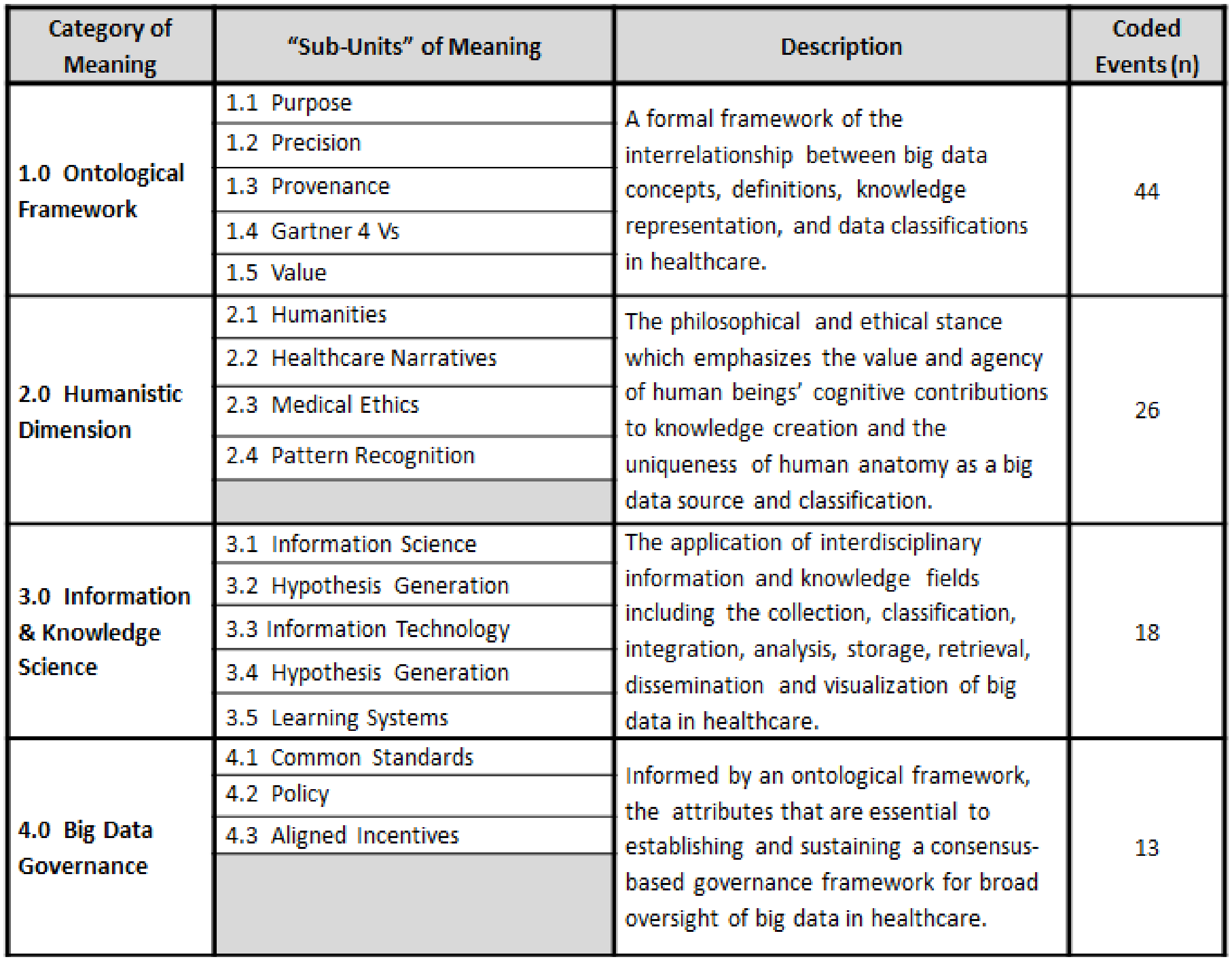
Explication of the Data

Groenewald (2004) revised Hycner's description of "data explication" (Hycner, 1985), suggesting "the term [analysis] usually means a 'breaking into parts' and therefore often means a loss of the whole phenomenon …whereas 'explication' implies an … investigation of the constituents of a phenomenon while keeping the context of the whole" (p. 17). Explication resonated with my edict to maintain the essence of the key healthcare stakeholder narratives. As such, explication of the data was yet another strategy to eliminate any predisposition of researcher bias on the research design.

Results

Four distinct, yet interrelated, important categories of meaning naturally emerged during the course of the data explication: (1) *Ontological Framework of Big Data in Healthcare*; (2) *Humanistic Dimension of Big Data in Healthcare*; (3) *Information and Knowledge Science for Big Data in Healthcare*; and, (4) *Governance of Big Data in Healthcare*. A description of each category of meaning and the number of times an event was coded within each "theme" is found in Table 7. Through the process of reading and rereading the text of the transcripts, contextualized data initially produced approximately 41 initial nodes. These nodes were reduced to 17 distinct "subunits of meaning," categorized into the four "important categories of meaning"

Each category of meaning constituted the essence of big data in healthcare as described by nine leaders from three key healthcare stakeholder classes. The general inductive approach allowed me to derive a description and interpretation of the key healthcare stakeholder narratives while also presenting a description of the categories of meaning within and ultimately across each key stakeholder class.



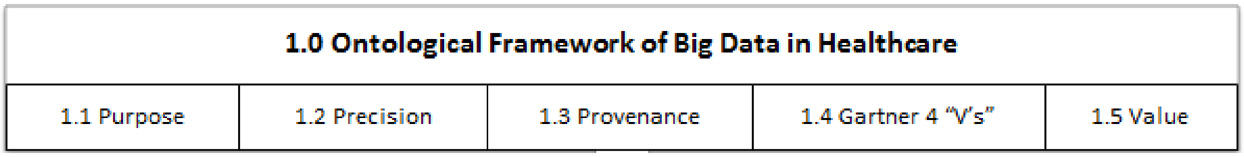
Category of Meaning	"Sub-Units" of Meaning	Description	Coded Events (n)
1.0 Ontological Framework	1.1 Purpose 1.2 Precision 1.3 Provenance 1.4 Gartner 4 Vs 1.5 Value	A formal framework of the interrelationship between big data concepts, definitions, knowledge representation, and data classifications in healthcare.	44
2.0 Humanistic Dimension	2.1 Humanities 2.2 Healthcare Narratives 2.3 Medical Ethics 2.4 Pattern Recognition	The philosophical and ethical stance which emphasizes the value and agency of human beings' cognitive contributions to knowledge creation and the uniqueness of human anatomy as a big data source and classification.	26
3.0 Information & Knowledge Science	3.1 Information Science 3.2 Hypothesis Generation 3.3 Information Technology 3.4 Hypothesis Generation 3.5 Learning Systems	The application of interdisciplinary information and knowledge fields including the collection, classification, integration, analysis, storage, retrieval, dissemination and visualization of big data in healthcare.	18
4.0 Big Data Governance	4.1 Common Standards 4.2 Policy 4.3 Aligned Incentives	Informed by an ontological framework, the attributes that are essential to establishing and sustaining a consensus-based governance framework for broad oversight of big data in healthcare.	13

Table 7. Description of four important categories of meaning of big data in healthcare reduced from over forty sub-meaning units.

From a Husserlian phenomenological perspective, the description of narratives, even though transcribed and possibly written, still remains a description (Amedeo Giorgi, 2009). Keeping with the construct of phenomenology, my objective was to also engage in the interpretation of the study participant's interpretation ("double hermeneutic") of big data in healthcare. I consciously set aside my own presuppositions so as not to bias the data explication and interpretation through bracketing (Groenewald, 2004) out my own experiences about big data in healthcare. Bracketing was yet another strategy to control for researcher bias. Using the raw transcripts as a primary reference, the clustering of important categories of meaning (“themes”) emerged iteratively through the study participant’s own words.

Phenomenology focuses on the common elements of a phenomenon, rather than on the individual. In keeping with this aspect of the chosen methodology for the study, I did not include participant names or pseudonyms in presenting excerpts from the interview transcripts. The header box before each category of meaning was extracted from Figure 7 to assist maintaining the reader’s orientation of each category of meaning and associated subunits of meaning. Findings were not intended to be generalizable across or within key healthcare stakeholder class.

Ontological Framework of Big Data in Healthcare



1.0 Ontological Framework of Big Data in Healthcare				
1.1 Purpose	1.2 Precision	1.3 Provenance	1.4 Gartner 4 "V's"	1.5 Value

Category Definition

This category refers to an informal representation of interrelated concepts, knowledge, words (and buzzwords) and phrases that describe the characteristics, attributes, and meaning of big data which produce information for wisdom and decision making in healthcare. In addition to Gartner's popular characterization (not definition) of big data as Volume, Velocity and Variety (3V's) is here augmented with another "V" - Value (Porter & Teisberg, 2006). Value is created by producing, through the disciplines of information and knowledge management, usable information for healthcare intelligence and decision making. These attributes and characteristics encapsulate big data's realized – not potential – intent. Many study participants exhibited a skeptical position on big data in healthcare by characterizing it as "*over-blurted*", "a *parody*" and "*the latest craze*" and observing the "*expression is overused*." Or as a government stakeholder said, "*it's a heap of 1's and 0's*." When study participants did define big data, intuitive references emerged such as "*smart data*" that has an ability to "*answer questions more deeply and get down to the causes as opposed to scratching the surface*" as described by another government stakeholder. This category refers to both human and organizational forces and events which drive the emergence of big data in healthcare. Big data drivers are those internal and external forces of the healthcare ecosystem which possess the ability influence or drive the use and advancement of big data in healthcare. Big data influencers could be a thing (e.g., technology), a policy (e.g., Affordable Care Act) or an attribute (e.g., better question generation). According to McKinsey, fiscal concerns, or healthcare costs are the primary driver of big data in healthcare.[9] Provider billing policies, public health surveillance, and marketing to individual patients were cited as other drivers of big data in healthcare.

1.1 Purpose

The analysis found that big data in healthcare must be collected with a purpose that is well defined during the initial planning stage for a project or initiative. Because of the massive data sets that are collected and the associated costs of designing systems to capture and analyze big data, a stakeholder posited:

> *You build data systems, they cost so much money to actually collect the data it's quite expensive. They're built for specific designated purposes and it's fairly important that you know what the purpose is before you start. I just did a carefully designed data system that is very purpose specific. They're designed for a specific utility a specific purpose and they just happen to collect massive amounts of data but they're always for a specific purpose (Provider).*

A government study participant pointed out the unintended consequences of purpose-driven big data by stating, "*Purely relying on machine learning without the application of subject matter expertise and the application of a clearly defined set of goals can lead to big data actually distracting an organization from its core goals and outcomes*." While potential nefarious uses of protected health information do exist, study participants overall welcomed repurposing big data for an array of uses:

[9] http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care

> *So public health surveillance, disaster preparedness, situational awareness, disease patterns, public policy, every single one of the beneficial activities or purposes can be accomplished if there is nothing in there that allows identification of individual patients (Consumer).*

Another consumer stakeholder agreed with the ideology of planning, with a purpose, for big data collection. Planning for big data usage at the onset of a project or initiative suggests that simply amassing large data sets as an organizational asset is only part of the strategy to realizing big data's true potential:

> *So my vision would be that big data sets first of all have strict purposes rather than let's create big data and then figure out what we might do with it. I think there needs to be a what are we addressing, what information do we need and how are we going to use that information (Consumer).*

1.2 Precision

Terms including precision medicine, personalized medicine, and resource-based medicine are interchangeable references to medicine that, at the very least, combines transactional data (e.g., claims) with computational biology, and genomics data based on an individual's genetic and social epidemiology profile. The meaning unit – precision – was not interpreted as the process of collecting and managing big data but big data's trustworthiness. The combined attributes of big data quality and trust creates confidence in the level of big data's preciseness. A provider government stakeholder states, "*I think one of the problems is trust in data quality and data fidelity."* He further expressed his view on a lack of confidence in the precision of big data connectedness:

> *I think people don't really know yet if this is something, you know, put everything in a little black box and it comes out the other end and it gives me a relationship, and people say I'm not so sure if that actually is true or not (Provider).*

While one provider study participant voiced his concern that, "*there's no proof point yet that is kind of real concrete,*" another study participant from the consumer stakeholder class pins his hopes on the premise that big data precision is necessary to support optimal hypothesis generation in patient episodes such as rare clinical cases:

> *So what big data could do is to help bring some precision to when a patient is presenting whether or not there's a strong likelihood of a particular rare disease (Consumer).*

A government stakeholder cautions, "*Big data is currently fraught with hype and over promising.*" He also believes big data "*is an area of incredible promise for healthcare*." Furthermore, there are approximately 500 petabytes of healthcare data in existence today and that number is expected to skyrocket to more than 25,000 petabytes within the next seven years (Savaiano, 2013).[10] According to several stakeholder narratives, these clinical and administrative data held in fragmented information systems will not produce the timely and accurate insights yield better questions for better answers. A government stakeholder adds:

> *I think everybody's hope is that in five years' time, there will be widespread integration of administrative, clinical and patient generated data that will be available through big data; it's assuring that the right person gets the right data at the right time and the right format for them (Government).*

[10] In the construct of Orders of Magnitude, a petabyte is the equivalent of 1,000 terabytes, or a quadrillion bytes. One terabyte is a thousand gigabytes. One gigabyte is made up of a thousand megabytes. There are a million petabytes in a zettabyte.

1.3 Provenance

The consumer stakeholder class elicited responses that focused on the individual and population health informational needs: "*we're really interested in how treatments work for individuals.*" Provenance of big data, which one study participant suggested is the vital commodity, notes, "*data is very nondescript*."

Data provenance refers to the information sources about data and includes data points such as contextual and physical metadata and Meaningful Use attestation data. In the government stakeholder class, one respondent spoke of the integration of genetics stating, "*We were getting lots and lots of data coming in from the genome sources from that point and we were injecting it and trying to be able to analyze it.*" Life science disciplines including biomedicine, neuroscience, genetics, and genomics were intentionally excluded in the definition of big data in healthcare. I intentionally wanted to let life sciences narratives naturally emerge from the narratives, provided study participants viewed the subject as an important theme. Genomics did naturally emerge from the narratives as an important source of big data *("New data sources including genomics, senomics,and other "-omecs" are out there. Metropolomics, for example, are generating huge amounts of data that need to be processed in ways different than we have in the past")*. In deference to Gartner's classification of big data, genetics and genomics fit into the High Variety classification group. Genetic and genomic data are fundamental to achieving precision in clinical hypothesis testing and is foundational to delivering personalized medicine. The richness of genetic information was championed across all stakeholder classes:

Government:

I think genomics and the ability to do sequencing is the first statement of it [as a driver of big data]. You know, there are a lot of things where we're getting ideas into genetics, we're getting them information about environment, and we're getting information on health. What affects us, what affects us mentally, you know, our health affects our mental state and our mental state affects our health, our emotional states (Government).

Providers:

I mean the integration of genomic data phenotypic data which is clinical data in the electronic medical record. It's something that every academic medical center is racing to do because the answers are not going to be in one or the other. The answers are going to be in the combination of both and so that is absolutely the future, a very reasonable approach (Provider).

I think the whole genome sequencing components so again some more of the genetics side of the equation over the patient's life span and helps predefine clinical context based on the genetic information. So I think that's hopefully going to help us with neurogenetic diagnosis decision support in the electronic medical record again so some of the things you see in 23andMe.com may be more sophisticated in terms of patient genomic test reports and that kind of thing (Provider).

Consumers:

Well some of it is the science, you know, now that we've sequenced the genome, now big data involves the computational biology, you need big data to just be able to make sense of all that genetic information (Consumer).

Something we didn't talk about but is going to be part of big data pretty soon is genetic information. I think it's just a matter of time before doctors are ordering genomic screens, the whole genome, and somebody is going to have to store that information somewhere and when it's stored then somebody is going to want to use it for research, so that's number one and that will take a lot of space, that will be really big data **(***Consumer).*

Although information for decision making as an output of big data was important to key stakeholder classes, almost all study participants were still plagued by the quality and trustworthiness of big data in healthcare, as one study participant put it, "*I think there are some really risky things about big data*." Another study participant gathered:

> *I think the cause of big data would be better served by characterizing it more clearly for a lot of us and their own constituencies if they want to be able to move forward. People can be afraid of it, so between not really understanding what's meant by it by being afraid of any big thing, because big things you know can take advantage of not big things, there needs to be, whoever big data is, whatever it is, it needs to be clarified and demystified I think, mainly clarified I'd say because I'm not sure what the hell they're talking about (Consumer).*

Several study participants emphasized the emergence of other new data sources into the big data equation, including "*biometric data from sensors which people tend to get excited about*," and the emergence of "*device interoperability and the data the comes from medical devices*." Among all of the transactional and biometric types of data mentioned, one respondent added a forgotten source of data - narratives:

> *I'm not against that whole thing on natural language processing and using narrative, it just has a different goal and the goal is to try and take somebody's really unstructured but maybe highly intelligent thinking and try to sense what general thing were they getting at, what can we discern from that … (Consumer).*

In two key stakeholder narratives, big data was characterized simply as "a *whole heap of 1's and 0's"* and "*at the end of the day you got a machine language that's all 1's and 0's"* without structure, governance, and purpose. As a government stakeholder speculated:

> *"I know that some people consider big data not to be 'big' until it's in the trillions, but we manage 400 billion discrete pieces of information … and it's growing by about four or five billion data records a year."*

Of note is Gartner’s High Variety characterization of big data. Healthcare is a transactional business that relies on “*administrative data as the foundational component*.” But stakeholders highlighted a host of data sources commonly used in their day-to-day routine, acknowledging that, “*it’s incomplete and inaccurate and there are no common standards.” “missinginess”* is a big problem, and there are an “*infinite number of data points I could collect.”* The following table (Table 8) provides an aggregated summary of the sources of big data cited by study participants.

Articulated Big Data Provenance		
Government	**Providers**	**Consumers**
Enrollment Data	Financial	Social Media
Hospital Data	Administrative	Public Data
Physician Data	Human Genome Project	Demographic Data
Assessment Data	Human Microbiome Project	Physical Measures
Laboratory Data	Meaningful Use Standards	Personal Medical History
Medicare Data	Management Treatments	Narratives (Stories)
Medicaid Data	**Sensors/Biometric**	Other

Table 8. Big Data provenance of referenced in key stakeholder narratives

1.4 Gartner 3 V’s

In the provider stakeholder class, two of the three study participants said big data “*means a lot of different things to a lot of different people.*” Gartner’s “3V’s” characteristics were referenced by several study participants across the three classes. One provider study participant said, “*I like the definition that Gartner coined … where big data is a high volume, high velocity, high variety information … used to enhance insight in decision making.*”

Provider stakeholders require speed, or in Gartner's characteristic of big data, High Velocity to provide the best clinical care at the lowest cost. But there is more work to be done to fully realize the "*hype*" of big data in healthcare. A provider stakeholder referenced the work of IBM and its capability to "*consume all the literature so they don't have to read it*." Another stakeholder noted:

> *So something that hasn't been fully mapped out into the analytical processing scheme that I want is something that's more real time feed that I can actually act upon much quicker. So there will be some components that are real time, some components that are like data warehouse and dashboard based, and then some components that are for big data for large data sets and for better insight. So I can go to my big data to find the hypothesis. I can go to a data warehouse to test that hypothesis and I can use my real time data to basically put that hypothesis into action with decision support (Provider).*

Summarizing the insight of the many study participant viewpoints on the Gartner's "3V's" characteristics of big data in healthcare that are currently in use and those which we can anticipate, one study participant asserted:

> *The sources of data that we use are pretty varied even though I'd say that administrative data is the foundational component. It actually meets the Volume and the Variety criteria we use records for multiple parts of the Medicare system, the Medicaid system, the enrollment data, hospital data, physician data, assessment data, laboratory data, Medicare data, and Medicaid data. I know we'd obviously be interested in adding other paired data to the mix. Then there's survey data, there is some pretty rudimentary Meaningful Use attestation data but we don't have any actual Meaningful Use data yet (Government).*

1.5 Value

Consistent with increasing healthcare costs, one provider healthcare stakeholder believes, "V*alue based purchasing and moving away from the lack of accountability of fee—for service-*

service" is key driver. Consumers of healthcare seek personalized medicine that is unique to their individual medical care and treatment plans. One consumer stakeholder gathered the "*rapid rise of performance measurement*" is a driver of big data in healthcare. Clinical performance measures (CQM) are developed by measurement developers to focus on patient-centered measures and the patient experience, creating value for patients. Another stakeholder believes big data is being driven by healthcare industry marketing strategies. The stakeholder proclaimed, "*Industry really sees big data as an opportunity to selectively market to the American people based on their healthcare individual needs*." Then he further elaborates it is possibly all unnecessary:

> *I think there's a sort of euphoria being built around it that is not necessarily, I think it's, and you get this sense there's this sort of train leaving the station and everybody is supposed to get on it and yet as I said I think where information has the greatest potential to improve health and raise cost is in the way it's used in individual provider's practices and in delivery systems (Consumer).*

National health spending has grown at historically low rates following the deep recession that ended in 2009. Whether this slowdown stems from broader economic factors, structural changes in the healthcare system, or some combination of the two,[11] big data in healthcare is seen as a commodity that if harnessed by technology and humans, can help make the delivery of healthcare even more cost-effective. But there was general disagreement on whether the costs of building healthcare systems and collecting and analyzing data is rising or falling. A government stakeholder says, "*It's much cheaper to collect big data and it's cheaper to store it*," and a

[11] Source: Kaiser Family Foundation (www.KFF.org)

provider stakeholder adds, "*Hardware is getting faster and is making itself available at low cost points*." But another healthcare leader suggests that the exorbitant costs of collecting big data have to justify their spending to manage big data:

> *First of all it costs a certain amount of money to extract the data and analyze it. So organizations spend billions building up these systems but they have a hard time justifying spending a million to analyze all that data, so it just is crazy but that is seen as a challenge (Consumer).*

Another consumer stakeholder is in agreement. He speculates that there is an association between deriving enough value from searching for optimal answers and good questions:

> *The third thing is just the extreme costs, you know, we got to look for ways to make this more efficient cost-wise and I think part of it is deciding that there's enough value in getting the answers to the questions that can be asked that you give up on the notion that your money is one of the ways, your data is one of the ways you make money. I think some of the big HMO's and others have seen their data as a commodity that they can capitalize on and that really gets in the way of data sharing and being in the learning healthcare system (Consumer).*

Summary

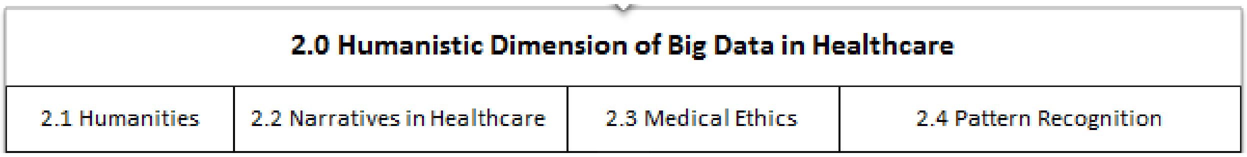
There is an awareness problem about big data in healthcare. A consumer study participant admitted, "*Nobody's ever defined it for me but I've heard it used a lot.*" Another frankly commented, "*I never thought about it until you asked me. I just assumed that I sort of knew what it was*." A government stakeholder narrative insightfully cautions us that because of the awareness issues associated with big data in healthcare, big data could potentially loose its momentum:

> *If I look at it today and big data has kind of lost some of the meaning that it had at the early part and maybe even gained, and eventually will probably gain maybe a*

> *specific definition to make it useful again. Some government stakeholder views of and formed opinions on big data in healthcare revealed feelings of cynicism and relevance. Its existence and possibly importance is undeniable. However, it requires a collaborative, momentous effort to define it and broadly diffuse its meaning – fast (Government).*

Another government stakeholder suggests that big data "*is currently fraught with hype and over promising*," and he also thinks big data "*is an area of incredible promise for healthcare*." Key healthcare stakeholders perceive big data as a buzzword or slogan that is not universally understood. Also other drivers of big data in healthcare were uncovered – a consumer stakeholder gathered that while the "*rapid rise of performance measurement*" is a driver of big data in healthcare, another stakeholder believes big data is being driven by healthcare industry marketing strategies ("*industry really sees big data as an opportunity to selectively market to the American people based on their healthcare individual needs*"). A government study participant confirmed McKinsey's assertion that big data can influence the spiraling costs of healthcare. He commented, "*We need to make the care we deliver less expensive and more efficient overall*." In addition, "*culture and leadership*" are important influencers of the explosion of big data in healthcare. Another study participant supported the notion that government rules drive big data in healthcare, "*Because it places data and the ability to harness and leverage data at multiple points throughout the healthcare ecosystem at the center as opposed to at the trenches.*"

Humanistic Dimension to Big Data in Healthcare



2.0 Humanistic Dimension of Big Data in Healthcare

2.1 Humanities

2.2 Narratives in Healthcare

2.3 Medical Ethics

2.4 Pattern Recognition

Category Definition

This category is a philosophical and ethical stance that, augmented with technology, emphasizes the value and agency of human beings' cognitive and critical thinking contributions towards optimizing the potential of big data in healthcare. Big data in healthcare is foundational to achieving precise decision making and refined hypothesis generation. The unparalleled ability of the human mind is crucial to realizing the potential of big data in healthcare. The anatomy of the human body which is a complex maze of interacting systems and organs that make big data in healthcare unlike any other big data generated in industries such as retail, transportation, and banking. The reduction of narratives in this category inductively generated four meaning units including humanities, narratives, bioethics, and pattern recognition.

2.1 Humanities

Achieving big data in healthcare, according to government stakeholders is an ability to link the capabilities of computers to the capabilities of humans. Where computers will facilitate the movement towards Singularity, analytics still requires humans to make decisions based on those findings. A study government participant pointed out, "*it's the human view of the data and it's transferring, it's gaining knowledge out of that data, transferring it to the human so that the* *human can actually do something useful with it*." Another study participant gathered there is a complementary role for humans and computers:

> *The computer is not a human brain and while we have computers that attempt to match many of what people do, much of what doctors do we don't have computers that can do all that doctors can do and that final step of processing, especially in complicated cases really needs to take place in the human mind, but it is processing of data for sure (Provider).*

Computers, and specifically highly portable tablets and smartphones, have become commonplace in inpatient hospital and outpatient clinic examination rooms. Technology is essential to facilitating access to complex drug databases and interoperating with patient's health histories and narratives of other clinicians almost instantaneously. Technology is also fundamental to establishing "health homes" for a physician practice's panel of patients. Yet, with the advent of health information technology, key healthcare stakeholders do not want to lose the spirit of the doctor-patient relationship. A study participant in the provider stakeholder class asserted that there is a "*human therapeutic relationship between patient and doctor*." From the analytical domain, providers see computers as being instrumental and necessary to provide personalized care demanded by patients. Study participants from this stakeholder class undoubtedly maintain that while the capabilities of computers and humans are vastly different, they are interrelated:

> *Again, I don't think it's going to necessarily replace what we currently do. I think you're still going to need people who are going to have to, you know, the big debate is will it be to the point where I don't even need extraction* (Provider).

Advocates for consumers were concerned about the erosion of the "*human therapeutic* *relationship*" between patient and doctor and cautions against the reliance on technology in the exam room and at the bedside. A study participant hopes, "*It doesn't lose its human side*." Another study participant commented on the over-reliance of technology and the sterilization of the patient-provider relationship:

> *There's way too much of a reliance on technology and information the part of healthcare professionals. Even just these clunky electronic health records, I mean the doc doesn't even look at you anymore. They've got their head in the computer, let alone take the stethoscope out and listen, or just take a look at you or listen to your story. Somehow don't create this automation of a healthcare professional, nurses, you name it, who forgets that there's a human-being sitting right there and just produces these weird robotic type of interactions where they don't think about the situation (Provider).*

Another consumer study participant expounded further on the differences between computers and humans. Human brain cognition is rooted in "*intuition*" and "*how humans are really good at figuring out the relative importance of different conflicting information and computers don't do that well*."

2.2 Narratives in Healthcare

Narratives in medicine are usually captured at the point of care and are often embedded in the patient's electronic medical record. While this unstructured source of big data is an important personal account of the patient's experience, big data in healthcare is not usually associated with "storytelling." While narrative medicine is the one-to-one interpersonal clinical conversation between provider and patient about illness, healthcare narratives captures the many-to-many conversations among healthcare stakeholders not just on illness, but about the experiential accounts of healthcare processes, insurance, access, and a host of other purposes related to the entire encounter with the healthcare system. A government study participant shared, **“***I can’t ever foresee a time when you won’t want to have the ability to collect narrative for at least some of the electronic record.”* One consumer stakeholder also suggests narratives are an integral part of the human experience and should not be lost as a data source that is not in the form of 1’s and 0’s:

> *This is what I suggested to my Watson friend there at IBM was you could add into this, they take a lot of the clinical data, scientific data, but I think you should add in the patient experience. That trajectory is meaningful. What docs do they go to first? Which ones sent them on their way? What was the sequence of docs they saw? What sequence of drugs they might have been given, what was the sequence of tests, are there certain things that you could make out of that that are consistent or at a high predictive value for a given disease that you could interrupt that odyssey early. That to me would be an important application of big data in rare disease (Consumer).*

2.3 Medical Ethics (Bioethics)

Published scholarly literature on bioethical analysis customarily focuses on human healthcare including issues of abortion, euthanasia, cloning, and health disparities. Big data and information in healthcare is an emerging topic in the field of medical humanities and bioethics. Big data bioethics was derived from the narratives of two consumer stakeholder’s experience. The discipline of medical ethics allows moral discernment to ground the understanding of illness and health. A study participant posits:

> *… once you get into the ethics and you start to understand the place of illness and the human condition really want to understand that, it’s the humanities that renders it much more clearly than any science does (Consumer).*

Another consumer stakeholder introduces the element of uncertainty about what is done in healthcare by policy and clinical professionals. By learning to talk to patients about big data and how it will be used to facilitate creation of individual treatment plans to cure illnesses, such conversations present a moral dilemma for the entire healthcare system:

> *I think a subpart of this is we have got to learn to talk to the patients about how these data, how and why these data are being used and why it's a good thing and certainly leave room for people to opt not to participate, but mainly beat that drum that, you know, we're practicing with uncertainty. We don't know what we're doing and we can learn from the data and you could contribute (Consumer).*

2.4 Pattern Recognition

In healthcare, pattern recognition, or assignment of labels to variables is a key statistical operation in population health and public health. There similar study participant views on whether the computer is more adept at pattern recognition than humans. One study participant proclaimed, "*There's a whole bunch of different things the human brain tends to work on intuition and pattern recognition on a speed that is far faster than computers actually,*" while another study participant spoke of the advantages of the human brain and its pattern recognition capabilities:

> *So if I want blood pressure to be set up as a graph so I can see whether it's getting better or worse with the individual numbers on a spreadsheet. They're very, very good at it, but we have to remember all the time that computers aren't anywhere near as smart as the person that's using them, so what the human brain is really good at is recognizing in patterns where we didn't know there was a pattern, computers can't do that at all (Consumer).*

Key healthcare stakeholders identified the human dimension as complimenting the capabilities of technology ("*I don't have a really strong faith that computers are going to somehow be smarter than people")*, working as an integrated unit to achieve big data's latency. And the emergence of big data is well documented in industries including aerospace, transportation, and banking. However, in healthcare, big data is a very different and complex endeavor which makes comparisons with other industries difficult. Comparisons of big data in healthcare to other industries that have mature big data capabilities are spurious and as one study participant remarked cannot be compared across industries:

> *I was giving an international speech in Brussels. While I was talking about measuring quality and safety somebody got up and asked me, 'well why don't you do it just like a bank has an ATM,' and I didn't laugh but I felt like it. I just said because everything isn't an integer, it's not as simple as a balance sheet or income statement, or a checking account. It's a whole different ball game and the relationship between processes and outcomes in healthcare is not totally defined. You can give the same drug to two patients in the same way, the same age, the same diagnosis and they'll react differently, and it's just we're not making patients (Government).*

Another consumer stakeholder delved into the complexity of human organisms and the connections across intricate bodily systems which constitutes the entire person. His narrative supports the ideology that comparisons with industries that produce "widgets" and defines their unit of analysis (e.g., retail) is unauthentic and that in the delivery of healthcare, the person has to be viewed holistically, making big data in healthcare unique:

> *We're looking at human-beings holistically. The reductionist approach I think as useful as it has been needs to be augmented, I won't say it needs to be replaced, but it needs to be augmented, a much more holistic approach. You know, I can look at cells all day long but until they're organized into tissues, into organs and into systems and then into whole species and individuals it really sort of doesn't matter. So it's understanding human health, understanding how choices we make in policy, shifting policy decisions so that we put investment where it matters the most as far as human condition and I think letting people realize their full potential as far as health and happiness as well. My vision is if it can lead to opening and the democratization of health (Government).*

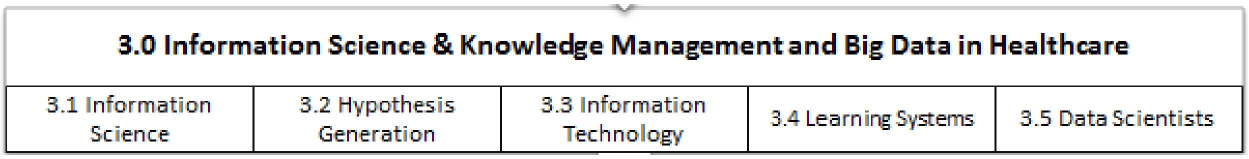
Summary

Harmonization between technology and humans was an essential perspective shared by many study participants of humanism in big data. While big data in healthcare is still an emerging phenomena, stakeholders in the three key stakeholder classes uniformly agree that the potential of big data will not be achieved without the complementary relationship between humans and technology. It was widely suggested that big data in healthcare is vastly different than big data in other industries because the complexity of human anatomy and physiology are not comparable to any “widget” that can be defined and produced by other industries. A consumer stakeholder summarizes this point succinctly:

> *One explanation is that one treatment is better than the other, that’s what you’d like to think, but first you have to resolve the possibility that it might be because the patients are different and that there’s confounding selection bias that patients who get one treatment are just different in ways that effect outcomes from the patients who get the other treatment (Consumer).*

As humans and technology combine to realize the potential of big data, big data information science and knowledge management is a framework that allows the vast “natural resource” of big data to produce precise insights and knowledge. The next category explores study participants experiential knowledge about the association and application of information science, knowledge management and the role of the data scientist in healthcare and big data.

Information Science & Knowledge Management and Big Data in Healthcare



3.0 Information Science & Knowledge Management and Big Data in Healthcare				
3.1 Information Science	3.2 Hypothesis Generation	3.3 Information Technology	3.4 Learning Systems	3.5 Data Scientists

Category Definition

This category consists of the interconnected fields of information science and knowledge management, facilitated by a team of scientists, primarily concerned with the analysis, storage, dissemination, and ontologies of big data and its knowledge engineering and visual representation. This category also includes study participant perspectives and insights on the skills and knowledge of the data scientist. The healthcare system, in part, is defined by its many disparate transactional (e.g., financial) and claims information technology systems which are created with the intent to derive healthcare intelligence.

3.1 Information Science

The intentionality of this important category of meaning arose from a hypothesis that there is an implied relationship between medicine and the discipline of information science. In saying true to phenomenological research, during the in-depth interviews, I did not frame a definition of information science – allowing the conversation to flow naturally based on the study participants experiential knowledge. Only one study participant inquired about what was meant by information science. All study participants were asked to specifically state “yes” or “no” and further elaborate either way. This study participant introduced the notion of cloud computing and the assumption of its privacy and security:

> *What [big data] means to me is that you assemble information from multiple sources but then get assembled in a large dataset and who knows where that actually resides in various servers around the country or even around the world I suppose which is sort of euphemistically referred to as the cloud and assuming that privacy and security constraints are being considerations for being adhered to which is I think a big question mark (Consumer).*

From the government stakeholder's experience, it was generally agreed that medicine is an information science, or at least it "*could be*." One government study participant responded, "*I'm not sure I'm qualified as a non-clinician, but yes*." In the information-rich field of medicine, when data was once a result of healthcare delivery, data and its resulting knowledge is now a prerequisite for delivering high quality, cost effective care. One study participant offered an argument regarding medicine as an information science is at the epicenter of care delivery:

> *It used to be you could almost argue it needs to be the center with providers and beneficiaries orbiting around it or at the very minimum it needs to be on the same level as what was previously considered the other core components in healthcare delivery, clinical knowledge (Government).*

One respondents' narrative challenged me to reflect even deeper on the question and how it was posed. Which discipline emerged first? Their responses elicited further exploration into the *history of medicine* and information by commenting, "*I would say that information science pre-dated medicine and allowed medicine to prosper*." While this statement is arguable, it piques a curiosity into further research and interpretation on the subject. A "double hermeneutic" was also accentuated, as I attempted to interpret the study participants interpretation of what was meant by an information science. The question objectively asked about medicine, but in reviewing my reflexive field notes, I wrote, "… *a definition of information needs to be included,* *and medicine could be misinterpreted as healthcare delivery*." A study participant highlighted this point pretty succinctly:

> *Yes, absolutely, medicine has always been an information science but whether or not that information has been at the individual patient level, at the group of patients a provider takes care of or at the population level has been the thing that has changed. So when we see the big data revolution we're seeing that transformation of the maturity of information science in medicine go from that individual patient, the anecdote to the types of patients, the whole constellation of patients I've seen my career, my experience, to understanding the relationship of clinical and biochemical data across population (Provider).*

The only stakeholder to definitively claim that medicine is not an information science resides in the government stakeholder class. Even still, the study participant suggests big data is "*also about smartly moving information around clinicians*," which is suggestive of the knowledge engineering and information sharing dimensions of information science. It is disputable whether a definition of information science would have biased this study participant's negative answer about big data as an information science. Looking back on my field notes, I documented the study participant was "*very sure of his response and gave no indication of uncertainty*." As sure as this study participant was certain medicine is not an information science, another study participant concluded:

> *Oh I've been saying that for 20 years. Medicine is an inherently an information science, the better data you have the better you can diagnose. The more effectively you can select treatment, the better you can actually see those treatments. It's unquestionably an information science (Provider).*

Unlike narratives from the government stakeholder class, a study participant from the provider stakeholder class acknowledges a reliance on just transactional data generated in the delivery of healthcare which was collected for that singular purpose. This study participant asserted:

> *I suppose you could add another of the Gartner's V's which is something that's been a problem has been Veracity meaning that all the transactional systems work properly when people enter things perfectly the way it's supposed to that doesn't always happen and even then just the way the data architecture is sorted in the transactional systems that has really designed for transactional processing (Provider).*

As a dimension of information science and a core attribute of healthcare delivery, including information technology and HIT interoperability, information sharing was discussed extensively by each study participant. There was a desire (possibly influenced by legislative mandates), among key stakeholder class to share data and information; however, a paradigm shift has occurred, according to one government respondent: "[*Data] is much cheaper to collect and it's cheaper to store … at the same time there's a cultural shift going on where people are much more willing to share it*." There was also disagreement about how much information is shared and the consequences for (or not) doing so. One government stakeholder wittily suggested:

> *We do share, but in my opinion we don't share enough. If I share something and lose out … then that kind of means I'm going to be less likely to share" or furthermore, "organizations don't want to share and we're not necessarily incentivized to share (Government).*

Whether these barriers such as incentives and competition are real or perceived, organizational culture and competition plays a central role in sharing information assets in healthcare.

Study participants from the provider stakeholder class shared a strong desire to share data, information and knowledge across the entire healthcare ecosystem. Empowering providers and consumers was mentioned as an objective, but one provider remarked "*especially in healthcare it's hard to share some of this data outside*" because of real and perceived barriers which includes "*transparency*", and "*people worried about sharing data or fighting for silos or turf.*" One provider stakeholder concluded that things are getting better because of the leadership role federal government has assumed and the historic precedent set by "liberating" big data and releasing it for research and innovation in sites like www.Healthdata.gov:

> *The good news is that federal government and state governments, are beginning to take the data that they're paying us for with Medicaid and Medicare and many other things and put it out there for us to use, for researchers to use and so when it comes to some clinical data, performance data you can actually begin to download big datasets publicly online (Provider).*

Providers are also collaborating to create cooperative big data sharing cooperatives. The perception is integrated delivery systems potentially have enormous amounts of big data and customarily keep it within the clinically or financially integrated health system for their own competitive advantage. But their big data combined with big data from other large delivery systems (e.g., "My big data – Your big data") creates an unprecedented amount of aggregated big data for precise decision making. Another study participant inferred:

> *We can start to collaborate together and as we work out the content of those data systems together so that we share the data back and forth it will accelerate the whole process. So the things that I could run a trial on that it would take me six months, I mean that's compared to ten years right now, the structure I might be able to get them done in six months, well under that structure we could do it in three weeks or at least that's the idea* (Provider).

3.2 Hypothesis Generation

While there was no attempt to generalize study participant's narratives, an ability to find answers to complex questions was a driver that resonated with each of the three stakeholder classes. As one provider study participant observed, "*If I can answer those questions, it can then lead me to more relevant questions of causes and the etiology of the disease*." Other study participants provided the following insights on hypothesis generation:

> *Having said that, the drivers that are pushing IT, getting us more into big data that will invite us to try and answer the questions that will allow computers to be more helpful are certainly driving costs (Government).*
>
> Also,
>
> *I think there are a whole lot of drivers to this. The data complexity, the data volume is increasing and the richness of what is there is increasing to throw out to ask questions we never could ask before. We're also collecting a lot of junk but clearly that's the big deal right, you go gold mining and it's not all gold (Government).*

Essentially, the three government study participants uniformly expressed the fact that big data allows for hypothesis generation and alternatively better question development in healthcare. In the provider stakeholder group, similar to the government stakeholder group, one stakeholder believes the ability to develop good questions is a byproduct of big data which has an ability to produce – good answers:

> *There's effectively no limit to them, so good answers come from good questions, nearly always to answer that good question you have to have data that matched that question, right, and so that's that idea back again (Provider).*

3.3 Information Technology

Information technology and clinical decision support are to conduits healthcare intelligence for providers ("*How do we make sense of all the data that we're getting from the science?*"). There is also relevant application in policymaking and healthcare consumerism. A consumer stakeholder suggests that clinical decision support is required to organize and generate contextually relevant information:

> *The second thing that's revolutionary is that that information then becomes the input for decision support engines and there's a whole bunch, there's a whole array of ways that clinical decision support can be set up, it can be in templates that prompt us to remember to do things we would otherwise forget. You know, chart order entry facilitators, again to help us to just make it easier and more efficient to order something because most of our orders are complicated and involve more than one thing or at least a lot of them are, you know, data presentation like graphs or spreadsheets and charts that's where we can see information over time, and then of course, prompts and alerts and things like that (Consumer).*

Several study participants found that clinical decision support is an important function of information technology which facilitates information organization and structure. A consumer study participant posits that this is the primary role of computers which are best suited for the task:

> *The other thing computers are really good at is when we have to actually slow down and think something through and figure something out then computers can organize information in ways that make it easy for us to solve difficult problems. So that's really the use of computers (Consumer).*

According to a couple of the government stakeholders, electronic health records must continually evolve to provide the clinical decision support and information structure that is necessary to organize clinical big data and its resultant information. National policy including

Meaningful Use Stage 2, which is a process designed to aid clinical decision support provides a standardized framework, but may not be achievable as a government study participant opines:

> *An example is the electronic health record where we're very poor at structuring clinical information so we come along and we turn everything into electronic form and we somehow expect that electronic records to solve all our problems and it doesn't do that unless you think through how you're going to structure the data before it goes in and what everybody else is doing. You're going to have big data and right now there are over 2,000 records that have been certified by CCHIT as meeting the Meaningful Use Stage One criteria and they're all written in different languages, different interfaces, different databases and they can't talk to each other. So it's kind of a mess* (Government).

Study participants point out an important task to clinical decision making that has not happened in healthcare: structuring the entire knowledge base of medicine ("W*e haven't structured the information in healthcare to the extent necessary to allow big data to have the kind of impact it will potentially have on the future")*. While the advent of new analytical methods and the Variety and Volume of big data in healthcare presents a tremendous opportunity to structure healthcare's vast body of knowledge in a meaningful way, a government study participant adds:

> *We need to go through the whole knowledge base of medicine that way and map it and it's, you know, nobody is even talking about doing that right now so we're a very long way from getting medicine to the point where we can do the kinds of things that they can do in other industries where the structure of data is simpler (Government).*

With the advent of advanced health information technology, including electronic health records, and personal health records, big data is an asset for provider organizations, such as Accountable Care Organizations, government agencies, and consumers, alike. Big data, which "*definitely correlates with technology*," has allowed the dimension of genomic data to enter into the equation of healthcare delivery, as a stakeholder pointed out:

> *Well, I think that the availability of big data is certainly something that's driving it, whether it's clinical technology to measure biochemical signals from people or sequence genes or sequence proteins or sample the air or whatever. The availability of data is one thing that's driving it (Government).*

Consistent with increasing healthcare costs, a provider stakeholder believes, "*Value based purchasing and moving away from the lack of accountability of fee—for service- service*" is another driver, while another spoke of the new electronic health record standards that are a result of new healthcare legislation:

> *I think another one is the Office of the National Coordinator has been pushing these Meaningful Use Standards and that's resulted in an abundance of data, and there's more demand on doing analytics with the data and they'll be actually in Stage Three more expectations around producing outcomes and you can't really produce outcomes without data (Provider).*

Two of the three study participants in the consumer class felt the unintended consequences of sharing big data were problematic. While sharing data and repurposing it for use by other stakeholders in the information value chain, consumers were concerned that if data ends up in the wrong hands, privacy will be potentially compromised. One study participant summed it by stating "*I might be concerned that the data that we shared might be used for some purpose other than the stated research questions*." While potential nefarious uses of protected health information do exist, the ability to link structured and unstructured data is the strength of technology:

> *I think obviously the rapidly emerging technologies in this area do allow people to crunch ever larger numbers of data in helping us bridge the gap between*

> *structured data analysis and unstructured data analysis, which I think is very important (Government).*

And the rise of "apps" developed to satisfy the demand for information access on mobile technology allows key healthcare stakeholders the ability visualize and assess information, often in real time, to make comparisons of peer activity, as one consumer stakeholder construed:

> *… here's a free iPhone app you can have so you can look in if you join our system and you can actually go in and look for individual patients, you can see how they're doing and what the gaps are for individual doctors, you can look at how they're doing in aggregate and they get their data from the EHR's (Consumer).*

Another consumer stakeholder study participant expressed an ability to make sense of the data and information they receive: "*Some of it is just trying to make sense of all the information that we have now so it's an organizing approach. Its how do we make sense of all the data that we're getting from the science?"* Another study participant suggested technology allows people the luxury to focus and think through complex problems rather than pour though intricate statistical operations and organizational exercises that once took weeks to accomplish can know be done in a matter of seconds:

> *The other thing computers are really good at is when we have to actually slow down and think something through and figure something out then computers can organize information in ways that make it easy for us to solve difficult problems. So that's really the use of computers (Consumer).*

3.4 Learning Systems

Participants in this study talked about creating healthcare learning systems which allow organizations involved the opportunity to "*ask and answer the questions, to share the findings broadly, and to drive quality up and cost down*." Learning systems in healthcare are similar to traditional clinical trials with the difference being learning systems allow provider organizations the ability to conduct clinical trial –like “research” on patient data warehoused in their information system networks, effectively generating thousands of published papers in a single year. The provider stakeholder revealed:

> *There’s this concept that effectively every patient goes on trial because of the way the data systems are structured. Now the jargon we used for that is a learning health care system where you build the learning, the knowledge management and it’s an information science tool that comes out of this you quickly learn is it’s perhaps the key capability in a system like this, it’s knowledge management. How do you identify best practice knowledge, how do you systematically and routinely deploy it into routine use (Provider).*

Another provider stakeholder mentioned his organization has created an immense data repository that warehouses patient claims data and demographic data, that while not standardized, allows multiple healthcare provider organizations to collaborate on a distributed learning network and learn from the data:

> *We are involved in a collaborative project that established a virtual data warehouse of basically it simplified data sharing by having a very reasonable similar data model that’s federated across all different organizations and it has demographic data, physical measures, personal medical history, management treatments, diagnosis, health claims and so forth and basically this data model retains control and stores data and stores kind of standardization across all sites (Provider).*

Generally, the consumer class produces massive amounts of source data from claims data, narratives and now sensors. Generally, as the participants for a host of public and private funded clinical trials, this class relies on others in the notional healthcare information value chain to reduce big data into credible information for its intended use. One study participant concluded:

> *To me it simply means lots of data, lots more than you're used to and you know, the reason big data is important is because without it you wind up with studies that are almost always too small. Smaller than ideal, because it's just simply is too costly to go out and collect all this data on the very large numbers of people that you need. So we're hopeful that the existence of these big sources of data allows us to do studies in a million people instead of 10,000 (Consumer).*

Members of the consumer stakeholder class are usually targeted to participate in clinical trials, or in this case learning systems in which "***e****verybody participates in some kind of learning,"* including patients. A provider study participant shared his vision of a learning system and shared the insight that learning can be distributed across all stakeholder classes. The study participant suggested healthcare organizations are in a central position to generate and spread clinical knowledge:

> *Once you have it you've created a learning environment and by a learning environment I mean a circumstance in which you can generate valid clinical knowledge by carefully structuring changes within that data environment, so I change a particular element of care and then causally figure out what that did to patient outcomes. So you see the idea? We call it a learning health care system (Provider).*

3.5 Data Scientist

I have always had a healthy curiosity about the role and skill that the "new" data scientist must possess with the advent of big data in healthcare. The provider stakeholder class offered a range of perspectives and insights into this profession. As clinical researchers, their training appeared to produce the richest insights into the knowledge and skills of a data scientist to manipulate big data. Consistent with findings from the literature review, the specialized skillset of the data scientist emerged as an important subunit of meaning across key stakeholder classes. I consciously set aside my own presuppositions about the skill and role of data scientists so not to influence the explication and interpretation of the data. Providers generate and consume large amounts of and require their employed or contracted data scientists to have "*an ability to think logically*," as one study participant surmised. Training in medicine, business, and the sciences were the trademark for this stakeholder group. As such, while each study participant has the analytical skills to lead data-rich environments, one study participant shared:

> *I have people who manage big data for me. I have a team of skilled data scientists who are part IT knowledge, part systems integrator, part subject matter experts, part analyst programmer; a data scientist isn't necessarily one person. You have a data scientist practice in which people specialize but talk to each other but you might have somebody doing the IT integration stuff and another separate subject matter expert and another programmer (Provider).*

Of note, there were a couple of colorful and profound insights elicited from the provider stakeholder class regarding the data scientist. One provider study participant proclaimed, "A *professional scientist or not … a scientist without data is a philosopher*" potentially as cautionary words of wisdom to scientists with such "sexy" titles.[12]

A couple of the study participants were aware of the potential limited labor supply of data scientists with the requisite skills to manage and analyze big data. Their narratives pointed out recruitment will be a barrier, as a study provider study participant acknowledged, "*There's still a very limited skill set out in the industry in terms of the people who know how to do this, so it's*

[12] In the book, Keeping Up with the Quants, Thomas Davenport and D.J. Patil proclaimed: "*Data Scientist: Sexist Job of the 21st Century*."

going to be hard to recruit a team of data scientists" and as another observed:

> *Finding somebody who's got 10 years of experience in big data is going to be pretty impossible to find. So getting experienced people, there's going to be a lot of on the job training and that's going to be a challenge for people (Provider).*

While providers typically use a third party administrator to perform the role of data scientist, particularly in small to medium sized provider organizations, the data scientists' competencies are more than "*creating reports and dashboards.*" It' also about being a trusted partner, managing large data sets with a degree of confidentiality, implementing analytics, , and "*creating the analytics and using the data and the knowledge and insights to translate them into decisions about how to improve*" the care of patients. One provider stakeholder gathered:

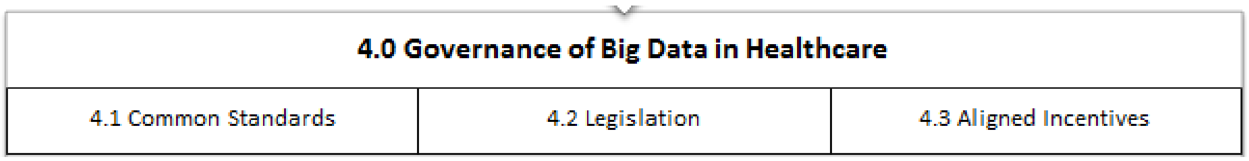
> *... now it's the funniest thing on this, most of my statisticians have some computer science background and regard themselves, they see themselves as fairly competent data architects. So as far as I can tell all of the data architects see themselves as analysts but when you're more than past familiarity with both fields you're different, and they're radically significantly different* (Provider).

It appears to be a fair assessment to say that, as another provider posits, "*a data scientist is more than one person*" and in order for knowledge to be optimally gleaned and analyzed – to fully thrive – another study participant suggested we need "*citizen scientists*" who might find insights overlooked by the relatively small cadre of bona fide data scientist.

Summary

This category examined the study participant's perspectives and insights into the place and role of information science and the skills of the data scientist. Information science and knowledge management are interdisciplinary fields that are essential to realizing the enormous potential of big data in healthcare. However, information science is typically implied as core a discipline in healthcare and rarely acknowledged as the foundation of healthcare delivery. The role of the data scientist is also critical to harnessing the potential of big data in healthcare. However, several study participants posit that the role of the data scientist is multi-faceted and usually does not consist of a single person. Even still, it’s recognized that the combination of knowledge and skill of the data science team are in short supply. The next session examines the results of a common objective across the key stakeholder classes: governance of big data in healthcare.

Governance of Big Data in Healthcare



Category Definition

This category examines study participant narratives about the attributes that are essential to establishing and sustaining a consensus-based framework for broad oversight and governance of policies and definitions of big data in healthcare. A set of common standards for big data could help improve data exchange among all healthcare stakeholders and would enable patients and providers to isolate parts of health and medical records, respectively, for refined analysis and information sharing. Classification systems called groupers, which include Episodic Care Groupers, (ECG) and Ambulatory Care Groupers (ACG) describe the "*illness-burden*" of populations (Weiner, Starfield, & Lieberman, 1992). While groupers are used within the healthcare industry for specific purposes (e.g., risk adjustment), they are not adopted as a universally accepted standard of big data.

4.1 Common Standards

Within the government stakeholder class it is known that a lack of governance and common standards, or a “*central use policy*” termed by one government stakeholder for big data in healthcare stifles big data growth and a realization of the true potential of big data. Such a deficiency appears to keep big data firmly entrenched in a spiral of big data “*hype*.” A familiar theme materialized in the government stakeholder class: establish a consensus-based common big data definition. Study participants concluded it will take time to refine and validate a working (or agile) definition, as precision in decision making and hypothesis testing are important attributes of big data's output. One study participant remarked:

> *I want to get people to use the same definition for enough time so that we can aggregate data, match it up against reality and then refine the definition so that the sensitivity and specificity of it, the accuracy, when the true positives and not a lot of false positives, not a lot of false negatives, so that all gets worked out by making the definition very precise and having people record it that way (Government).*

Study participants suggested activation of a common standard for major clinical problems represented in patients would include "*both processes and outcomes of care*." Of particular importance, the United States was on the threshold of a conversion from the International Classification of Diseases – Version 9 (ICD-9) to ICD-10 which facilitates data better analysis of disease patterns and treatment outcomes among a host of other healthcare benefits. While implementation of ICD-10 has been delayed, the updated code set with requires detailed clinical documentation could be the impetus to cultural change to include both processes and outcomes of care that the study participant suggests. However, there are segments within the healthcare industry that oppose conversion to ICD-10 including costs of implementation, a lack of an infrastructure to conduct end-to-end testing, and simply an aversion to change.

With the multitude of applications and software vendor products, such standards would be fundamental to comparisons across a uniform set of big data. As it stands, even with the advent of electronic health records, no such standards are ready for testing and validation. Study participants in the provider stakeholder class recognize the lack of a common standard for big data in healthcare. In research, which is a data-intensive endeavor, there are institutional review

Boards (IRB) which govern data and data collection standards that "*say what people can and can't do with the data and who can and can't see things to protect our patients*." In the delivery of healthcare, a study participant felt that without big data standards and governance, data aggregation would not be possible and questions whether big data in healthcare is big enough due to the absence of a common set of standards:

> *We need our big data to get bigger. We need to actually aggregate this stuff. So to do that what does the future need – it needs standards. It needs standards for clinical data, standards for research data, standards for educational data, that's beginning to emerge but it's definitely not there yet. We need reasonable and rational policies around how to protect these data but also how we can flexibly use and release the data (Provider).*

The absence of big data governance and common standards is a perceived barrier to big data's untapped potential. This raises an important question: without big data standards in healthcare, is big data truly big?

4.2 Legislation

Healthcare legislation over the last twenty years has been a focal point of political debates at the national and state levels. As such, legislation was another influencer identified by several of the study participants. A government key stakeholder supported the notion that government rules drive big data in healthcare, "*Because it places data and the ability to harness and leverage data at multiple points throughout the healthcare ecosystem at the center as opposed to at the trenches.*" This stakeholder further added:

> *I think another one is the Office of the National Coordinator has been pushing these meaningful use standards and that's resulted in an abundance of data, and there's more demand on doing analytics with the data and they'll be actually in*

> *Stage Three more expectations around producing outcomes and you can't really produce outcomes without data (Provider).*

A core issue for each of the three stakeholder classes is repurposing data for different uses which could lead to breaches of privacy. In the absence of big data governance in healthcare, study participants identified several unintended consequences that potentially could come to fruition based on the current informal structure and a lack of governance associated with big data in healthcare. A stated unintentional consequence was a lack of oversight and adherence to data privacy policies, which one study participant noted, "*There's a lot of work to be done on patient privacy and oversight.*" Eventually without governance, the healthcare industries will "*just kind of give up on privacy*." A consumer study participant remarked:

> *I think it's unrealistic to expect providers and patients to really adhere to strict privacy and security standards which for the past 10 years we've taken very, very seriously and then have it a very, very highest governmental level completely all of those standards just find out that a government agency is writing rough shot over them and just sort of shrug our shoulders and say you know, who knew, but I mean maybe that's you know, maybe that's possible but I think it really is a crisis that has to be addressed (Consumer).*

Study participants in the consumer class advocated for rigorous patient privacy policies. Two consumer study participants suggested that while there are standard patient privacy rules in effect, the industry should rethink these rules because, "*we just kind of give up on privacy and say well that's kind of over or we have to say no look we actually do take privacy and security seriously.*" Study participants pondered the questioned current federal rules regarding Institutional Review Boards and human study subject oversight as a barrier to effectively employ big data. A consumer study participant agrees there is much more work to be done, even potentially easing the framework of current patient privacy rules:

> *Yes, everything has to be done to keep this data secured and protect people's privacy, on the other hand, when these studies are not posing any physical harms to patients because we're just looking at data, you know, you do not need to require that a patient sign a 10 page or 20 page consent form. Even in certain randomized trials, yes, you need a consent form but it doesn't need to be 20 pages long if it's a very low risk question. So I think figuring out these issues about now that we've got big data, how do we work with IRB's and human subjects oversight to rationalize how we use it and how we talk to patients about use (Consumer).*

4.3 Aligned Incentives

A consumer stakeholder believes that within the current unstructured approach to big data in the healthcare industry, "*incentives aren't necessarily aligned to make it easy to change*." Another study participant articulated, "*Raising the risk of what I consider to be misuses of data in ways that are not necessarily in the public interest*" in the absence of big data incentive alignment. These risks create different standards for different stakeholder classes; this results in unaligned incentives. The confluence of a lack of both data standards and transparency ("*I think we're going to have to have a lot of transparency*") creates a culture of mistrust among healthcare stakeholders. A study participant commented:

> *If it's holding providers to a different standard then you can't tell what's going on, so an ideal system ... you get to transparently manage your population according to outcomes that everybody agrees upon both inside the delivery system and among the ones are paying for it (Consumer).*

Uniform data standards like ICD-10 exist to classify illness. However, ICD-10 is one data set among potentially hundreds or thousands used in the delivery of healthcare. The absence of a common classification system, ontologies, policies and aligned incentives for big data emerged as a real barrier to realizing the potential of big data.

Summary

Common standards and privacy are commonly referenced subunits of meaning identified by the three key stakeholder classes. However, study participants highlighted the unintended consequences of increased competition to develop and publish healthcare intelligence and unaligned incentives are barriers to effectively achieving big data's potential in healthcare. There is evidence of common standards on data through government policies, including Meaningful Use. But big data taxonomies and ontologies are nonexistent. This is especially troublesome given the emergence of genomics data as an integral source of data that enables precision medicine and informed decision making. Governance of big data in healthcare is an objective of the three key stakeholder classes and must include "*the patients who need to be kept in the governance*."

CHAPTER VII. DISCUSSION AND INSIGHTS

Overview

This section discusses the findings of the research and compares study participant insights to observations found in the scholarly and grey literature. The study's research question, *Within and across the narratives of three key healthcare stakeholder classes, what are the important categories of meaning or current themes about big data in healthcare,* was designed to elicit a priori insights into the attributes, definitions, and uses of big data in healthcare. As a reference, I restated the observations found from the modified systematic review of the literature to make comparisons between findings from the literature and study participant narratives. The research uncovered important categories of meaning or themes within and across three key healthcare stakeholder classes. The aim was not to generalize the study's findings. Rather, the explication of study participant narratives was intended to delineate categories of meaning to find common themes within and across study participant narratives and construct a cohesive 'story' or framework of big data in healthcare. Unique themes were also included as an important source of data. Also, a "main takeaway" is offered at the beginning of each category as a fundamental fact or point of reference for all stakeholder classes to adopt.

Category of Meaning #1: Ontological Framework of Big Data in Healthcare

Main Takeaway: Without a consensus-based "framework" of big data in healthcare, 'buzzwords' and slogans will continue to play an important role in describing big data's meaning in healthcare.

Many characteristics, definitions and references to big data across various industries were mentioned by study participants. A review of the scholarly literature found a host of definitions on "big data" including "*it's [big data] a characterization of the never-ending accumulation of all kinds of data*" (EMC2, 2012), and big data "*is the ability to mine and integrate data, extracting new knowledge from it*" (Roney, 2012). Or "*big data is the belief that any sufficiently large pile of sh** contains a pony"* (Arbesman, 2013). Begley (2011) defined big data in healthcare as "*the healthcare context to longitudinal medical claims data for millions of patients linked to their EHR (p .50)*" Begley's definition conservatively quantifies big data in the "millions" where petabytes, even terabytes are now the gold standard of healthcare big data quantification. This definition illuminates a common problem of attempting to quantify big data in healthcare: data are counted by patient encounters, not petabytes.

The scholarly and grey literature on "big data" and "healthcare" also confirmed there is no consensus on what big data means in healthcare (Dumbill, 2013; Villars et al., 2011). Findings from the stakeholder narratives were consistent with the literature. Study participants generally did not know what big data in healthcare meant ("*nobody's ever defined it for me*"; "*it's like a parody*"; and, "*it needs to be clarified and demystified*"). While Gartner's credible "3V's" of High Volume, High Velocity, and High Variety were referenced across the three key stakeholder classes, the oft-cited 'characteristics' of big data is not a definition. Gartner's characteristics of big data have entered into the lexicon of big data in healthcare as buzzwords (T. Borangiu & V. Purcarea, 2008; Davenport et al., 2012; Rooney, 2012) that continue to play an important role in the absence of a vetted consensus-based definition.

Depending on the scholarly communication or source of grey literature, one could find at least two additional “V’s” – Value (Porter & Teisberg, 2006) and Variability (Gartner, 2013) that have emerged as some of the important buzzwords that characterize modern big data.
Across all stakeholder classes, the multitude of big data definitions do not sufficiently address the enormous complexity of healthcare’s aim of delivering precision medicine, commonly referred to as personalized medicine. In a recent paper by Ward and Barker (2013), they collated four common definitions of big data which “gained some degree of traction” (p. 1) agnostic to industry and market sectors. The definitions were extrapolated from big technology and consulting firms, including Gartner, Intel, Oracle, and Microsoft. After generalizing characteristics of each company’s interpretation of big data, they constructed their own definition: *Big Data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including but not limited to NoSQL, MapReduce and machine learning* (Ward & Baker, 2013). This is a progressive definition of big data that intentionally omits Gartner’s 3”V’s”. Instead, Ward and Baker take into account the tools and technology required to evolve big data into information and knowledge. Yet, based on key stakeholder perspectives, their definition seemingly falls short of recognizing the complexities of the “*black box*” of big data in healthcare - people.

Ward and Baker’s insight into big data is consistent with the perspectives of key healthcare stakeholders: big data is only a single dimension of a larger framework whose end goal is precise information derived with a purpose. Shaw (2014) points out “historically, … scientists would plan for an experiment to collect and analyze data … because of the price of storing a bit of data has dropped 60 percent … people now collect everything and then search for significant patterns in the data" (p. 34). Several study participants disagreed with such a theory. A provider study participant pronounced "*it's important that you know what the purpose is before you start.*" The vision of big data from the perspective of a consumer study participant was to "*have strict purposes rather than let's create big data and then figure out what we might do with it.*" Shaw even agrees his perspective has its inherent risks, which includes data dredging – data which is statistically significant by chance resulting in poor "scientific output from throwing everything against the wall and seeing what sticks*"* (p. 34).

Study participants offer the wisdom that though a consensus-based big data definition is necessary, its maturity and widespread adaption will not happen overnight. A government stakeholder postulates the industry needs to "*use the same definition for enough time so that we can aggregate data*" for initiatives like healthcare learning networks.

Category of Meaning #2: Humanistic Dimension of Big Data in Healthcare

Main Takeaway: There is a dual 'humanistic dimension' to big data in healthcare that takes into (1) account people's cognitive contributions; and, (2) the uniqueness of human data as a unit of analysis.

Big data in healthcare, in part, is about empowering people with information and knowledge to make evidence-based decisions about policy, clinical treatment plans, and healthcare consumer choices. Study participants agree, "*Empowering the people whose data it is should be an important value for all of us as we go forward.*" Medical humanities and medical ethics as a potential practical application in the policymaking process (Greenhalgh & Russell, 2009) is an intensely explored subject. Yet, key healthcare stakeholders generally agree narratives and the ability to listen to stories remain an essential skill that adds to the body of knowledge in evidence-based healthcare – augmented with the power of big data analytics. Study participants fear the art of listening to stories and the "*human therapeutic relationship*" will be lost with much of the industry focus on conquering the big "data deluge." A government stakeholder further elaborates that "*purely relying on machine learning without the application of subject matter expertise"* does not foster precise knowledge for decision making.

While computers are a necessary requisite and tool of big data in healthcare, the human dimension of big data cannot be lost in the "hype" of defining big data in healthcare. Study participants agree that computers organize information extremely well, but the human mind is calibrated for unparalleled intuition, speed and pattern-based recognition. Absent from any big data definition, characteristic or attribute found in the scholarly and grey literature was the importance of the humanistic dimension of big data in healthcare. Data and information in healthcare is still imprecise. Whether used for development of new healthcare legislation or patients sharing stories about health and healthcare, narrative provides meaning, context, and perspective (Greenhalgh & Hurwitz, 1999). Gardner (2013) found "individual instances [narratives] are part of an ever-growing study of pedagogy of a health humanities approach that focus on narrative, sometimes called 'narrative medicine'… and involves narrative in a number of ways, including qualitative analysis" (p. 4). The essence of this research advocates for introducing narratives beyond exam rooms, but across the healthcare information value chain, especially as a data collection methodology that is part of big data as a source of clinical and policy making data.

In addition, the human body is uniquely complex. Study participants surmised that big data in healthcare is unlike big data generated in other service industries ("*genomics, senomics, and other "-omics" that are out there. Metropolomics for example are generating huge amounts of data that need to be processed in ways different than we have in the past" - Provider*). Transactional data used in day-to-day healthcare delivery, human genome data, and human microbiomic data, if integrated with social epidemiological data, will eventually create an unthinkable amount of big data from just a single person.

Category of Meaning #3: Information and Knowledge Science and Big Data in Healthcare

Main Takeaway: The ability to link and visualize genomic, environmental, and other heterogeneous sources of complex data positions the disciplines of information science and knowledge management at the center of the delivery of healthcare and medicine.

The literature review produced the following observations: (1) consumers of healthcare do not have enough trustworthy, credible information to understand the scope and depth of big data and its impact on their health and healthcare, and (2) the gulf between "life sciences" and "healthcare" is closing – fast. Big data is entrenched in life sciences research, including genetics, biomedical research, computational biology, and nanomaterial science. However, these sciences are quickly making its way into point-of-care decisions (e.g., shared physician and consumer decisions about treatment plans).

In its strategic plan for the Department of Medical Information Science at the University of Illinois administrator's confirmed, "*in the 21st Century, Medicine will be viewed as an Information Science*" (Schatz, 2006). Shaw (2014) proclaimed "information science promises to change the world" (p. 1). This statement is consistent with most study participant's ideology about medicine as an information science, or as one provider stakeholder stated, "*medicine and maybe you mean healthcare is an information science.*" In fact, a government study participant speculated, "*Information science pre-dated medicine and allowed medicine to prosper.*" While his proclamation is debatable, there is no question among most key healthcare stakeholders, medicine, or healthcare, is an information rich endeavor (Villars et al., 2011) that is "*also about smartly moving information around clinicians.*" A government stakeholder gathered, "*Until you were able to classify that in terms for differential diagnosis and how I should treat it, medicine was basically voodoo and witchcraft ... the better data you have the better you can diagnose*."

Indirectly, another provider summarized information science as being foundational to medicine: "*the more effectively you can select treatment, the better you can actually see those treatments*." At the very core of medicine is science and evidence (J. Bellamy & Bledsoe, 2006; Sackett et al., 1996; Thyer & Myers, 2011), proliferated by the disciplines of information science and knowledge management. But according to several study participants across the key healthcare stakeholder classes, medicine is much more than an information science.

In 2006, the National Science Foundation identified a core set of capabilities that are fundamental to the role of the data scientist including: collaboration, coordination, and the ability to conduct research and education using digital data collections; serve as a mentor; and, design and implement education and outreach programs. These capabilities are consist with Davenport et al. (2012) who wrote, "data scientists understand analytics, but they also are well versed in IT, often having advanced degrees in computer science, computational physics or biology- or network-oriented social sciences. Their upgraded data management skill set — including programming, mathematical and statistical skills, as well as business acumen and the ability to communicate effectively with decision-makers — goes well beyond what was necessary for data analysts in the past" (p. 23).

Study participants want data scientists to also be able to "*think logically*" with "*profound specialty knowledge*" and perform "as a *competent data architect*." Many of the skills identified by the National Science Foundation were noteworthy among study participants across the classes. Healthcare consists of many domains, (e.g., quality, payment, policy) and in order to effectively create information from big data in healthcare, specialty domain skills and knowledge are an essential capability key healthcare stakeholders. Study participants identified a list of skills and knowledge necessary for the data scientist to become an integral member of the care delivery team. Participants of this study advise simply calling yourself a data scientist does not necessarily make you a data scientist, as one government stakeholder points out, "*A scientist without data is a philosopher*."

Harvard Business Review touted the data scientist as the sexist job in the 21st Century (Davenport & Patil, 2012), with demand for data scientists sharply on the rise. The U.S. alone will need 140,000 to 190,000 people with deep analytical skills by 2018 just to keep up with the pace of innovation (Brown et al., 2011) and the explosion of big data. The problem as two provider study participants observed, "*There's still a very limited skill set out in the industry in terms of the people who know how to do this … finding somebody who's got 10 years of experience in big data is going to be pretty impossible to find*."

The healthcare industry is inherently one of the most information-rich market sectors. Study participants surmise the entire healthcare ecosystem would be well served by uniformly employing the disciplines of both information science and core. This vision can only be realized with governance and a common set of standards. The next section explores governance of big data in healthcare.

Category of Meaning #4: Governance of Big Data in Healthcare

Main Takeaway: Data stewardship, modern and refined privacy rules, and a set of common standards are required for all healthcare stakeholders to realize the benefits of big data in healthcare.

The NCVHS is an eighteen member statutory public advisory committee to HHS that has created selection criteria for interoperable clinical data standards and standards for e-prescribing body (Grossmann, 2010) and other national standards for federal rule-making. No standards have been passed or are currently under consideration for big data in healthcare (Pavolotsky, 2012) – a vision of several key healthcare stakeholders. Study participants from both provider and consumer stakeholder classes envision "*widespread integration of administrative, clinical and patient generated data that will be available through big data*." But the literature suggests a fundamental barrier to widespread big data integration: health system fragmentation (L. R. Burns et al., 2002) of heterogeneous health and healthcare data (Grossmann, 2010).

Consistent with the literature, participants in this study identified competition (Cukier, 2010; Frangenberg, 2013; Grossmann, 2010) as a problem in the healthcare industry ("*There's a "desire to maintain a competitive edge" – Provider and "that means having a competitive advantage over somebody else and in today's world that is information." - Government*). Study participants across all key stakeholder classes generally agree the lack of a governing body and organizing framework for big data in healthcare prevents the industry from realizing the true benefits of big data in healthcare. Several study participants called for a "*common set of standards and user policies*." In the absence of such a framework, unintended consequences such as barriers to wide-spread sharing will continue to plague the healthcare industry. Study participants offer the wisdom that though a consensus-based big data definition is necessary, its maturity and wide-spread adaption will not happen overnight. A government stakeholder postulates the industry needs to "*use the same definition for enough time so that we can aggregate data*."

Study participants believed privacy was an issue as several pointed out current federal rules are not appropriate for big data in healthcare. In order for privacy to be effective, HIPAA rules must be revisited, as patients are sharing increasing amounts of data about themselves and their health. McGraw (2012) asserts "federal privacy regulations do not set clear and consistent rules for access to health information to improve health care quality" (p. 75). The linkage of life sciences data (e.g., genomics) alone to traditional transactional healthcare data dramatically changes the privacy landscape, effectively requiring an overhaul of healthcare privacy laws. Genomic information is fundamentally identifiable and the privacy implications are profound (Shaw, 2014).

Contributions and Implications for Future Research

This research is significant because it: (1) produced new thematic insights about the meaning of big data in healthcare through narrative inquiry; (2) offered an agile definition of big data that can be deployed across all industries; and, (3) made a unique contribution to scholarly qualitative literature about the phenomena of big data in healthcare for future research on topics including the diffusion and spread of health information across networks, mixed methods studies about big data, standards development, and health policy.

In Burns (2013) feature article, Healthcare's Big Data Tsunami, the author postulated, "the big data tsunami in healthcare is washing ashore today and few healthcare organizations are effectively dealing with it" (p. 59). The next logical question is: why are healthcare organizations not be prepared to effectively deal with what is widely presumed to be an organizational asset (and in some circles, healthcare's "natural resource")? Through qualitative and phenomenological research using narrative, this study provided new knowledge about the important categories of meaning of big data in healthcare through the insights and perspectives of nine key healthcare stakeholders. The results found big data in healthcare remains poorly defined – relying almost exclusively on axioms to explain its purpose, provenance, and meaning. Dr. Myles Rennaker, director of a governmental agency admits, "*Nobody ever defined for me.*" While Gartner's widely-publicized (updated from 3) "4V's" of High Volume, High Velocity, High Varity and High Veracity is entrenched into the lexicon of healthcare organizations, Dr. John Boyken, associate dean at a major medical school adds, "*It's a popular term that means a lot of different things to a lot of different people*." Buzzwords are deeply-rooted as important descriptors of big data. They provide sorely needed context to a potentially transformative organizational asset. Nonetheless, Dr. Rennaker concludes healthcare standard's organizations must "*clarify and demystify*" big data in healthcare.

Findings from this qualitative study also uncovered a critical dimension of big data that perilously has been overlooked, or dismissed, in the many well-intended offers to "characterize" big data in healthcare: the humanistic dimension of big data in healthcare. The humanistic dimension of big data emphasizes the cognitive prowess and contributions of the human mind, the extraordinary complexity of the human body as a source of big data, and the lost narratives and relationships forged between people. And as a government stakeholder shared after reading the executive summary on the study, "*I think you have articulated the attributes that make healthcare different. This paper represents a contribution to resetting expectations more in line with reality, which can facilitate more effective use of computers and large databases to contribute to research, diagnosis, treatment, and quality measurement."*

The widespread integration of vast amounts of genomics data, environmental data, and new sources and diversity of data generated by wearable devices and sensors with traditional transactional healthcare datasets requires improved statistical, computational methods, and visualization tools (Shaw, 2014). The healthcare industry is at the threshold of such widespread big data integration, fueled by the Triple Aim of improving the experience of care, improving the health of populations, and reducing per capita costs of health. Such a vision is why the interdisciplinary fields of information science and knowledge management play a crucial role in the delivery of 21st Century medicine.

Health and healthcare data provenance include metadata and Meaningful Use attribution data, not to mention public health surveillance data and global health data. With the never ending possibilities of adding to healthcare data provenance, there was near unanimous consensus that big data in healthcare requires a common ontology for healthcare organizations to effectively utilize this “natural resource.” With truly massive amounts of heterogeneous big data being collected now in disparate databases, there is a concrete need for standards advisory organizations like the National Committee on Vital and Health Statistics (NCVHS) and the National Institute of Standards (NIST) in partnership with private sector companies and federal organizations to recommend a consensus-based definition and ontology of big data in healthcare. Big data generation and integration in healthcare is best served by defining its provenance, privacy and precision, and purpose (4”P’s”). Study participants concluded governance of big data in healthcare will allow healthcare organizations to not only “effectively deal with the data tsunami,” but generate and share sought after knowledge and wisdom for healthcare intelligence across the healthcare information value chain.

Big data in healthcare is not customarily discussed in qualitative terms. While not intended to be generalizable, this phenomenological research uncovered foundational insights and perspectives capable of augmentation with basic research in disciplines to include social network analysis and health policy development. For example, a phenomenology study using narrative can inform policy makers and researcher which barriers impede the flow of information between key healthcare stakeholders and how healthcare stakeholders influence the fidelity of information that is shared within networks? The findings from this rigorous qualitative study that uncovered the “know about” big data can be used as the foundation to conduct further mixed methods research hypotheses that explores the “know that” about big data. Such a study using regression or path analysis can then generalize the themes and subunits of meaning found in this study. Furthermore, these findings can also provide standards advisory organizations with experiential insights and knowledge about defining a big data definition germane to health and healthcare.

Finally, derived from the nine key healthcare stakeholder narratives, I offer the following agile “definition” of big data, which could serve as a spring board for a consensus-based framework for big data in any industry:

> *“Big data” is both an organizational philosophy and strategy, enabled by information science discipline, to purposefully collect, link and analyze a variety of heterogeneous data resources and data ontologies, requiring the confluence of people and computers to generate precise information displayed through advanced visualization tools.*

Lessons Learned

There were many valuable lessons learned from conducting this phenomenological study.

First, among the many qualitative methods available to me to conduct this important research, a phenomenological study using narrative was appropriately chosen to answer the research question. This research is an important foundational qualitative study to fully understanding the meaning about big data in healthcare. The experiential knowledge of key healthcare leaders provided timely, thick descriptions the big data phenomena in healthcare. Perhaps a mixed methods study design would add further rigor to the findings in this study. Using modern quantitative data analysis methods adds tremendous insight and value (Shaw, 2014). Weber (1990) points out that the “best content-analytic studies use both qualitative and quantitative operations” (p. 2). Future research using a mixed methods approach would certainly yield new insights and rigor to the research topic, particularly as big data and information sharing practices are explored.

Second, interviewing patients, caregivers, and other healthcare consumers would have been ideal – achieving an unparalleled richness and truth about healthcare consumer’s views. The consumer advocates provided outstanding narratives; however, the voice of the patient is rarely integrated into policymaking. I have developed a passion for capturing the narratives of healthcare consumers and look forward to pursuing such work in future academic and professional endeavors.

Finally, the phrase “large data sets” was often found in the literature but was not included in this study so to maintain consistency with the study term, “big data.” In retrospect, including “large data sets” might have added additional sources of scholarly literature to the study. Several study participants mentioned, “*Managing large data sets of implementing analytics*.”

CHAPTER VIII. CONCLUSION

Within and across each of the three key healthcare stakeholder classes, big data in healthcare remains a misunderstood phenomenon. Unfortunately, the absence of a consensus-based, industry-wide definition of big data enables buzzwords to maintain a prominent and important descriptor of the phenomena. While key healthcare stakeholders accentuated a keen awareness of big data, most lacked a concise understanding of its meaning and relied on either Gartner's 4 V's characteristics of High Volume, High Variety, High Velocity, and High Veracity as a definition or conceding to not understanding what it really means. One consumer stakeholder frankly admitted:

> *I think the cause of big data would be better served by characterizing it more clearly for a lot of us if they want to be able to move forward. People can be afraid of it. So between not really understanding what's meant by it, whatever it is, it needs to be clarified and demystified I think, mainly clarified I'd say because I'm not sure what the hell they're talking about* (Consumer).

Big data is employed extensively in other industries in which a multitude of lessons learned can be applied. However, there persists a shortsighted supposition that big data in healthcare is the same as or even nearly identical to big data in industries that *define* their products. Stakeholders agree that the human dimension of big data is what makes big data in healthcare unique from every other industry sector – from human's cognitive ability to recognize patterns to our complex physiology and genetic makeup. A common unit of analysis in healthcare is a human who's phenotypic and microbiomic makeup is unique from one individual to the next.

Information science is an interdisciplinary field that is a fundamental core to delivering evidence-based medicine and healthcare intelligence. The information science framework includes the 3 ”C’s” of big data collection, classification, and curation as well as linking and creatively visualizing big data sets (Shaw, 2014) over its lifecycle. The information science field enables the transformation of “big data” into “smart data,” which satisfies stakeholders thirst for precision and trust, to be used for a variety of healthcare intelligence uses. The reformed healthcare industry which demands exceptional value for care delivered is in the midst of an emerging health information economy which requires a new big data governance framework where health information technology interoperability, metadata provenance, usage policies, and common standards will allow big data to be analyzed and shared across a connected, “many-to-many” healthcare information value chain.

In summary, this research provided four main categories of meaning and four takeaways for key healthcare stakeholders to consider:

1. Without a consensus-based “framework” of big data in healthcare, ‘buzzwords’ and slogans will continue to play an important role in describing big data’s meaning in healthcare.
2. There is a dual ‘humanistic dimension’ to big data in healthcare that takes into account (1) people’s cognitive contributions and (2) the uniqueness of human data as a unit of analysis.
3. The ability to link and visualize genomic, environmental, and other heterogeneous sources of complex data positions the disciplines of information science and knowledge management at the center of the delivery of healthcare and medicine.
4. Data stewardship, modern and refined privacy rules, and a set of common standards are required for all healthcare stakeholders to realize the benefits of big data in healthcare.

Finally, medicine is not only rooted in information science. It is a confluence of many other sciences and arts, including the medical humanities, which include capturing patient narratives and their unique ‘stories’ in an ethical manner. Such big data need not sit stagnant in electronic health records, but be used as a credible source of ‘big data’ that generates knowledge about personalized healthcare. This is the disruptive innovation in a reformed, patient-centered healthcare system that healthcare policymakers must seriously employ as a credible data source in the development of healthcare policy. As one provider stakeholder fittingly summed up big data in healthcare:

> *Big data doesn’t mean unstructured data. You always create data for a purpose, right. That’s the human creation. It always has purpose, you have to understand the purpose if it’s going to be effective.*
>
> *And then everything else is just a tool.*

APPENDIX A. STUDY DESIGN

Phenomenological Study using Narratives

Phenomenology is a philosophy that had its beginnings in the early years of the 20th Century and became explicitly aware of itself in 1913 (Husserl, 1970). Phenomenology became popular in the social and health sciences, especially in sociology (Borgatta & Borgatta, 1992), psychology (A. Giorgi, 1985), and education (M Van Manen, 1980). While phenomenology has a rather ambiguous history, as late as the 1970's, its popularity in the social and health sciences has potential applicability to current healthcare issues, including the persistent phenomena of healthcare disparities, social epidemiology of social networks and population health. Phenomenological research tends to converge with qualitative research strategies (Amedeo Giorgi, 2009) in which narratives are used as data (Clandinin, 2013).

Phenomenological and narrative-based methodologies have a modest history in public policy. These methodologies embrace an assortment of epistemologies ranging from interpretative methods to empirically-oriented narrative policy frameworks. While narratives are indeed used in the exploration and practice of policy, my practical experience in healthcare policy development lead me to believe general lay person narratives offered in the policymaking context are frequently treated as purely persuasive mechanisms, not as part of the body of evidence (Steiner, 2005) relevant to phenomena, policy-making or public administration (Borins, 2012). A Cornell University e-Rulemaking Initiative (Epstein, Heidt, & Farina, 2013) perhaps frames the void of multidisciplinary collaboration between the general lay person and government policy-makers best:

> *Given the disparity in power between government decision-makers and the public, ways of arguing for a particular policy position and perceptions of valid evidence constitute important boundary objects that make many civic engagement efforts ineffectual. Members of the lay public largely do not have the skills and the culture necessary to engage in formal argumentation based on empirical data. Yet, they possess the unique situated knowledge of living with existing policy or proposed policy changes. Helping the two communities to establish a shared repertoire may help in creating better policy solutions (Epstein et al, 2013, p.20).*

Epstein et al (2013) also provides a coherent perspective for capturing the narratives of both policymakers and the general public with the creation of a narrative framework that embraces the "value of narratives as input in the policymaking process" (p. 1). In today's modern healthcare delivery system, there remains a dearth of phenomenological studies encompassing narrative (Clandinin, 2013; Amedeo Giorgi, 2009; M Van Manen, 1980). Scholarly evidence supports my decision to approach the inquiry of big data in healthcare through semi-structured interviews with ten leaders from three key healthcare stakeholder classes: government, providers, and consumers. A narrative describes the lived experience of a single individual*;* a phenomenological study describes the meaning for several individuals of their *lived experiences* of a concept or a phenomenon. Phenomenologists focus on describing what all participants have in common as they experience a phenomenon. The basic purpose of phenomenology is to reduce individual experiences with a phenomenon to a description of the universal essence (M Van Manen, 1980). The following (Table 9) provides a comparative summary of potential study design options considered to conduct this study.

CHARACTERISTICS	**NARRATIVE RESEARCH** (DENZIN & LINCOLN, 2002)	**PHENOMENOLOGY** (MOUSTAKAS, 1994)	**CASE STUDY** (STAKE, 1995)
Focus	Exploring the life of an individual	Understanding the essence of the experience	Developing an in-depth description and analysis of a case or multiple cases
Type of Problem Best Suited for Design	Needing to tell stories of Individual experiences	Needing to describe the essence of a lived phenomenon	Providing an in-depth understanding of a case or cases
Discipline Background	Drawing from the humanities including anthropology, literature, history, psychology, and sociology	Drawing from philosophy, psychology, and Education	Drawing from psychology, law, political science, Medicine
Unit of Analysis	Studying one or more individuals	Studying several individuals that have shared the Experience	Studying an event, a program, an activity, more than one individual
Data Collection Forms	Using primarily interviews and Documents	Using primarily interviews with individuals, although documents, observations, and art may also be considered	Using multiple sources, such as interviews, observations, documents, artifacts
Data Analysis Strategies	Analyzing data for stories, "restorying" stories, developing themes, often using a chronology	Analyzing data for significant statements, meaning units, textural and structural description, description of the "essence"	Analyzing data through description of the case and themes of the case as well as cross-case themes

Table 9. Comparative summary of narrative inquiry, narrative research, phenomenology and case study

Either of the study designs evaluated in Table 9 was adequate to conduct this research study. Healthcare has historically used a shallow toolbox of research practices to elicit knowledge and insights. Experimental (e.g., randomized trials) and quasi-experimental designs have been overused in clinical practice, in part, because the science (and art) of medicine is grounded in developing a credible evidence-base that informs clinicians and patients. Evidence-based medicine is defined as the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients (Sackett, 1997). Evidence-based medicine is also founded on the principle that scientific inquiry is superior to expert opinion and testimonials. It is not often narrative is used to inform decisions in healthcare – making this phenomenological study a timely scholarly contribution. The following are examples of a phenomenological study encompassing narrative with a similarly-sized study population.

In Gabrielson’s (2009) dissertation, a qualitative study using narrative analysis of interviews with ten older lesbians (aged 55 and over) who made a financial commitment to live in a continuous care retirement center (CCRC) specializing in lesbian, gay, bisexual and transgender (LGBT) care was conducted. The specific aims of the study were to:

- Describe what has impacted older lesbians' decisions to live in an LGBT-specific CCRC;
- Describe factors that both positively and negatively impact older lesbians’ perceptions of elder care (Gabrielson, 2009).

The study combined two qualitative strategies (across-case, thematic analysis and narrative analysis) and used a convenience sample.

Another comparable study in phenomenology using narrative was conducted by Baily and Tilly (2002) in which their study was a constructivist approach to narrative in which ten stories about death, physical and emotional vulnerability from three key informant groups: patients, caregivers and nurses were analyzed. Bailey and Tilly (2002) suggest that the events were not recounted to convey objective reality but to convey meaning, concluding that these stories were reconstructed in a way to convey their perspective of an event, rather than decontextualized truths (Bailey & Tilley, 2002).

An important depth-related study that provided a framework for the analysis of big data was conducted by Halevi and Moed (2012) where they explored the term big data as it evolved in the peer-reviewed literature. They sought to understand big data as a topic of study and the scientific problems, methodologies and solutions that researchers focused on in relation to it. Through a modified systematic review of literature in *Scopus*, an abstract and citation database of peer-reviewed literature (http://www.info.sciverse.com/scopus), Halevi and Moed (2012) uncovered three important themes from their research:

- The first appearance of term big data in scholarly literature appears in a 1970 article on atmospheric and oceanic soundings;[13]
- Early papers (1970 until 2000) were led by computer engineering, building materials, electric generators, electrical engineering, telecommunication equipment, cellular telephone systems and electronics disciplines; and,

[13] This is an important finding, as many sources of 'grey literature' credit the first references to the term 'big data' circa 2000.

- From 2000 onwards, the field is led by computer science followed by engineering and mathematics disciplines (Halevi & Moed, 2012).

These findings are significant. It suggests a strong correlation between the rise of big data in direct parallel to advances in technology, science and mathematics. Intuitively, with the advent of HIT in healthcare, there has been a direct upsurge in the notion of big data, too, along with the renaissance of the data scientist.

Worldview Paradigm

At the foundation of any research project are epistemologies, or philosophical worldviews (Creswell, 2009) which include postpositive, social construction, advocacy/participatory and pragmatism. These types of worldviews influence the type of research design (qualitative, quantitative, or mixed methods) the researcher selects as the most effective method to study the intended topic (Table 10).

FOUR WORLDVIEWS	
Postpositivism	Constructivism
• Determination • Reductionism • Empirical Observation and Measurement • Theory Verification	• Understanding • Multiple Participant Meaning • Social and Historical Construction • Theory Generation
Advocacy/Participatory	Pragmatism
• Political • Empowerment • Collaborative • Change-Oriented	• Consequences of Action • Problem-Centered • Pluralistic • Real-World Practice Oriented

Table 10. Four Philosophical Worldviews

As data is gathered and synthesized, assumptions are formed to test claims or hypotheses. Depending on the type of research conducted, a researcher can begin (and end) *with either philosophical worldview that presents the best fit to the intended course of study.* Theoretically, a researcher, as well, could seamlessly traverse each of the four ontologies described by Creswell (2009):

- Postpositivism worldview, which is considered the traditional form of learning and is grounded in measurement of observations and outcomes;
- Constructivism worldview, where researchers seek to understand the world in which they work and live by collecting data personally, interpreting the results and forming conclusions;
- Advocacy/Participatory worldview, which holds that politics are intertwined in the research and that there is a political or advocacy action agenda for change; and,
- Pragmatism worldview, which holds that there is no singular system or philosophy that researchers are committed to and employ mixed methods of study and multiple methods that happen to work at that time (Creswell, 2009, p. 6).

Early in my professional and academic career, my philosophical position manifested from having worked in diverse healthcare settings, including federal and state government, academia and private sector managed care organizations which require a practical and academic perspective for which to understand the work. After contemplating and absorbing each of these worldview beliefs for at least two academic school years, it is clear that my epistemological position about the phenomena of big data in healthcare can be constructed as follows: a) data, information and knowledge are contained within the perspectives of people that are experienced in healthcare and big data, either as a policymaker, provider or consumer; and b) my academic and professional experience is unique and allowed me to collaboratively engage with the study participants in collecting and constructing meaning about big data in healthcare.

Denzin and Lincoln (2002) postulate a relevant vignette which undoubtedly influences my study design approach:

> *Constructivism - Knowledge consists of those constructions about which there is relative consensus (or at least some movement toward consensus) among those competent (and, in the case of more arcane material, trusted) to interpret the substance of the construction. Multiple "knowledges" can coexist when equally competent (or trusted) interpreters disagree, and/or depending on social, political, cultural, economic, ethnic, and gender factors that differentiate the interpreters. These constructions are subject to continuous revision, with changes most likely to occur when relatively different constructions are brought into juxtaposition in a dialectical context (p. 113).*

By virtue of conducting a phenomenological study, which has a very deep history in philosophy (Groenewald, 2004), my philosophical grounding in constructivism was reinforced.

This study is about discovery of important categories of meaning about big data in healthcare through the experiential knowledge of nine key healthcare stakeholders. By listening to, writing, describing and interpreting text of an individual's "lived experience," I also successfully elicited original, first-hand data about rich social, cultural, and institutional narratives (Clandinin, 2013) that are potentially lost in a quantitative approach. Make no mistake, Amedeo Giorgi (2009) was clear that "a completely full experiment requires both aspects" (p. 39) of quantitative and qualitative approaches. I did not consider a mixed method analytical technique, drawing on my training as an epidemiologist, Leinweber (2011) points out that the "best content-analytic studies use both qualitative and quantitative operations" (p. 2). However, researching the qualitative aspects of the healthcare big data phenomena yielded a timely and much richer description about the phenomena. Phenomenology using narratives is appropriate and timely for the phenomena under study.

Interview Procedures

Each study participant was given the list of the interview questions as part of the Interview Guide (Appendix C) at least one week prior to the scheduled initial interview. While there was no formal preparation required, sending the questions ahead of time allowed each study participant to think through a sequence of events and topics that were possibly forgotten to memory. An executive summary no longer than three pages was sent to each study participant in a PDF format.

While the study topic was positively received by potential study participants and industry leaders, a moderate-level risk loomed: potential candidates who verbally and informally agreed to participate in this study could recuse themselves for a number of factors, including, schedules, new commitments, time-lapsed between initial contact and interview, and the end of the current federal fiscal year (September 30, 2013). To mitigate this risk, I kept potential study participants informed of the progress of the study's development through email. This was important because at the onset of my data collection period, the federal government historically shut down its operations between October 1 and October 16, 2013.

Of the study participants selected, six were geographically located in the Mid-Atlantic region (Figure 5) of the United States because of the density of federal healthcare agencies and integrated delivery systems. The region is also a hub for national patient advocacy organizations.

The remaining three study participants were selected from the Midwest and Pacific Coast regions.

After ten study participants were selected from the purposive sample, each signed and returned the original copy of the Participant Study Consent Form (Appendix B). A one-hour interview was subsequently scheduled. No potential study participant declined verbally or by email.

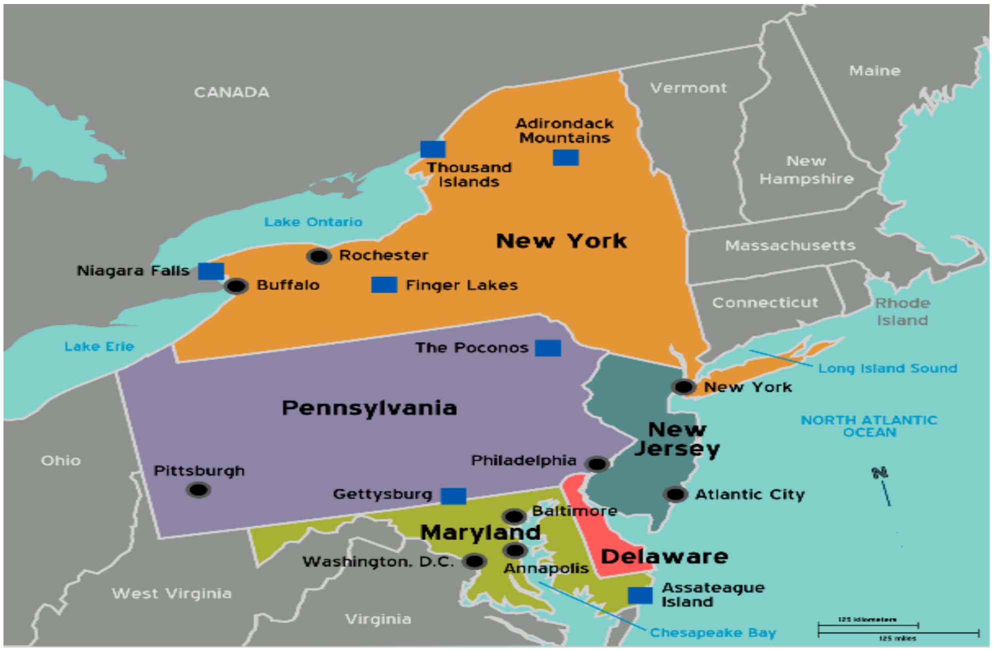


Figure 5. Geographic regional sampling frame from which purposive sample will be drawn (source: Internet: Google Images www.google.images.com)

Semi – Structured Interviews

A semi-structured interview has a freewheeling (Bernard, 2006) quality – the flow of the interview, rather than the order in the interview guide (Bailey & Tilley, 2002) which provides explicit directions about how the interview will be conducted, guides the healthcare “stakeholder

– researcher" collaboration. A copy of the Interview Guide that provided clear instructions to guide this study is in Appendix C.

Each initial interview lasted approximately 43 to 60 minutes, with one interview lasting one hour and 16 minutes. Five interviews were held in the study participant's place of work, three were held over the phone, and two were held through Skype. Each study participant conducted their interview from their place of work with the exception of two who took the interviews from their homes.

The semi-structured interviews served as the primary data collection method; my written field notes were a secondary source of data along with additional supplemental data. Four of the participants provided additional sources of data, including Microsoft PowerPoint slides from previous presentations on big data, a book co-authored by a study participant, and a URL to a personal website.

Study participants responded to 11 open-ended questions and one yes-no question that elicited further elaboration. I solely conducted each interview and recorded the "conversation" on an Apple iPhone 5S. The data were immediately loaded into a secure password-protected data management account and uploaded for transcription and analysis. I augmented the recordings with personal field notes kept in a dedicated journal. Follow up interviews occurred face-to-face in the study participant's place of work, via Skype and on the phone. The intent was to maintain the most comfortable setting for study participants to share their in-depth narratives about big data in healthcare. No other research interviewers were used in this study. The interview guide was about the most structured part of the interviews.

Bracketing

For a major federal project, I conducted twenty-five semi-structured interviews with middle to senior-level managers in a large healthcare agency to chronicle and synthesize their requirements for and insights into an enterprise-wide portfolio management initiative. I developed a study guide to help facilitate the interviews; however, this elite group of federal staff relied on my ability to navigate an informal conversation, keep them engaged and respect their limited time.

During interviews, I maintained a collaborative rather than an objective or neutral relationship with each study participant. One of the lessons learned from the aforementioned experience was to engage in an informal conversation with lots of flexibility, but maintain a degree of structure bound by the interview guide. From this in-depth, six-month long project, I also learned that semi-structured interviewing works very well in projects where researchers engage with high-level bureaucrats and elite members of a community—people who are accustomed to efficient use of their time (Bernard, 2006).

I have reflected a lot on my role during this research study. My research has uncovered the fact that there are a couple of prominent ideologies on the level of involvement of the researcher. Dahlberg's (2006) notion of 'bridling' provided a reference that guided my interactions with each study participant. Bracketing, or putting aside my experiences beliefs and opinions, is a commonly used approach in phenomenology studies. It was very difficult to simply set aside my presuppositions, opinions and ideas about a topic I am very close to. However, to get to the "truth" of the story, I successfully set aside my personal knowledge and ideologies on big data in healthcare and remained conscious of each study participant's lived experience, employing my excellent listening skills. There were three aspects of bridling that guided my presupposition as described by Bremer (2009):

- Like “bracketing,” bridling is “the restraining of one’s pre-understanding in the form of personal beliefs, theories, and other assumptions that otherwise would mislead the understanding of meaning and thus limit the research options” (pp. 129 – 130).
- It is also about the “understanding as a whole” not just the “pre-understanding”-this is done so as to not “understand too quickly, too carelessly” (p. 130). It is an “open and alert attitude of activity waiting for the phenomenon to show up and display itself within the relationship” (p. 130); and,
- It is forward looking rather than backward looking, allowing “the phenomenon to present itself” (p. 130)(Bremer, Dahlberg, & Sandman, 2009).

Data Management

There were many types of data that required management: documents, interview transcripts, field notes, websites and books. During the first semester of the doctoral program, I began ‘memoing’(Miles & Huberman, 1994), or journaling. Journaling is a process of maintaining a written record of my experiences, activities, thoughts, and ideas on regular basis. It is a practice that I maintained throughout my studies and research. I used Evernote as the primary electronic document management system to manage and secure websites and other documents except the raw transcripts. As a supplement to the electronic media, I maintained a dedicated written journal to document reflections and thoughts about this research process.

All interview audio files were stored and managed in a dedicated, secure, password-protected Apple iTunes account which I only had access. Six (6) months after the date of the final study analysis, all iTunes audio files associated with this research study will be destroyed and not be available for use in further research, articles or publications.

Transcription

- Only after permission was granted in writing and verbally approved by each study participant, each interview was recorded using an Apple iPhone 5s. I took hand-written field notes to supplement each recording. Field notes were kept in a confidential journal;
- After recording each interview, the audio file was converted into a written transcript through a technique called "parroting:"
  - Download the audio file to an Apple iTunes secure cloud platform using a Mac Air laptop;
  - Through Dragon NaturallySpeaking 12 Premium Student/Teacher edition software, a recording of the interview was heard through the Dragon headset;
  - No later than one day after each interviews I listened to the recorded text;
  - For quality control, the audio file was re-checked against transcription.
- I used Microsoft Word as the word processor to manage text data recorded from each audio interview. A password protected file for each interview was created to ensure privacy and eventually merged for analysis.
- Files were saved based on the coding scheme in Table 13. To maintain confidentiality, no study participant names were associated with any file. I assigned a web-generated pseudonym to each participant. Rather than use an impersonal identification code, I chose to maintain authenticity of narratives realism by assigning traditional names.

Timeframe

All data collected from the initial semi-structured interviews and subsequent follow up interviews were conducted between September 23, 2013 and December 10, 2013.

Data Analysis Procedures

For this research study, I employed a commonly used content analysis framework: a general inductive approach to qualitative analysis (Elo & Kyngäs, 2008; D. Thomas, 2003; Zhang & Wildemuth, 2009)[14]. A general inductive approach to qualitative content analysis is a valuable alternative to more traditional methods when attempting to identify important themes or categories within a body of text (Zhang & Wildemuth, 2009).The technique is drawn from a variety of related techniques used in exploratory qualitative research, qualitative content analysis and constructivist grounded theory (Pope, Ziebland, & Mays, 2000) which if a theory were to be used is the closest theory that relates to this research study.[15] Thomas (2003) purports that the primary purpose of the inductive approach is "to allow research findings to emerge from the frequent, dominant or significant themes inherent in raw interview data, without the restraints imposed by structured methodologies" (p. 2). I chose this framework because the general inductive approach is frequently reported in health and social science research (D. R. Thomas, 2006; D. Thomas, 2003) and information & library sciences (Zhang & Wildemuth, 2009) and has a close counterpart, quantitative content analysis.

[14] David R. Thomas is professor at the School of Population Health, University of Auckland

[15] This is a phenomenological study. Dewey's Theory of Experience (1938) is most often cited as a philosophical underpinning of narrative inquiry.

Qualitative data analysis involves searching for emerging themes, first within an interview and then across a series of interviews. The search for emerging themes is common practice in qualitative research and involves the interplay between data and the emerging themes (Tan & Hunter, 2003). There is no one method to analyze narrative data, and arguably, there are a host of appropriate analytical methods for a qualitative study in information studies (Table 11).

Trustworthiness

Though as novice researcher and rising scholar-practitioner, my personal goal was to conduct an ethical high quality research study on big data in healthcare. Qualitative researchers, who frame their studies in an interpretive paradigm, think in terms of trustworthiness as opposed to the conventional, positivistic criteria of internal and external validity, reliability, and objectivity (Denzin & Lincoln, 2002; Guba & Lincoln, 1985). To ensure trustworthiness, I relied on two methods: triangulation of stakeholder participation of three key healthcare stakeholder classes and stakeholder checks, which were important to ensure I maintained the essence of each stakeholder's narrative. Stakeholder checks were also an invaluable method to capture additional new information from study participants post initial interview. Many of the study participants provided additional data and clarified inaudible or erroneous interpretations of their words.

My objective was to not merely connect "thick descriptions" of narrative, but to create a trusted, meaningful account about big data in healthcare through the insights of those who know the subject best.

Study Limitations

This study posed three limitations that could have potentially impacted this study. The first limitation was the construct of a phenomenological encompassing narrative study design. Small qualitative studies yield very limited information about a phenomenon from a limited sampling frame. The study participants selected from the purposive sampling strategy produced credible and reliable original data. Second, I had no expectations of achieving saturation of themes that were generalizable to the entire healthcare ecosystem. This study focused on three key healthcare stakeholder classes out of many that constitute the healthcare ecosystem. “Key” healthcare stakeholders could be defined differently by other researchers. I chose not to poll other healthcare experts to validate if the three classes identified in this study as “key.” Third, patient privacy is protected by federal laws that would jeopardize this study. Patient privacy is not a risk related to this study as it has been mitigated by purposively selecting responsible consumer advocates who are well positioned to assist patients in decision making about their health issues (Petronio, Sargent, Andea, Reganis, & Cichocki, 2004).

	GENERAL INDUCTIVE APPROACH (THOMAS 2003)	GROUNDED THEORY (CHARMAZ, 2006)	INTERPRETATIVE PHENOMENOLOGICAL ANALYSIS (IPA) (SMITH ET AL., 2009)	DISCOURSE ANALYSIS (POTTER, 1996)
Study Aim & Research Question	To examine topics and themes, as well as the inferences drawn from them, in the data and to generate theory	To generate theory from empirical data (e.g. stigma in mental health)	To understand individual in-depth experience; rooted in psychology	To capture nuances of text or public discourse (e.g., understanding political theory)
Sampling & Methods	Samples usually consist of selected texts which can inform the research questions being investigated. Purposive sampling	Range of perspectives and stay true to research question; unstructured questionnaire Theoretical sampling	Homogenous sample and stay true to participants' stories; unstructured questionnaire Purposive sampling	Documents, speeches, newspapers, mass media Purposive/ Theoretical sampling
Analysis	Identification of descriptive and interpretative themes that actively engages the researcher and participants	Data-driven Constant comparison and iterative approach	Identification of descriptive and interpretative themes that actively engages the researcher and participants	Detailed, thorough analysis of discourses – speeches, written text, conversations
Researcher's Position	Immerse in the data and allow themes to emerge from the data	Potential 'bias' is managed	Paramount; importance of reflexivity	High level of interpretation or abstraction expected

Table 11. Comparison of common qualitative data analysis methods

Some of the assumptions of a general inductive approach are described below:

- Data analysis was determined by both the research objectives (deductive) and multiple readings and interpretations of the raw data (inductive).
- The primary mode of analysis was the development of categories from the raw data into a model or framework that captures key themes and processes judged to be important.
- The research findings result from multiple interpretations made from the raw data by the researcher who codes the data. Inevitably, I independently made decisions about what was more important and less important in the data.
- Trustworthiness of findings was assessed (a) triangulation within across key healthcare stakeholders and (b) feedback from participants in the research (D. R. Thomas, 2006; D. Thomas, 2003; Zhang & Wildemuth, 2009).

I did consider four alternative approaches commonly used in the social sciences: general inductive approach, grounded theory, Interpretative Phenomenological Analysis (IPA) and discourse analysis. Because of the time it took to develop an adequate working knowledge of qualitative content analysis, I chose a credible data analysis procedure that allowed me systematically apply important categories of meaning necessary to ‘restory’ study participant narratives.

Presentation of Findings and Conclusions

The framework of a general inductive approach provided a vetted approach to presenting research study findings. I must note that while this data analysis approach was a good starting point, the final presentation of the findings is undetermined. In the case of a general inductive approach to content analysis, the coding process played a central part in how data the data was reported; I am thankful for NVIVO 10. The general inductive approach did not produce counts and statistical significance; instead, it effectively uncovered patterns, themes, and categories important to a social reality. I let the themes emerge from the coding scheme before I defined how the data was to be presented. While I visualized many, many approaches to presenting the data, with the guidance of my committee, the study's finding as they are presented felt like the most appropriate way to present these important 'stories' on the phenomena of big data in healthcare. So that study is replicable, I monitored and reported analytical procedures and processes as completely and truthfully as possible (Patton, 2005). Where possible, I included tables, graphs, and charts (Miles & Huberman, 1994), and did not deviate from the true objective of completing a qualitative phenomenological study.

I attempted to maintain a balance between both interpretation and description of themes, and important categories of meaning. Description gives readers background and context (Denzin & Lincoln, 2002). An interesting and readable report provides sufficient description to allow the reader to understand the basis for an interpretation, and sufficient interpretation to allow the reader to understand the description (Patton, 2005).

My curriculum vita (CV) is included at the end of this dissertation.

APPENDIX B. STUDY PARTICIPANT CONSENT FORM

SCHOOL OF INFORMATION STUDIES

343 Hinds Hall Syracuse, NY 13210

An Epidemiology of Big Data

My name is John Young and I am a professional doctorate student at Syracuse University, School of Information Studies. I am inviting you to participate in a research study. Involvement in the study is simple, voluntary and with very little risk, so you may choose to participate or not. This document will explain the study to you and please feel free to ask questions about the research if you have any. I will be happy to explain anything in detail if you wish.

I am interested in learning more about the important categories of meaning about big data in healthcare – through the experiences of ten leaders representing three key healthcare stakeholder classes: government, providers and consumers. You will be asked to provide your insights by participating in a face-to-face interview at your place of work. Interviews will take approximately up to two hours of your time, beginning with an initial one hour interview. A subsequent follow-up interview either face-to-face or by phone will be used to validate and enhance your narrative. Your participation will be a contribution towards providing new knowledge about important categories of meaning about big data in healthcare through an intertwined ‘story’ of ten key healthcare stakeholders.

Your privacy is important and your responses will remain confidential. I will assign a unique number to your responses, and only I and my faculty advisor will have the key to indicate which number belongs to which participant. In any articles I write or any presentations that I make, I will use a made-up name for you and I will not reveal details or I will change details about where you work.

I would like to audio record this face-to-face interview using an Apple iPhone 5 so that I can use it for reference while proceeding with this study. I will be the only one who will hear the audio recordings, which will be transcribed by me. I will not record this interview without your permission. If you do grant permission for this conversation to be recorded, you have the right to end the interview at any time.

This project will be completed by February 15, 2014. All interview recordings will be stored in a secure, password protected Apple iTunes account that I will only have access to until six (6) months after that date. The audio files will then be destroyed. Your study data will be kept as confidential as possible, with the exception of certain information we must report for legal or ethical reasons.

Contact Information:

If you have any questions, concerns, complaints about the research, contact my faculty advisor and professor, Dr. Jian Qin at (315) 443 - 5642. If you have any questions about your rights as a research participant, you have questions, concerns, or complaints that you wish to address to someone other than the investigator, if you cannot reach the investigator, contact the Syracuse University Institutional Review Board at 315-443-3013.

All of my questions have been answered, I am 18 years of age or older, and I wish to participate in this research study. I have received a copy of this consent form (please keep a copy of this consent form for your records).

___ I agree to be audio recorded.

___ I do not agree to be audio recorded.

Signature of participant Date

Printed name of participant

Signature of researcher Date

Printed name of researcher

APPENDIX C. INTERVIEW GUIDE

School of Information Studies

343 Hinds Hall Syracuse, NY 13210

An Epidemiology of Big Data

Interview Guide

Script:

Thank you for inviting me to your office and agreeing to participate in this research study. My name is John Young and I am a graduate student in the doctorate of professional studies – information management program at Syracuse University, School of Information Studies in Syracuse, NY. This initial interview will take about 60 minutes and will include 11 questions regarding your experiences and insights about big data in healthcare. I would like your permission to audio record this interview, so I may accurately document the information you convey. I will also keep hand-written notes to supplement the audio recording. I will schedule another follow-up face-to-face or telephone interview to check if you have additional insights to share and to ensure my draft transcription accurately reflects your narrative. If at any time during the interview you wish to discontinue the use of the recorder or the interview itself, please feel free to let me know. Your privacy is important; all of your responses will remain confidential.

Your confidential responses will be used to contribute to new knowledge about themes, challenges and meaning about big data in healthcare using a narrative-based data collection method. A coherent 'story' from three key healthcare stakeholder classes: government, *providers, and consumers will be the outcome of the study. The purpose of this study is to discovery important categories of meaning about big data in healthcare.*

At this time I would like to remind you of your written consent to participate in this study. I am the responsible researcher for this research project: An Epidemiology of Big Data. You and I have both signed and dated each copy, certifying that we agree to begin this interview. You will receive one copy and I will keep the other under lock and key, separate from your reported responses. Thank you.

Your participation in this interview is completely voluntary. If at any time you need to stop to take a break, please let me know. You may also withdraw your participation at any time without consequence. Do you have any questions or concerns before we begin? Then with your permission we will begin the interview.

A phenomenological study encompassing narrative captures a holistic account of people's experiences related to a phenomenon. The objective of this phase of the research is to capture study participant's insights and perspectives about big data in healthcare in their own words. The following questions are guide for the interview to ensure I have collected the intended information. The trustworthiness and credibility of this study relies on study participant's to talk openly and objectively about various aspects of big data in their daily routine and within their healthcare organization. There are no right (or wrong) answers and no preparation beyond your subject matter knowledge and experience is required.

	DESCRIPTION	RATIONALE	SOURCE
IQ1	What does big data mean to you? Your organization? What about big data in healthcare specifically? How did you arrive to this conclusion?	I am looking for categories of meaning derived from experiential knowledge which could inform a cohesive definition of big data.	(Dumbill, 2013; Villars et al., 2011)
IQ2	Describe some of the important professional and academic experiences that prepared you for your current position. Please emphasize any academic training or practical preparation.	I am seeking to understand how study participant evolved professionally which could provide insight into professional development of big data in healthcare.	(Borgman, 2012)
IQ3	What makes 'big data' different from 'data?' Are there certain attributes? This is the only 'yes' or 'no' question, but please elaborate: Is medicine an information science?	Big data is a "buzzword" that is poorly defined.	(Borangiu & Purcărea, 2008; Davenport et al., 2012; Rooney, 2012; Sackett et al., 1996; Smith, 1996)
IQ4	Describe the drivers and influencers that impact 'big data' in healthcare?	Big data has been slow to catch on in healthcare. IQ3 provides professional and organizational insights into drivers and influencers of big data	(Bollier & Firestone, 2010; L. R. Burns et al., 2002; Sullivan, 2011)
IQ5	Describe the 'big data' sources (e.g., data sets) you use. How do you get access to these data sources? Does someone else manage access to and analysis/interpretation of 'big data?'	Big data requires computing platforms and analytics that are not customarily available on a desktop. Provides content and context into the capabilities, support, tools needed to manage and use big data.	(Anderson, 2004; Davenport & Patil, 2012; Eysenbach, 2008; Pryor & Donnelly, 2009; Rhoads & Ferrara)
IQ6	Describe the organizational challenges of making data driven. Can 'big data' help address these challenges?	These challenges might provide insight into why big data has been slow to evolve in healthcare.	(Porter & Teisberg, 2006; Weisbrod, 1991)
IQ7	Describe what 'big data' you share? How do you share it? With whom do	My thought here is by understanding multidisciplinary perspectives	(Theodor Borangiu & Victor Purcarea,

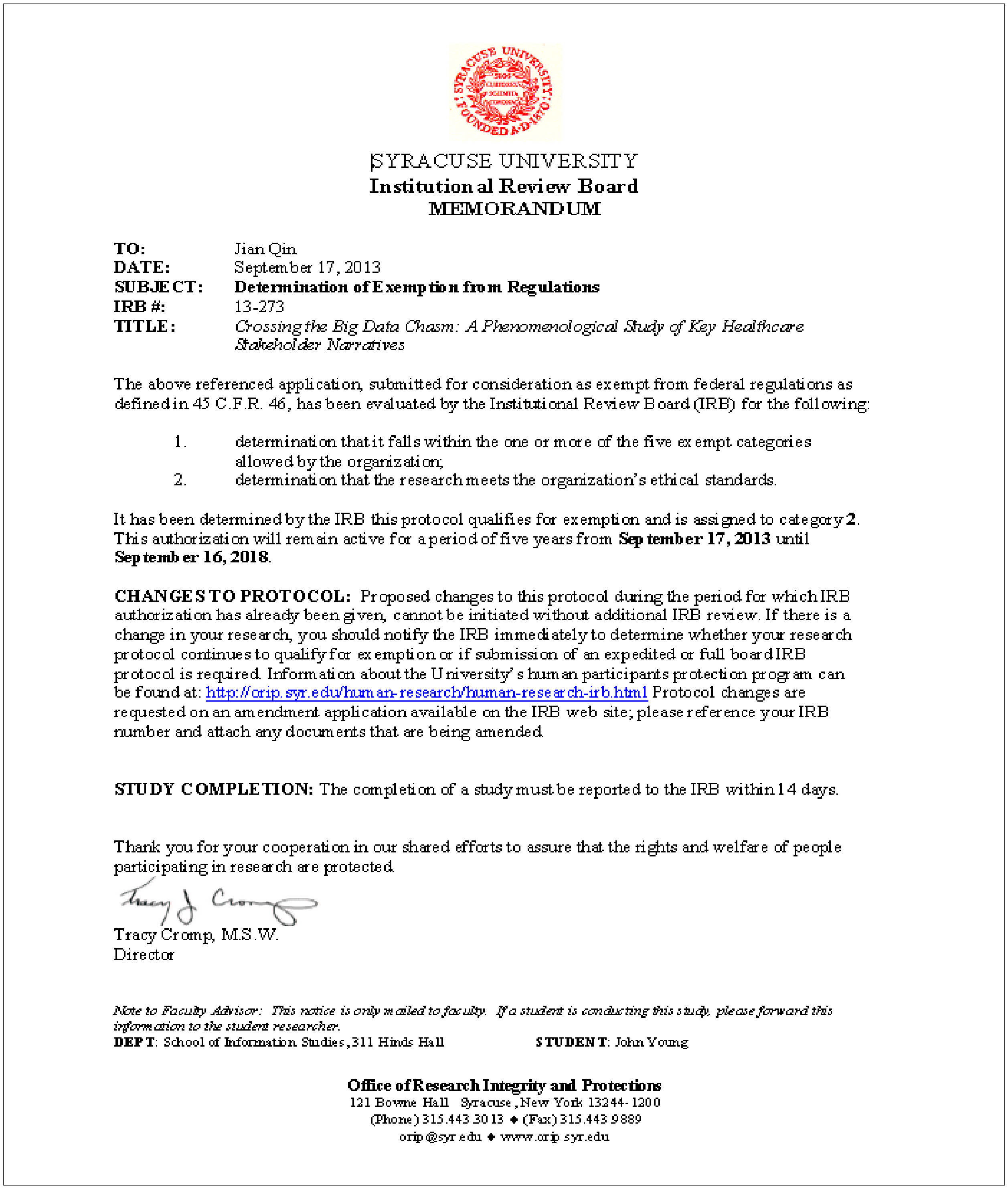
	you share your ‘big data?’	about big data, this study could be a small step towards informing further studies in health data sharing policy.	2008; Gorman, 1995; Porter & Teisberg, 2006)
IQ8	Describe any uses, unintended consequences, or reuses of big data. Should big data be repurposed for secondary use by each of the three stakeholder classes? Please Elaborate.	Here, I am hoping to capture data on any unintended consequences of big data and whether data prepared for government can be used for consumers.	(Borgman, 2012; Kerr, Norris, & Stockdale, 2007)
IQ9	Describe your vision of a future state of big data in healthcare. What are your hopes for big data?	Provides content and framework for current gaps between “as is” and “to be” big data.	(Borangiu & Purcărea, 2008; Feldman et al., 2012)
IQ10	Metaphors and symbols are prevalent in healthcare. Can you describe any big data metaphors or symbols that resonate with you or your organization? Why?	Metaphors like “data deluge”, the new oil,” and “data tsunami” all attempt to describe big data and highlight the challenges of doing so.	(Burns, 2011)
IQ11	Please elaborate on any points about big data not covered in these questions that make sense for you and add other points that are unique to you and your organization.	Always end with an open-ended question in the event I missed something.	(Borins, 2012; Boyce & Neale, 2006; Ryan & Bernard, 2003)

Table 12. Semi-Structured Interview Questions and Rationale

Script continued-

This concludes the initial interview. I will follow up with next steps about the follow up interview. Thank you very much for taking time from you busy schedule to participate in this research study.

APPENDIX D. IRB DETERMINATION OF EXEMPTION



SYRACUSE UNIVERSITY
Institutional Review Board
MEMORANDUM

TO: Jian Qin
DATE: September 17, 2013
SUBJECT: **Determination of Exemption from Regulations**
IRB #: 13-273
TITLE: *Crossing the Big Data Chasm: A Phenomenological Study of Key Healthcare Stakeholder Narratives*

The above referenced application, submitted for consideration as exempt from federal regulations as defined in 45 C.F.R. 46, has been evaluated by the Institutional Review Board (IRB) for the following:

1. determination that it falls within the one or more of the five exempt categories allowed by the organization;
2. determination that the research meets the organization's ethical standards.

It has been determined by the IRB this protocol qualifies for exemption and is assigned to category **2**. This authorization will remain active for a period of five years from **September 17, 2013** until **September 16, 2018**.

CHANGES TO PROTOCOL: Proposed changes to this protocol during the period for which IRB authorization has already been given, cannot be initiated without additional IRB review. If there is a change in your research, you should notify the IRB immediately to determine whether your research protocol continues to qualify for exemption or if submission of an expedited or full board IRB protocol is required. Information about the University's human participants protection program can be found at: http://orip.syr.edu/human-research/human-research-irb.html Protocol changes are requested on an amendment application available on the IRB web site; please reference your IRB number and attach any documents that are being amended.

STUDY COMPLETION: The completion of a study must be reported to the IRB within 14 days.

Thank you for your cooperation in our shared efforts to assure that the rights and welfare of people participating in research are protected.

Tracy J Cromp

Tracy Cromp, M.S.W.
Director

Note to Faculty Advisor: This notice is only mailed to faculty. If a student is conducting this study, please forward this information to the student researcher.
DEPT: School of Information Studies, 311 Hinds Hall **STUDENT**: John Young

Office of Research Integrity and Protections
121 Bowne Hall Syracuse, New York 13244-1200
(Phone) 315.443.3013 ♦ (Fax) 315.443.9889
orip@syr.edu ♦ www.orip.syr.edu

APPENDIX E. INTERVIEW DATA EXPLICATION SCHEME

STEP	ACTIVITY	DESCRIPTION
Step 1	Prepare the Data	After transcription from audio to text, I formatted the raw data files into a common format (e.g., font size, margins, questions or interviewer comments highlighted). I printed and made backups of each raw data file and kept hard copies each interview in a single binder.
Step 2	Close Reading (and Rereading) of the text.	The raw text files were read in detail to become familiar with the content and gain an understanding of the categories of meaning or "themes" and details in the text.
Step 3	Develop Categories and a Coding Scheme	Categories and a coding scheme were derived primarily from the semi-structured interview data. Other data sources including scholarly and grey literature, study participant supporting materials, (e.g., books, resumes) were also analyzed. This study did not require a theoretical framework; categories were inductively generated from the interview data.
Step 4	Overlapping Coding and Un-coded Text	Among the commonly assumed rules that underlie qualitative coding, two are different from the rules typically used in quantitative coding: (a) segmentation of text was coded into more than one category and (b) a considerable amount of the text was not assigned to any category.
Step 5	Code All the Text and Continuing Revision and Refinement of Category System	Within each category, I searched for subunits of meaning and included contradictory points of view and new insights. I select appropriate quotes that conveyed the core theme or essence of a category. The categories were often combined and linked when the meanings are similar.
Step 6	Draw Conclusions from the Coded Data	This step involved making sense of the themes or categories identified, and their properties. I began making inferences and presented the reconstructions of categories of meaning derived from the data, including exploring different dimensions of categories, identifying relationships between categories, uncovering patterns within and across healthcare stakeholder classes, and testing categories against the full range of data.

Table 13. Coding Scheme: A General Inductive Approach

REFERENCES

Acar, U., Buneman, P., Cheney, J., Van Den Bussche, J., Kwasnikowska, N., & Vansummeren, S. (2010). *A graph model of data and workflow provenance.* Paper presented at the Proceedings of the 2nd conference on Theory and practice of provenance, TAPP.

AHRQ. (2014). Stakeholder guide 2014. *Effective Healthcare Program, AHRQ Publication No. 14-EHC010-EF*, 1-44. Retrieved from http://www.ahrq.gov/research/findings/evidence-based-reports/stakeholderguide/stakeholdr.pdf website:

Anderson, W.L. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. *Data Science Journal, 3*, 191-201.

Arbesman, S. . (2013). Five myths about big data *Washington Post*.

Bailey, P.H., & Tilley, S. (2002). Storytelling and the interpretation of meaning in qualitative research. *Journal of Advanced Nursing, 38*(6), 574-583.

Begley, S. (2011). The best medicine. *Scientific American, 305*(6), 50-55.

Bell, E. (2006). Reseach for health policy. In Oxford (Ed.), *Research for Health Policy* (pp. 48-49).

Bellamy, J., & Bledsoe, S.E. (2006). The current state of evidence-based practice in social work: A review of the literature and qualitative analysis of expert interviews. *Journal of Evidence-Based Social Work, 3*, 1.

Bellamy, Jennifer L, Bledsoe, Sarah E, & Traube, Dorian E. (2006). The current state of evidence-based practice in social work: A review of the literature and qualitative analysis of expert interviews. *Journal of Evidence-Based Social Work, 3*(1), 23-48.

Benzies, K.M., Premji, S., Hayden, K.A., & Serrett, K. (2006). State-of-the-evidence reviews: Advantages and challenges of including grey literature. *Worldviews on Evidence-Based Nursing, 3*(2), 55-61.

Bernard, H.R. (2006). Interviewing: Unstructured and semistructured. *Research methods in anthropology: Qualitative and quantitative approaches*, 190-233.

Berwick, D.M., Nolan, T.W., & Whittington, J. (2008). The triple aim: care, health, and cost. *Health Affairs, 27*(3), 759-769.

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research, 10*(2), 141-163.

Boehm, Barry. (1975). Session II Structured Programming: A Quantitative Assessment. *Computer, 8*(6), 38-40.

Bollier, D., & Firestone, C.M. (2010). *The Promise and Peril of Big Data*: Aspen Institute, Communications and Society Program.

Borangiu, T., & Purcarea, V. (2008). The Future of Healthcare–Information Based Medicine

Borangiu, Theodor, & Purcărea, V. (2008). The Future of Healthcare–Information Based Medicine. *Journal of medicine and life, 1*(2), 233.

Borangiu, Theodor, & Purcarea, Victor. (2008). The Future of Healthcare–Information Based Medicine The Future of Healthcare–Information Based Medicine.

Borgatta, E.F., & Borgatta, M.L. (1992). *Encyclopedia of Sociology* (Vol. 2): Macmillan Nova York.

Borgman, C.L. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078.

Borins, S.F. (2012). Making Narrative Count: A Narratological Approach to Public Management Innovation. *J Public Adm Res Theory, 22* ((1)), 165-189.

Boyce, C., & Neale, P. (2006). *Conducting In-Depth Interviews: A Guide for Designing and Conducting In-Depth Interviews for Evaluation Input*: Pathfinder International Watertown, MA.

Bremer, A., Dahlberg, K., & Sandman, L. (2009). To survive out-of-hospital cardiac arrest: A search for meaning and coherence. *Qualitative Health Research, 19*(3), 323-338.

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of 'big data'? *McKinsey Quarterly, 4*, 24-35.

Brynjolfsson, E., & McAfee, A. (2011). The Big Data Boom is the Innovation Story of Our Time. *The Atlantic*.

Buchan, I. . (2009). *A Unified Modeling Approach to Data Intensive Healthcare*.

Burns. (2011). Healthcare's data tsunami. *日立評論*

Burns, L.R., DeGraaff, R.A., Danzon, P.M., Kimberly, J.R., Kissick, W.L., & Pauly, M.V. (2002). The Wharton School study of the health care value chain. *The health care value chain: producers, purchasers and providers. San Francisco: Jossey-Bass,* 3-26.

Butte, A.J., & Shah, N.H. (2011). Computationally translating molecular discoveries into tools for medicine: translational bioinformatics articles now featured in JAMIA. *Journal of the American Medical Informatics Association, 18*(4), 352-353.

Charon, R. (2006). *Narrative Medicine: Honoring the Stories of Illness*: Oxford University Press.

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.

Cheng, B., Luo, C., Chiu, W., & Chen, H. (2009). Applying cluster analysis to build a patient-centric healthcare service strategy for elderly patients. *International Journal of Technology Management, 47*(1), 145-160.

Clancy, C.M. (2006). Informing quality healthcare. *Healthcare Financial Management, 60*(3), 64-68.

Clandinin, D.J. (2013). *Engaging in narrative inquiry*: Left Coast Press.

CMS. (2012). *National healthcare expenditures-highlights*. Retrieved from http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/highlights.pdf.

Creswell, J.W. (2009). *Research design: qualitative, quantitative, and mixed methods approaches*: SAGE Publications, Incorporated.

Cukier, K. (2010). Data, data everywhere. *The economist, 394*(8671), 3-16.

Cyr, D.J., & Reich, B.H. (1996). *Scaling the ivory tower: Stories from women in business school faculties*: Praeger Publishers.

Davenport, T.H., Barth, P., & Bean, R. (2012). How ‘big data’ is different. *MIT Sloan Management Review*.

Davenport, T.H., & Jarvenpaa, S.L. (2008). *Strategic use of analytics in government*: IBM Center for the Business of Government.

Davenport, T.H., & Patil, D.J. (2012). Data scientist. *Harvard business review*.

Denzin, N.K., & Lincoln, Y.S. (2002). *The qualitative inquiry reader*: Sage.

Dumbill, E. (2013). Making Sense of Big Data. *Data*(1).

Edgman-Levitan, S., & Cleary, P.D. (1996). What information do consumers want and need? *Health Affairs, 15*(4), 42-56.

Elo, Satu, & Kyngäs, Helvi. (2008). The qualitative content analysis process. *Journal of advanced nursing, 62*(1), 107-115.

EMC2. (2012). Big Data: Big Opportunites to Create Business Value.

Epstein, D., Heidt, J.B., & Farina, C.R. (2013). The value of words: Narrative as evidence in policymaking.

Ernst, George W. (1976). A definition-driven theorem prover. *Computers, IEEE Transactions on, 100*(4), 317-322.

Eysenbach, G. (2002). Infodemiology: the epidemiology of (mis) information. *The American journal of medicine, 113*(9), 763-765.

Eysenbach, G. (2008). Medicine 2.0: Social networking, collaboration, participation, apomediation, and openness. *J Med Internet Res, 10*(3), e22.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining.

Feldman, B., Martin, E.M., & Skotnes, T. (2012). Big Data in Healthcare Hype and Hope.

Frangenberg, E.H. (2013). An autonomous model of health care: Are third parties really needed? *Health, 5*, 1590.

Frehywot, Seble, Vovides, Yianna, Talib, Zohray, Mikhail, Nadia, Ross, Heather, Wohltjen, Hannah, . . . Scott, James. (2013). E-learning in medical education in resource constrained low-and middle-income countries. *Human resources for health, 11*(1), 1-15.

Gabrielson, M.L. (2009). The long-term care decision making of older lesbians: a narrative analysis.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC iView*, 1-12.

Gardner, R. (2013). The humaniteis, narrative, and the social context of the patient-professional relationship. *The Health and Humanities Reader*, 1-23.

Gartner. (2013). Gartner IT Glossary. from http://www.gartner.com/it-glossary/big-data/

Geertz, C. (1973). *The interpretation of cultures: Selected essays* (Vol. 5019): Basic books.

Giorgi, A. (1985). *Phenomenology and psychological research*: Duquesne Univ Pr.

Giorgi, Amedeo. (2009). *The descriptive phenomenological method in psychology: A modified Husserlian approach*: Duquesne University Press.

Gorman, P.N. (1995). Information needs of physicians. *Journal of the American Society for Information Science, 46*(10), 729-736.

Greenhalgh, T., & Hurwitz, B. (1999). Why study narrative? *Bmj, 318*(7175), 48-50.

Greenhalgh, T., & Russell, J. (2009). Evidence-based policymaking: a critique. *Perspectives in biology and medicine, 52*(2), 304-318.

Groenewald, T. (2004). A phenomenological research design illustrated. *International Journal of Qualitative Methods, 3*, 1.

Grossmann, C. (2010). *Clinical data as the basic staple of health learning: Creating and protecting a public good: Workshop summary*: National Academies Press.

Guba, E.G., & Lincoln, Y.S. (1985). Competing paradigms in qualitative research.

Gudea, S. (2005). Data, information, knowledge: A healthcare enterprise case study. *Perspectives in Health Information Management/AHIMA, American Health Information Management Association, 2*.

Halevi, G., & Moed, H. (2012). The evolution of big data as a research and scientific topic: Overview of the literature. *Research Trends*, *30*, 3-6.

Hardy, Q. (2012, March 24, 2012). Just the facts, yes, all of them, *The New York Times*.

Heudecker, N. (2013). Hype cycle for big data, 2013 from http://www.gartner.com/id=2574616

Hey, A., Tansley, S., & Tolle, K.M. (2009). The fourth paradigm: data-intensive scientific discovery.

Hood, L., & Friend, S.H. (2011). Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nature Reviews Clinical Oncology, 8*(3), 184-187.

Hopkins, B. (Producer). (2011). Beyond the Hype of Big Data. *CIO*. Retrieved from http://www.cio.com/article/692724/Beyond_the_Hype_of_Big_Data

Hopkins, B., & Evelson, B. . (2012). Forrester: Big Data-Start Small, but Scale Quickly. *Forrester Research*.

Hornbrook, M.C., Hurtado, A.V., & Johnson, R.E. (1985). Health care episodes: definition, measurement and use. *Medical Care research and review, 42*(2), 163-218.

Horowitz, B. (2012). Big data, personalized medicine to trend in healthcare in 2012. from http://www.eweek.com/c/a/Health-Care-IT/Big-Data-Personalized-Medicine-to-Trend-in-Health-Care-in-2012-364022/

Husserl, E.G. (1970). *The crisis of European sciences and transcendental phenomenology: An introduction to phenomenological philosophy*: Northwestern Univ Press.

Hutton, G. (2009). *Scientific grey literature in a digital age: measuring its use and influence in an evolving information economy.* Paper presented at the Proceedings of the 2009 Canadian Association of Information Science Conference, Ottawa, Ontario.

Hycner, R.H. (1985). Some guidelines for th phenomenological analysis of interview data. *Human Studies, 8*, 279-303.

Kates, Robert W. (1969). Mirror or Monitor for Man? *Antipode, 1*(1), 47-53.

Kerr, K., Norris, T., & Stockdale, R. (2007). *Data quality information and decision making: a healthcare case study.* Paper presented at the Proc. 18th Australasian Conference on Information Systems.

Kinnstaetter, K, Lohmann, Adolf W, Schwider, Johannes, & Streibl, Norbert. (1988). Accuracy of phase shifting interferometry. *Applied Optics, 27*(24), 5082-5089.

Konkel, F. (2013). Big data's federal hurdle: federal policy. *FCW*. Retrieved from http://fcw.com/articles/2013/03/11/big-data-policy.aspx website:

Kuner, C., Cate, F.H., Millard, C., & Svantesson, D.J.B. (2012). The challenge of 'big data'for data protection. *International Data Privacy Law, 2*(2), 47-49.

Lameire, N., Joffe, P., & Wiedemann, M. (1999). Healthcare systems—an international review: an overview. *Nephrology Dialysis Transplantation, 14*(suppl 6), 3-9.

Lavis, J., Davies, H., Oxman, A., Denis, J., Golden-Biddle, K., & Ferlie, E. (2005). Towards systematic reviews that inform health care management and policy-making. *Journal of Health Services Research & Policy, 10*(1), 35-48.

Leinweber, David. (2011). Avoiding a billion dollar federal financial technology rat hole. *The Journal of Portfolio Management, 37*(3), 1-2.

Lester, W.T., Zai, A.H., Grant, R.W., & Chueh, H.C. (2008). Designing healthcare information technology to catalyse change in clinical care. *Informatics in primary care, 16*(1), 9-19.

Litvin, C.B., Cavanaugh, J.S., Callanan, M. , & Tenner, C.T. . (2008). To err is human continued: A failure of follow-up. *Journal of Clinical Outcomes Manage, 1,* 21-23.

Lomas, J. (2005). Using research to inform healthcare managers' and policy makers' questions: from summative to interpretive synthesis. *Healthcare Policy, 1*(1), 55-71.

Mandl, K..D, & Kohane, I.S. (2008). Tectonic shifts in the health information economy. *New England Journal of Medicine, 39,* 39-52.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A.H. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, 1-137.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A revolution that will transform how wWe live, work, and think*: Eamon Dolan/Houghton Mifflin Harcourt.

Mays, Nicholas, Pope, Catherine, & Popay, Jennie. (2005). Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of health services research & policy, 10*(suppl 1), 6-20.

McGraw, D. (2012). Paving the regulatory road to the" learning health care system. *Stanford Law Review Online, 64,* 75.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook*: Sage Publications, Incorporated.

Moustakas, C. (1994). *Phenomenological research methods*: Sage.

Neches, Philip M. (1983). *Hardware support for advanced data management systems.* California Institute of Technology.

Paredes, D. (2012). The big career shift: Big Data. *Computer World*.

Parsons, M.A., Godøy, Ø., LeDrew, E., De Bruin, T.F., Danis, B., Tomlinson, S., & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science, 37*(6), 555-569.

Patton, M.Q. (2005). *Qualitative research*: Wiley Online Library.

Pavolotsky, J. (2012). Demystifying big data. *Business Law TODAY*.

Petronio, S., Sargent, J., Andea, L., Reganis, P., & Cichocki, D. (2004). Family and friends as healthcare advocates: Dilemmas of confidentiality and privacy. *Journal of Social and Personal Relationships*, *21*(1), 33-52.

Pope, C., Ziebland, S., & Mays, N. (2000). Qualitative research in health care: Analysing qualitative data. *BMJ: British Medical Journal*, *320*(7227), 114.

Porter, M.E., & Teisberg, E. (2006). *Redefining healthcare*: Harvard business school press.

Porth, AJ, Badke, C, & Mieth, I. (1982). A fundamental data base design for clinical laboratory information systems *Medical Informatics Europe 82* (pp. 57-63): Springer.

Pryor, G., & Donnelly, M. (2009). Skilling up to do data: whose role, whose responsibility, whose career? *International Journal of Digital Curation*, *4*(2), 158-170.

Re, C., Nter, U., & Mill, E. (2012). Agencies rally to tackle big data. *Science*, *336*.

Rhoads, J., & Ferrara, L. Transforming healthcare through better use of data.

Ridley, D. (2009). The literature review: A step by step guide for students. *Sage Study Skills*.

Robertson, J., Dehart, D., Tolle, K., & Heckerman, D. (2009). Health delivery in developing countries: challenges and potential solutions *The fourth paradigm: data-intensive scientific discovery*, 65.

Roney, K. (2012). The rise of big data in hospitals: Opportunities behind the phenomenon. *Beckers Hospital Review*.

Rooney, B. (Producer). (2012). Healthcare Is next frontier for big data. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424052970204468004577169073508073892.html

Roper, W.L., Winkenwerder, W., Hackbarth, G.M., & Krakauer, H. (1988). Effectiveness in health care. An initiative to evaluate and improve medical practice. *The New England journal of medicine*, *319*(18), 1197.

Roscoe, K.D. (2009). Critical social work practice a narrative approach. *Centre for Health and Community Research*, 15.

Ryan, G.W., & Bernard, H.R. (2003). Techniques to identify themes. *Field methods, 15*(1), 85-109.

Sackett, D.L. (1997). *Evidence-based medicine.* Paper presented at the Seminars in perinatology.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B., & Richardson, W.S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ: British Medical Journal, 312*(7023), 71.

Sahoo, S.S., Sheth, A., & Henson, C. (2008). Semantic provenance for escience: Managing the deluge of scientific data. *Internet Computing, IEEE, 12*(4), 46-54.

Savaiano, J. (2013). Bring heatlhcare's dark data to light.

Schatz, B. (2006). A Strategic plan for the department of medical information science. *UIUC College of Medicine*.

Shaw, J. (2014). Why big data is a big deal. *Harvard Magazine, March-April 2014*.

Simmhan, Y.L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM Sigmod Record, 34*(3), 31-36.

Smith, R. (1996). What clinical information do doctors need? *BMJ: British Medical Journal, 313:1062*.

Stake, R.E. (1995). The art of case study research.

Steiner, J.F. (2005). The use of stories in clinical research and health policy. *JAMA: the journal of the American Medical Association, 294*(22), 2901-2904.

Sullivan, Frost &. (2011). Enterprise content management.

Tan, F.B., & Hunter, M.G. (2003). *Using narrative inquiry in a study of information systems professionals.* Paper presented at the System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on.

Teddlie, C., & Yu, F. (2007). Mixed methods sampling a typology with examples. *Journal of mixed methods research, 1*(1), 77-100.

Thomas, D.R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation, 27*(2), 237-246.

Thomas, DR. (2003). A general inductive approach for qualitative data analysis. *School of Population Health, University of Auckland*.

Thomas, I. (2013). An emerging era of big data.

Thyer, B.A., & Myers, L.L. (2011). The quest for evidence-based practice: A view from the United States. *Journal of Social Work, 11*(1), 8-25.

Tongco, M.D.C. (2007). Purposive sampling as a tool for informant selection.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management, 14*(3), 207-222.

Van Manen, M (1980). Researching Lived Experience: Human science for an action sensitive pedagogy., from http://www.researchproposalsforhealthprofessionals.com/phenomenology.htm

Van Manen, M. (1997). From meaning to method. *Qualitative health research, 7*(3), 345-369.

Villars, R.L., Olofson, C.W., & Eastwood, M. (2011). Big data: What it is and why you should care. *White Paper, IDC*.

Weiner, J.P., Starfield, B.H., & Lieberman, R.N. (1992). Johns Hopkins Ambulatory Care Groups (ACGs). A case-mix system for UR, QA and capitation adjustment. *HMO practice/HMO Group, 6*(1), 13-19.

Weisbrod, B. (1991). The health care quadrilemma: an essay on technological change, insurance, quality of care, and cost containment. *Journal of economic literature, 29*(2), 523-552.

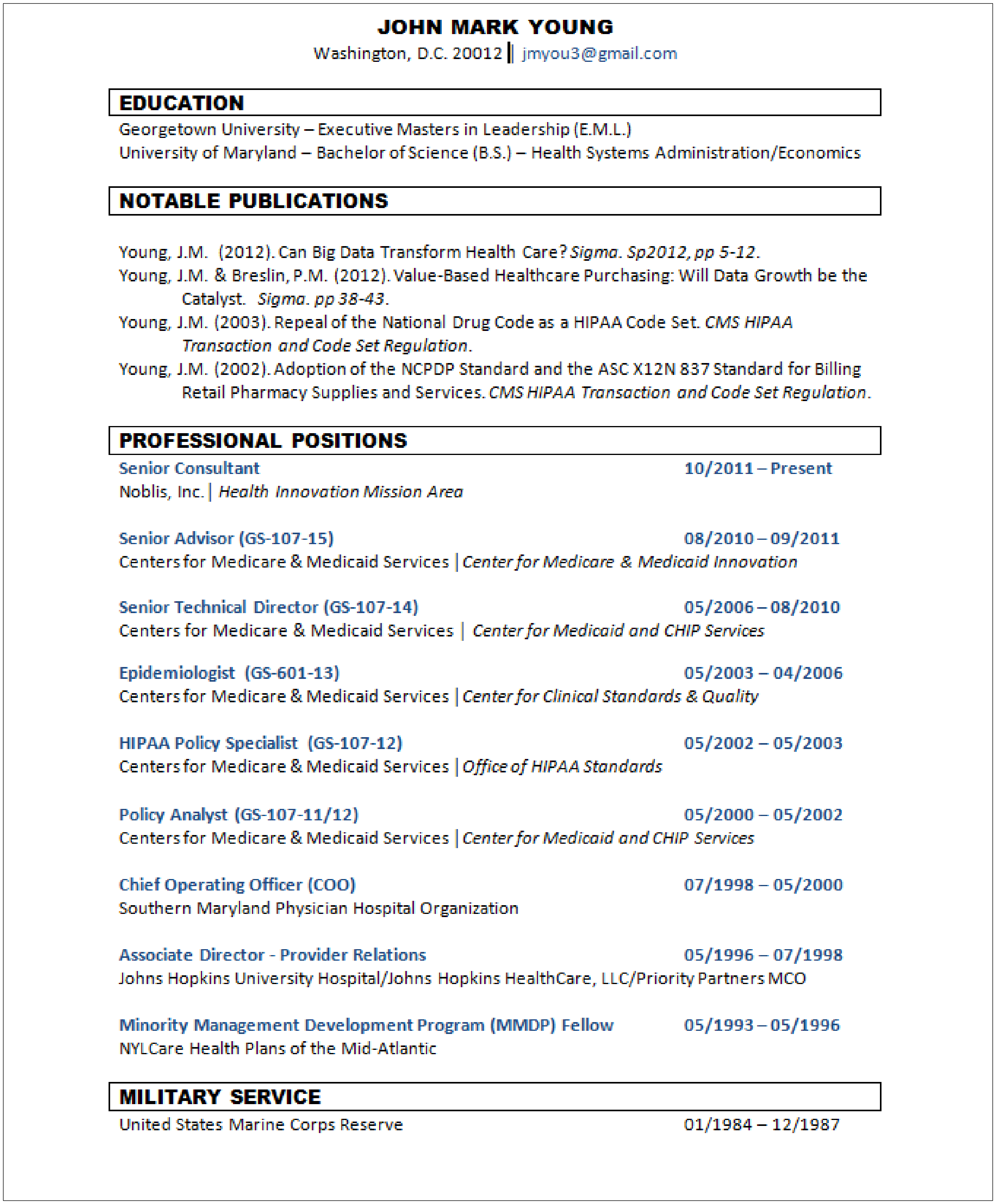
Young, J.M. (2012). What does data-driven healthcare look like? *Sigma*, 8.

Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2013). Sensing as a service and big data. *arXiv preprint arXiv:1301.0159*.

Zeng, M., & Qin, J. (2008). Metadata, by Marcia Lei Zeng and Jian Qin. New York: Neal Schuman Publishers, Inc., 2008. 365 p. $65.00. ISBN 978-1-55570-635-7. *Serials Review, 36*(4), 271-272.

Zhang, Y., & Wildemuth, B.M. (2009). Qualitative analysis of content. *Applications of social research methods to questions in information and library science*, 308-319.

CURRICULUM VITAE



JOHN MARK YOUNG

Washington, D.C. 20012 | jmyou3@gmail.com

EDUCATION

Georgetown University – Executive Masters in Leadership (E.M.L.)
University of Maryland – Bachelor of Science (B.S.) – Health Systems Administration/Economics

NOTABLE PUBLICATIONS

Young, J.M. (2012). Can Big Data Transform Health Care? *Sigma. Sp2012, pp 5-12*.

Young, J.M. & Breslin, P.M. (2012). Value-Based Healthcare Purchasing: Will Data Growth be the Catalyst. *Sigma. pp 38-43*.

Young, J.M. (2003). Repeal of the National Drug Code as a HIPAA Code Set. *CMS HIPAA Transaction and Code Set Regulation*.

Young, J.M. (2002). Adoption of the NCPDP Standard and the ASC X12N 837 Standard for Billing Retail Pharmacy Supplies and Services. *CMS HIPAA Transaction and Code Set Regulation*.

PROFESSIONAL POSITIONS

Senior Consultant 10/2011 – Present
Noblis, Inc. | *Health Innovation Mission Area*

Senior Advisor (GS-107-15) 08/2010 – 09/2011
Centers for Medicare & Medicaid Services | *Center for Medicare & Medicaid Innovation*

Senior Technical Director (GS-107-14) 05/2006 – 08/2010
Centers for Medicare & Medicaid Services | *Center for Medicaid and CHIP Services*

Epidemiologist (GS-601-13) 05/2003 – 04/2006
Centers for Medicare & Medicaid Services | *Center for Clinical Standards & Quality*

HIPAA Policy Specialist (GS-107-12) 05/2002 – 05/2003
Centers for Medicare & Medicaid Services | *Office of HIPAA Standards*

Policy Analyst (GS-107-11/12) 05/2000 – 05/2002
Centers for Medicare & Medicaid Services | *Center for Medicaid and CHIP Services*

Chief Operating Officer (COO) 07/1998 – 05/2000
Southern Maryland Physician Hospital Organization

Associate Director - Provider Relations 05/1996 – 07/1998
Johns Hopkins University Hospital/Johns Hopkins HealthCare, LLC/Priority Partners MCO

Minority Management Development Program (MMDP) Fellow 05/1993 – 05/1996
NYLCare Health Plans of the Mid-Atlantic

MILITARY SERVICE

United States Marine Corps Reserve 01/1984 – 12/1987