Governing Medical Big Data, Protecting Patient Privacy

Bilal Jibai, Hassan Najdi

Abstract: With the power that big data provide to managers and organizations to yield insights for optimal institutional decision making, much concerns on the privacy of individuals rise on how collected information are reused. Individuals, here, are patients in the healthcare context. The study proposes a model for governing Big Data mapping traditional data governance practices while integrating security and privacy practices ensuring its continuity past Big Data processing and analytics.

Index Terms*: *Keywords: Big Data, Data Governance, Information Security, Information Privacy, Compliance, Healthcare Sector.

I. INTRODUCTION

In the information era, as scholars around the world still debate the definition of Big Data, they agree on the power that Big Data promises through analytics, predictive analysis, machine learning and artificial intelligence. Big Data represents a new leading frontier for science in the technological big ban, considering it as a core economic strength or strategic tool for organizations and nations alike to nurture growth and innovation, as Rubinfeld and Gal (2017)[1] "In such a world, access to data and to the information based on it becomes a strategic and valuable asset". From Alphabet, Google's parent company, Chairman Eric Schmidt describes it "I think that big data is so powerful that nation-states will fight over how much data matters"[30]. With headlines as IBM's 1 Billion dollar in Watson investment focusing on Big Data and artificial intelligence, as well as prediction by the leading global market intelligence provider International Data Corporation (IDC) that big data analytics industry will foster 203 billion dollars in revenues by 2020, all of which reflects the size of matter at study. Big Data has gained the attention of governments over the globe; the U.S. Obama administration in March 2012 launched a 200-million-dollar investment in "Big Data Research". Moreover, the U.S. Internal Revenue Service (IRS), assigned the tax collection monitory role, employs Big Data analytics to detect uncompliant taxpayers.

defines "Big Data is an abstract concept." A more coherent Big Data core definition rests with its original 3Vs characteristics as described by Doug Laney from Gartner[3][37]. Volume, the main, represents the massive huge size and amount of data, Velocity the speed at which data flows from different data sources, and Variety for different sources and types of data. Some scholars have stretched this attribution to include Veracity describing data's accuracy and more importantly, Value for the hidden values explored from data's analysis and correlation. The technological advancements paved the way for real time operations at all levels of the data lifecycle, allowing instant collection and storage, and instantly employing mining, associating and analyzing abilities. These advancements are algorithms that extend its abilities for an efficient and quick analysis of correlations between variables, patterns discovery, predictions creation and past actions learning.

Big Data ultimately provides insights that in conventional method would be rather be tougher for researchers to extract, as experts shed lights to the importance of its implementation in the healthcare sector. Organizations have stretched the infrastructure of data collection from databases to data warehouse integrating different sources of datasets. Data that can be structured or unstructured, data can be of different formats, types and schemas. Hence, we have exceeded the capacity of data warehouses and reached a contemporary concept "Data Lakes" that has the ability for unlimited storage of all types of data. In his article "5 Big Data Trends in Healthcare in 2017", illustrates "One major healthcare provider leveraged a data lake approach as it aggregated massive volumes of data as a data hub for various departments, including fraud prevention. As a result, the provider is on the way to capturing an incremental 20% of fraud, waste, and abuse in its claims department." As the practice in healthcare providers nowadays are moving toward patient-centric and evidence-based medicine, it is crucial to capture all data types and sources such as claims, clinical, pharmacy, electronic health records for the outcome of a combination of health care services. The ability of patient real time monitoring from Health IoT wearable devices collected data combined with real time processing algorithms of predictive analytics and machine learning provides physicians with insights for lifesaving decisions.

Revised Manuscript Received on December 22, 2018.

Bilal Jibai, Department of Management Information, International University of Beirut,Lebanon bilal.jibai@liu.edu.lb

Hassan Najdi, MBA in MIS, Department of Management Information Systems, International University of Beirut,Lebanon 11532071@students.liu.edu.lb

Fig. 1. Information at the heart of privacy then security pillars (CIA)

"Not all privacy scholars are as concerned, however: Omer Tene and Jules Polonetsky have advocated a loosening of privacy regulations in order to unleash the full power of big data for economic and social growth." It goes, without saying, that some scholars had privacy concerns for Big Data. However regardless, it is necessary to recognize the relation between information security and information privacy. The U.S. Health Insurance Portability and Accountability Act (HIPAA) Title II is the best advocate of this relation as it has developed two rules for healthcare providers to comply with; Privacy Rule and Security Rule. Information privacy covers the combination of what an individual expects private, means of collecting and disseminating information and applicable laws and regulations; whereas information security through safeguards/ controls ensure its pillars confidentiality, integrity and availability of information. These safeguards contribute to developing data governance. Therefore, achieving privacy requires security.

The study proposes a model mapping traditional data governance practices to big data ensuring the continuity of security and privacy practices past big data analytics and processing. We will be looking at exploring the area where Big Data, Privacy and Data Governance meet. Figure.2 Illustrates.

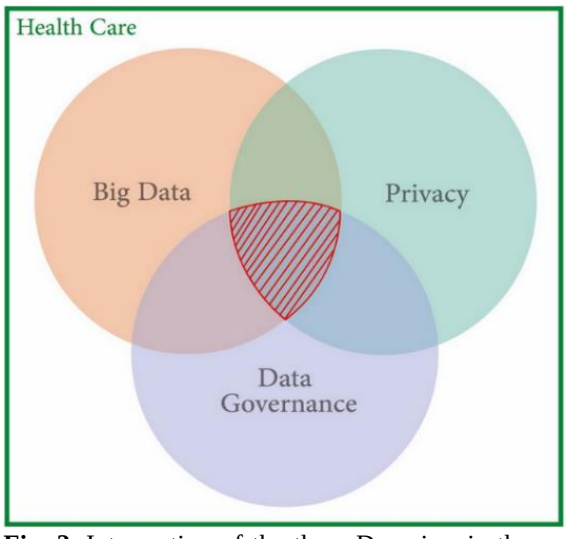

Fig. 2. Intersection of the three Domains, in the context of healthcare.

II. LITERATURE REVIEW

A. ig Data

Chen et al. (2014) states that "Big Data typically includes masses of unstructured data than need more real-time analysis." They declared that Big Data opens opportunities for insights to discover new hidden values when datasets are effectively organized and managed. The importance of Big Data has captured the interest of industries and government agencies as they set forth major plans for research and application of Big Data while the public media speculates around its issues, however many companies like Google and Facebook generate and process tens of Petabyte of data daily from online transactions. The challenging problems that demand express solutions include the collection and integration of widely distributed data sources, increased dependency on cloud computing and rapid growth of Internet of Things (IoT) that employs sensors globally to collect and transmit data to the cloud for storing and processing, all of which has surpassed the computing capabilities of enterprise IT architecture and infrastructure. Big Data's foundation could be traced back to Doug Laney's study in 2001 with the 3Vs model. The model describes three characteristics volume meaning the mass generating of data, velocity to collecting and analyzing, and variety indicating the various types of structured and unstructured data. However, an influential leader in Big Data, an International Data Corporation (IDC) report defined the 4Vs of Big Data that includes value for exploring the meaning and necessity over the original 3Vs. reflects on a report from McKinsey & Company that conducted and in-depth study on five core industries including U.S. healthcare, administration of E.U. public sector and U.S. retail. The report found that Big Data when utilized effectively and creatively improves productivity and competitiveness among enterprises and the public sector. The report's results also suggest an increased 300 Billion USD in the value of the medical industry in U.S. Healthcare sector and an 8% reduction in U.S. expenditure.

defines "Big data is a generic name for data that share several characteristics with regard to their aggregation, rather than content". Their observations however include *(1)* the wide range of the content of data, *(2)* markets requiring specific data such as a sports car dealer's interest in income and spending habits, *(3)* data having different use and meaning to different users…The authors investigated the access barriers to Big-Data markets along the data value chain. Technological barriers include abundance of unique data that are not publicly available, high fixed costs of building and infrastructure for data collection versus low cost of data extraction, intermediary platform between two groups of users such as Google's role between the advertisers and the advertising sites, and information location and ownership. Legal barriers include data protection and privacy laws; many authorities have forced limitation of data collection activities especially to personal data. European Union and United States has a "Safe Harbor" agreement to protect privacy their

citizens in terms of their personal data storage outside its jurisdiction. Other barriers include the way data are organized in databases with numerous parameters and the constant updates, the power analytical tool used to make correlation through advanced algorithms empowering techniques such as artificial intelligence, machine learning and pattern recognition. Intellectual property in some cases, considered by the authors as a direct barrier of Big Data reuse. concluded that Big Data markets "have the potential to create durable market power in data-related markets or to serve as a basis for anticompetitive conduct."

In the context of Medical Bioinformatics, Merelli et al. (2014) defines, "Big Data is based on the concept of data sets whose size is beyond the management capabilities of typical relational database software[6]." With the complexity of biological systems in mind, the authors discuss the consequent need for complex algorithms for mining bio-information. Big Data architectures requires reliable storage capacity, consistent access to data from integrated systems and excessive transfer operations. Managing and accessing Big Data is one critical area that has challenges determining the technical requirements of the file system, network bandwidth, and storage and computational parallelism, and server/client models. Data parallelism in cluster computing is one aspect of analysis facilities. Cloud computing, especially in bioinformatics field is advancing as it addresses issues relating to storage and analysis. Semantics encourages standardization in data integration and annotations. Ontologies are very important in the bioinformatics context as it removes ambiguity over terminology of words. Whatever technology, features and capabilities it provides, the need for secure authentication and measures to ensure authorized access to data maintaining its integrity. The authors conclude with open technical problems and solutions such as relational databases lacking scalability and consistency in maintaining complex Big Data relations, where as non-relational databases can combine both. Virtualization technology that can be key in facilitating the in-memory in databases, among others as well.

Mendonc et al. (2016) paper proposes a risk management model for big data, integrating risk assessment exercises in the Big Data process at different crucial stages using NASA's approach for specific FMEA (Failure Mode and Effects Analysis) method for discovering failure causes[10], effects and problem affecting a system. The model also employ the Grey theory introduced by J. L. Deng for grey systems, a methodology for solving uncertainty problem dealing with system of incomplete and lack in information. The proposed model includes several steps starting from past data or an expert's knowledge in previous failure, followed by determining and evaluating potential failure modes while classification under the four dimensions of Big Data Security (identification and access management, device and application registration, infrastructure management and data governance). Step 3 is the grey analysis of relation for possible accidents where grey coefficient and degree of relation is calculated to rank the priority of risk. The researchers concluded that Big Data Security's main difficulties in risk analysis is the volume and variety of data from different data sources requiring structured analysis to address risks. The proposed model is justified to identify and assess quantitatively Big Data risks. They also highlighted the importance of Big Data governance as to ensure that appropriate controls exist.

paper introduces the Internal Revenue Service (IRS) branch in U.S department of treasury, and their implementation and application of Big Data to detect fraud and tax evasions[11]. Using Big Data Analytics, data from commercial (Facebook, Instagram) and public pools were mined. Potential non-compliant taxpayers were identified by running pattern recognition algorithms.

B. Data Governance

discusses the corporate level of data governance while reflecting its importance through data integrity being mentioned in the guidance documents of several agencies such as the World Health Organization. McDowall states that it takes longer time for compliance but remains cheaper than adopting corrective action in case of non-compliance[20]. Despite that, data governance has been recognized for the past decade, however is not in good practice, starting from differences between regulatory and non-regulatory definitions. He emphasized the involvement of senior management to maintain effective data governance by referring to several regulators (U.S. and E.U.) from a quality system and in the context of pharmaceutical. Accordingly, when considering several guidance documents and a publication from the Society of Pharmaceutical Engineering, we can drive a data governance structure for sponsoring data governance program, change management controls, developing policy and procedures, identifying data roles and responsibilities, and establishing quality assurance. Certain applications of role and responsibilities in a data governance program may include designating Chief Data Integrity Officer, data governance steering committee and subcommittees, as well as certain responsibilities at level line management, information technology and quality assurance. Furthermore, policy is required to guide implementation and compliance of the program, as well as metrics to help managers monitor it.

developed whitepapers for guide to data governance from Microsoft. The authors examined different aspects of data governance recognizing the increased reliance of confidential data that may include intellectual property, corporate secrets, market insights and customers' information all at which to achieve compliance obligations. Data governance as defines it[17], "an approach that public and private entities can use to organize one or more aspects of their data management efforts, including business intelligence (BI), data security and privacy, master data management (MDM), and data quality (DQ) management." Challenges to information security from criminal enterprises, espionage and warfare; identity theft and misuse of personal information for information privacy; and legislations such as U.S. HIPAA (Health Insurance Portability and Accountability Act), Canada's PIPEDA (Personal Information Protection and Electronic Document

Act) and E.U DPD (Data Protection Directive) makes the regulation landscape more and more complex.Data Management International provides a guide for core- data management functions. The authors proposed a framework for DGPC (Data Governance Privacy, Confidentiality and Compliance). The framework ensures implementation of the data governance practices reviewing business strategy, regulations and standards; allows refinement of requirements for GRC (Governance, Risk and Compliance) to define strategy followed by controls. The framework provides four guiding principles for honoring confidentiality, reducing risk of unauthorized access/ misuse, reducing impact of data loss, documenting and demonstrating applicable controls and their effectiveness. Furthermore, basic DGCP policies are information security, privacy, data classification and data stewardship.

article on "The Changing Health Data Governance Landscape". The author observed the change on three major areas, *(1)* the challenges from the growth of importance of data governance must be addressed; *(2)* the governance process in healthcare organization must be prioritized based on design, implementation and functions *(3)*; and that governance is naturally standardizing. The author reflected the shortage in literature for risk-based strategies for managing privacy of new electronic sources of health data from the field of health data governance. The author stated the works of several of his colleagues and associates. This included revised data access policies for funded data sets from both state and federals, the differentiation between data of different activities of treatment and research, a case study for extraction of EHR (electronic health record) data as well as engaging patients in using their data and learning privacy for better decision all of which has impact to governance structure. In this sense, more governance policies and procedures are required forced by emergence of distributed research networks, healthcare organization require more specialized individuals such as privacy, security and regulatory experts; and governance structure required institutional review boards and data governance committees.

From eGEMs, we also look at Petersen (2016) article on "The Future of Patient Engagement in the Governance of Shared Data." The author describes excitement in rapid evolving filed of precision medicine and patient-centered medical care at home model. Precision medicine tailors medical care, practices, decision and products to patients individually. The great quantity of PHI (Personal Health Information) brought with such models allows the use of patients' data for research for new treatment advances. However, this brings forth concerns regarding security, privacy, ethics, legal and social issues, which requires evolving policies towards governing data sharing and usage. The author suggests that in the same manner where patients sign written consent for the use of their demographics and health information for providing medical care, they are asked additional questions on how their data are used, and how it can contribute to data sharing for research, quality improvements… This process requires technical assurance through developing and monitoring a data use policy, data management and data protection mechanisms that are compliant with state and federal regulations such as HIPAA (Health Insurance Portability and Accountability Act).

In the January edition, senior writer of wrote, "Organizations are increasingly recognizing the value of corporate data as a resource and consequently are focusing more on governance of that data[16]." The title of the article reflects governance as being a mandate driver for data enterprises. The author gave an example of a gaming company that sought GRC (Governance, Risk, and Compliance) solutions and its choice of Keylight. Keylight provided four programs such documentation controls for policies, risk assessment for third party vendors, incident management and security management. The applications provided a structure approach for managing policies and their reviews, a more systematic way to assessing risks, as well as integrations with other security solutions for capturing analytics and drawings dashboard. Such solutions capture the attention of C-level management such as the Chief Information Officer from its governance nature, since components of governance are settings business policies pertaining data security and privacy; and enforcing those policies.

discuss, "Critical Factors in Data Governance for Learning Analytics". The term "learning analytics" was defined with the support of previous studies as all activities in the context of an environment to optimize the understanding and learning process. Activities, here, include collecting, analyzing, reporting and measuring data about the learners. While "Governance" is described in the study as being the process of influence, power and authority distribution among involved parties including students, faculty, trustee board and even committees and sub committees. The researcher emphasizes relation between the definition and governance, where a broad definition states more conflict to the accountability and control of stakeholders. The model takes into account both governance from IT and institutional (business) perspectives with careful observation of authority distribution.

The critical factors, as the researcher puts is, includes *(1)* data ownership, its nature and distribution such as faculty owning data from learning process while administrative facts owning part of that data. *(2)* Organizational data silos resulting from interpreting learning analytics data, and accordingly (*3)* Strategic decisions that require evidence and facts –based analysis. Suggestions made in the study to build a comprehensive model is stating small, key stakeholder support and empowerment, considering the information from learning analytics process, power distribution setting roles and responsibilities, provide operational principles and sanctions to manage conflicts and struggles, assessing and sharing the maturity of the model among stakeholder and understanding the requirements of ethical and legal aspects.

C. Privacy

discuss, "Dynamic access control model for privacy preserving personalized healthcare in cloud environment". PHR (personal health record) unlike EHR (electronic health

record) or EMR (electronic medical record) contains all health information about an individual such as blood pressure, weight, medical history, surgeries, vaccination… with data even pulled from wearable devices. The researchers here propose a DAC (dynamic access control) model, which substitutes existing models such as traditional RBAC (role based access control) that is most convenient for health care setting however has privacy preservation limitations. Components of the model structure as suggested by researchers, Son et al. (2016),"ProfileManager, ContextManager, Access ControlManager, and PHR Data Storage." Just like any access control having a Subject (patient, physician or any individual) requesting is requesting access to an Object (data, application, system…), the process of accessing PHR on the cloud are follows

1. ProfileManager validates subject's information and role request, (2)ContextManager analyzes the context (question around who, when, where, why…) of the subject (3) Access ControlManager checks the access conditions and goals of the object (4) Map context information and access conditions.

The proposed model is intended for use in the healthcare domain with data on the cloud. DeAngles (2015) debates, "EHR systems improve the quality of care from providers, reduce mistakes, allow for more timely research in public health, and greatly reduce costs." EHR stands for Electronic Health Records, a system that is implemented in healthcare organization to manage clinical data about their patients. While providing the increasing numbers of EHR implementation between the years of 2005 and 2009, the writer reflects on the notions from both former US presidents George W. Bush, and Barack H. Obama to adopt EHR at all healthcare providers through compliance with Health Insurance Portability and Accountability Act (HIPAA). Office of Inspector General (OIG) of the U.S. Department of Health and Human Services (HHS) called for its concerns around security and privacy issues in the absence of EHR. OIG conducted regular reviews for security controls of EHR as well as medical device and machine that can be connected to organizational network. The writer expressed opinion in federal government nationalizing EHR regulations, establishing an EHR network on a nation-wide scale. Technical constraints include interfacing induced loss of data transmissions, data formats rendering and cloud computing among others require some standardization.

Existing legislative and laws present barriers for transferring EHR, examples include privacy laws of some states are more restrictive than federal laws as well as HIPPA privacy regulations. With more details around the limitation of U.S. laws and regulations, suggests, "the federal government should regulate a national network of EHRs to minimize waste associated with the current EHR framework of fragmented networks".

discusses, "A tidal wave that could turn our healthcare system upside down" that there is possibility of putting use of the evolving number of CHDs (connected health devices) referring to applications and gadgets that provide individuals with health indicators[24]. Using CHDs, individuals' behaviors such as physical activity, diet, smoking, consumption of alcohol, use of medication can be tracked. Looking at future possibilities, doctors will use this abundance of information for their practices, however a need for expert systems and artificial intelligence to avoid overload information. This mandates re-organizing healthcare systems to import such data sources, however this may be tripped by the fact that some healthcare activities may fade with time. Another challenge with the development and adoption of CHDs in healthcare is maintaining the availability and access to everyone, suggesting that authorized government or state agency should govern this function. The writer, therefore, suggests that instead of resisting this change, embrace it by channeling it towards positive impact.

Key concern here is CHDs data quality vs. legal issues; where in some critical areas are risky considering their control over insulin pump for example. A proper assessment should conducted to evaluate the quality of CHD. Another concern regarding data and information collected and processed from CHDs is privacy. Individuals who access to such information must be authorized to do so. Technical restrictions must be in place to monitor and protect unauthorized access. Laws is another control for preserving individual's privacy. All European products of such devices must comply with the obligations of protecting privacy of individuals. The writers reflect the importance of "Privacy by Design", where privacy is considered in the early stages of the project taking into account information collection, use and user's access rights. Another concern is the security and hence the safety of patients, when an attacker abuses a weakness in the flow capturing a message from CHD specifying high glucose level instructing the insulin pump to deliver insulin dosage. The attacker can then replay attacks putting patient safety at risk. Therefore, there is need for mechanisms that can ensure the security goals (Confidentiality, Integrity and Availability). The writers propose that there is a need for ethical and medical evaluation, guidelines for security, as well as sanction to enforce compliance from manufacturers and producers of such CHDs

D. Big Data, Data Governance and Privacy

exploring "it is the legal framework of global privacy that gives new color to this concept as applied to Big Data." Everson was referring to the concept rooting from system engineering known as "Privacy by Design". The framework is a comprehensive approach that takes privacy practices prior to the system development lifecycle, anticipating how data will be collected, used and accessed. The research explores the principles of this framework being proactive, preventive, the default setting of privacy, privacy embedded design, full lifecycle security protection, visibility and transparency and user-centric privacy.

The researcher emphasizes the failure neglecting the adoption of the framework in system/solution development leading to limited controls. Repercussions on cost of time and money,

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

reputation, performance falling back and security is doing due to privacy weaknesses.

Privacy by Design framework assists organizations in breaking down the complex legal, ethical and social concerns by fostering privacy practices into systems requirements.

explores the principles mentioned herein and mapping its practices to Big Data while providing real life examples. *(1)* Proactive consideration of the broad context of data during its lifecycle (collection, storage and use) and the impacts and perceptions that may arise. *(2)* Fostering a privacy aware culture through privacy policies enforced and monitored by management privacy governance. *(3)* Design the Big Data infrastructure building the data environment up while privacy is in mind. *(4)* With privacy in mind, surrogating certain data elements removes privacy risk when provided to creators of new data element. *(5)* Privacy expectations in the public context such as Facebook's self-enable privacy settings represents best practices for data-driven businesses. *(6)* Visibility and transparency throughout Data lifecycle remove obscurity around data source; Cloud-Computing represents a good example as a case study. *(7)* Building a privacy aware data-rich solutions require a tiered approach segregating end user's access to data. *(8)* Privacy laws and regulations that cater for the emerging Big Data era. Everson (2016) concluded, "the sheer volume and velocity of future Big Data environments will continue to expand" expressing opinion in making the decisions that will shape he Big Data landscape future based on Privacy by design.

their processes and procedures because of the definition if holds for information from UTSA (Uniform Trace Secret Act).

Industry practices for dealing with Big Data includes practitioners of big data to mask PII (personally identifiable information) within the raw data. Removing information relating to names, addresses, and phone numbers that

However, many international organizations are working on standardizing the data format to enable aggregation of Big Data. Furthermore, procedures around how data are collected and prepared are being required from researchers. The author explains how intellectual property law in the U.S. influences public disclosure through can identify individual anonymizing the dataset. Classification is another approach for altering data before it becomes public available. Some practitioners perform such functions with risk of subjectivity, while advanced software does that for them. However, it remains crucial the knowledge of the context for such human judgment.

Author discusses implications and concludes suggesting a law proposal. Concerning patent law, it is unlikely that it will encourage disclosure for Big Data due to the absence of an official definition of Big Data. Copyright law provides Big Data minimal protection as courts consider subjective judgment as complications in process of data selection and arrangement. A new law is required for Big Data under intellectual property with characteristics of subject matter, exclusivity for publishers while conditioned by set of acquisition rules.

"Big Data and security policies" debate the development of a framework for regulating all phases of Big Data process (collection, analysis and use) in the field of security for law enforcements and intelligence agencies[28]. This includes adding new layers of protection, "duty of care" audit reviews and legislations for responsibility of accurate analysis of data processing. Big Data provides a sense of threat towards individuals and societal freedom, therefore mechanisms for accountability should be in place. Government agencies use Big Data for combating fraud, due to its predictive analysis in the field of intelligence. Some indications provide insight to the future of Big Data. Smart devices and Internet of Things (IoT) have exponentially exploded the availability of data. Emerging new techniques and algorithms such as self and machine learning. Predictive analysis is another emerging concept with the availability of both historical and real-time data, which in turns provide preemptive capabilities in terms of security to law enforcement and government agencies.

More benefits in the field of security include refined and precise risk analysis better forensics analysis in investigations. Hence the dilemma here between its benefit and the collection of a mix of both private and public data, and its impact on the perception of freedom. The authors debate their framework listing formal practices of regulation on the analysis and use phases of the lifecycle of Big Data, considering that existing regulations already force rules on data collection phase. In regulating analysis, principles such as duty of care are suggested to address concerns. Concerns include legally acquiring data, data consistently up to date, and that biases may exist within their dataset. Moreover, the methodology and algorithm must follow scientific criteria and is open for reviews. Another suggestion includes external reviews or audit by an oversight authority. Evaluation of such projects to include goals and impacts to privacy, freedom and security as well as cost. In regulating use however, benchmarks are required to link the security organization and the impact to individual and his rights, more details here are required to remove margins of error. Additionally, consider banning of automated decision-making, as well as the responsibility that falls in the hand of the party that acts based on the analysis showing the considerations and factors.

In a separate complimentary area, of "Who watches the watcher?" where the authors emphasized the term "intensified oversight". Transparency, accountability and judicial reviews all fall under this category.

III. METHODOLOGY:

Big Data still maintains its position as a trending concept, researchers have yet to quantify and harness their broad applicative theories, as information are not yet saturated. The study here will be an exploratory qualitative research study to gain an understanding of data governance and privacy practices and how they can be mapped to Big Data. This should provide some insights into the problem of privacy concerns in the Big Data landscape.

Data collection tools includes interviews as primary source of data with a convenient sampling of key selected personnel from an organization working in IT department where their

experience within the field and positions of operations management and data warehouse administration that will expand the knowledge into the study problem. Observation is another means of qualitative research technique, with experience in the field of information security and privacy, and researching literature as secondary data sources. This provides insights that will assist in developing the proposed governance model.

Some limitations exist to the study that are barriers. This demeans the ability for conclusive ideas and suggestions with the scarcity of available information for a comprehensive view to the area where Big Data, Data Governance and Privacy meet. Interviews, despite convenience, were conducted in one organization disregarding other organizations' views of implementing security controls, privacy practices and data governance with possibilities of biases. The analysis of data collected are done on the assumption that it is accurate from primary and secondary sources.

IV. FINDINGS:

A. Interviews

The interviews conducted with operations manager reflects the business aspect, while database warehouse administrator reflects the technical aspect of the study:

- Privacy practices through several activities such access monitor and review, patient consent and documentation, policy on confidentiality and release of information…
- Security practices such as continuous information risk assessments and evaluations, identity and access provisioning, log management and reviews, disaster recovery exercises…
- Audit and reviews both internally and externally conducted on yearly basis
- Data governance practices and structures exists, roles and responsibilities about data management surfaces, high management committee exists making decision for organizational data…

B. Observation from Literature

1) Privacy Practices

There are general privacy principles that several laws describe. Such laws are Canadian Personal Information Protection and Electronic Documents Act (PIPEDA), U.K. Data Protection Act (DPA) and U.S. Health Insurance Portability and Accountability Act (HIPAA). Objective reflected to our study here is that it highlights the importance of key principle factors that are

- Data obtained with the consent of patients
- Data are accurate, complete and is up to date; patients have the right to request changes
- Data collected must be limited towards the specific purpose of collections
- Data access, authorization and disclosure follows the concept of "Need-to-Know" basis for providers of care services; patients have the right to be aware when their data are disclosed.

2) Security Practices

Security practices are to ensure that the security pillars are maintained, aiming at avoiding security risks impact to information systems and data… and with Bigger Data comes bigger magnitude of the impact.

Risk Analysis

Risk analysis is an ongoing continuous process. It allows identifying risk and associated threat and vulnerability, analyzing the measure required for appropriate action to risk mitigation. Risks observed here may include unauthorized access or disclosure all of which may lead to security breaches or privacy violations. Several methodologies exist for a risk assessment exercise; most common methodology is the "NIST SP 800-30" from National Institute of Science and Technology with more emphasis to healthcare and HIPAA in precise in their special publication NIST SP 800-66.

Access Monitor and Review

Access monitoring and review is another process that requires continuous attention. HIPAA forces monetary penalties in case covered healthcare organizations are not compliant to its practices, one of which is access provisioning and continuous monitoring by checking systems logs and audit trail for successful logins and authorizations within the scope and duration of the work whereas otherwise would raise flags.

Compliance

Compliance typically refers to the adherence to rules, regulations, laws, policies and procedures in a certain written document. Up to this date, and while the concept is trending, no law exists to define nor to govern Big Data. However, traditional policies allow organizations to expand the framework of applicability to Big Data. It is by default essential for below data governance policies to guide this applicability.

- Information Security Policy, high-level policy supported by others that guide the overall implementation of an information security program.
- Data Classification Policy, lays out a baseline for the levels of data based on their nature and criticality for proper allocation of security measures.

Proposed Model

Following the above findings on privacy and security practices, compliance and data governance practices; we introduce the proposed model. The proposed model assumes a broader definition for data roles to satisfy the need to govern Big Data while ensuring patient privacy in our case. The below proposed definition can be functions or roles expanding traditional data governance data roles.

Data Council represents both data owners and stewards at one instance, who have experience with data decisions, data types and data management. In our model, they represent the entity that defines the scope of data and its types required for data analytics activities whether in a routine or at instance processing.

Data Immunizer represents in part the data custodians,

the role or function in charge of applying security practices to Big Data to monitor accesses, identify risks for proper measure allocations, educate individuals involved in the Big Data process to security risks and monitors compliance. It is natural to consider this role/function falling within the scope of an IT unit or department.

Data Scavenger is the role/function in charge of the data collection process. It may also fall under IT responsibilities due data collection techniques including creating data warehouse and data marts as well as data mining techniques to create new data.

Data Anonymizer is the role/function that monitors the privacy principles and conditions when processing Big Data as well as applying de-identification techniques, which eliminates the chance of patient re-identification in the scope Big Data processing.

Data Researcher is the role/function that runs tools and techniques for discovering hidden information within data. Tools may include business intelligence tools and business analytics techniques such as machine learning, predictive analysis…

The above proposed roles/functions as well as policies and procedures and applicable law and regulations should formulate the following model. It is worth mentioning the need to expand the coverage of laws and regulations and policies and procedures to cover Big Data in the governance proposed model.

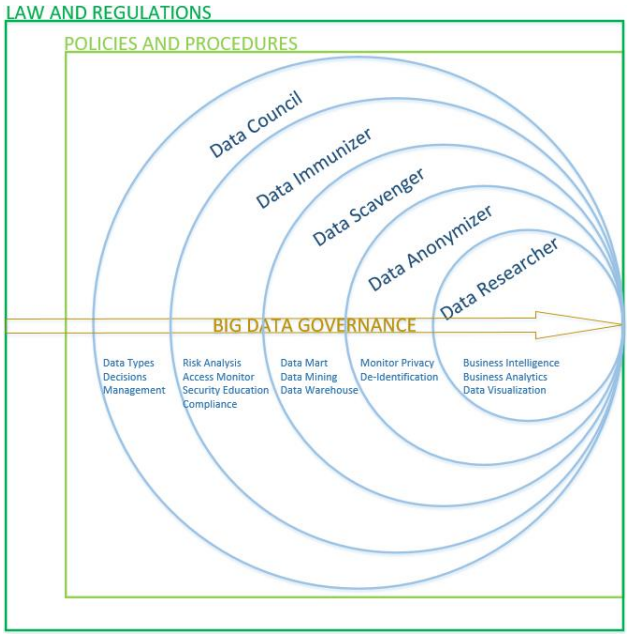

Fig. 3. Proposed Model for Big Data Governance

To better illustrate our model, let us consider in our healthcare setting that there is a new H1N1 outbreak and that a specific medical center has a Big Data implementation to make data analytics. Management of the medical center wants to forecast and predict the threshold of this pandemic to make available the required vaccines and medication. Data Council meets to decide on what type of data they are looking for. These data may include symptoms found in datasets within patient records, in addition for data relating to residence and recent location visits traveled. Data immunizer makes sure that the Data scavenger's access is provisioned and monitored for the required data extraction from data repositories. Data scavenger creates data mart from extracted data and makes it available for the Data Anonymizer that applies safe harbor de-identification technique to eliminate the chance of patient re-identification to avoid privacy violations. Data researcher applies business intelligence and analytics and gets useful information. In our example, this information can yield predictive information such as that Africa is place of origin of the new H1N1 outbreak as well as that the medical center should expect 100 cases to be reported within the coming the month.

V. CONCLUSION :

Traditionally organization would not store data beyond the capacity of its operations. It is not until it was legally required to retain certain data that organizations started collecting and storing them for periods. In the case of enterprises, it was an unlimited retain strategy rather than face non-compliance issues. Having that history of records and data, as well as the advancements in technology, more accurately, data storage and analytics allowed the big bang of Big Data. However, Big Data is still at large for reaching a concrete definition that allows development of international laws providing legal frameworks for organization and institutions alike to work with. This is one end of the problem; it is common and collectively agreed upon the power that Big Data provide. However, the means of processing them beyond the purpose that they were originally collected for raises privacy concerns.

Within the context of the healthcare industry, we proposed the title "Governing Medical Big Data" out of which the area of study is the intersection of three major topics, Big Data, Privacy and Data Governance. The exploration of this area allowed the proposal of a governance model for Big Data. The discovery of certain materials were required to assemble the proposed model. Privacy practices include general conditions and principles from international rules and regulations. The essential privacy practice for the model was de-identification techniques of patients. Security practices essentials included risk analysis and access monitoring, compliance to predefined organizational policies and procedures as well as rules and regulations. Data governance practices include plans, policies and structure. The main concern is structure of the data governance body that includes data owner, data steward, data custodian and data user.

The proposed model expanded the roles of traditional data governance best fitting the argument. The model introduced roles and functions such as data council who is the acting authority of data, data immunizer who employs security practices, and data scavenger who collects and creates data, data anonymizer that employs privacy practices and researcher who looks for hidden information within data.

The contribution to this domain is a basic model that requires more research for clearer descriptions for one part of

an overall framework. With this said, we have not fully explored other aspects required for a standardized Big Data framework; this includes technical, social and legal aspects. Researchers are urged to explore Big Data from other aspects and other settings while attempting to provide solutions that can leverage the potential to develop hypotheses for quantitative researchers. Such solutions can assist in developing the rules and regulations for a legal framework as well as develop best practices for standardized Big Data operational framework. Organizations wishing to pursue Big Data projects should employ the highest levels of security and privacy practices. This includes educating, in our case for example, patients about how their data will be used and what measures exists to protect their privacy thus obtaining their consent for further use. The combined cooperation among organizations will provide the public interest with the greater good. Data sharing and exchange, leveraging Big Data analytics to provide better services and research capabilities. This is most precise and promising for healthcare organizations.

REFERENCES

1. Rubinfeld, D. and Gal, M. (2017), *Access barriers to big data*, Arizona Law Review. 2017, Vol. 59 Issue 2, p339-381. 43p.
2. Jung, A., Ardihi S. and Inoubli, W. (2017), *An experimental survey on Big Data frameworks*, Cornell University Library, available online: https://arxiv.org/abs/1610.09962
3. Chen, M., Mao, S. and Liu, Y. (2014), *Big Data A Survey*, Publisher: Springer US, Mobile Netw Appl (2014) 19:171–209
4. Fatimah Lateef (2016), *Big Data: Applications in Healthcare and Medical Education*, Education in Medicine Journal. 2016, Vol. 8 Issue 1, p85-89. 5p.
5. Saxena, S. and Al-Tamimi, T. (2017), *Big data and Internet of Things (IoT) technologies in Omani banks a case study*, foresight, Vol. 19 Issue: 4, pp.409-420, available online: https://doi.org/10.1108/FS-03-2017-0010
6. Merelli, I., Sanchez, H., Gesing, S. and D'Agostino, D. (2014), *Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives*, BioMed Research International. 2014, Vol. 2014, p1-13. 13p.
7. Mittelstadt, B. and Floridi, L. (2016), *The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts*, Science & Engineering Ethics. Apr2016, Vol. 22 Issue 2, p303-341. 39p.
8. Sun, Y. and Mao, H. (2017), *A Way to Understand Inpatients Based on the Electronic Medical Records in the Big Data Environment*, International Journal of Telemedicine & Applications. 2/9/2017, p1-9. 9p.
9. Mark Rothstein (2017), *Structural Challenges of Precision Medicine: Currents in Contemporary Bioethics*, Journal of Law, Medicine & Ethics. Jun2017, Vol. 45 Issue 2, p274-279. 6p.
10. Mendonc, M., Poleto, T., Silva, L., De Gusamo, A. and Costa, A. (2016), *A Grey Theory Based Approach to Big Data Risk Management Using FMEA*, Mathematical Problems in Engineering. 8/28/2016, p1-15. 15p.
11. Houser, K. and Sanders, D. (2017), *The Use of Big Data Analytics by the IRS: Efficient Solutions or the End of Privacy as We Know It*, Vanderbilt Journal of Entertainment & Technology Law. Summer2017, Vol. 19 Issue 4, p817-872. 56p.
12. Nazar, N. and Senthilkumar, R. (2017), *An Online Approach or Feature Selecion for classification of big data*, Turkish Journal of Electrical Engineering & Computer Sciences. 2017, Vol. 25 Issue 1, p163-171. 9p.
13. Noureddine Elouazizi (2014), *Critical Factors in Data Governance for Learning Analytics*, Journal of Learning Analytics, v1 n3 p211-222 2014. 12 pp.
14. Young, A. and McConkey, K. (2012), *Data Governance and Data Quality: Is It on Your Agenda*, Journal of Institutional Research, v17 n1 p69-77 Oct 2012. 9 pp.
15. Sara Rosenbaum (2010), *Data Governance and Stewardship Designing Data Stewardship Entities and Advancing Data Access*, Health Services Research. Oct2010, Vol. 45 Issue 5p2, p1442-1455. 14p.
16. Judith Lamont (2017), *Governance: a mandate for the data-driven enterprise*, KM World. Jan2017, Vol. 26 Issue 1, p28-30. 3p. 1 Color Photograph.
17. Salido, J. and Voon, P. (2010), *A Guide to Data Governance for Privacy, Confidentiality, and Compliance*, IAPP [Online] available from: https://iapp.org/media/pdf/knowledge_center/Guide_to_Data_Governance_Part1_The_Case_for_Data_Governance_whitepaper.pdf
18. John H. Holmes (2016), *Privacy, Security, and Patient Engagement: The Changing Health Data Governance Landscape*, eGEMS (Generating Evidence & Methods to Improve Patient Outcomes). 2016, Vol. 4 Issue 2, p1-4. 4p.
19. Carolyn Petersen (2016), *The Future of Patient Engagement in the Governance of Shared Data*, eGEMS (Generating Evidence & Methods to Improve Patient Outcomes). 2016, Vol. 4 Issue 2, p1-7. 7p.
20. R.D McDowall (2017, p.32), *Understanding Data Governance, Part I*, Spectroscopy. Feb2017, Vol. 32 Issue 2, p32-38. 7p.
21. Allen, C., Des Jardins, T., Heider, A., Lyman, K., McWilliams, L., Rein, A., Schachter, A., Singh, R., Sorondo, B., Topper, J. and Turske, S. (2014), *Data Governance and Data Sharing Agreements for Community-Wide Health Information Exchange Lessons from the Beacon Communities*, *e*GEMS (Generating Evidence & Methods to Improve Patient Outcomes) 2014, Vol. 2 Issue 1, p1-9. 9p.
22. Son, J., Kim, J., Na, H. and Baik, D. (2016), *Dynamic access control model for privacy preserving personalized healthcare in cloud environment*, Technology & Health Care. 2016 Supplement1, Vol. 24, pS123-S129. 7p.
23. Maureen DeAngles (2015), *National electronic health record network regulation and synchronization of national and state privacy laws to increase efficiency and reduce costs in healthcare*, Journal of Legal Medicine, 36:413–419.
24. Allaert, F., Mazen, N., Legrand, L. and Quantin, C. (2017), *The tidal waves of connected health devices with healthcare applications consequences on privacy and care management in European healthcare systems*, BMC Medical Informatics & Decision Making. 1/17/2017, Vol. 17, p1-6. 6p.
25. Soon, K., Jong Mo, K., Deok, S., Gwang, H. and Yoon, S. (2014), *Privacy Protection for Personal Health Device Communication and Healthcare Building Applications*, Journal of Applied Mathematics. 2014, p1-5. 5p.
26. Ira S. Rubinstein (2012), *Big Data: The End of Privacy or a New Beginning*, New York University Public Law and Legal Theory Working Papers. Paper 357 available from: http://lsr.nellco.org/nyu_plltwp/357
27. Xin, F., Wojak, A., Neagu, D., Ridley, M. and Travis, K. (2011), *Data governance in predictive toxicology: A review*, Journal of Cheminformatics [online]. Available from: https://jcheminf.springeropen.com/articles/10.1186/1758-2946-3-24
28. Broeders D., Schrijvers, E., Van der Sloot, B., van Barkel, R., de Hoog, J. and Ballin, E. (2017), *Big Data and security policies: Towards a framework for regulating the phases of analytics and use of Big Data*, Computer Law & Security Review, Vol. 33 (3): 309-323
29. Michael Mattioli (2014), *Disclosing Big Data*, Minnesota Law Review. 2014 Vol. 99 Issue 2, p535-583. 49p. available from: http://www.minnesotalawreview.org/wp-content/uploads/2015/02/REVISEDMattioli_MLR1.pdf
30. Eric Everson (2016), *PRIVACY BY DESIGN: TAKING CTRL OF BIG DATA*, Cleveland State Law Review. 2017, Vol. 65 Issue 1, p27-43. 17p.
31. Bryant Bell (2014), *Where Big Data and Information Governance Meet*, KM World. Sep2014 Supplement, pS5-S5. 1p.
32. Ronald J. Krotoszynski (2015), *Reconciling Privacy and Speech in the Era of Big Data: A Comparative Legal Analysis, 56 Wm. & Mary L. Rev. 1279*, available from: http://scholarship.law.wm.edu/wmlr/vol56/iss4/8
33. UN World Health Organization (2006), *Health System Profile Lebanon*, WHO report [online], available from: http://apps.who.int/medicinedocs/documents/s17301e/s17301e.pdf

34. Carol McDonald (2017), *5 Big Data Trends in Healthcare for 2017*, MAPR report [online], available from: https://mapr.com/blog/5-big-data-trends-healthcare-2017
35. Stephen E. Arnold (2015), *Big begets big: The information governance challenge*, KM World. Jul/Aug2015, Vol. 24 Issue 7, p1-2. 2p.
36. Claps, M. and O'Brien, A. (2017), *The Strategic Value of Big Data and Analytics in the Public Sector*, IDC report [online], available from: https://www.sap.com/documents/2017/05/083593b6-ba7c-0010-82c7-eda71af511fa.html
37. Eric Schadt (2015), *The role of big data in medicine* [online], available from: http://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/the-role-of-big-data-in-medicine
38. Nancy Davis Kho (2017), *The State of Big Data*, EContent. Jan/Feb2017, Vol. 40 Issue 1, p10-12. 3p.
39. Giacalone, M. and Scippacercola, S. (2016), *Big Data: Issues and an Overview in Some Strategic Sectors*, Journal of Applied Quantitative Methods. Fall2016, Vol. 11 Issue 3, p1-17. 17p. 4 Diagrams, 3 Charts.
40. Shabani, M., Dyke, S., Joly, Y. and Borry, P. (2015), *Controlled Access under Review: Improving the Governance of Genomic Data Access*, PLoS Biology. 12/31/2015, Vol. 13 Issue 12, p1-6. 6p.
41. Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S., Goodman, A., Hollander, R., Koenig, B., Metcalf, J., Narayanan, A., Nelson, A. and Pasquale, F (2017), *Ten simple rules for responsible big data research*, PLoS Comput Biol 13(3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399
42. Trotter, F. and Uhlman, D. (2013), *Hacking Healthcare*, Publisher: O'Reilly Media Inc.
43. Hernandez, S., Inserro, J., Kelly, P., Monsees and M., Orlove, J. (2014), *Official (ISC)2 Guide to the HCISPP CBK*, Publisher: CRC Press Shon Harris (2013), *CISSP All-in-One Exam Guide 6th Edition*, Publisher: McGraw Hill Education