



JÖNKÖPING UNIVERSITY

*Jönköping International
Business School*

Opportunities and challenges of Big Data Analytics in healthcare

An exploratory study on the adoption of big data analytics in the Management of Sickle Cell Anaemia.

MASTER THESIS WITHIN: *Informatics*

NUMBER OF CREDITS: 30

PROGRAMME OF STUDY: *IT, Management and Innovation*

AUTHOR: *Betty Saenyi*

JÖNKÖPING November 2018

Master Thesis in Informatics

Title: Opportunities and challenges of Big Data Analytics in healthcare
Authors: Betty Saenyi
Tutor: Osama Mansour
Date: 2018-11-30

Key terms: Big data, Analytics, Sickle cell Anaemia, Healthcare,

Background: With increasing technological advancements, healthcare providers are adopting electronic health records (EHRs) and new health information technology systems. Consequently, data from these systems is accumulating at a faster rate creating a need for more robust ways of capturing, storing and processing the data. Big data analytics is used in extracting insight form such large amounts of medical data and is increasingly becoming a valuable practice for healthcare organisations. Could these strategies be applied in disease management? Especially in rare conditions like Sickle Cell Disease (SCD)? The study answers the following research questions;

1. What Data Management practices are used in Sickle Cell Anaemia management?
2. What areas in the management of sickle cell anaemia could benefit from use of big data Analytics?
3. What are the challenges of applying big data analytics in the management of sickle cell anaemia?

Purpose: The purpose of this research was to serve as pre-study in establishing the opportunities and challenges of applying big data analytics in the management of SCD

Method: The study adopted both deductive and inductive approaches. Data was collected through interviews based on a framework which was modified specifically for this study. It was then inductively analysed to answer the research questions.

Conclusion: Although there is a lot of potential for big data analytics in SCD in areas like population health management, evidence-based medicine and personalised care, its adoption is not a surety. This is because of lack of interoperability between the existing systems and strenuous legal compliant processes in data acquisition

Acknowledgement

To my family and friends, I owe it all to you! Thank you for your unwavering support and encouragement through it all.

I am grateful to my supervisor Osama Mansoor for his guidance during my writing, his constant feedback and most especially his criticism that helped shape my thesis. To Prof. Christina Keller for being such a rock and sounding board, thank you!

I would also like to pass my gratitude to my interviewees Dr. Susan Paulanokis, Dr. Susan Murumba, Dr. Mary Hullihan, Dr. Tom Williams, Dr. Nirmish Shah, Dr. Sheriff Badawy and Dr. Jane Hankins who graciously gave me their insights on data management in Sickle cell anaemia.

To the Swedish institute for giving me the scholarship and the opportunity to undertake my master studies in Sweden, tack så mycket!

Last and most important, to the sickle cell warriors whose fight inspired the writing of this thesis, keep fighting on. To Dan and Khanjila, your fighting spirits live on.

Table of contents

1	Introduction	1
1.1	Background	1
1.2	Problem.....	3
1.3	Purpose	4
1.4	Research Questions	5
1.5	Delimitations of the study.....	5
1.6	Definitions.....	5
2	Literature Review	7
2.1	Big data analytics	7
2.2	Features of Big data	8
2.3	Potential of Big Data Analytics	9
2.4	Big Data Analytics in Health Care.....	10
2.4.1	Features of healthcare Big Data.....	11
2.4.2	Big data Set-up in healthcare	14
2.4.3	Big Data Potential in Healthcare	16
2.5	Sickle Cell Anaemia	17
2.5.1	Prevalence of Sickle Cell Anaemia.....	17
2.5.2	Socio-Economic and Clinical Impacts of Sickle Cell-Anaemia	18
2.5.3	Big Data in Managing Sickle Cell Anaemia	19
3	Theoretical framework for the Study	21
3.1	Big Data Theory model	21
3.2	Resource based view-Big Data (RBV-BD)theory.....	22
3.3	Modified big data framework.....	24
4	Methodology	27
4.1	Research approach.....	27
4.2	Research Design.....	28
4.2.1	Research purpose	28
4.2.2	Research method	29
4.2.3	Research strategy	29

4.3	Data collection	30
4.3.1	Sampling process	30
4.3.2	Primary Data collection.....	32
4.3.3	Secondary data Collection	34
4.4	Data analysis.....	36
4.5	Qualitative validity	37
5	Results.....	38
5.1	Mary Hullahan- Centre for Disease Control (CDC), Atlanta Georgia	38
5.1.1	Causal Conditions for the CDC data collection projects	39
5.1.2	Strategic implementation of the CDC projects	39
5.1.3	Big Data enabled- Capabilities.....	40
5.2	Susan Paulukonis-California Sickle Cell Disease Longitudinal Data Collection Project (SCDC)	41
5.2.1	Causal conditions for SCDC project in California	42
5.2.2	Big data Context for the SCDC project in California	43
5.2.3	Strategy for the SCDC California project	44
5.2.4	BD Enabled capabilities for the SCDC project.....	46
5.3	Susan Murumba, Kory Family Hospital, Bungoma, Kenya	49
5.3.1	Big data context in Kory Sickle Cell Support group.	49
5.4	Dr. Nirmish Shah, Principal Investigator TRU-Pain App and director of Sickle Cell Transition programme.....	51
5.4.1	Causal Conditions for TRU-Pain App and the Sickle cell Registry.....	51
5.4.2	Big data context for the TRU-Pain App and the SCD Registry	52
5.4.3	Strategy for the TRU-Pain App.....	54
5.4.4	BD Enabled capabilities for TRU-Pain App	54
5.5	Dr. Sheriff Badawy and Dr. Jane Hankins, Principal Investigators Hydroxyurea Adherence App.	57
5.5.1	Causal Conditions for Hydroxyurea Adherence app.	57
5.5.2	Big data context and Strategy for the Hydroxyurea App	58
5.6	Tom Williams, Kenya Medical Research Institute (KEMRI) Wellcome Trust	59
6	Analysis.....	62
6.1	Deductive analysis.....	62

6.1.1	Causal Conditions	62
6.1.2	Context	63
6.1.3	Big data.....	63
6.1.4	Strategy.....	64
6.1.5	Big data-enabled capabilities	65
6.2	Inductive analysis	66
6.2.1	Possible opportunities for big data analytics in SCD management.....	66
6.2.2	Challenges facing the adoption of big data analytics in SCD management	70
7	Conclusion.....	74
8	Discussion.....	76
8.1	Results Discussion.....	76
8.2	Methods discussion	78
8.3	Implications to research and Practice	78
8.4	Future recommendation.....	78
9	References	80

Figures

Figure 1	The 4Vs of big data.....	9
Figure 2	File split process in Hadoop.....	14
Figure 3	Conceptual Framework of big data in healthcare	15
Figure 4	The potential of big data analytics in healthcare	17
Figure 5	Prevalence of Sickle Anaemia.	18
Figure 6	Big Data descriptive model.....	22
Figure 7	Big data conceptual research framework.....	23
Figure 8	Modified Big Data framework.....	24

Figure 9 California Newborn Screening Identified SCD Births, 2004-2008 47

Tables

Table 1 Features of big data 8
Table 2 Interview questions guide 32
Table 3 Interview Schedule..... 33
Table 4 Secondary data 34

1 Introduction

This chapter examines the background of the research, the objectives and purpose of the study. Additionally, it discusses the research problem and its limitations. The study questions as well as the definition of key terms are also presented.

1.1 Background

Big data analytics (BDA) has recently become a popular topic. Adams et al. (2009); Sivaraman, Kamal, Irani and Weerakkody (2017) state that a mention of the big data analytics topic evokes diverse reactions from persons with solid data analysis skills and lay people alike. The authors argue that this growing interest is based on the reality that big data analytics has been labelled as the de facto panacea for many data management challenges facing a wide range of sectors. On their part, Baseman, Revere and Painter (2017) give an example of the healthcare sector where huge investments have been made as organisations attempt to create capabilities such as centralised or decentralised databases that can be mined to create rich and refined set of information. McAfee and Brynjolfsson (2012) wade in the discussion by showing that such databases and data mining capabilities could capture, analyse and track crucial health data touching on areas such as patients' health history, medical supplies trends, disease prevention trends, and effectiveness of disease treatment plans.

Such rich data should be manipulated in a manner that lays bare major trends, themes, insights, and correlations in a faster and more efficient manner. Henke, Libarikian, and Wiseman (2016) show that while data manipulation and interpretation is not a new endeavour in most sectors including the health industry, the elements of volume, variety, and velocity (3Vs) are unique to the big data analytics field. Moreover, some researchers have further presented veracity or 'data assurance' as the fourth element of big data analytics (Raghupathi & Raghupathi, 2014). Whilst the above examples and arguments do not extensively cover the application of big data analytics in the health sector, Lee and Yoon (2017) give an insight on how health organisations can positively manipulate large volume of raw data to create actionable information in real-time. Indeed, Kruse,

Goswamy, Raval, and Marawi (2016) argue that big data analytics is becoming a popular investment in the field of health, with more governments and private sector players dedicating huge sums of capital towards setting up sophisticated data analytics systems that can gather, analyse, interpret and report data on their own.

But why is big data analytics in healthcare becoming so popular? Belle et al. (2015) point out that because of digitalization and advanced technologies, huge amounts of heterogenous data from different sources like hospitals, insurers, pharmaceuticals, researchers and government agencies have become accessible. The authors however, state that this data is siloed. Insight generated from integrating such different types of data would facilitate the design of programmes that would result in improved patient outcomes, and possibly reduce incidences of chronic diseases (Task Force 7 Health subgroup [TF7.SG3], 2016). Henke et al. (2016) show that big data analytics tools such as NoSQL, YARN, and Hadoop and actions such as statistical algorithms, what-if analyses, and predictive modelling help organisations to draw meaning from this rich data in ways that traditional statistical methods are not able to.

This promise of improving healthcare through generating insights from data has seen organisations like the American Society of Haematology (ASH) launch its vision for chronic and genetic haematologic big data (The Hematologist, 2017). ASH's main goal is to build an all-inclusive knowledge base and create a platform for information exchange in rare haematologic diseases like Sickle Cell Anaemia and Multiple Myeloma which have a high impact on healthcare. During its 59th Assembly, the World Health Organization (WHO) reported that 5% of the world's population is affected by Sickle cell Anaemia (World Health Organization [WHO], 2006). Therefore, big data analytics could be leveraged on data generated from Sickle Cell Disease (SCD) stakeholders including patients and healthcare providers towards improving its management. This study will be looking at opportunities that exist for SCD, a rare condition.

Despite its growing adoption, big data analytics is prone to several challenges. Luna, Mayan, García, Almerares, and Househ, (2014) warn that big data analytics is prone to at least three core challenges. The authors point out that the first challenge has to do with the structure and accessibility of raw data – most raw data captured by organisations is

usually scattered in several silos which are sometimes hard to consolidate and integrate. Further, the authors show that most organisations lack clear business cases to guide in the process of harnessing raw data using big data analytics. Belle et al., (2015) wade in to support the authors third challenge by showing that organisations experience a lack of robust coordination among big data analytics teams when attempting to manipulate and interpret raw data in their custody.

In their study to explore the opportunities and challenges that may emerge when applying big data analytics in the health sector, Kruse et al., (2016) found that organisations are likely to experience difficulties in the areas of “... data structure, security, data standardization, storage and transfers, and managerial skills such as data governance” (p.38). At the same time however, the authors found that opportunities manifest in the form of “... quality improvement, population management and health, early detection of disease, data quality, structure, and accessibility, improved decision making, and cost reduction” (p. 38). These findings echo those of TF7.SG3 (2016) to the effect that health sector organisations must be willing to overcome challenges to optimise the benefits and opportunities that come with big data analytics. It is therefore not surprising that the adoption of big data approaches in the healthcare sector is still low (TF7.SG3, 2016).

1.2 Problem

As established in section 1.0 above, several studies have been undertaken on the application of big data analytics in the health sector. See for example, Baseman *et al.* (2017); Belle *et al.* (2015); Caban and Gotz (2015); Gaitanou, Garoufallou and Balatsoukas (2014); Lee and Yoon (2017); Kruse *et al.* (2016). While these studies offer crucial hindsight, insights and foresights on the opportunities and challenges that exist in the application of big data analytics in managing rare health conditions (Caban & Gotz, 2015; (TF7.SG3, 2016), their scope and focus is too general to be applicable to the case of Sickle Cell anaemia. Moreover, the findings of these studies highly focus on the opportunities of big data analytics with little emphasis on the challenges of such frameworks (Lee & Yoon, 2017).

Yet reality shows that there are many challenges hindering the successful application of big data analytics in the health sector especially in rare genetic diseases like SCD especially in developing countries. For example, the researcher for this study established that there is no proper management of sickle cell related information in Kenya when she volunteered with the country's National Sickle Cell Foundation. Specifically, the researcher discovered that there are many sickle cell organisations in the country playing almost the same or complementary roles. Interestingly, it emerged that these organisations keep haphazard records, do not share information, and run their affairs independently. This realisation persuaded the researcher to arrive at the conclusion that a proper information management system that harnesses the gains of big data analytics could help to collect, analyse and interpret data for sickle cell anaemia trends.

In Geneva 2006, the WHO made several resolutions on the management of Sickle Cell Anaemia (WHO, 2006). Key among them was to have the national governments of the high prevalent regions adopt national policies and to systematically gather information on the most cost-effective approaches for prevention and treatment. So far, very scanty details exist on public and private bodies exploiting such information using modern data management approaches such as big data analytic tools to manage this condition. This creates a huge policy and literary gap that should be filled by a comprehensive research study.

The proposed study will highlight the opportunities for managing sickle cell anaemia opened up by big data analytics as well as explore the challenges that could be hindering its adoption. While narrowing the scope of the study on SCD, the study will explore whether there are any such big data projects focused on SCD and if not, if there are pre-existing conditions that will make it possible for the implementation of big data analytics. As evidence adduced by TF7.SG3 (2016) shows, the study will lead to crucial hindsight, insight, and foresight for managing sickle cell anaemia.

1.3 Purpose

The aim of this study is to serve as a pre-study in establishing the opportunities and challenges of applying big data analytics in the management of SCD. This broad aim

forms the basis of the research questions. However, the ultimate purpose is to contribute knowledge towards the possible options in improving patient care and management in Sickle cell anaemia.

1.4 Research Questions

4. What Data Management practices are used in Sickle Cell Anaemia management?
5. What areas in the management of sickle cell anaemia could benefit from use of big data Analytics?
6. What are the challenges of applying big data analytics in the management of sickle cell anaemia?

1.5 Delimitations of the study

While this study looks at opportunities and challenges of big data analytics in healthcare, it'll take a narrow stance and only focus on their application in sickle cell Anaemia management. This means that its findings will be limited to SCD.

Secondly, the field of health analytics is quite broad and is categorised into business and clinical analytics. Business analytics deals with the business side of healthcare while clinical analytics deals with patient care. This study, however, only focuses on clinical analytics. Additionally, the study will employ a purposive sampling approach, meaning it will only use the expert opinions of persons working in sickle cell affiliated organisations.

Lastly, the research will only be carried out as a pre-study and will be limited to the exploration of the analytic frameworks without implementing any system nor providing a detailed implementation process.

1.6 Definitions

Sickle cell anaemia:

“also known as sickle-cell disorder or sickle-cell disease is a common genetic condition due to a haemoglobin disorder – inheritance of mutant haemoglobin genes from both parents”. It has high

morbidity and mortality rates especially among sub-Saharan countries.

Sickle cell anaemia management: this refers to the activities involved in the research, monitoring, prevention, and treatment of sickle cell anaemia.

Big Data Analytics: this term defines the collection, analysis and interpretation of huge volume of dynamic and varied data that is updating in a high velocity.

Disease management: this term relates to all activities involved in the research, monitoring, prevention, and treatment of diseases such as sickle cell anaemia.

Healthcare: This represents the maintenance as well as the promotion of health through diagnosis, treatment of illness and prevention in humans. Healthcare professionals in various health niches provide healthcare.

Hadoop: This represents software utilities that are open-source in nature and are used with various computer networks to solve issues with massive huge data and information computations

Data Management *“the method of recording, organizing, and storing information; for instance, handwritten or typed on alphabetically arranged charts, or direct keyboard entry into a computer.”* (Last, 2007, p.59).

2 Literature Review

The purpose of this chapter is to examine various literature of scholarly works by other researchers on big data analytics and Sickle cell Anaemia as well as offer a significant theoretical background to the discussion on the adoption of big data analytics in healthcare and more specially in the management of Sickle Cell disease.

2.1 Big data analytics

Big data analytics refers to the process of studying large sets of different data to find out the hidden patterns, trends in the market, people's preferences and other critical information in organizations for quality decision-making (Pouyanfar, Yang, Chen, Shyu, & Iyengar, 2018). With the increased aspect of technological innovation today, convectional database management systems are ineffective in managing huge data. The convectional software tools do not have the ability to capture or process the data in order to store and manage them in human time (Kubick, 2012). This aspect makes the entire process tedious and quite challenging. According to Kubick (2012), the size of the big data currently ranges from some dozens of Terabytes to Petabytes per data set and they are continuously increasing. Thus, it is hard to visualize the data. Besides, the analytics of the data has become quite challenging with the implementation of the traditional frameworks (Baseman et al., 2017). Similarly, the traditional methods and techniques have made it difficult to store, share, search and capture the data. The reason for huge data is that enterprises are currently gathering key user details to generate data thus leading to huge volumes of data being generated (Baseman et al., 2017). Russom (2011) further points out that these organizations intend to analyse the data to discover new facts. The analysis of big data therefore requires advanced techniques that can easily manipulate the huge data sets. These frameworks allow businesses to sample large data and evaluate the same to derive different information that facilitates the sustainable operation of the business. Since the big data is complex, Baseman *et al.* (2017) posited that real-time analysis is likely to help generate significant information.

2.2 Features of Big data

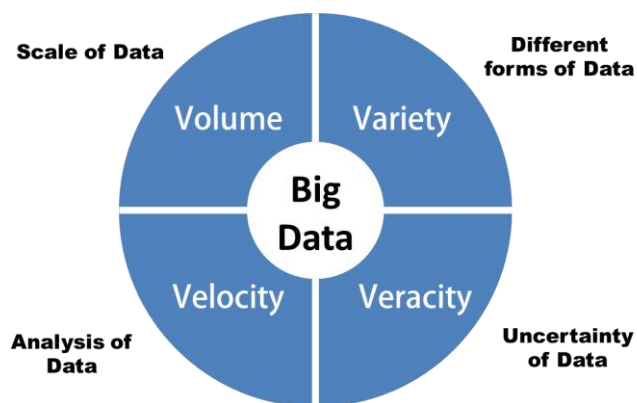
According to Gaitanou et al., (2014), big data is characterised by diversities and scales. The underlying architecture, tools and analytics need to respect the timely generation of the information in order to produce useful information for the end-users. This meaningful information creates significant business values. Table 1 below highlights the Vs that are the distinguishing features of big data including veracity as the fourth element. Although this concept remains as an objective and not a reality, Raghupathi and Raghupathi (2014) argue that veracity of data usually influences critical decisions.

Table 1 Features of big data

No	V Feature	Description
1	Volume	Advanced communications technologies such as Smartphones and social networks have led to the generation of large volumes of data from different devices and applications. This data is now in Petabytes and is growing every second. It is quantified using sophisticated data sets using metrics of the order of TB or PB. (McAfee & Brynjolfsson, 2012)
2	Variety	The data being generated is varied in terms of data type and application as well as the analysis framework (TechAmerica Foundation, 2012). The source of the data generated also varies and could take the form of videos, comments or documents. This could take the form of Structured data which can be stored in a traditional row-column database or unstructured data which cannot reside in such a database (McAfee & Brynjolfsson, 2012).
3	Velocity	Data is created at high speeds and is supposed to be analysed in good time to gain meaningful insight from it. For some applications, the speed of generating the data is more vital than its volume. Furthermore, businesses gain a competitive advantage when they

		have access to real time or near real-time information (Oussous, Benjelloun, Ait Lahcen, & Belfkih, 2017)
4	Veracity:	This aspect indicates the quality of data, it shows whether the data is incomplete, approximate, deceptive, ambiguous, inconsistent, latent or active. Therefore, for data to be meaningful, it must be from a reliable source, accurate, and analysed within its context (TechAmerica Foundation, 2012).

Figure 1 The 4Vs of big data



Source: Adapted from Zanabria and Mlokozi (2018)

2.3 Potential of Big Data Analytics

While considering the application of big data analytics across different sectors like the manufacturing, retail, healthcare, telecommunications and the public sector, Mckinsey's global survey revealed that Big Data is an essential concept in understanding the productivity, competition and innovation of a business or organisation (McKinsey Global Institute, 2011). Gaitanou et al. (2014) further argue that by integrating, digitising and utilising big data, various companies ranging from small businesses to multi-national firms as well as multi-providers to huge healthcare providers stand to gain significant advantages.

In the health sector alone, McKinsey estimated that Big Data analytics have allowed over \$300 billion in healthcare savings annually in the United States with research and design as well as clinical operations representing the largest niches for potential savings with \$108 and \$165 billion in waste (McKinsey Global Institute, 2011).

With the increasing amount of data that is being created and stored across the globe, there is therefore a lot of potential to garner key insights from this information using big data analytics in various sectors such as marketing, automation, defence and healthcare. Application of big data analytics in healthcare will be discussed further in the next sub-topic as the focus of this research.

2.4 Big Data Analytics in Health Care

With the increasing adoption of Electronic health records (EHRs) and patient's monitoring systems, there has been a continuous flow and pile up of large volumes of data and physiological data that calls for mining and analysis (Simpao, Ahumada, & Rehman, 2015). The authors further point out that this rising acknowledgement of the potential of big data in healthcare has created an interest in collecting and pooling EHR and other patient related data in national databases which provide information on rare diseases that would otherwise have been difficult to analyse without huge sample sizes. Raghupathi and Raghupathi (2014) also state that as a move to comply with government regulations or to simply improve their healthcare delivery, health providers have accumulated large amounts of data while digitising their records, and that this data has the potential of being used in clinical decision support, management of population health and many other functions.

Studies have also pointed out that the volume of data and information in health care is likely to increase with time as technology is incorporated to facilitate healthcare performances via the utilisation of significant and relevant information within the healthcare sectors (Gaitanou et al., 2014; Kruse et al., 2016). This information is being accumulated from different sources and in her book, "The patient revolution", Tailor (2016) identifies and categorises the sources of patient data as below;

- i. Clinical data which includes among others structured EHR records, unstructured clinical notes, medical images, videos and audio recordings.
- ii. Active or passive self-generated data through patient monitoring systems and social media
- iii. Patient satisfaction and patient-reported outcomes data through surveys
- iv. Medical claims data.

Jeba and Srividhya (2016) further categorises another important source of healthcare data as research and development which encompasses data from genomics, DNA and clinical trials. These sources have led to increased availability of large amounts of data which calls for an efficient database that can integrate the data to adopt to the increased evolution of information (Gaitanou et al., 2014). Zastrow (2015) agrees with this argument by stating that the most critical aspect of handling and managing data is to establish how and where the data is stored after it is collected. However, as Gaitanou et al. (2014) points out, the traditional frameworks of retrieving and storing data are not currently efficient because they function on relational databases which cannot handle the varied nature of healthcare data. This structured and semi-structured data can be efficiently analysed by Big data analytics systems (Wang, Kung, & Byrd, 2018).

2.4.1 Features of healthcare Big Data.

As is the case in other fields, the 4V's of big data are applicable in analysing healthcare data;

Volume

The availability of large volumes of medical data that is already highly categorised needs advanced management systems. These data volumes are available in large varieties from medical records to submissions on clinical trials (Sivarajah et al., 2017). Feldman, Martin, and Skotnes (2012) also state that emerging forms of big data such as the biometric sensor reading, the 3D imaging and the genomics have spearheaded the growth of data management techniques which have made it easy to handle data. Furthermore, they point out that virtualization of data and cloud computing is an element that has made it possible

for development of more effective methods of manipulating and storing large amounts of data.

Velocity

Feldman et al. (2012) argue that the shift from traditional manipulation of medical data inherently based on paper is a challenge especially since it deals with the analysis of data that is generated in real time and at increased turnover rates than before as well as in different unexpected and increased speeds. Developing of advanced platforms to capture and store this data effectively as well as the ability to retrieve and analyse the data with the aim of making medical decisions based on the findings have also been significantly improved (Gaitanou et al., 2014).

Real time data handling such as bed heart monitors, trauma monitors for blood pressure and operating room anaesthesia monitors needs to be highly monitored and handled effectively. This is because they are likely to cause fatal outcomes including death cases among patients (Pouyanfar et al., 2018). For instance, the already existing real time monitors in the ICU rooms are likely to continue helping limit life threatening infections at their early stages (Institute for Health technology Transformation [IHTT], 2013). The possibility to rapidly analyse real time data is likely to spur revolution in the healthcare sector and help apply the most effective treatment options (Feldman et al., 2012).

Variety

Evolution of health data means that the data can no longer be analysed exclusively in electronic health records since larger types of data are available in categories such as structured and unstructured as well as semi-structured data. This aspect means that the analytic techniques have enhanced the evolution of health information. The main challenging yet interesting aspect of health data is its availability in largely varietal forms such as in multimedia formats (Feldman et al., 2012).

Structured data is the data that can easily be retrieved, integrated and stored by machines with the aim of manipulating it to produce actionable information from the data (Sivarajah et al., 2017). Structured data has historically been derived in the form of nurses' and doctors' notes, MRI and radiograph films as well as CT scans and other images. Feldman et al. (2012) however, points out that the need for field-coding information at this point of care for electronic management represents significant challenges to adopting EMRs by the doctors and nurses as they lose the normal comfort of language as well as understanding provided by handwritten notes. On the contrary, a significant number of

providers do agree that digital entries can decrease prescription errors as opposed to handwritten notes. The big data capability in the healthcare lies largely in combining the old data with modern data systems at both personal and population levels (Lee & Yoon, 2017).

Currently, there are sets of data gathered from various sources that support more rapid and reliable findings. For instance, pharmaceutical developers are likely to incorporate significant clinical data groups with genomics statistics. This development in return could help the developers in gaining approvals on improved drug treatments more efficiently and more significantly, expedite supply to the correct patients. The forecasts for all the healthcare areas are unlimited (Feldman et al., 2012).

Veracity

(Raghupathi and Raghupathi (2014) point out that problems concerning data quality are of serious concern in healthcare for two key reasons. First, it involves life and death choices that greatly depend on using precise information and secondly, the quality of healthcare data is highly inconsistent particularly that of its unstructured data.

Veracity undertakes the instantaneous scaling up of developers' performances and system approaches as well as algorithms to match the challenges linked to big data. A typical data management framework therefore considers stored data as clean, certain and accurate, however, the veracity of healthcare-related data may still face various issues (Lee and Yoon, 2017).

Refining care management, avoiding mistakes, decreasing costs and improvements in drug care and efficiencies are determined by high-quality data (Raghupathi & Raghupathi, 2014). However, Feldman et al. (2012) points out that the variety and velocity of big data may impede the capability of cleansing information before the analysis and decision-making process thus magnifying the aspect of data reliance.

Accordingly, the 4Vs represents a suitable starting point for the debate concerning big data analytics in healthcare. As (TF7.SG3, 2016) argues, the effective utilisation and performance of such frameworks have increasingly emerged in the current healthcare sector. While profit is not the core motivator, it is critical for the healthcare organizations to obtain critical frameworks as well as mechanisms to effectively integrate big data. The implementation of such analytic frameworks to the wide category of data related to medical records and patient-oriented health has allowed for the in-depth understanding of

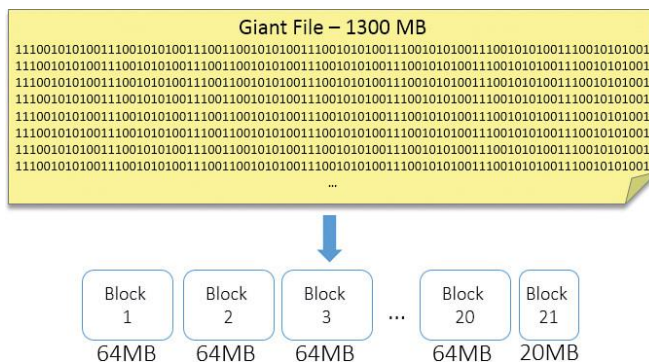
the results that when implemented at the point of care helps in informing the healthcare providers and is critical in the decision-making process for both patients and providers (Kruse et al., 2016).

2.4.2 Big data Set-up in healthcare

Raghupathi and Raghupathi (2014) state that the conceptual framework linked to healthcare big data analytics is the same as the conventional frameworks in analytic projects. According to the researchers, the most significant variation is observed in the manner in which data is processed. They argue that while evaluation in regular health analytic frameworks can be carried out with a single analytic mechanism integrated into a stand-alone machine, big data is processed and executed via multiple servers due to its large volume.

Additionally, the volume of data in big data is unpredictable and as such, its physical infrastructure is based on a distributed computing model where data is usually stored in various places and linkage is allowed through the networks, big data analytic tools and Apps as well as the use of a distributed file system (Biswas & Sen, 2016). Kumaraguru and Chakravarthy (2017) further state that open source platforms like Hadoop MapReduce have promoted the integration of big data analytics in healthcare as the sector taps into the available huge data sets to obtain insight with the aim of formulating objective decisions. According to Borkar, Carey and Li (2012), Hadoop can process large data sets of both structured and unstructured data by partitioning and allocating them to multiple servers which independently solve pieces of the problem and later assembles them for the final solution.

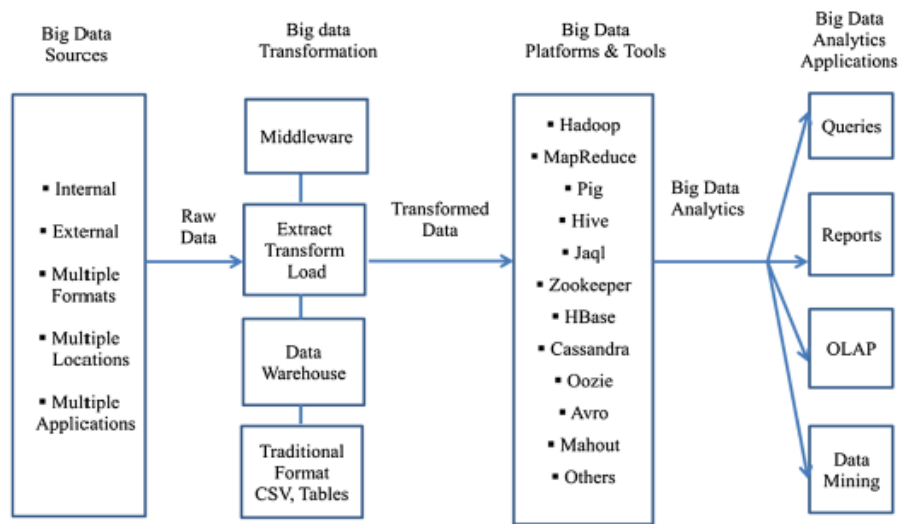
Figure 2 File split process in Hadoop



Source: Adapted from Singh and Ali, (2015)

However, since these frameworks have emerged in an ad hoc style and from open-source development frameworks, Kumaraguru and Chakravarthy (2017) points out that they are not user- friendly nor vendor supported, are complicated and require intensive skills and knowledge to manage. Raghupathi and Raghupathi (2014) further argue that the complexity within the big data analytics starts with the data as shown in Fig. 3 below.

Figure 3 Conceptual Framework of big data in healthcare



Source: Adapted from Raghupathi and Raghupathi (2014)

According to the authors, raw data is aggregated from different sources, formats and locations then processed. A few options are also available for the processing which could be through middleware web services or through data warehousing where data is not collected in real time but instead it is collected and warehoused for processing. After processing, choices on the appropriate big data platform and tools for the project are made and finally the visualisation of the big data analytics application is considered. This visualization could be through reports or Online analytical processing (OLAP).

Security is another vital aspect to be considered when considering a big data analytics infrastructure. According to Boja, Pocovnicu and Batagan (2012), it is crucial to secure data where big data becomes part of the workflow, for instance, big data applications may be very useful to identify changes that may come up in terms of patients' needs and such information needs to be secured to meet the patient's privacy requirements and compliance requirements as well. Moura and Serrão (2015) state that it is obligatory for

an organization to put in measures to ensure all legal requirements on handling data are addressed and that confidential healthcare data is encrypted, and access policies established. However, TechAmerica (2012) points out that for these measures to be effective, they must be transparent to the end user and at the same time not affect the performance and scalability of the systems.

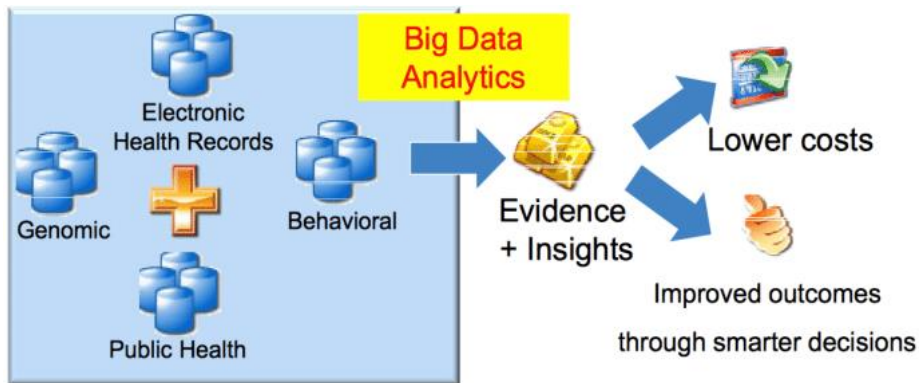
2.4.3 Big Data Potential in Healthcare

Healthcare organizations have benefited significantly from the use of big data. The benefits have been felt from the smallest of the single-physician clinics to the largest of the hospital networks and systems (Burghard, 2012). Kruse et al., (2016) further argues that many issues within the healthcare sector will likely to be greatly solved and others significantly optimized using big data analytics. In general, the objectives for big data analytics in health care as illustrated in fig 4 are to gain insight and provide precise and timely interventions to patients, provide personalised patient care and gain a competitive advantage for the healthcare providers (Khalid & Abdelwahab, 2016).

Areas in healthcare that could benefit from the application of big data analytics include among others;

- 1) Reduction of healthcare costs
- 2) Early detection of diseases
- 3) Research and development (R&D)
- 4) Specialized care
- 5) Managing population health and
- 6) Fraud detection

Figure 4 The potential of big data analytics in healthcare



Source: Adapted from Khalid and AbdelWahab (2016)

2.5 Sickle Cell Anaemia

As noted by Lopez, Cacoub, Macdougall, and Peyrin-Biroulet, (2016), Sickle Cell Anaemia is a blood disorder that is attributed to an inherited abnormal haemoglobin. This abnormal haemoglobin results to distorted red blood cells, which are fragile and can be ruptured easily. When the red blood cells are ruptured and decrease in number, the process results to anaemia. The irregularly shaped sickle cells have the potential to block the blood vessels resulting to organ and tissue damage and significant pain in patients.

Lopez et al. (2016) further explains that for this condition to be experienced, the sickle cell gene must be inherited from both parents. Therefore, children of two carrier parents have one in four chances of becoming anaemic, however when a child inherits only a single gene, the child becomes a carrier. It is important to note that a carrier does not experience similar impacts as individuals with Anaemia. Carriers usually have limited symptoms, however there have been reports of abrupt deaths and medical complications when carriers are subjected to extreme low oxygen conditions (Lervolino et al., 2011).

2.5.1 Prevalence of Sickle Cell Anaemia

The World Health Organization states that Sickle-cell anaemia is predominantly common amid people whose ancestors originated from sub-Saharan Africa, India, Saudi Arabia and Mediterranean countries (WHO, 2006). It estimates that 5% of the world's population

carries genes that cause haemoglobinopathies, and that 300,000 children are born every year with these disorders with 200,000 cases of sickle cell anaemia coming from Africa. According to (Amendah, Mukamah, Komba, Ndila, & Williams, 2013), this number is expected to increase to four hundred thousand in the coming decades. Although it might have its roots in the aforementioned countries, Lervolino et al. (2011) points out that migration raised its gene frequency in other continents. For instance, the forced immigration of African slaves to the Americas greatly increased its prevalence in America especially in regions with large African population.

Figure 5 Prevalence of Sickle Anaemia.



Source: Adapted from https://www.cdc.gov/ncbddd/sicklecell/documents/SickleCell_infographic_5_Facts.pdf

2.5.2 Socio-Economic and Clinical Impacts of Sickle Cell-Anaemia

Clinical impacts

In their research, (Neto, Lyra, Reis, & Goncalves, 2011) found that sickle cell anaemia is a serious illness that imposes profound effects to various individuals as well as key impacts to the health sector and quality of care for the affected people. When this illness is linked to another illness such as kidney or heart complications, it is likely to accelerate the severity of the illness through worsening the experienced symptoms. While the disease is manageable, most clinicians focus on other elements of the illness thus failing to acknowledge the need to cure the anaemia linked to the illness. Yardley-Jones, (1999)

further points out that a repeated series of vascular occlusions will lead to chronic complaints such as vision impairment, proliferative retinopathy and pulmonary fibrosis

Social and Economic Impacts

SCD raises the cost of accessing quality care for the individuals with the illness especially in developing nations (Choubey, Mishra, Soni, & Patra, 2016; Kubick, 2012). Huge claims evaluations have indicated the increased healthcare use as well as expenditures when sickle cell anaemia coexist with various key illnesses. The Centre for Disease Control (CDC) in the USA points out that in 2005, children with SCD under the Medicaid coverage spend approximately \$11,702 on medical expenses while those under private insurance spent about \$14,772 (CDC, 2018). Additionally, a study carried out by Amendah et al. (2013) in Rural Kenya puts the estimated annual medical expenses per patient at \$138 in 2010. This is quite an economic burden on people in this rural part of a developing nation.

Fatigue and other constraints linked to SCD have effects on the indirect cost for various employed people. The case is even worse among disabled individuals (Plessow et al., 2015). The indirect costs incorporate reduced productivity, disability payments as well as the cost of travelling to access the medical care. Furthermore, Yardley-Jones (1999) argues that frequent pain episodes and hospitalization could lead to psychological problems in patients.

2.5.3 Big Data in Managing Sickle Cell Anaemia

While there has been a decrease in mortality rate in children with SCD in developed countries, Sickle cell is still a high-risk disease to the survival of children across most developing nations and has been largely abandoned leading to high rate of childhood mortality (50% to 90%) (Grosse et al., 2011). Pilot programs are necessary to collect and analyse data to find out the outcomes among the affected child population. However, little has been done to quantify the public health problems and burden of Sickle cell disease. The state of California in the US has embarked on an ambitious project to carry out a population-based surveillance on Sickle cell Anaemia to analyse and measure outcomes (California Sickle cell Resources, 2018). This project will be among cases studied in this research.

In their research, (Baseman et al., 2017) provided an illustration of enterprises within the healthcare sectors that have employed huge investments with the attempt to creating capabilities such as centralised or decentralised databases that can be mined to create rich and refined set of information. Furthermore, several healthcare providers are adopting systems such as EHR in the management of SCD , for instance, the government of Chhattisgarh, India adopted an EMR system to manage large data captured in its SCD screening program (Choubey et al., 2016). The authors further posited that government could use the accumulated data in plotting the prevalence of SCD in all its 27 districts. This research will be looking into such opportunities as well as challenges that could be facing such projects. This will be done through analysing interviews and information gathered from SCD related projects or efforts that have been put in place to improve the management of SCD through technology and data management.

3 Theoretical framework for the Study

This chapter outlines the theoretical framework to be adopted in the study.

The big data research field is yet to be fully defined and most of the research done is use case driven and multi-disciplinary (Pospiech & Felden, 2013). Some big data studies especially in medicine and biology have failed to provide conceptual contexts to which they are applied and coupled with the rising interest in big data analytics, an impression has been created that such problems can be solved without predictable scientific methods of inquiry (Coveney, Dougherty, & Highfield, 2016). They however argue that it's crucial to use a theory as a guideline to any study for optimum efficiency in the collection of data to produce reliable results.

In an attempt to create a holistic theoretical framework for Big Data research, several studies have been done so far. For instance, Sanyal, Bhadra, and Das, (2016) proposes a conceptual framework to analyse ideas and value derived from the use of big data approaches. The framework does not however, provide a guide for determining benefits or value derived from the use of big data analytics. For this reason, it shall not be used in this study, instead Pospiech and Felden's (2013) big data theory model which outlines value derivation through definition of big data constructs, shall be adopted.

3.1 Big Data Theory model

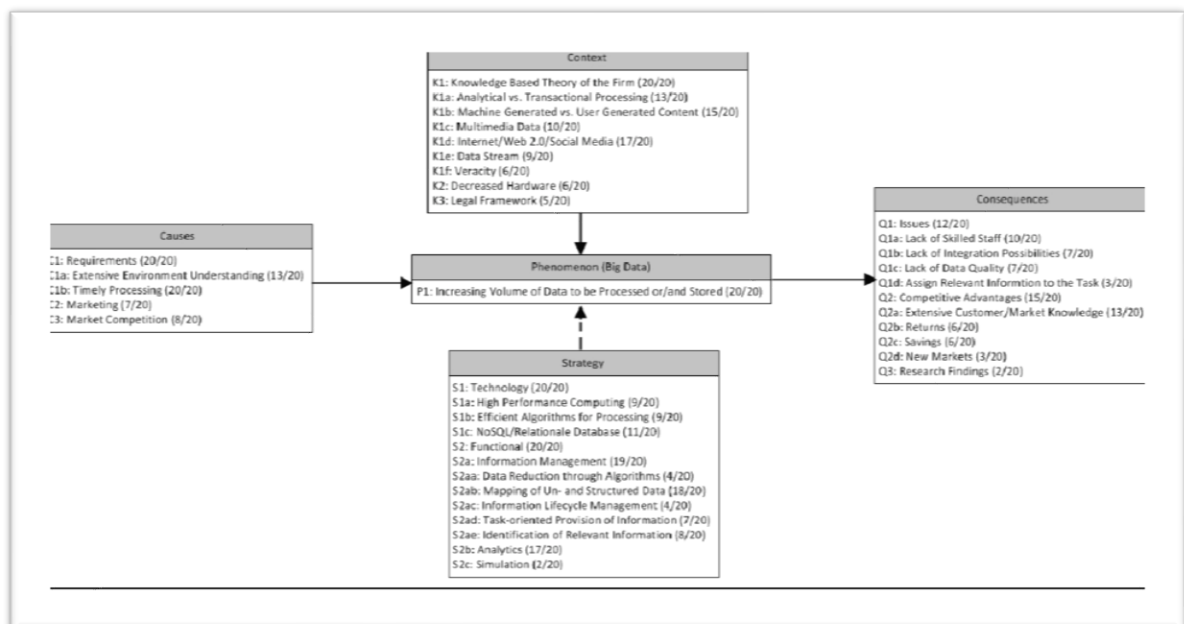
Using Grounded theory, Pospiech and Felden, (2013) undertook a study to conceptualise a descriptive big data model. They conducted and transcribed expert interviews which were used as a basis for a grounded theory design. Five themes emerged which were categorised into “cause, context, phenomenon, strategy and consequences” until a theoretical saturation was attained. In their subsequent paper Pospiech and Felden, (2015), they quantitatively proved the theory to give rise to a Big data theory model that has been used to;

- 1) Show the inherent characteristics of Big data such as volume, variety and velocity,
- 2) Illustrate how big data strategies are set
- 3) Show the value of using big data in businesses.

Fig 4 shows the five constructs that were identified by (Pospiech & Felden, 2013) as being the key concepts in big data and the relationships between these concepts. They defined *Big data* as a phenomenon itself, *causal conditions* as the happenings that led to the occurrence of big data or the reasons behind the accumulation of big data such as the need for market understanding, *context* as the conditions under which big data evolves or rather the different forms of big data that could either be user generated or machine generated, *strategy* as the necessary technological and functional steps taken to address the phenomenon and *consequences* as the outcome of applying these strategies.

In this study, the model will be used to determine the presence of big data in data from the management of SCD by establishing its features, it will also be used to find out the strategies that were applied in SCD management projects and to evaluate the outcome of these projects. Although the model was initially meant to be applied in business analytics, this study will borrow its concepts but modify the specific indicators under the constructs to reflect the concept of the study in health analytics as illustrated in section 3.3.

Figure 6 Big Data descriptive model

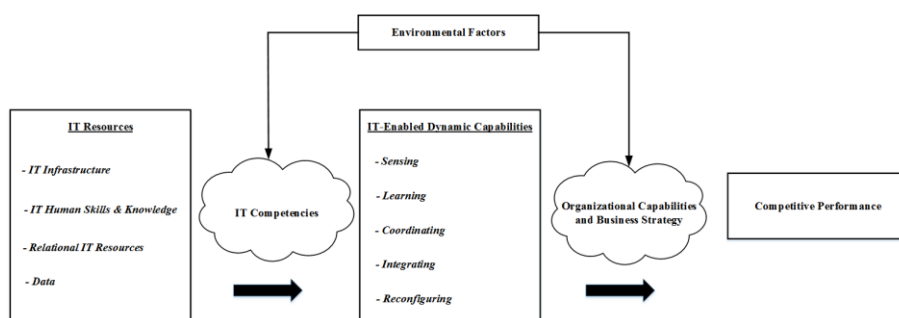


Source: Adapted from Pospiech and Felden (2013).

3.2 Resource based view-Big Data (RBV-BD)theory

To supplement the big data theory model, a resource-based view- Big data (RBV-BD) model postulated by (Mikalef, Pappas, Giannakos, Krogstie, & Lekakos, 2016) will also be incorporated in the study. Resource-based view of a firm is a theory in management strategy that determines valuable and inimitable resources with the capability to deliver a competitive advantage to a firm (Bharadwaj, 2000). Several information Systems (IS) researchers have adopted this perspective to establish IT-related resources that could potentially provide this competitive edge (Bharadwaj, 2000; Mikalef et al., 2016). A Resource-based view of IT proposes that organizations can indeed gain a competitive advantage based on their IT resources such as big data (Mikalef et al., 2016). Bharadwaj, (2000) further points out that an organizations’s IT infrastructure, human IT skills and its ability to harness IT for benefit forms its unique resources which collectively will lead to its IT capability.

Figure 7 Big data conceptual research framework



Source: Adapted from Mikalef et al. (2016)

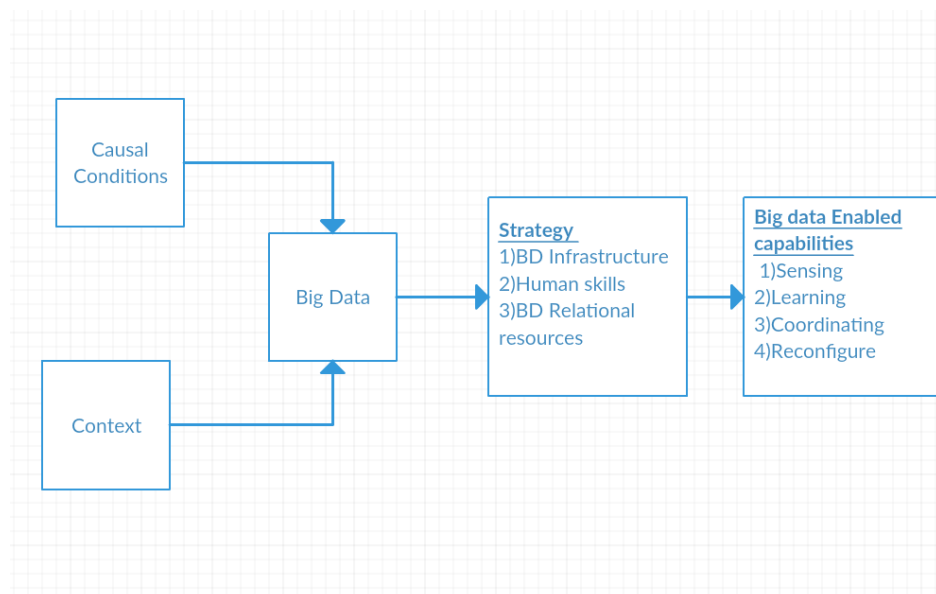
Considering big data as a unique IT resource, Mikalef et al. (2016) put forth a theoretical framework to define the fundamental areas to be considered when laying out strategies for big data initiatives. Their theory gives the foundation for understanding how to deal with big data initiative for business reasons. Analyzing RBV in the IT context shows the differences between “*IT infrastructure, IT human skills and knowledge and relational IT resources*”. In dealing with big data a fourth aspect was added into the model, the 4Vs of data; volume, variety, velocity and veracity. (Mikalef et al., 2016) argue that all these aspects collectively lead to an organization’s big data capability which must be put into action to transform the big data projects to deliver a competitive advantage or what they refer to as “*IT-enabled dynamic capabilities*”. The authors further point out that this is

the most crucial stage in the model for establishing value gained from big data projects. In this paper, this stage will be referred to as BD-enabled capabilities and will be analysed in establishing value gained from SCD data projects

3.3 Modified big data framework

Supplementing the Big data theory model with the Resource-based view-Big data theory will therefore result into a much more focused framework for exploring the research questions in this paper. This is because it can be argued that both theories suggest that the value of big data does not exclusively rely on the technologies applied, but instead through the relationships of its constructs and therefore strengthening its constructs will lead to better outcomes. Furthermore, both theories point out the same constructs which could easily be merged. Ali and Birley (1999) argue that using constructs in models is more flexible to the needs of respondents and could also give a researcher access to findings that they did not intend to at the start of their study.

Figure 8 Modified Big Data framework



In the modified framework, the construct of big data is seen as a phenomenon which emerges from both causes and context and is characterised by an increased volume of a variety of data. The causes are attributed to the reasons behind the emergence of big data, which

could be due to adoption of EHRs, or data collection for patient monitoring and decision support. Context on the other hand describes the circumstances through which big data came to be, for instance, was the data machine or user generated? Additionally, the constructs of big data infrastructure, human skills and big data relational resources established through the RBV-BD will be analysed as strategy as it encompasses both the technological and functional strategies established by Pospiech & Felden (2013).

Empirical studies done by Pospiech & Felden (2016), established that an increase in both context and causal conditions led to an increase in the phenomenon of big data and its intrinsic measures, moreover, the rise of big data also saw an increase in the strategies applied. Although the indicators established under these constructs were based on business analytics, this study will be applying the general findings to establish the specific indicators for the various constructs that have led to the presence and rise of the big data phenomenon under SCD data projects and how the organizations have strategised to leverage it.

The strategies applied will then be evaluated for the *BD-enabled capabilities* of “*sensing, learning, coordinating, integrating, and reconfiguring*” as follows;

Sensing- The ability of an organization to use big data analytics to spot customer needs, get feedback and keep up with competition. In the context of this study, this will be viewed as the ability of a sickle cell organization to use BD analytics in spotting the most important needs for the patients, monitoring them and getting their feedback.

Learning- The ability of an organization to leverage big data to explore and apply new knowledge in decision making. In the context of this study, this will be viewed as an organization's ability to use big data analytics in exploring new information that should be used in decision making. And this does not solely depend on analysing for trends but also creating new information for improvement opportunities

Coordinating- The ability of an organization to leverage big data to distribute responsibilities and resources, and synchronize actions with all the concerned stakeholders (Pavlou & El Sawy, 2011). In the context of this paper, an organization's ability to coordinate its activities at different implementation stages with sickle cell stakeholders will be viewed, any development made on the processes or any product or new management products, or programmes started will be determined.

Integrating- ability of an organization to leverage big data to assess external resources and capabilities and incorporate them to improve on their workings. This will be looking

at how a sickle cell organization has co-operated if any with external resources and how well that has worked for them

Reconfiguring- ability of an organization of an organization to leverage big data to adapt to new strategies when need arises and in the context of the study will be looking at how this orgnaizations have used big data, the constant flow of dat to respond to emergencies or new problems that have emerged.

By differentiating the various BD constructs from BD-enabled dynamic capabilities, it becomes easy to understand the link between relationships through which BD ventures can be evaluated.

Therefore, this modified framework shall be used as a guideline in the interview questions.

4 Methodology

This chapter outlines how the research was conducted and provides the reader with a clear understanding of the motivation for the chosen methodology.

4.1 Research approach

Saunders, Lewis, and Thornhill (2012) state that in order to form a basis of their research design and approach, a researcher has to give clear theory at the beginning of their study. One's choice of a research approach is vital because it enables a researcher make well-informed choices on their research design, choose the right research strategies and accommodate the shortcomings of the chosen strategies (Easterby-Smith, Thorpe, & Jackson, 2015).

Saunders et al. (2012) further discuss two research approaches; deductive and inductive. In deductive approach, a researcher first develops a theory or hypothesis and sets out a research design to test the hypothesis of prove the theory, while as for inductive approach, a researcher first collects data then develops a theory after analysing the data. They however point out that it is possible to use both methods in one single research, this study shall also be using both approaches.

In their paper, "*Integrating deductive and inductive approaches in a study--*" Ali and Birley (1999) argue that integrating both approaches is advantageous especially in research areas where extensive literature exists but a solid theory might be missing where such literature is used in formulating a deductive research framework. Although there exists a lot of literature on big data analytics with several frameworks being postulated, most of them are use- case scenarios (Pospiech & Felden, 2013). The modified framework to be used in this study has not been empirically tested as discussed in chapter 2 and this study therefore focused on spotting the consistent constructs from the literature. According to Ali and Birley (1999), one can develop a theoretical framework based on key themes, develop questions and discuss them in details or even beyond the constructs during data collection. Analysis of such data can either be based on the framework or

inductively. The modified framework shall be used as a guide for the interview questions and the analysis in this study. However, the analysis shall not solely rely on the framework but also inductively based on the data collected. This is because inductive reasoning is crucial when it comes to data-driven approaches, a researcher needs to look beyond the surface for relationships between seemingly unrelated data. Ross (2010) states “*some data nuggets never hint at their worth as predictors or indicators when considered in isolation and as we harvest more data faster, the challenge of making sense of it all becomes ever more pressing*”

4.2 Research Design

A research design is a researcher’s overall plan of answering their research questions. It specifies the source of data, how the data will be collected, analysed and how ethical issues and constraints will be addressed (Saunders et al., 2012). Priyadharshini (2012) further points out that a good research design should be adaptable, suitable and effective. That it should eliminate bias, promote the reliability of the data collected while at the same time yielding as much information as possible to answer the intended research questions.

4.2.1 Research purpose

Saunders et al., (2012) states that research questions usually reflect the purpose of a research and they can either be *descriptive*, *explanatory* or *exploratory*. However, they also point out that a research project with more than one purpose could be a combination of these. The research questions listed in chapter 1 are of exploratory nature, this is because an *exploratory study* is used to illuminate a researcher’s understanding of a phenomenon, to find out the happenings and search for new insights (Saunders et al., 2012). Considering big data as phenomenon, the first question seeks to find out the existing data practices in Sickle cell Anaemia management while question 2 and 3 seek to explore the areas in Sickle cell management that could benefit from using BD analytics and the challenges they could be facing.

4.2.2 Research method

A research can be carried out qualitatively or quantitatively. According to (Saunders et al., 2012), quantitative research methods are used in studies that generate or utilise numerical data to quantify defined variables while qualitative research methods are applied in studies that do not generate or use numerical data to uncover trends and seek insights. Although Saunders et al. (2012) also point out that a single study could use both methods, this thesis adopts a mono-method qualitative design.

As argued by (Monfared & Derakhshan, 2015), a qualitative research is basically exploratory research which gives an understanding of a problem and helps to form new hypothesis for subsequent quantitative research. In order to explore the feasibility of adopting big data analytics in SCD, semi-structured interviews are used to collect data from project managers of SCD data projects and leads of Sickle-cell affiliated organizations. To further establish activities and stages involved in the management of SCD, an in-depth interview is used to collect data from an experienced physician.

4.2.3 Research strategy

Research questions, purpose, availability of time and resources and availability of prior knowledge on the subject of study usually determine the research strategy that is adopted (Saunders et al., 2012).

Qualitative interview was chosen as the data collection tool to answer the research questions for this study. Interviews seek and provide understanding of the main concept of study and it's a researcher's responsibility to derive facts and meaning from the respondent's answers (Kvale, 1996). Interviews can be structured, semi-structured or in-depth, this exploratory study adopts semi-structured interviews. In semi-structured interviews, guiding questions are prepared based on themes on the subject of study but may differ from one interview to another depending on the context of the organization or the respondent (Saunders et al., 2012). Blumberg, Cooper, & Schindler (2008) argue that this is to allow a researcher to make inferences on causal relationships between constructs. By establishing the causal relationships between the big data constructs identified in the

modified framework, this study aims at evaluating value to be derived from big data approaches.

The semi-structured interviews shall also seek to understand opinions of leaders in Sickle cell organizations regarding data management, and also find out the professional opinions on Sickle cell data by Big data health practitioners. Their opinions will also be used to make inferences during data analysis (Saunders et al., 2012).

Furthermore, both *semi-structured* and *in-depth interviews* usually allow a respondent to expound on their answers when nudged by the interviewer. A respondent can also lead a discussion in a direction that the interviewer had not previously intended but is of importance (Saunders et al., 2012). This being an exploratory study, it is keen on tapping on the interviewee's knowledge and expertise in both Sickle cell management and big data analytics, thus the flexibility of semi-structured and in-depth interviews is desired.

4.3 Data collection

4.3.1 Sampling process

Applying *purposive* sampling technique, this study set out to interview people with practical knowledge in both Sickle cell anaemia management and Big data analytics. Expertise was the desired level of knowledge. Saunders et al. (2012) states that *purposive sampling* allows a researcher to use their wisdom in selecting cases that will be highly insightful and illuminating in answering their research questions.

Considering Big data analytics as a field that is advanced in its applications, it was expected that there is a fairly large number of practitioners, however with the study focused on the management of Sickle cell anaemia, it sought to specifically interview big data practitioners in the disease management. With this as a preliminary focus on the google search and LinkedIn, it soon became apparent that there weren't any big data practitioners working specifically on Sickle cell management. The scope was then broadened to project managers and team leaders in Sickle cell affiliated organizations such as Specialist Sickle cell hospitals, Sickle cell research centres, government agencies, regional organisations and patient organisations. A check was then done on the websites

and social media pages of the targeted organizations to ensure that they purposefully collected and analysed patient data or were running special data projects.

A total of 10 organisations or team leaders met the criteria and were contacted via a mail explaining the intent of the study and requesting for an interview. Only five responded, three were willing to give an interview, one could not give a recorded interview due to ethical binding by their organization while the other felt that their knowledge could not be substantial in the study. One of the respondents further recommended another interviewee.

To establish projects that were involved in data collection, whether deliberate or not, a further search was conducted for SCD publications that were technology oriented. Search databases PRIMO and Google scholar were used in the search as well as specific journals like Blood, Pubmed and IEEE Xplore. The search started off by using the using the keywords “Sickle cell” and “big data” where only one relevant paper, (Khalaf et al., 2015) was established. It also became apparent that there was lack of relevant papers on big data in Sickle cell disease and the search was once again redirected to technology related papers accumulating big data in their applications in improving the lives of SCD patients. The keywords “Sickle cell”, “mhealth”, “ehealth” and “ICT” were used in this search. A total of 14 papers were relevant, but out of these, 10 were literature reviews with only four writing about actual projects, the contacts from these papers were looked up and sent for request for interviews where three responded. All the interviewees were principal investigators for mobile phone applications in Sickle cell management, one for Pain management and two for hydroxyurea adherence.

In total, seven (7) interviews were conducted for the study.

To gain a professional input of big data practitioners, a broader search was done on LinkedIn focusing on big data analysts or data scientists in disease management and disease surveillance, with an experience of over 7 years. A total of five were contacted with only one granting an interview (a senior data analyst at Kaiser Permanente) but that interview will not be discussed as it did not meet the confines of this study.

4.3.2 Primary Data collection

Semi- structured interviews usually comprise of questions in a few specific themes or constructs identified in the study. Their main aim is to help the interviewer steer the discussion towards the areas in which they want to learn. This interview guides are generally open and could vary from very detailed to simple (Kvale, 1996). To answer the research questions in this study, an interview guide was prepared based on the Big data constructs in the modified framework as shown in table 1. Leading questions on BD causal conditions, context and strategy were prepared. The questions were however modified depending on the interviewee, for instance questions relating to *BD-enabled capabilities* were only asked to interviewees who were running specific data projects in their states or whose projects had been running for a considerable amount of time and could therefore have experienced these capabilities in some way. Furthermore, some questions were developed during the process of the interviews, while keeping in mind the fundamental assumption of structured interviews that questions have to be logical to the interviewee (Kvale, 1996).

Although structured questions were not prepared for the in-depth interview, a lot of reading had to be done in both fields to prepare insightful questions.

Table 2 Interview questions guide

<p>Causal Conditions: Establish the events that led to development of big Data.</p> <ul style="list-style-type: none"> • Why were the data collection projects initiated? • Why is patient data collected? 	<p>Context: Describe the circumstances in which big data evolved (4Vs).</p> <ul style="list-style-type: none"> • Find out the type of data they keep -transactional data EHR • Establish existence of multi-media sources of data (examples) • Consider an unknown data quality within the data itself (Veracity)
---	---

<p>Strategy: Technological and functional strategies.</p> <p>How did they go about implementing the projects? And the reasons behind their choice of technology and analytical methods</p>	<p>BD-enabled capabilities: What inferences have been made from the data regarding sensing, learning, coordinating, integrating and reconfiguring.?</p>
<p>Environmental factors;</p> <ul style="list-style-type: none"> • Public policy • Legal considerations 	<p>Consequences</p> <p>What has been the outcome? Good or bad?</p>

The respondents were initially notified of the intended subject of the interviews and that their expertise and knowledge in the areas would be sought. They were also informed of the intention to record the interviews and had to give their consent prior to the recording. All interviews were conducted via Skype and recorded via Call note, a recording software that allows transcription from speech to text as encouraged by (Saunders et al., 2012). Table two shows the interview schedule.

Table 3 Interview Schedule

Date and Time	Name	Organisation	Role	Duration
30/3/2018	Dr. Susan Paulanokis	State of California, USA	Project Manager, Sickle Cell Data Collection project	1hr 04mins
4/4/2018	Dr.Susan Murumba	Kory Family Hospital, Kenya	Executive Director	41Mins
17/4/2018	Dr. Mary Hullian	CDC, Atlanta, Georgia, USA	Project Manager PRESH and Rush	27Mins

20/4/2018	Dr. Tom Williams	KEMRI Wellcome Trust, Kenya and UK	Senior Researcher in Blood diseases	55Mins
28/09/2018	Dr. Nirmish Shah	Division of haematology, Duke University	Principal Investigator, TRU-Pain App	57 Mins
12/10/2018	Dr. Sheriff Badawy	Lurie Children's Hospital of Chicago	Principal Investigator, Hydroxyurea adherence App	28 mins
16/10/2018	Dr. Jane Hankins	Haematology, Jude Children's Research Hospital	Principal Investigator, Hydroxyurea adherence App	21 mins

4.3.3 Secondary data Collection

The interviewees from CDC, State of California and KEMRI Wellcome Trust have published their findings from analysing Sickle cell data in several papers. To properly discuss the findings they shared during the interviews and to dig for more insight they might have missed, this study analysed fourteen (14) of their published papers as supplementary. Saunders et al., (2012) states that secondary data can be resourceful if it enables a researcher to answer their research questions.

Table 4 Secondary data

	Paper	Author
1.	Longitudinal Data Collection for Sickle Cell Disease in California: History, Goals and Challenges	(Paulukonis, Raider, & Hulihan, 2015)
2	California Sickle Cell Disease Longitudinal Data Collection Project: Findings	(Paulukonis & Hulihan, 2017)

3	RuSH-Strategies from the field: Data collection	((U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities, & Division of Blood Disorders, 2015)
4.	Sickle Cell Data Collection Program: Three-Year Dissemination and Analysis Plan for Georgia	(Georgia Health Policy Center, 2017)
5.	Emergency department utilization by Californians with sickle cell disease, 2005–2014	(Paulukonis et al., 2017)
6.	Defining Sickle Cell Disease Mortality Using a Population-Based Surveillance System, 2004 through 2008	(Paulukonis et al., 2016)
7.	Hydroxyurea Therapy for Children with Sickle Cell Anaemia in Sub-Saharan Africa: Rationale and Design of the REACH Trial	(McGann et al., 2016)
8.	The accuracy of hospital ICD-9-CM codes for determining Sickle Cell Disease genotype	(Snyder, Lane, Zhou, Paulukonis, & Hulihan, 2017)
9.	Population Based Surveillance in Sickle Cell Disease: Methods, Findings and Implications from the California Registry and Surveillance System in Hemoglobinopathies Project (RuSH)	(Paulukonis et al., 2014)
10.	State-based surveillance for selected hemoglobinopathies	(Hulihan et al., 2015)
11.	Technology Access and Smartphone App Preferences for Medication Adherence in Adolescents and Young Adults with Sickle Cell Disease	(Badawy, Thompson, & Liem, 2016)

12.	Usability and Feasibility of an mHealth Intervention for Monitoring and Managing Pain Symptoms in Sickle Cell Disease: The Sickle Cell Disease Mobile Application to Record Symptoms via Technology (SMART)	(Jonassaint, Shah, Jonassaint, & Castro, 2015)
12	Understanding patterns and correlates of daily pain using the Sickle cell disease Mobile Application to Record Symptoms via Technology (SMART)	(Jonassaint et al., 2018)
13	Hybrid statistical and mechanistic mathematical model guides mobile health intervention for chronic pain.	(Clifton et al., 2017)

4.4 Data analysis

Data analysis was conducted through analytic induction as described by (Patton, 2015). The six semi-structured interviews were deductively analysed against the modified big data framework. The interviews were collectively analysed for the big data constructs of causal conditions, context, big data, strategies and big data-enabled capabilities. The constructs were then further analysed to establish specific indicators for each construct regarding healthcare big data projects. It is important to note that not all of the big data constructs were identified in all the interviews, some could only establish one or two constructs.

The In-depth interview by Pro. Williams was analysed to establish the most crucial areas for management of SCD patients. This was to gain further insight on areas that could gain from applying big data management.

After having established the presence of big data constructs in the first round of analysis, all the interviews were then inductively analysed to answer research questions two and three. This was done through the conventional content analysis as described by (Elo & Kyngäs, 2008). The interviews were at first analysed for meaning units, then condensed

into codes, categories and finally into broad themes. Three themes emerged for each of the questions and were further discussed and analysed with reference to existing practices in other diseases and the medical field at large.

The coding process was done with the help of the NVivo software.

4.5 Qualitative validity

To be regarded as good measures, data collection tools in every research must meet the validity, reliability and generalisability requirements (Sekaran, 2003). Noble and Smith (2015) defines validity as the accuracy with which the results reflect the data, reliability as consistency of the procedures used without bias such that similar results can be duplicated by a different researcher and generalisability as the transferability of the findings to other contexts.

Qualitative interview method applied in this study most certainly raises questions of reliability, however, while quantitative researchers use statistical methods to ensure validity and reliability in their results, qualitative researchers strive to ensure that their findings are trustworthy (Noble & Smith, 2015). To ensure trustworthiness of the findings, this study endeavoured to eliminate bias by being objective with the questions and ensuring that they were loosely related to the constructs of big data identified in theory. The full interviews were then wholly transcribed, and their analysis process systematically explained to ensure the interpretations were transparent and consistent.

Saunders et al. (2012) argue that generalisability of qualitative research depends on its significance to existing theory. Given that a concrete framework does not exist in this study, the findings shall be tested against the modified framework and analysed in the context of existing Big data practices. The purposive sampling technique used in this study does not statistically represent the whole population (Saunders et al., 2012). However, the interviewees were selected on the basis of their expert knowledge and the findings can be used as an illustration of the typical sickle cell management practices that could apply big data analytics.

5 Results

This chapter presents the results of each of the conducted interviews. The interviews have however, not been presented in a chronological order to provide a comprehensive flow of the topic of discussion. Secondary data is also referenced to supplement the interviews

5.1 Mary Hullihan- Centre for Disease Control (CDC), Atlanta Georgia

Mary is a health Scientist working in the division of blood disorders at the CDC. She focuses on hemoglobinopathy projects and has been the lead for the design and implementation of a framework for a state- based longitudinal surveillance system for SCD.

Background on Sickle Cell Disease Longitudinal Data Collection Project (SCDC)

The team started off in 2010 with the Registry and Surveillance System for Hemoglobinopathies (RuSH) project which ran in seven (7) states. The main objective for RuSH was to determine the number of people living with SCD or Thalassemia and thereafter develop plans for a national surveillance system to increase the understanding of this population's health status and practice.

RuSH ran until 2012 where a follow- up project, Public Health Research, Epidemiology, and Surveillance for Hemoglobinopathies (PHRESH) kicked off in three (3) states with focus on monitoring, health promotion and prevention of health complications, this project also ran for two years until 2014. After the RuSH and PHRESH projects, a new project focusing on Sickle cell Data collection was launched in 2015 in California and Georgia.

The interview with Mary shall be reported based on the modified framework through the big data constructs identified during the interview. It is also important to note that as the project coordinator based at CDC, she has no access to the daily implementation activities as these were individually carried out by participating member states (Georgia and California).

5.1.1 Causal Conditions for the CDC data collection projects

The American Society of Paediatric Haematology Sickle Cell Summit in 2008 made several recommendations in bridging the gap in Sickle cell healthcare. Using a population-based surveillance to measure the disease outcomes was their third recommendation (U.S. Department of Health and Human Services et al., 2015).

SCD stakeholders including clinicians, patients and researchers expressed the existence of a huge knowledge gap regarding SCD and the lack of data. So, these programs were established with the goal of amassing and combining all SCD related data from different sources into a comprehensive population-based data system.

“Really the only data that was available, and for the most part what is used by researchers is from clinical centres or data that comes from single administrative sources, like data from a particular insurance payer and so forth. So, the program that was developed was meant to overcome some of the challenges-----”.

However, the main goal of these projects had been to map the demographic and geographical representation of the SCD in the participating states;

“---to find out how many people were living with the disease, where they were receiving care, which doctors were providing care, at the start of the project those were really the main goals of the project in the states that were participating in the project”

5.1.2 Strategic implementation of the CDC projects

The organizing team mandated the participating states to carry out the data collection projects independently. CDC’s main role was to administer the project, to provide technical assistance and oversight, propose and oversee collaboration among the groups working on the project and to disseminate findings from the project.

“So, all the data collection takes place at the site of the state, so CDC is not involved in the actual data collection. The data is collected inhouse and maintained by the states participating in the work”

This was also due to lack of system integration and interoperability among the participating states.

“I mean you couldn't write a manual and say follow these 10 steps -----those steps differ by states, different data agreements by states, different partners who house the data by state, different data systems, information available, different integration systems”

5.1.3 Big Data enabled- Capabilities

Reconfiguring

From the data collection projects, CDC learned a valuable lesson, that a methodology was not simply transferable in these states. That participating states presented inherent properties and therefore different data and each project had to be approached state wise.

“I mean you couldn't write a manual and say follow these 10 steps-----so each state did things very differently----- we realised that one methodology did not work across different states”

This was mainly attributed to geographical differences between the states, for instance California is a large state with SCD patients distributed throughout the state while Georgia is smaller with a majority of the SCD patients localised in the urban regions.

Sensing and Learning

CDC has also gained a lot of insight from data collection projects, the number of SCD patients in participating states and their utilization of health services was established, it has increased awareness and availability of educational material for people living with SCD, their families and caregivers. (U.S. Department of Health and Human Services et al., 2015)

“--and as we continued working with the data now, we're understanding different health outcomes, healthcare utilization patterns, about the SCD population than what has previously been known about it.”

Learning and Integrating

Data collected from the projects has also been used by different partners, for instance, the federal government has increased health centres dealing with SCD after data collected

showed lack of access to healthcare by patients, healthcare providers in identified patient populations have also been provided with education on the management of SCD. Researchers seeking grants have also used this data to justify their studies.

“—so, the main reason we were collecting the data is to make changes-----it has been used to show lack of care for SCD patients -----has been used to have conversations with healthcare administrators to better understand how resources could be better utilised for the SCD population-----so our data has been used by researchers while applying for grants and funding to justify maybe as to why the grant should be provided for certain sites as opposed to others----- So I think it really helps to highlight where patients are being seen and where care could be improved”

Learning

Consequently, CDC developed a monitoring process involving data collection for future projects (Rush Paper) which was used in the subsequent SDC data collection projects in California and Georgia.

5.2 Susan Paulukonis-California Sickle Cell Disease Longitudinal Data Collection Project (SCDC)

Susan is based at the state of California Department of Public Health (CDPH), California Rare Disease Surveillance. She is the principal investigator and program director for the Sickle cell Disease data collection project in the state of California, a project sponsored by Pfizer and the Centre for Disease Control (CDC).

The SCDC project started in April 2015 and is expected to run over a period of 5 years. Currently it's in its third year and on course. The objective of the project was to create a surveillance system for patients with Sickle Cell Disease that contained their demographic, laboratory, clinical, treatment and outcome information and eventually monitor them.

To start off, they needed to coordinate information from different databases such as EHRs and enhance interoperability from various sites.

Using the constructs from the Modified big data framework, this thesis broke down the project as follows;

5.2.1 Causal conditions for SCDC project in California

The initial objective for the SCDC project which was borne on the third recommendation was to map the demographic and geographical representation of the SCD in the state of California. This was to gain insight on the probable geographic challenges in accessing healthcare.

“So, the initial idea was we just count people and figure out how many people there were and secondary to try and describe what was happening with that population, how long they were living, how often they were going to the Emergency Department, where they were receiving care, -----

We truly had no idea how many people had this disease, we did not know how long they were living, we knew within certain settings like clinical settings, all the people that one hospital sees , they might keep track of that but that's a very biased population and we truly did not know how many people there were in the US, or within the state, we did not have any way of looking at what was happening with the population ”

However, with time as more information is being gathered from the data collected, these objectives have evolved from simply carrying out a population mapping to trying to help solve the challenges faced by the SCD population.

“So now the objective is to try, and both publish that information and also to try and work to change some of these problems -----we give them data that will help support their missions so many of the hospitals around the state apply for grants to help them care for patients and help them improve care so we supply them with all the data -----we supply data to community based organisations who are trying to do outreach and policy changes. So now our mission is really to disseminate our data to make it useful”

Another objective has been to provide meaningful data as evidence in treatment options for patients with SCD, key among them being the use of narcotics and pain management.

“we’re also trying to provide data on how important it is for people to do pain management for SCD, that the outcome if people don't have access to these kinds of drugs are very negative, which can be death or long-term hospitalization. -----for now, we don't have as much data as we would like -----around opioid issues.”

5.2.2 Big data Context for the SCDC project in California

Data from this project presents the main characteristics of Big data; volume, veracity and variety. Velocity, however is a feature that is not well exploited since none of this data is uploaded to their system in real time.

Data being collected and analysed in the SCDC project comes from different sources such as the state hospitals, private hospitals and multiple health insurers. However, majority of SCDC’s data comes from the insurance claims filed by patients.

“---as I mentioned we have a huge data set from the govt. health insurance, then we have data sets from all the hospitalizations in the state, all the Emergency Department visits, we have clinical data that comes to us in a spreadsheet, we have the new-born data set that comes in a spreadsheet too”

“So, it comes to us as claims-----we have many different data sources but that's our biggest one.”

However only the data from the state agencies is freely accessible’

“some of the data we have to pay for, some of them we just request, some of them are from within the state govt while some of them are from outside so we have to get a special permission.”

All information pertaining to a SCD patient in the state of California is accumulated by the SCDC project. The project collects information on newborn screening, hospital visits, treatment given, the date of the incident and the costs incurred. They further link the patient data from the different data sources to build a comprehensive patient profile.

“So that means we have data on all of their office visits, all of their prescriptions, all of their hospitalizations and emergency departments that of level of information---”

“which helps us to find them in the other data sources and link them together and also helps us see what's happening with the patients who are not on our govt. plan”

Due to having multiple sources, these data sets are presented in differing formats, with each source having its own format or code for presentation. Only state sources like the state hospital and the government insurer, Medicaid follow the standard format. Some of the data is even represented in the form of insurance claims.

“All these different data sources have completely different variables----- and each one has different patient identifiers, some of them have names, some of them don't have names, some of them don't have the Social security number.

-----it's not nice and clean, there's either data variables missing so we can't make the link, sometimes there's an error, often times there's an error in one or more of the fields.”

“So it comes to us as claims -----and that's the level of information that we see, so that includes little bit of information about the patient, it includes the diagnosis, the ICD coding, it includes whatever happened in the visit so the procedure codes which can be CPT codes----”

5.2.3 Strategy for the SCDC California project

Infrastructure

The California SCD surveillance system was built on the original system set up by the RuSH and PHRESH projects. The original data linking, matching and analysing structure was to be reviewed and a new matching and linking process set-up based on the lessons learnt. It was expected to also make use of a relational database format instead of the flat file and to enable agile data analyses as opposed to the then sorting/matching of data for the analyses. (Paulukonis et al., 2015)

After two years of development, a new system called the Dynamic Health Data Linkage (DHDL) was created in April 2018.

“We developed it because there wasn’t anything on the market that met our needs – we wanted something that maintained an index of different file to file links so that when new data are brought in annually, it remembers that ‘Joe Dokes’ = social security number ‘123456789’ and also county insurance program ‘ABCD4321’”

Human resources

The project encompasses a health educator, data analyst, biostatistician, GIS/Mapping executive and an IT and data linkage overseer (Paulukonis & Hulihan, 2017). The DHDL system was also developed by an inhouse team of software developers working for the state department.

“we talked with them a lot about whether there was anything off the shelf that would give us what we needed, they felt that in order to meet the need of the project they needed to develop something.”

Implementation strategy

To start off, the project purposed to utilize the state-wide data sources that had been previously collected (Paulukonis et al., 2015).

Subsequently their database is updated as more data is received from their working partners. The DHDL system links up the data to create and update patient profiles

“DHDL Begins with some basic logic that says ok, the dates are the same in these two data sets and the sex are the same then this might be a link and then depending on the data source, it pulls different data pieces together. And everything matches up really nicely -----and it creates an index so that the future data that comes in that looks the same will automatically be linked to that person.”

However, the DHDL is not as smart as the staff have to manually go through problematic data sets to sort them out.

“So sometimes things are wrong and what the DHDL system does is it flags all the ones with the problems then I and one of my staff try and figure it out for each person. We do it manually, we have to call the clinic to confirm the variables or we might look in another data source, it's manual, we really just try to do detective work.”

After the data is sorted and matched by the DHDL, all the analysis is done in Statistical Analysis System (SAS), however the team is currently training on doing analyses in the R language for future analyses.

“So, we do all the analysis currently in SAS, two of us are training on R, so we'll probably start R frequently because it has some capabilities that will be helpful.”

Most of the analyses, however have majorly been descriptive and they're currently moving towards carrying out more inferential analyses.

“this year we're getting into inferential and more complicated analysis, -----”

5.2.4 BD Enabled capabilities for the SCDC project

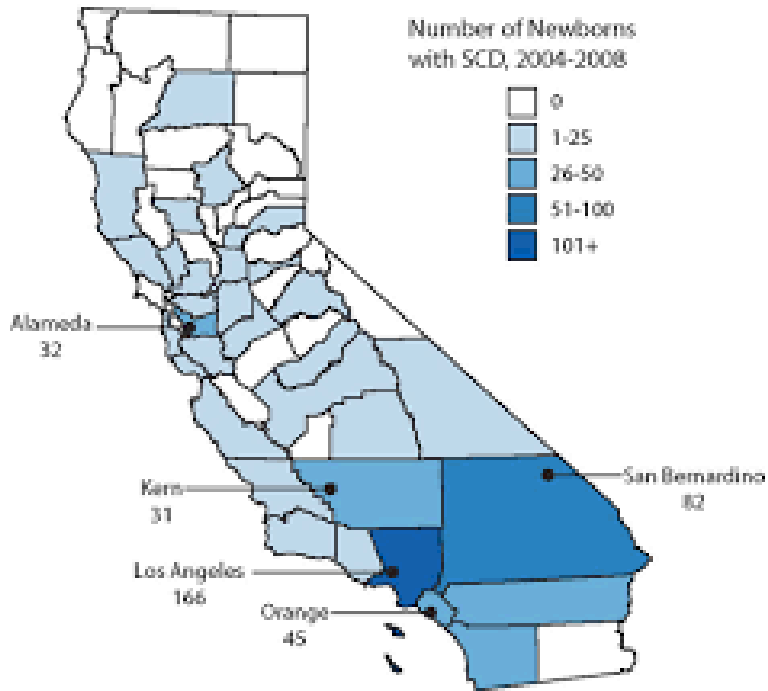
As result of collecting and analysing Sickle Cell data from the state of California, the project has been able to provide insight into several issues;

Sensing

Having set out to carry a demographic mapping for people with SCD in the state of California, the project has established that SCD cases are scattered across the state as shown below in Fig 6. Consequently, specific questions have been and are still being analysed, for instance, healthcare centres seeing SCD patients and their accessibility to patients have been established.

“So, we've done a lot of that work and the answers especially to help the utilization were discouraging, to the majority of people in California. California is not a good state to have SCD in, the majority of adults with SCD are not being seen with a haematologist and they're seen primarily in the emergency room. So now the objective is to try and both publish that information and also to try and work to change some of these problems so we are working with a number of other groups on developing a plan with California

Figure 9 California Newborn Screening Identified SCD Births, 2004-2008



California Newborn Screening Identified SCD Births, 2004-2008

Learning and reconfiguring

Susan admits that they have gained much more insight from the data than they anticipated

“Almost everything is unexpected! because so little was known , one of the recent ones that was discovered entirely by accident in a conversation with Georgia-----It's embarrassing that we've never looked at this, I looked at the number of the ED visits for SCD by year and it doubled in 10 years in California and we've not had a doubling of patients, the demographics have not changed. So, when we saw that we went back and checked, double checked, triple checked that we were not doing something wrong, I contacted Georgia and asked them if they were seeing the same thing and they were seeing a huge increase in the no. of ED visits. So that kind of thing happens regularly so we're now writing a very brief paper, probably just like a letter and publish. So, we will try and figure out what could be the reason, we do not know what the reason is but it was an alarming and we think it's important.”

Reconfiguring and Integrating

Realizing that the department alone could not influence the policy changes needed within SCD healthcare, the department has reached out and partnered with other stakeholders while leveraging insight derived from analysing their data.

“I don't think we can change the federal government right now, - - - -so we're partnering with other organisations to try and give them data that will help support their missions, so many of the hospitals around the state apply for grants to help them care for patients and help them improve care so we supply them with all the data about how many patients are in their catchment area, what the baseline in terms of care and outcome, we supply data to community based organisations who are trying to do outreach and policy changes. So now our mission is really to disseminate our data to make it useful”

5.3 Susan Murumba, Kory Family Hospital, Bungoma, Kenya

Susan is the Medical director for Kory Family Hospital, a private hospital in Bungoma County, Kenya. Kory Family hospital also runs a Sickle cell Support Programme that caters for children with SCD. It has been running the programme since 2015 in one of its two branches. The program now caters to one hundred and fifty (150) children where they are attended to when sick and provided with the required medication and their vitals closely monitored every month at the clinical support groups. The children and their caregivers are also provided with Counselling and basic home management skills for SCD.

The interview with Susan sought to establish SCD data management practices in hospitals and clinics. Kory Hospital does not however, run any elaborate data analysis programme nor own any proper data management systems. Using the modified framework, only the constructs of big data and its context were established as reported below.

5.3.1 Big data context in Kory Sickle Cell Support group.

Kenya does not run a new-born screening programme, so there exist no prior records of the children and therefore, all the children enrolled in the programme must first undergo the diagnosis test, regardless of their age and history. So, the initial data recorded of a child in the programme is from the HB electrophoresis test that is done to ascertain the presence of the disease.

“So, we usually take samples then we do HB electrophoresis so that we shall confirm if they indeed have the SCD, then now we enrol them in our support group. -----usually most of the ones in our clinics are from maybe 2yrs old up to 18 years old.”

The children’s age, sex, date of diagnosis, their medical and vaccination history and the parents’ background is recorded. Then all subsequent visits, both emergency and routine are recorded. Their vitals, any complains, and complications presented, Laboratory test results, medical imaging results and the attending doctor’s notes are recorded in the patient profile.

“Of course, the usual like age, sex, usually you also want to learn more about the parents and the date of diagnosis -----also the prophylactic treatment that has been taken prior to the diagnosis. So usually you find that most of them have been on folic acid and

Paludrine for Malaria protection -----we find that most of them don't have any vaccines like the pneumococcal, ----- and if they're already on hydroxyurea, we note the date when they started-----we note how many crises they've had in the past one year, prior to joining our support group-----any complications they might have had. So, with that information, when you follow up, you're able to see if there's an improvement or deterioration over time.”

This data is recorded threefold, in Kory Hospital’s Health Management system (HMS), in a hardcopy book at the hospital and in the patient’s booklet. The patient keeps a booklet just in case they fall sick and are admitted to a different hospital. This is because there is no system integration among the healthcare providers. Kory’s HMS is only available in its two branches.

“we have the hard copy at the hospital, -----our health management system which we input all these information----- a booklet which the patient carries around, because we're aware that they might get admitted somewhere else so this history needs to be known, so we give them a booklet which we usually update with every visit”

Kory Family Health Management System (HMS)

The HMS keeps record of patient information, it’s only locally accessible and integrates information from both branches of Kory Hospital. The system was outsourced from a third-party provider who has also forwarded personnel that maintains and supports the system from the hospital premises.

“Yeah, it's from a third party-----the provider has given us one person who is charge and sits with us in our hospital”

Not so much is done regarding data analysis, Kory’s focus is to evaluate the impact of the support group. This is measured through the number of admissions, blood transfusions and complications before and after joining the group. A reduction in these factors shows a positive impact while an increase is a cause for concern.

Patients with more than three admissions, major complications or those who suffer stroke are referred to a specialist for further review.

“So, our focus really is looking at the impact of the support group, -----One is there a decrease in the number of admissions? two, the number of transmissions before and after joining the support group and three the complications before and after---”

5.4 Dr. Nirmish Shah, Principal Investigator TRU-Pain App and director of Sickle Cell Transition programme.

Dr. Shah is an Assistant Professor of Medicine at the division of Haematology at Duke University and the Director of Sickle Cell Transition Program. For the past six years, he has been involved in technology-oriented studies and more specifically development of smart phone applications in the management of diseases. He is currently the principal investigator for the TRU-Pain app (Technology recordings to understand pain), an application used by Sickle cell patients in the management of pain. TRU-Pain is version 2.0 of the original app, SMART app (Sickle Cell Mobile Application to Record symptoms via Technology) which was initially intended to have patients log and monitor all their symptoms on the app, but because of several reasons, the app pivoted to the now TRU-Pain App that is focused on pain management.

TRU-Pain app is currently in research settings and has a total of 50 patients enrolled at the Duke clinic, but the study is set to expand with the opening of a new clinic in Fort Lauderdale. The clinic is also part of the larger 8 U01 sites that have received funding and are currently setting up a registry that will collect comprehensive data from over 2400 patients over a period for future studies and analyses.

The interview with Dr. Shah is reported based on the modified framework through the big data constructs identified during the interview;

5.4.1 Causal Conditions for TRU-Pain App and the Sickle cell Registry.

The main goal of the tru-pain App which is a pivoted version of the original SMART app, is to have physicians remotely manage a SCD patient's pain. The app was to make it possible for patients to record their symptoms and have them seen by a physician without necessarily going to the ER.

“So both of these apps focused on the idea that we need to have patients to record how they're feeling and it's not just pain, so maybe they're fatigued or they're itching from

pain medication or whatever the symptom is, we wanted an ability to allow the patients to use the app”

“----- I watch how the patient's pain is doing on a dashboard and if I see their pain is going up then I can text them and say your pain is getting worse and what can I do to help ”

As for the registry which is going to include over 2400 patients from the 8 sites across the United states, its main objective is to create a central database for a condition that has otherwise been segmented. A consensus was built from the stakeholders on the most important data variables to be collected in this registry.

“So 2400 patients that will have patient reported outcomes which is in all these surveys about how they're doing in regards to pain, and sleep interference with quality of life as well as patient report of different complications that they had based on their own reports”

“we make a big effort of having 8 sites around the country come to a consensus of all the things that everyone though they are important and so as new people are coming up with interest to do registries we want them to use our platform to do their registry, then again even though they may be separate , we can put our data together. So that's one of our goals”

5.4.2 Big data context for the TRU-Pain App and the SCD Registry

Patients using the TRU-Pain app are required to record their symptoms twice daily. They get a pop-up notification on their phones reminding them to do log their symptoms. Although the original SMART app had them recording all their symptoms including the type of medications they took, this was deemed to be complicated for patients especially when they were in pain as they did not have the energy nor patience to answer all the questions. So, the TRU-Pain app was simplified to have patients only recording their pain scores on a scale of 0 to 10. This a challenge for the collection of sufficient data as one has to balance between the actual information collected and the feasibility of the patient

generating the said data.

TRU-Pain App has also been built on Apple's CareKit platform which makes it easier for patients to navigate through the app.

"-----we do remind patients to record their symptoms twice a day, there's a pop-up that comes on their screens that says please record your symptoms"

The patient's data is stored in a cloud server and the physician can be able to view it on a dashboard in his office. If the physician sees a need for intervention, he initiates communication through text messaging.

"That's all they need to do, then we can monitor and see that, in regard to texting, it's literally texting back and forth and so if I see that their pain score is high, I will text them and see your pain is high, what's going on?"

As for the registry, the U01 sites intend to collect a variety of data from the 2400 patients over a period of time, the agreed variables are medical records, diagnosis and laboratory reports and bloodwork.

"--so, we shall pull information out of the medical records to get the diagnosis and labs, hospitalizations, and transfusions, all these data points-----the last thing we're doing is then drawing blood and putting that as bio-banking so we do genetic testing and any other testing that we wanna do"

This extraction of data will however not be carried out in real time but once in every two years.

"-----and another two years, we will go back and do another medical extraction again."

5.4.3 Strategy for the TRU-Pain App

The TRU-Pain app was developed by Dr. Shah's research nurse, who is a self-taught programmer and has developed 20 different apps for other conditions. Therefore, the team did not need to hire a developer.

The patients log in their pain scores on the app which is automatically updated on the dashboard. This information is updated in real time but since the app is still in a research setting, the study has been designed to have one person do the monitoring of the 40 plus patients. This is done once a day, usually in the mornings by either a nurse, a nurse practitioner or physician. If they see the need to intervene, they contact the patients through text messaging.

“So, once they record in their app that their pain is let's say 7/10-----then we can monitor and see that,-----so if I see that their pain score is high, I will text them and see your pain is high, what's going on? and we had a huge range of texting with one patient that texted for 30 days we texted back and forth 19 times----.”

Data collected is stored in a cloud-based database provided by CITRIX Sharefile

“We make sure that the cloud-based database is HIPAA compliant before we send/store data there. This currently is CITRIX Sharefile”

The app has also pivoted from the original app after getting feedback from the patients, mainly because the user experience is always being improved to ensure patients stick to it.

“So technically this is a little bit complicated, we keep updating the app and upgrading the version as we test it and use it like I said”

5.4.4 BD Enabled capabilities for TRU-Pain App

Dr Shah points out that the rate of re-utilization of care for patients using the TRU-Pain app has significantly reduced since using the app.

“So, re-utilization was 60 % but we did our study, the patients who got the app and are texting the re-utilization of care in 30 days was 20% so we were significantly blessed in that angle.”

Although this is the ultimate goal for many of the projects, we can analyse big data enabled capabilities of the TRU-Pain app as further illustrated by Dr. Shah.

Sensing

He states that they now have more information than he would normally when it comes to making decisions about a patient’s course of treatment

“I can report things I normally wouldn't have data very easily to get so when patients go home, when they're in hospital, instead of asking them all the time how they're doing, they record it then it immediately gets send to us”

“-----as a physician's context is to have additional data to make me feel better about my decision about what to do next”

Learning

Data from the original SMART app was used to develop a hybrid statistical and mechanistic mathematical model to understand the patients’ experience of subjective pain and how their pain levels respond to certain medication. This was published in the paper (Clifton et al., 2017). It was actually established that the pain levels fit the response to medication.

“we basically took the data from the app, in the first SMART app, we had patients also recording when they took the medications so if they took like an opioid narcotic drug, they recorded it, if they took ibuprofen not strong, they recorded it----”

“So the model was trying to predict what pain will do in response to pain medications. So we basically with pain I know that pain should come down this much with a minor pain medication and with opioid it should really come down, and it did, it actually fit ok”

Figure 10 Sample images of SMART app for iPhone/Android smartphone devices



Adapted from Clifton et al. (2017)

Reconfiguring

Having been able to model a patient’s pain score in response to the medication, the team has just been awarded a grant to further study the applicability and use of the available data from the Pain apps in predicting a patient’s outcome. The team is excited about setting up this project

“----and we actually just last week were awarded a grant to do just that-----I'm the clinician on it but we shall have a mathematician and statistician to put a group together and look at this data and see if we can predict, forward predict what's gonna happen.”

5.5 Dr. Sheriff Badawy and Dr. Jane Hankins, Principal Investigators Hydroxyurea Adherence App.

Drs Badawy and Hankins are Principle investigators for Hydroxyurea adherence app at Lurie Children's Hospital and St. Jude Children's Research Hospital respectively.

Their interviews are reported as one since they are both working on similar applications with similar vision. The applications are yet to be launched.

Hydroxyurea is a drug that has been approved for use by SCD patients and it greatly reduces morbidity and mortality(Badawy et al., 2016). However, this medication is underutilised and not adhered to especially by the adolescents and Young adults. So, the hydroxyurea adherence apps were made as an effort to increase its use among these target group.

Dr Hankins pointed out possible reasons for the lack of adherence to hydroxyurea to be due to memory loss, misconceptions about the side effects and fear of stigmatization.

“remembering to take the medicine everyday can be difficult, and the we also know that SCD patients have memory problems because the SCD can damage the brain”

----sometimes patients don't have enough information, they are not aware of what the medicine will do for them or they have misconception, they think the medicine can hurt them, that it has more side effects than it actually has, so there are a lot of incorrect assumptions that people make that gets in the way of taking the medicine -----amongst adolescents and young adults -----and then they have the issue of "I don't want my friends to know", the issue of disclosure.

The big data constructs established from their interviews are reported below;

5.5.1 Causal Conditions for Hydroxyurea Adherence app.

The key objective of these apps is to increase adherence to hydroxyurea by overcoming the challenges and barriers created for not taking the medication.

The apps will also allow patients to track their progress over time.

Dr. Hankins;

“So the app we're developing is to increase the adherence to hydroxyurea,----- to help overcome the challenges all of these barriers, the memory deficit, so it reminds the patients by sending text messages, you can browse through the app and read about SCD, treatments, research and really understand what are the real side effects and benefits of the drug and then you can talk to other people, other patients through the app, you can track your progress, you can track your pain and then graph your pain Vs adherence so you don't take the medicine and maybe hurting more and you can see that in a graph”

Dr. Badawy;

“so that's really the main goal of the app so sending patients reminders, telling them how they're doing, you know with like taking their hydroxyurea and just giving them feedback on how they're doing overall over time.”

5.5.2 Big data context and Strategy for the Hydroxyurea App

Patients will receive a push notification reminding them to take their medication which they will have to respond with a Yes or No to indicate whether they took the medication or not. The app then further asks them how they're feeling and gives them an option to rate their pain. Opting to rate their pain will allow them to track its response to the medication over a 7-day or 30-day period.

“So they get their reminders and then, I should say they log, the hydroxyurea as well , so they just click on something called log, and then hydroxyurea and this is when we get the feedback so they have to log not just get the message so they have to document that they took it.”

The patients will also be able to communicate with other members of the SCD society through a link on the app that connects to the SCD Voices, an online platform for SCD patients, caregivers, families and other resources.

“by communicating with other patients will mean outside the app, so the app will link out

to the SCD voices which is a group of patients that exist and talk through their website, so you click on it and then you get out of the app and then you go on the internet and talk to other patients.”

The patient’s usage of the app shall be captured and monitored to ensure the app is optimised and the relevant information resources availed.

“So, we're going to capture how many times they access the app per day which places in the app they click on the most, did they click on information about hydroxyurea, or did they click to talk to other patients? and the daily answer to the question of taking the medicine or not so we're gonna track that.”

Finally, to ensure the validity of the process, the evaluation of adherence shall not only rely on the app but on other measures such as pharmacy records, laboratory markers and electronic pill bottles which shall be incorporated.

“So, will evaluate adherence using different measures, through self-report, pharmacy records, look at laboratory markers and then we'll see the use of the app over time, and then we're also using electronic pill bottles so we can actually capture adherence and see when patients actually open the bottle which can be kind of a sign that they have taken the medicine”

5.6 Tom Williams, Kenya Medical Research Institute (KEMRI) Wellcome Trust

Tom is a Professor of Haemoglobinopathy Research at Imperial College, London and trained in Paediatrics. He has worked in in tropical medicine research for over 20 years in Kilifi, Kenya. His main projects relate to haematologic disorders including SCD, its distribution and its clinical consequences in Africa. He is currently running an extensive cohort of over 16,000 children, who are monitored through a continuous populational surveillance, linked to clinical and laboratory data through the integrated data management system at Kilifi.

Having had a lot of hands on experience in the management of SCD in Kilifi, Kenya and having done extensive studies on SCD, the interview with Prof. Williams sought to

inductively establish the areas, stages and activities involved in Sickle cell management particularly in patient care.

Diagnosis

Patients are diagnosed using several options available, while the developed countries run new-born screening where infants are screened for the presence of the SCD genes, most patients in developing countries are opportunistically diagnosed meaning they are only tested when they show symptoms of the SCD. Through research, KEMRI Wellcome trust ran a newborn (first year of life) screening where 16,000 children were tested.

“so, the first thing to start before you think of doing anything is diagnose. And that's quite difficult, it's not cheap, it's not widely available----- So we're just mainly testing opportunistically and usually when parents come with children with symptoms of SCD or when people come with previously affected children, the parents are advised to test the other children.”

Social support and Education

Once positively diagnosed, the children are immediately placed on prophylactic treatment. The affected families are then counselled and educated on how to effectively care and improve their children's quality of life and how to deal with the social stigma that is associated with the condition.

“---educating the parents so they know they've got an affected child, and they know the important things to look out for, ---- particularly in the very newborn, them being aware of sickle stration which used to be a very big cause of mortality where the spleen suddenly enlarges, and the children become suddenly anaemic.

----- it's very highly stigmatised in lots of places, people are ostracised by the communities, parents have marital problems once they have children with SCD.”

Preventive Care

All patients are then put on preventive care where they're given folic acid, Penicillin and anti-Malarial, and pneumococcal vaccine. Currently some patients in Kilifi are on the clinical trial on the Hydroxyurea drug which requires regular screening and blood checks.

“so, in the clinic for preventive, we keep them all on penicillin and anti-malarial, make sure they're vaccinated, put them on folic acid. Those are the main things. Then we've got a portion of the children that we're doing some clinical trials on hydroxyurea but we're not using it in our clinic routinely because the drug itself is expensive and it requires a lot of extra screening which is expensive for most patients.”

Treatment

Most patients have episodes of pain or what is called vaso-occlusive crisis, they may also present with fever, chest syndrome and anaemia. They are then observed and kept well hydrated and those requiring specific attention like blood transfusion attended to.

“So, they get to come to hospital for any event, could be pain, severe anaemia, sickle stration on the spleen, -----your normal blood count in HB in most sicklers in that part of the world is about 7. So, you only have to drop by 2 or 3 g to become dangerously low, so we need to keep a careful eye on the HBC as well when they get sick.”

6 Analysis

This chapter collectively analyses the interviews based on the modified big data framework to establish indicators for each of the constructs. The interviews are further inductively analysed to answer the research questions two and three. Secondary data is also referenced during this analysis.

6.1 Deductive analysis

Construct indicators established from the data collection have been discussed based on the modified framework.

6.1.1 Causal Conditions

According to Pospiech and Felden (2016) causal conditions are the preliminary reasons for the rise of the phenomenon that is big data. The CDC project needed a new system that could allow for integration and processing of data from different sources as the old systems could no longer serve the purpose. On the other hand, the mobile applications were intended to facilitate and improve patient-physician interactions through technology. This is congruent with Pospiech and Felden's (2016)'s indicator for the need of timely processing, an indicator they point out to have the highest impact on the construct.

Another key indicator that was established was the need for population mapping and management. The initial objective for the CDC project was to map the demographic and representation of SCD and measure its outcome on the population. This indicator can be equated to Pospiech and Felden's (2016) need for processing data for market understanding for the business world which they also stated to have a high impact on the construct.

Although Pospiech & Felden's (2016) does not list it, accessibility of healthcare to patients was also established as an indicator. Bringing healthcare near the patients was a key motivation when the CDC started collecting data on the SCD population. Moreover, with the evolution of technology, healthcare is moving away from the traditional health centres to patients' homes. This has facilitated self-management through mobile applications like the TRU-App and the Hydroxyurea adherence App.

6.1.2 Context

Pospiech and Felden (2016) define context as the conditions or environment which leads to the evolution of big data. The authors further state that the key context leading to the phenomenon of big data is the strategic acquisition of knowledge as a resource, where a firm integrates knowledge from different data sources to gain a competitive advantage. The CDC project set out to deliberately acquire data from different sources in order to gain insight into the management of SCD in the participating states. They may not have set out to gain a competitive advantage but their quest for further insight into the SCD condition led to accumulation of large sets of data.

The rising IT pervasiveness has also seen to the growth of machine generated content. Interviewees talked of having moved from traditional patient records to digitalised EHR platforms and medical claim systems. Additionally, smart phone applications like TRU-App and Hydroxyurea adherence app are also contributing to the accumulation of big data in SCD. The increasing volume of data is also coming in different forms or multimedia, for instance, Susan talked of analysing medical reports and lab reports which could come in different formats like images, text, video or audio and she also pointed out that often, this data is not always clean but full errors that have to be first cleaned. These indicators are consistent with Pospiech and Felden (2016) where they state that the continuing increase in machine generated content, multimedia data formats have created an environment for big data which has also brought forth questions of data veracity.

Another key indicator that the authors discuss is user generated content through social media, which they state is a major source of information on customers' preferences. This was however missing from the interviews, Susan P pointed out that they do not listen in on patients' social media pages because they lack the necessary tools to do so while respecting patients' privacy. This further raises another key indicator discussed by Pospiech and Felden (2016), the legal context which they point out that big data applications raises questions of data privacy.

6.1.3 Big data

Pospiech and Felden (2015) describes big data as the phenomenon itself and is seen through rising volumes of data to be stored and processed. Additionally, they state that this data could also come in different and unstructured formats. Data management in SCD like any other condition is swiftly moving towards IT enabled systems. CDC has gone a

step further to aggregate all different forms of SCD data for storage and analysis. They are updating the DHDL daily with data from private and state hospitals, medical claim forms and Doctor's report. The volumes of data available to them is therefore bound to increase with time. On the other hand, advancements in patient-centric ehealth innovations as pointed out in a systematic literature review carried out by (Badawy et al., 2018) have also led to the emergence of big data in SCD. Smart phone applications like the SMART, TRU-App and Hydroxyurea adherence apps are accumulating data daily and thus big data in SCD will keep rising.

6.1.4 Strategy

The emergence of big data requires deliberate steps to manage and handle it (Pospiech & Felden, 2013, 2015). They refer to these steps as strategy and that it's therefore not a component of big data itself but is vital for its implementation. The modified big data framework divides it into three categories; infrastructure, human resource and relational resources. Although the presence of big data in SCD could be established from the interviews, it was rather obvious that strategy was a construct that had not been properly addressed in SCD data management. This is may be an indicator of how information management has been adopted in SCD, much less the adoption of big data. However, the strategies established from the field are discussed through the framework as below;

Infrastructure,

Mikalef et al., (2016) states that big data infrastructure should comprise of both hardware and software and analytical tools. Pospiech & Felden (2015) further breaks it down into indicators like high performance computers, cloud computing, efficient programming models and relational databases. To achieve their goal for integrating and analysing data from varying sources, the SCDC project in the state of California had to invest in new hardware and system, the DHDL was built to link, match and analyse data through a relational database. While the SCDC project chose to invest and build its own systems, Mobile phone applications chose to make use of cloud-based services like Amazon web services (AWS) instead. Cloud-computing is an effective strategy for big data as these technologies provide organisations with a scalable infrastructure at an affordable cost (Botta, Donato, Persico, & Pescapé, 2016).

Human resources

Personnel working in big data are expected to be knowledgeable in analytical techniques and have the skills for using data-driven approaches to solve problems (Mikalef et al., 2016). In SCD, data management is being managed by either physicians or researchers who are not necessarily professional data scientists but take up the roles as secondary. This could be because big data is still a novelty and the stakeholders do not see the need for investing in professionals. However, as Dr. Shah pointed out, the increasing need and pile up of data for storage and analysis is putting a strain on them.

“As a physician I can only handle so much, so now you're telling me I have to look at the patient's data all the time when they communicate with me? and so I really do think that it doesn't have to be a physician, could be a nurse, it could be nurse assistant, it just needs to be a layer of someone to triage.”

To create a skilled workforce, the (European Union [EU], 2016) report suggested for the healthcare workforce to be equipped with technological and data modelling capabilities. However, incentives need to be given for more data scientists to invest in healthcare and more specifically in rare conditions like SCD.

6.1.5 Big data-enabled capabilities

To establish value gained, an organization evaluates the impact of the strategies applied to the phenomenon big data, these are referred to as big data-enabled capabilities. Although the projects covered in this study were not set out to be big data projects, big data has been established to be present through the manifestation of its constructs in the projects and has consequently led to the big data-enabled capabilities. Sensing and learning were the most notable capabilities. This could be because the organizations have not fully embraced big data approaches and are missing out on leveraging it in reconfiguring, coordinating and integrating.

Sensing; from the collection and analysis of SCD related data, the CDC has been able to demographically map the condition across the participating states as well establish the

challenges being faced. As a result new health centres were established in areas lacking and more educational resources distributed to the both patients and caregivers. Furthermore, data from other e-health technologies like TRU-App is now providing physicians with more information to support their decisions.

Learning; According to Mikalef et al. (2016), learning is the ability of an organization of to leverage big data to explore and apply new knowledge in decision making. While the main objective for the SCDC was to carry out a population count and mapping, much of the insight from the data has been unexpected, for instance, they've realised a double increase in emergency room visits over the past ten years and have launched investigations to find out why this is so. Such revelations could lead to interesting outcomes. Additionally, patient's pain response to medication was modelled from data accumulated from the pain management app , SMART.

6.2 Inductive analysis

Borrowing from the words of Ross (2010) "*some data nuggets never hint at their worth as predictors or indicators when considered in isolation*", primary data from all the interviews and secondary data were inductively analysed and three key areas that could benefit from the use of big data analytics emerged. They are discussed in the next section.

6.2.1 Possible opportunities for big data analytics in SCD management

Population Health management

(McKinsey Global Institute, 2011) stated that the availability of large data from population surveillance translates into actionable information and implementation of critical management frameworks through the acquired information. The successful completion of the RUSH and PHRESH projects saw individuals with SCD identified, their health-care utilization and clinical outcomes monitored. Most importantly however, both these projects laid ground for the development of a linked population surveillance system that has now been established in the state of California which also serves as a patient registry. The system (DHDL) aggregates data from multiple sources and ad hoc analysis done. Some of the analysis done included;

- 1) Determination of all-cause mortality rate among SCD patients
- 2) Utilization of the emergency department by the SCD patients
- 3) The availability of good healthcare to SCD patients.

Availability of this information has brought attention to the state government of the healthcare disparities associated with SCD. Moreover, Feldman et al. (2012) argues that big data will reduce wastage in public health by identifying and providing needs to the most vulnerable population. Systematic collection of data will also ensure that future incidences are explained as it will make it easier to analyse for any existing patterns. Additionally, it could be used in predictive modelling as argued by (Khalid & Abdelwahab, 2016). The team is currently working on establishing triggers for frequent emergency department visits during certain periods. Factors such as weather patterns are being studied and when done, this information is expected to contribute towards the management of the SCD.

“but what we see from the data is that some people go to the ED a lot, maybe a four or 6 month period they might go every day, every 3 or 4 days and then they stop and their life seems to return to normal for a while, maybe for a year or two and then return to another period where they're back in the ED a lot. And so, we're trying to do a time analysis of those events so that we can describe that and also try to look at what happened before that time of a lot of activity and see if we can figure out what some of the triggers are. -- ----- we have an analysis that looks up at weather patterns and whether that triggers SC crisis.”

With the availability of data, policies enacted by public health agencies can be evaluated for their impact. Data from the RuSH project in Georgia was analysed for adherence to quality metrics indicators and it was found that a significant number of SCD patients did not receive the recommended PPV immunization and TCD screening (Neunert et al., 2016). Pneumococcal polysaccharide vaccine (PPV) immunization is prophylactic antibiotic therapy given to children while transcranial Doppler ultrasonography (TCD) screening is done to establish the probability of patients getting a stroke and are therefore crucial in improving the outcomes in SCD

Adoption of big data analytics in real time will not only improve the outcome but expedite the processing of information for the decision makers.

Evidence- based Medicine

Interviewees pointed out that paucity of information and guidelines in the management of SCD poses a great challenge in providing quality care to patients. This challenge is felt across all the key areas of management of the disease including preventive care, acute complications and chronic complications.

Big data analytic systems could be utilised in gathering and analysing information across the board from existing literature to public health management in a bid to develop guidelines. This use of current evidence in making decisions about patient care is referred to as evidence-based medicine (EBM). (El-Gayar & Timsina, 2014) define Evidence-based medicine as the point where clinical expertise, external evidence and value to the patient meet. They further propose that EBM could be used in reviewing extant literature by automating article selection, mining of EHRs and clinician notes, turning them into computer executable notes and finally determining their statistical implication. The DHDL system applied in the state of California is not being efficiently utilised as it does not automatically pull data from its different sources but is instead fed manually which leads to delayed analyses. Furthermore, the analysis is only done when need arises to answer specific research questions.

While it has been argued that opioid is an effective and essential pain management drug for SCD, patients have often been denied the drug or labelled as ‘drug seeking’ (Paulukonis et al., 2015). The use of narcotics for pain management in SCD patients is thus controversial. Durkin (2018) reports that the death rate from opioid overdose in the USA rose by 45% between 2016 and 2017. Consequently, the government has restricted access to prescription opioids in a bid to curb this crisis. In this well-meaning war against narcotics abuse, SCD patients are the ‘collateral damage’ as failure to receive this kind of drug has adverse effects.

“-----the outcome if people don't have access to these kinds of drugs are very negative, which can be death or long-term hospitalization.” -----Susan P.

Guidelines in pain management and more specifically on the use of narcotics should be adopted. To supplement research being done in this area, data from EHRs, insurance

claims, medical reports and self-management should be analysed to provide an illumination on the necessity of such drugs. Whereas the TRU-App, team was able to model data from their App to prove the necessity of Opioids in extreme pain management, there is need for more evidence. The state of California is currently carrying out a similar analysis on their data.

Using big data analytics will facilitate the translation of all these forms of SCD data (extant literature, EHRs, clinical notes, imaging diagnostics and Insurance claims) into actionable evidence.

Specialized patient- care

In as much as SCD patients generally present similar symptoms across the divide, several factors come in where specialised and individual care for patients is recommended. The general guidelines being sought are not meant to be restrictive but to act as a reinforcement to the available care. Different patients present different clinical phenotypes and sometimes these phenotypes keep changing as observed by Prof. Williams, therefore, tailoring treatment to an individual patient is crucial in such cases.

“--when it comes to clinical phenotype there are very few things that are stable, so you get some children who are very very well and some cases sick all the time, and then the well ones sometimes die very suddenly”.

The IBM (2012) white paper pointed out that through the use of specialised data analytics on patient profiles using the predictive and segmentation models, big Data analytics makes it easier to identify patients that are likely to benefit from specialised care and also offer alternative treatment available to them or preventive options. In this case, big data analytics can be applied to determine SCD patients that are prone to suffering stroke which has been identified as a significant complication of SCD. A study carried out by (Ni et al., 2018) on leveraging large scale data from the Greater Cincinnati/Northern Kentucky Stroke Study (GCNKSS) in stroke detection found out that using ICD-9 for phenotyping stroke was not adequate due to complications in stroke diagnosis. It was however, shown that applying analytics to comprehensive data from EHRs and other patient records enabled the accurate prediction and distinction of the stroke types.

Applying prediction models on comprehensive patient data including data from Transcranial Doppler ultrasonography (TCD) could therefore lead to better clinical outcomes. Dr. Shah further pointed out that physician's do need extra data to make informed choices when attending to patients and the possibility of being able to predict a patient's state using various sources of data is exciting.

Another factor that should be considered in the universal management of SCD is the demographic differences. Different regions present different economic, social and cultural practices. All these have an impact on the management of the disease and providers must take this in consideration when tailoring treatments. CDC discovered that different states within the United states presented different findings and therefore different approaches were adopted in the SCDC project. Prof. Williams also noted that there are different issues dealt with in different regions. For instance, chronic transfusion therapy that is recommended in the west does not work in most parts of Africa as there isn't steady supply of safe blood. Using big data approaches in such cases could simply map out the differences and provide insight into the best possible solutions for different regions.

Despite the aforementioned potential, not much has been done in terms of leveraging BDA in the management of Sickle cell anaemia be it from a patient care or public health perspective. This could be attributed to the challenges that were noted during the study and mentioned by the respondents. The next subsection discusses these challenges in three categories;

6.2.2 Challenges facing the adoption of big data analytics in SCD management

Interoperability

The Healthcare Information and management systems society (HIMSS) of the US defines healthcare interoperability as the ability of healthcare information systems (HIS) to interconnect, exchange and use the exchanged the information (HIMSS, 2013). From the results, it was clear that they faced a major challenge with information silos. The SCDC

team collects data from different systems and manually feeds it to the DHDL. Furthermore, this data not always clean and is usually presented in different formats that make sense to each supplier. This means that more effort is required in sorting it before storage and analysis. Having interoperable systems could hasten this information sharing process and facilitate exchange of clean data. Sadeghi, Benyoucef and Kuziemy (2012) also argue that apart from raising data quality, interoperable systems could lead to the development of common interfaces and standardisation of data sets.

The mobile application teams also pointed out their dissatisfaction with the lack of common platforms for storage and sharing of information for similar applications. They fear that unless the issue is addressed, it will continue being a major problem.

“----everyone's data is siloed, everyone has different backend storage of data, different elements that are being used to record and store the same data element. So, in trying to put these data together, it's gonna be a mess until there is a constant way, although there's already efforts to do that for medical records and so on but that's way further along to have that happen but for mobile app data, we were reviewing this -----so until there is a more constant way of getting our data to talk to each other, be more continuous, we're gonna stay in these silos”.-----Dr. Shah

This is clearly a major challenge for the SCD community as not much effort has been shown towards achieving interoperability of their systems. Additionally, having such a baseline could facilitate the development of international data platforms for healthcare as argued by Adebesin, Foster, Kotzé, and van Greunen (2013).

Access to Data and legal compliance

Another key challenge that was noted from the results was access to data. This is mainly due to siloed information systems and organisations. Of note however, is that interoperability is not the only impeding factor but lack of legal structures in data sharing. While the SCDC had a fast access to data from public bodies, it was challenging for them to gain access to private sources.

Furthermore, the approval process takes a long time (18 months to 2 years) as legal compliance processes are time consuming. These processes involve negotiations in privacy, legal and ethical implications. The approval process becomes even ridiculous when one has many data sets to be approved as each partner goes through their compliance routines. However as much this could be a deterrent to access to data, these processes are necessary in building trust among all the parties involved. As Salas-Vega, Haimann and Mossialos (2016) point out, patients fear that mishandling of their information could harm their personal circumstances. On the other hand, healthcare providers want to maintain that trust and still gain a competitive advantage through analysis of patient information.

While the USA and Europe have privacy and security rules that govern the protection of personal data through the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) GDPR respectively, most of the developing world has not even started this conversation so accessing and sharing of such data poses a great challenge. Privacy not only builds trust but is intrinsically centred on values of dignity and integrity and therefore frameworks must be put in place to ensure privacy is safeguarded with efficient access to data.

The emerging options of cloud computing has also made it easier for organisations to run HIPAA/GDPR compliant databases by simply using cloud databases that are compliant. The mobile app teams are using Amazon Web services (AWS) and Citrix Sharefile which are HIPAA compliant. Botta et al. (2016) further argue that cloud computing has not only reduced organisations' cost of access to data but its scalability makes it flexible.

Lack of awareness and government Policy.

From the interviews, most of the respondents were of the opinion that big data was a buzzword that might not be useful to them. In as much as big data approaches are becoming increasingly popular in healthcare, the SCD society does not seem to be adopting. Moreover, a search on LinkedIn with the words 'big data' and 'sickle cell' could only identify 3 relevant big data practitioners in SCD related projects. A search on the same platform with 'big data' and 'cancer' yielded hundreds of profiles showing more professionals are actually involved in cancer study than SCD. Furthermore, the people who were more than willing to give interviews were physicians who also pointed out that the burden of information management has been left to them. This could be an indication of the lack of interest or awareness in SCD. These organisations could be failing to implement big data analytics because they do not appreciate how data analysis could be a useful strategy in SCD management.

Moreover, there have been no concerted efforts by stakeholders and government policies to enact such strategies. According to Salas-Vega et al. (2016), governance and management of health data in the EU has also not been well documented and neither have big data approaches been promoted.

There was also the general feeling across the respondents that SCD is a disease that is among the least prioritised by governments and underfunded. Interviewees from Kenya pointed out that there exists no national coordination but instead several bodies carrying out activities that make sense to them. While the USA might appear to be better organised, Mary pointed out that sustaining the projects was one of their biggest challenge as funding has stopped several times in the past. The US senate has this December passed a bipartisan bill that will see funds channelled towards the surveillance of SCD through collecting, analysing and interpreting data regarding sickle cell for accurate information on SCD the United States (NBC News, 2018).

SCD management is a public health issue and any meaningful data management policy can only be realised through effective government policies. Lack of deliberate policies therefore postures a huge challenge in its management much less the adoption of data management practices.

7 Conclusion

This chapter provides reflective answers to the purpose and research questions for the study

The purpose of this research was to serve as a pre-study in establishing the opportunities and challenges of applying big data analytics in the management of SCD. Through interviews, it sought to establish the current data management practices in SCD to determine the feasibility of adopting big data analytics. The research questions are answered as follows;

What Data Management practices are used in Sickle Cell Anaemia management?

While most organisations have moved towards keeping digital records, some especially in the developing world, still store data in hard copies. Furthermore, there exists no unified data management systems in SCD, different organisations keep different data sets and variables important to them. This data is also siloed within the organisations and there have not been much effort towards integration.

What areas in the management of sickle cell anaemia could benefit from use of big data Analytics?

There is a lot of potential for big data analytics in SCD. From the interviews, it was established that big data is already present in the SCD data. The data was also found to encompass big data's 4Vs of volume, variety, velocity and veracity making it feasible to apply big data analytics on it. It could be applied in the area of population health management as a public health initiative. It could also be applied in the medical area by providing evidence for the recommended treatment options such as the use narcotics for pain management. Lastly, big data analytics will provide more information for tailoring treatment for individual patients, for instance those at high risks of developing strokes.

What are the challenges of applying big data analytics in the management of sickle cell anaemia?

Although there is potential in big data analytics, it's mainly talk at this point. Its implementation is faced with challenges mainly because there are no government policies in place and not many people in the SCD world are aware or appreciate its importance. Secondly, most of the existing systems are not interoperable making data sharing a challenge. Lastly, the legal implications for data privacy and security act as a deterrent to access to a lot of data as the compliance processes are long and costly.

8 Discussion

In this chapter, the results are discussed with reference to the research questions. New understanding and insight from the research conducted is inductively discussed together with proposals for future research concerning the application of Big data analytics in Sickle cell Anaemia.

8.1 Results Discussion

With the advanced technologies and data management practices, big strides have been achieved in the healthcare industry, however, data management practices in SCD are still lagging in the industry. Interviews with different stakeholders within the Sickle cell community showed a lot of basic record keeping. Only small amount of data is being electronically captured and systematised for analysis. Healthcare applications require more effective compilation frameworks as well as conversion of various data, which includes automated conversion of structured and unstructured data.

Also noted was that different parts of the world are at different levels in their data management practices. The interviewees in the study were from two different parts of the world, one that is seen as the developed world while the other as a developing nation. Whereas the developing world, in this case Kenya is still manually keeping data in hard copies with a small amount in electronic format, the developed world, the USA has gone a step further and adopted electronic data storage. However, in both cases, not much is done once the data has been collected. Data is being collected for records sake and future references. This can be considered archaic and slow if compared to where the rest of the world is in terms of data management. Collecting data is simply not enough, what matters is how it is harnessed to create value for patients and the other stakeholders.

It is evident that data from SCD possesses the major characteristics of big data like variety, volume, veracity and velocity but not much has been done towards harnessing this information. The few organizations that have actively tried to gain insight from aggregating and analysing this information agree that they've learnt much more than what they expected.

Although most of the interviewees also pointed out that the sickle cell population is very small compared to the general population, the volume of data collected by this population cannot be ignored. Moreover, it is imperative to note that big data resources are not considered through their individual aspects but instead through all the 4Vs. A typical big data analysis process often involves iteration of several stages from data collection to analysis. A lot of varying information is usually collected from a SCD patient including structured data like their age and gender, semi-structured data like diagnosis codes, to unstructured data like their doctor's notes, laboratory and imaging results. Furthermore, Sickle-cell patients frequently visit the emergency departments compared to non- Sickle cell people and hence the SCD population is constantly generating data. Keeping proper data management systems will therefore ensure a smooth integration when needed as these databases provide information on rare diseases that would otherwise have been difficult to analyse without huge sample sizes.

Any modern technology, if correctly applied can provide numerous potential improvements and benefits. Big data analytics have the potential to transform the approach health care providers use by applying sophisticated technologies to their medical and other data sources subsequently making informed decisions. In SCD, it can be useful in population health management, personalisation of care to patients and provision of evidence for treatment options. However, it is very challenging as it calls for proper information processing, management, storage, integration and analysing. As healthcare grows, complications facing the old data management systems are only set to increase with the accumulation of large data sets due to added sizes, speeds, and different data types in addition to numerous sources and hence the need for advanced data management systems like big data analytics.

On the other hand, its adoption by the mainstream has also raised concerns in safeguarding security, guaranteeing privacy, creating standards and control but the biggest concern in SCD healthcare is the apparent lack of its adoption which can largely be attributed to lack of awareness and policies.

Lastly, the changes in the use of data and analytics in healthcare and especially in rare conditions such as SCD can only be driven by the need to save money and offer better care to the patients.

8.2 Methods discussion

As this study was purposed to determine the feasibility of adopting big data analytics in SCD management, qualitative interview was chosen as the methodological strategy. Because of the rarity of SCD and its low adoption of big data approaches, sampling techniques had to be shifted in order to get respondents who could actually contribute to the purpose of the study. The sampling frame changed from big data experts in SCD to any data management personnel within SCD management. Both semi- structured and in-depth interviews were befitting the exploratory nature of the study. The semi-structured interviews followed a modified framework which allowed the study to analyse the interviews for big data constructs.

This framework was however limiting as most of the projects did not cover all the constructs while some covered just one construct. However, this was because none of the projects were big data projects but instead they contributed towards establishing the presence big data constructs in simple data management projects.

8.3 Implications to research and Practice

This study contributes to the extant literature on big data analytics in healthcare, however it provides new insights into the possible opportunities for leveraging big data approaches specifically in SCD. Furthermore, it discusses the challenges that could be holding back its implementation. Moreover, the study has proposed and adopted a modified framework for evaluation of big data projects in healthcare.

It is also expected that this study will contribute towards effecting policies on data management and subsequent adoption of big data approaches in SCD by providing policy makers with the benefits of implementing such measures.

8.4 Future recommendation

The focus of this study has been patient-centric as opposed to health business analytics and thus further studies could investigate the feasibility of adopting big data analytics on

the business side of healthcare. This could also serve as an incentive for the financiers to invest in big data approaches. Future studies could also go further and investigate specific big data analytic practices to be applied to the different datasets

The model used in this study borrowed heavily from business analytics and hence there existed no empirically tested indicators for healthcare big data. Further studies could take the steps for adopting the model in healthcare by empirically testing the indicators established in this study.

9 References

- Adams, K. F., Piña, I. L., Ghali, J. K., Wagoner, L. E., Dunlap, S. H., Schwartz, T. A., . . . Oren, R. M. (2009). Prospective evaluation of the association between hemoglobin concentration and quality of life in patients with heart failure. *American Heart Journal*, *158*(6), 965–971. <https://doi.org/10.1016/j.ahj.2009.10.015>
- Adebesin, F., Foster, R., Kotzé, P., & van Greunen, D. (2013). A Review of Interoperability Standards in E-health and Imperatives for their Adoption in Africa. *South African Computer Journal*, *50*(1). <https://doi.org/10.18489/sacj.v50i1.176>
- Adebesin, F., Foster, R., Kotzé, P., & van Greunen, D. (2013). A Review of Interoperability Standards in E-health and Imperatives for their Adoption in Africa. *South African Computer Journal*, *50*(1). <https://doi.org/10.18489/sacj.v50i1.176>
- Ali, H., & Birley, S. (1999). Integrating deductive and inductive approaches in a study of new ventures and customer perceived risk. *Qualitative Market Research: an International Journal*, *2*(2), 103–110. <https://doi.org/10.1108/13522759910270016>
- Amendah, D. D., Mukamah, G., Komba, A., Ndila, C., & Williams, T. N. (2013). Routine paediatric sickle cell disease (SCD) outpatient care in a rural Kenyan hospital: Utilization and costs. *PloS One*, *8*(4), e61130. <https://doi.org/10.1371/journal.pone.0061130>
- Badawy, S. M., Cronin, R. M., Hankins, J., Crosby, L., DeBaun, M., Thompson, A. A., & Shah, N. (2018). Patient-Centered eHealth Interventions for Children, Adolescents, and Adults With Sickle Cell Disease: Systematic Review. *Journal of Medical Internet Research*, *20*(7), e10940. <https://doi.org/10.2196/10940>
- Badawy, S. M., Thompson, A. A., & Liem, R. I. (2016). Technology Access and Smartphone App Preferences for Medication Adherence in Adolescents and Young Adults With Sickle Cell Disease. *Pediatric Blood & Cancer*, *63*(5), 848–852. <https://doi.org/10.1002/pbc.25905>
- Baseman, J., Revere, D., & Painter, I. (2017). Big Data in the Era of Health Information Exchanges: Challenges and Opportunities for Public Health. *Informatics*, *4*(4), 39. <https://doi.org/10.3390/informatics4040039>
- Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big Data Analytics in Healthcare. *BioMed Research International*, *2015*, 370194. <https://doi.org/10.1155/2015/370194>
- Bharadwaj, A. S. (2000). A Resource-Based Perspective on Information Technology Capability and Firm Performance: An Empirical Investigation. *MIS Quarterly*, *24*(1), 169–196.

- Bipartisan bill aimed at fighting sickle cell disease signed into law. (2018, December). Retrieved from <https://www.nbcnews.com/news/nbcblk/bipartisan-bill-aimed-fight-sickle-cell-disease-signed-law-trump-n949691>
- Biswas, S., & Sen, J. (2016). A Proposed Architecture for Big Data Driven Supply Chain Analytics. *SSRN Electronic Journal*. Advance online publication. <https://doi.org/10.2139/ssrn.2795884>
- Blumberg, B., Cooper, D., & Schindler, P. (2008). *Business research methods: second european edition* (2nd ed.): Maidenhead: McGraw-Hill Higher Education.
- Boja, C., Pocovnicu, A., & Batagan, L. (2012). Distributed Parallel Architecture for" Big Data. *Informatica Economica*, *16*(2), 116.
- Borkar, V., Carey, M. J., & Li, C. (2012). Inside Big Data management: Ogres, onions, or parfaits? In *Proceedings of the 15th international conference on extending database technology* (pp. 3–14).
- Botta, A., Donato, W. de, Persico, V., & Pescapé, A. (2016). Integration of Cloud computing and Internet of Things: A survey. *Future Generation Computer Systems*, *56*, 684–700. <https://doi.org/10.1016/j.future.2015.09.021>
- Burghard, C. (2012). Big data and analytics key to accountable care success. *IDC Health Insights*, 1–9.
- Caban, J. J., & Gotz, D. (2015). Visual analytics in healthcare--opportunities and research challenges. *Journal of the American Medical Informatics Association : JAMIA*, *22*(2), 260–262. <https://doi.org/10.1093/jamia/ocv006>
- California Sickle cell Resources. (2018). <http://casicklecell.org/>.
- Choubey, M., Mishra, H., Soni, K., & Patra, P. K. (2016). Implementation of Indigenous Electronic Medical Record System to Facilitate Care of Sickle Cell Disease Patients in Chhattisgarh. *Journal of Clinical and Diagnostic Research : JCDR*, *10*(2), LC01-6. <https://doi.org/10.7860/JCDR/2016/16047.7186>
- Clifton, S. M., Kang, C., Li, J. J., Long, Q., Shah, N., & Abrams, D. M. (2017). Hybrid statistical and mechanistic mathematical model guides mobile health intervention for chronic pain. *Journal of Computational Biology*, *24*(7), 675–688. <https://doi.org/10.1089/cmb.2017.0059>
- Coveney, P. V., Dougherty, E. R., & Highfield, R. R. (2016). Big data need big theory too. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2080). <https://doi.org/10.1098/rsta.2016.0153>

- Easterby-Smith, M., Thorpe, R., & Jackson, P. (2015). *Management and business research* (5th ed.): Los Angeles: SAGE.
- El-Gayar, O., & Timsina, P. (2014). Opportunities for Business Intelligence and Big Data Analytics in Evidence Based Medicine. In *2014 47th Hawaii International Conference on System Sciences* (pp. 749–757). IEEE. <https://doi.org/10.1109/HICSS.2014.100>
- Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107–115. <https://doi.org/10.1111/j.1365-2648.2007.04569.x>
- European Union [EU]. (2016). Study on Big Data in Public Health, Telemedicine and Healthcare.
- Feldman, B., Martin, E. M., & Skotnes, T. (2012). Big data in healthcare hype and hope. *Dr. Bonnie*, *360*.
- Gaitanou, P., Garoufallou, E., & Balatsoukas, P. (2014). The Effectiveness of Big Data in Health Care, A Systematic Review. *Research Conference on Metadata and Semantics Research*, 141–153.
- Georgia Health Policy Center. (2017). SCDC_Report_FINAL: Sickle Cell Data Collection Program: Three-Year Dissemination and Analysis Plan for Georgia.
- Grosse, S. D., Odame, I., Atrash, H. K., Amendah, D. D., Piel, F. B., & Williams, T. N. (2011). Sickle cell disease in Africa: A neglected cause of early childhood mortality. *American Journal of Preventive Medicine*, *41*(6 Suppl 4), S398-405. <https://doi.org/10.1016/j.amepre.2011.09.013>
- Henke, N., Libarikian, A., & Wiseman, B. (2016). Straight talk about big data. *McKinsey Quarterly*. (4), 42–51.
- Hulihan, M. M., Feuchtbaum, L., Jordan, L., Kirby, R. S., Snyder, A., Young, W., . . . Grant, A. M. (2015). State-based surveillance for selected hemoglobinopathies. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *17*(2), 125–130. <https://doi.org/10.1038/gim.2014.81>
- IBM. (2012). Large Gene interaction Analytics at University at Buffalo, SUNY. <Http://public.Dhe.Ibm.Com/common/ssi/ecm/en/imc14675usen/IMC14675USEN.PDF>.
- Institute for Health technology Transformation [IHTT]. (2013). Transforming Healthcare through big data: Strategies for leveraging big data in the healthcare industry.
- Jeba, P. J., & Srividhya, V. (2016). Big Data Analytics in Healthcare. In proceedings of 9th National Level Science Symposium. *Christ Publications*. <Www.Ss.Christcollegerajkot.Edu.in>, *3*, 212–215.
- Jonassaint, C. R., Kang, C., Abrams, D. M., Li, J. J., Mao, J., Jia, Y., . . . Shah, N. (2018). Understanding patterns and correlates of daily pain using the Sickle cell disease Mobile

- Application to Record Symptoms via Technology (SMART). *British Journal of Haematology*, 183(2), 306–308. <https://doi.org/10.1111/bjh.14956>
- Jonassaint, C. R., Shah, N., Jonassaint, J., & Castro, L. de. (2015). Usability and Feasibility of an mHealth Intervention for Monitoring and Managing Pain Symptoms in Sickle Cell Disease: The Sickle Cell Disease Mobile Application to Record Symptoms via Technology (SMART). *Hemoglobin*, 39(3), 162–168. <https://doi.org/10.3109/03630269.2015.1025141>
- Khalaf, M., Hussain, A. J., Al-Jumeily, D., Keenan, R., Fergus, P., & Idowu, I. O. (2015). Robust Approach for Medical Data Classification and Deploying Self-Care Management System for Sickle Cell Disease. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing* (pp. 575–580). IEEE. <https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.82>
- Khalid, B., & Abdelwahab, N. (2016). Big Data and Predictive Analytics: Application in Public Health Field.
- Kruse, C. S., Goswamy, R., Raval, Y., & Marawi, S. (2016). Challenges and Opportunities of Big Data in Health Care: Systematic Review. *JMIR Medical Informatics*, 4(4), e38.
- Kubick, W. R. (2012). Big Data, Information and Meaning in Clinical Trial Insights. *Applied Clinical Trials*, 26–28.
- Kumaraguru, P. V., & Chakravarthy, V. J. (2017). A Study of Big Data Definition, Layered Architecture and Challenges of Big Data Analytics. *Indian Journal of Forensic Medicine & Toxicology*, 11(2), 635–641.
- Kvale, S. (1996). *Interviews : An introduction to qualitative research interviewing*. (6th ed.): Thousand Oaks, Calif.: Sage.
- Last, J. M. (2007). A dictionary of public health.. *Oxford University Press, USA*. Advance online publication. <https://doi.org/10.1155/2017/6120820>
- Lee, C. H., & Yoon, H.-J. (2017). Medical big data: Promise and challenges. *Kidney Research and Clinical Practice*, 36(1), 3–11. <https://doi.org/10.23876/j.krcp.2017.36.1.3>
- Lervolino, L. G., Baldin, P. E. A., Picado, S. M., Calil, K. B., Viel, A. A., & Campos, L. A. F. (2011). Prevalence of sickle cell disease and sickle cell trait in national neonatal screening studies. *Revista Brasileira De Hematologia E Hemoterapia*, 33(1), 49–54. <https://doi.org/10.5581/1516-8484.20110015>
- Lopez, A., Cacoub, P., Macdougall, I. C., & Peyrin-Biroulet, L. (2016). Iron deficiency anaemia. *The Lancet*, 387(10021), 907–916. [https://doi.org/10.1016/S0140-6736\(15\)60865-0](https://doi.org/10.1016/S0140-6736(15)60865-0)

- Luna, D., Mayan, J. C., García, M. J., Almerares, A. A., & Househ, M. (2014). Challenges and potential solutions for big data implementations in developing countries. *Yearbook of Medical Informatics*, 9, 36–41. <https://doi.org/10.15265/IY-2014-0012>
- McAfee, A., & Brynjolfsson, E. (2012). Big Data The management Revolution. *Harvard Business Review*, 90(10), 60–68.
- McGann, P. T., Tshilolo, L., Santos, B., Tomlinson, G. A., Stuber, S., Latham, T., . . . Ware, R. E. (2016). Hydroxyurea Therapy for Children With Sickle Cell Anemia in Sub-Saharan Africa: Rationale and Design of the REACH Trial. *Pediatric Blood & Cancer*, 63(1), 98–104. <https://doi.org/10.1002/pbc.25705>
- McKinsey Global Institute. (2011). Big data: the next frontier for innovation, competition, and productivity. *McKinsey Global Institute*.
- Mikalef, P., Pappas, I., Giannakos, M., Krogstie, J., & Lekakos, G. (2016). *Big Data and Strategy: A Research Framework*.
- Monfared, J. H., & Derakhshan, H. (2015). THE COMPARISON QUALITATIVE AND QUANTITATIVE RESEARCH. *Indian Journal of Fundamental and Applied Life Sciences*, 5(2231– 6345), 1111–1117.
- Moura, J., & Serrão, C. (2015). Security and privacy issues of big data. In *Handbook of research on trends and future directions in big data and web intelligence* (pp. 20–52). IGI Global.
- Neto, J. P. M., Lyra, I. M., Reis, M. G., & Goncalves, M. S. (2011). The association of infection and clinical severity in sickle cell anaemia patients. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 105(3), 121–126. <https://doi.org/10.1016/j.trstmh.2010.11.007>
- Neunert, C. E., Gibson, R. W., Lane, P. A., Verma-Bhatnagar, P., Barry, V., Zhou, M., & Snyder, A. (2016). Determining Adherence to Quality Indicators in Sickle Cell Anemia Using Multiple Data Sources. *American Journal of Preventive Medicine*, 51(1 Suppl 1), S24–30. <https://doi.org/10.1016/j.amepre.2016.02.011>
- Ni, Y., Alwell, K., Moomaw, C. J., Woo, D., Adeoye, O., Flaherty, M. L., . . . Kissela, B. M. (2018). Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis. *PloS One*, 13(2), e0192586. <https://doi.org/10.1371/journal.pone.0192586>
- Noble, H., & Smith, J. (2015). Issues of validity and reliability in qualitative research. *Evidence-Based Nursing*, 18(2), 34–35. <https://doi.org/10.1136/eb-2015-102054>

- Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A., & Belfkih, S. (2017). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*. Advance online publication. <https://doi.org/10.1016/j.jksuci.2017.06.001>
- Patton, M. (2015). *Qualitative research & evaluation methods : Integrating theory and practice : The definitive text of qualitative inquiry frameworks and options* (4th ed.). Thousand Oaks, California: SAGE Publications.
- Paulukonis, S., & Hulihan, M. (2017). California Sickle Cell Disease Longitudinal Data Collection Project: Findings.
- Paulukonis, S., Raider, F., & Hulihan, M. (2015). SCDC Report California: The Sickle Cell Disease Longitudinal Data Collection System project in California.
- Paulukonis, S. T., Eckman, J. R., Snyder, A. B., Hagar, W., Feuchtbaum, L. B., Zhou, M., . . . Hulihan, M. M. (2016). Defining Sickle Cell Disease Mortality Using a Population-Based Surveillance System, 2004 through 2008. *Public Health Reports (Washington, D.C. : 1974)*, *131*(2), 367–375. <https://doi.org/10.1177/003335491613100221>
- Paulukonis, S. T., Feuchtbaum, L. B., Coates, T. D., Neumayr, L. D., Treadwell, M. J., Vichinsky, E. P., & Hulihan, M. M. (2017). Emergency department utilization by Californians with sickle cell disease, 2005-2014. *Pediatric Blood & Cancer*, *64*(6). <https://doi.org/10.1002/pbc.26390>
- Paulukonis, S. T., Harris, W. T., Coates, T. D., Neumayr, L., Treadwell, M., Vichinsky, E., & Feuchtbaum, L. B. (2014). Population based surveillance in sickle cell disease: Methods, findings and implications from the California registry and surveillance system in hemoglobinopathies project (RuSH). *Pediatric Blood & Cancer*, *61*(12), 2271–2276. <https://doi.org/10.1002/pbc.25208>
- Pavlou, P. A., & El Sawy, O. A. (2011). Understanding the elusive black box of dynamic capabilities. *Decision Sciences*, *42*(1), 239–273.
- Plessow, R., Arora, N. K., Brunner, B., Tzogiou, C., Eichler, K., Brügger, U., & Wieser, S. (2015). Social Costs of Iron Deficiency Anemia in 6-59-Month-Old Children in India. *PloS One*, *10*(8), e0136581. <https://doi.org/10.1371/journal.pone.0136581>
- Pospiech, M., & Felden, C. (2013). A Descriptive Big Data Model Using Grounded Theory. In *2013 IEEE 16th International Conference on Computational Science and Engineering* (pp. 878–885). IEEE. <https://doi.org/10.1109/CSE.2013.132>
- Pospiech, M., & Felden, C. (Eds.). (2015). *Towards A Big Data Theory Model*. Piscataway, NJ: IEEE. Retrieved from <http://ieeexplore.ieee.org/servlet/opac?punumber=7347101>

- Pospiech, M., & Felden, C. (2016). Big Data -- A Theory Model. In *2016 49th Hawaii International Conference on System Sciences (HICSS)* (pp. 5012–5021). IEEE.
<https://doi.org/10.1109/HICSS.2016.622>
- Pouyanfar, S., Yang, Y., Chen, S.-C., Shyu, M.-L., & Iyengar, S. S. (2018). Multimedia Big Data Analytics. *ACM Computing Surveys*, *51*(1), 1–34. <https://doi.org/10.1145/3150226>
- Priyadharshini, M. (2012). Research design2012.
- Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare promise and potential. *Health Information Science and Systems*.
- Ross, J. M. (2010). Informatics creativity. *Communications of the ACM*, *53*(2), 144.
<https://doi.org/10.1145/1646353.1646390>
- Russom, P. (2011). Big Data Analytics. *TDWI Best Practices Report*. (1-40).
- Sadeghi, P., Benyoucef, M., & Kuziemsky, C. E. (2012). A mashup based framework for multi level healthcare interoperability. *Information Systems Frontiers*, *14*(1), 57–72.
<https://doi.org/10.1007/s10796-011-9306-0>
- Salas-Vega, S., Haimann, A., & Mossialos, E. (2016). Big Data and Health Care: Challenges and Opportunities for Coordinated Policy Development in the EU. *Health Systems & Reform*, *1*(4), 285–300. <https://doi.org/10.1080/23288604.2015.1091538>
- Sanyal, M. K., Bhadra, S. K., & Das, S. (2016). A Conceptual Framework for Big Data Implementation to Handle Large Volume of Complex Data. In S. C. Satapathy, J. K. Mandal, S. K. Udgata, & V. Bhateja (Eds.), *Advances in Intelligent Systems and Computing. Information Systems Design and Intelligent Applications* (Vol. 433, pp. 455–465). New Delhi: Springer India. https://doi.org/10.1007/978-81-322-2755-7_47
- Saunders, M., Lewis, P., & Thornhill, A. (2012). *Research methods for business students* (6th ed.): New York: Pearson.
- Sekaran, U. (2003). *Research methods for business: A skill building approach* (4th ed.). Hoboken, NJ: John Wiley and Sons.
- Simpao, A. F., Ahumada, L. M., & Rehman, M. A. (2015). Big data and visual analytics in anaesthesia and health care. *British Journal of Anaesthesia*, *115*(3), 350–356.
<https://doi.org/10.1093/bja/aeu552>
- Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286.
<https://doi.org/10.1016/j.jbusres.2016.08.001>
- Snyder, Lane, Zhou, Paulukonis, & Hulihan, M. (2017). The accuracy of hospital ICD-9-CM codes for determining Sickle Cell Disease genotype.

- Taylor, K. (2016). *The patient revolution: How big data and analytics are transforming the health care experience*: John Wiley & Sons, Incorporated.
- Task Force 7 Health subgroup. (2016). Big data technologies in healthcare: Needs, Opportunities and challenges. *Big Data Value Association (BDVA)*.
- TechAmerica Foundation. (2012). Demystifying big data: A practical guide to transforming the business of government.
- U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities, & Division of Blood Disorders. (2015). Registry and Surveillance System for Hemoglobinopathies -- RuSH: Strategies from the Field: Data Collection.
- Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3–13. <https://doi.org/10.1016/j.techfore.2015.12.019>
- World Health Organization [WHO]. (2006). Sickle-cell anaemia: Report by the Secretariat.
- Yardley-Jones, A. (1999). What are the implications of sickle Anaemia. *Occupational Mod*, *49*, 55–56.
- Zastrow, M. (2015). (2015). Data visualization: Science on the map. Nature News,,: Science on the map. *Nature News*.