Journal of Computer Science and Information Technology

https://jcsitech-upiyptk.org/ojs

2021 Volume 7 Issue 4 Hal: 82-93 e-ISSN: 2502-1486

A Mathematical Approach to Healthcare Insurance Data Analytics

Terungwa Simon Yange[1]✉, Ishaya Peni Gambo[2], Theresa Omodunbi[2] and Hettie Abimbola Soriyan[2]
[1]Department of Computer Science, Joseph Sarwuan Tarka University, Makurdi, Nigeria
[2]Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
lordesty2k7@ymail.com

Abstract

The emergence of big data analytics as a way of deriving insights from data has brought excitement to mathematicians, statisticians, computer scientists and other professionals. However, the near absence of a mathematical foundation for analytics has become a real challenge amidst the flock of big data marketing activities, especially in healthcare insurance. This paper developed a mathematical model for the analytics of healthcare insurance data using set theory. A prototype for the model was implemented using Java Programming Language, MapReduce Framework, Association Rule Mining and MongoDB. Also, it was tested for accuracy using data from the National Health Insurance Scheme in Nigeria with a view to reducing delays in the processes of the Scheme. The result showed that the accuracy level was 97.14% on average, which depicts a higher performance for the model. This result implies that delays affecting the processing of data submitted by the providers and enrollees to the HMOs reduced drastically leading to the improvement in the flow of resources.

Keywords: Healthcare, Insurance, Data, Big Data, Analytics, Mathematical, Set Theory.

1. Introduction

One important resource that drives organizations through the use of information technology is data. It represents facts on people, things, ideas and events using symbols such as letters of the alphabets, numerals or other special symbols [1][2]. The advancement in technologies to capture and store both structured and unstructured data have led to tremendous increase in the quantity available known as big data [3]. This vast volume of data needs to be transformed into information and knowledge to derive benefits from it, (i.e. big data analytics). The advent of big data analytics has brought exciting aspiration in the field of computing. This is one of the leading edges for innovative research and development in computer science, industry and business [4]. This has also becomes assets for organizations, industries, entrepreneurs, businesses, individual lives and the security of a nation as it derives meaningful insights which hitherto was difficult and time consuming. Even with the advent of big data analytics, many organizations are still drowning in data due to the lack of appropriate tools for the analytics [5]. Although data analytics hinges on mathematics and statistics, there has not been a sound universal mathematical framework for for the analytics of healthcare insurance data [4]. Consequently, having a mathematical model could go a long way in the abstraction of the technologies, systems, tools for data management and processing that transforms data into useful insights for better outcomes and decision making [6].

Healthcare insurance is one out of the many sectors that are currently drowning in data [7]. It covers the entire or fragment of the risk of a patient incurring medical expenses, spreading the risk over a large number of persons [8][9]. By estimating the total risk of healthcare and other expenses in the healthcare system over the pool of risk, an insurer could develop a routine finance structure, such as a monthly premium or payroll tax, to provide the money to pay for the healthcare benefits specified in the insurance treaty. The benefit is administered by a central body such as a government agency, private business, or charity organisations [8][10]. The benefits of the healthcare insurance are managed by a central body such as government agencies, private businesses, or not-for-profit entities. The primary aim of data analytics here is to investigate the cost of healthcare services rendered (i.e., payment made for healthcare services by the health insurance) [11].

In Nigeria, the National Health Insurance Scheme (NHIS) is the agency that is saddled with the responsibility to administer health insurance. It was set up by the Federal Government of Nigeria to provide universal access to quality healthcare service in the country. It covers civil servants, the armed forces, the police, the organized private sector, students in tertiary institutions, the self-employed, vulnerable persons, and the unemployed, among others [9]. The beneficiaries are required to pay a premium to NHIS, which is used to pay for their healthcare services once they visit the facility. This Scheme, though at its infancy, is faced with several challenges that some researches have established the dissatisfaction of beneficiaries, especially delay in services delivery [12][13]. This has significant adverse effects on the quality of care and patient safety.

Accepted: 21-11-2021 | Revision: 21-10-2021 | Publication: 23-11-2021 | doi: 10.35134/jcsitech.v7i4.18

In this paper, a mathematical approach to healthcare data analytics was developed using Set Theory for the processing of healthcare insurance data. The NHIS was used as the case study. This has increased the effectiveness and efficiency of the scheme, and also guard against the danger of sustaining medical expenses among individuals, thus, plays the most vital role in the improvement of healthcare quality. The improvement of healthcare delivery would aid the healthcare insurance in Nigeria to achieve its set targets for ensuring the universal health coverage and the provision of quality healthcare towards the realization of the National Health ICT Strategic Framework 2015-2020 of the Federal Republic of Nigeria.

The rest of the paper is organized in the following order. Section two gives the review of related works while Section three presents the methods deployed in achieving the purpose of the paper. The results are presented in Section four. In Section five, the discussion of the results is presented, and Section six gives the unique contributions of this article, limitations of the research and some future research directions as well as the conclusion for the paper.

2. Review of Related Works

Any data that cannot be easily handled by traditional data processing tools like relational databases is known as big data [11]. It is characterized mainly by its large volume, high velocity and increased variety which makes it difficult to be processed using orthodox data management technologies. The Analytics of Big Data involves searching through huge dataset with the aid of sophisticated tools to identify undiscovered patterns and establish the existence of hidden relationships within the dataset. In other words, it is the discovery, interpretation, and communication of meaningful patterns in data [15].

Big data analytics provide the understanding of the relationships underlying vast amounts of data from different datasets and consequently provides useful insights for decision-makers in organizations to create new business value that would also improve the quality of services they render so as to satisfy their clients. This has increased the demand by most organizations for big data analytics tools to predict market trends, optimize warranty and insurance claims, and mobilize customers [16].

The healthcare data can be broadly categorized into four groups [17]: clinical data (patient health records, medical images, laboratory and surgery reports etc.); patient behaviour data (collected through monitors and wearable devices); pharmaceutical research data (clinical trial reports, high throughput screening results); and the health insurance data which is the main source of the cost associated with healthcare which is vital to addressing the economic challenges associated with the modern healthcare system [18]. In this paper, our interest is on the healthcare insurance data, thus, we are not going to discuss the other types of healthcare data.

2.1 Healthcare Insurance Data

Healthcare insurance companies and other medical insurance providers have a wealth of data at their disposal [19]. Operations in this area produce varying amount of information, especially operational data, and also other more persistent forms of data, such as demographics, healthcare provider's location, and other essential data to accurately carry forward all business processes [20]. The non-operational data can be operational at some points in their lifetime. Therefore, our interest in this research is on the operational data.

The operational data in the healthcare insurance includes: referrals, enrollment, claims, capitation and other forms of payments which could sometimes be regarded as a report from the physician or healthcare provider to the insurance company [19]. According to [20], claims are managed primarily for the administration of payment for healthcare services delivered by healthcare providers and facilities. A claim informs all the details of a patient's visit or medical procedure. Even though the data about claims may vary, it generally contains an ID of the healthcare professional involved in the procedure (it may also be a group of professionals), and ID of the patient who was treated and a timestamp corresponding to the moment the event took place. In other words, claims data is designed to hold only those pieces of information that are required to facilitate payment by an insurance company: what service was provided, the diagnosis, who was the service provider, how much money is owed for that service. Further vital information is usually added, such as which types of health services were delivered, and the associated costs owed for the insurance company to process, among others [11][21].

The information on each claim is categorized into four classes: beneficiary information, provider information, claim header and claim details [22]. Features of the claim are mined from these categories of information. Beneficiary and provider information span across the whole claim, and it also provide information about the patient, and the provider (hospital, doctor, or medical facility). The claim header gives information about the entire claim: contact information, amount billed, diagnosis codes, dates of service and many others. The claim details give the specifics of each line in the claim. It is used to itemize each procedure that was conducted on the patient in the claim. For each procedure, the claim line includes the amount billed, the procedure code, the counter for the procedure [21].

Healthcare insurance data are complex, and generating useful insights from them cannot be done without a sophisticated clean-up. This data is also challenging to collect because of outdated and misaligned incentives, leading to prohibitively high prices to obtain and store the data. They are notoriously dirty [20]. Finding meaningful events from this data is very hard to do without building complex models to group these specific codes, and attributing claims to specific doctors or hospitals can be inconsistent without sophisticated data science detective work.

2.2 Challenges with Healthcare Data

In order to get a clearer understanding about the data analytics techniques that are needed for processing healthcare data, it is of utmost importance to comprehend the challenges presented by healthcare data [23]. These challenges are discussed as follows.

Large Volume: Healthcare repositories are often very huge, requiring large storage space which ranges from petabyte to zettabyte. The sizes of these repositories are not just in the number of patients but also in terms of the data elements stored in them [7]. With the numerous features of these databases, there comes the issue of “curse of dimensionality” most especially when dividing the patient populations using multiple filters. Besides, in most instances, data analytics techniques come with a finite array of features (e.g., x-position, y-position, colour etc.) which represent this multiplicity of variables. This data is distributed across the electronic records of different departments including billing, administration, clinical care and even medical devices. Processing these enormous, amorphous and complex datasets can be tasking and time-consuming [21].

Variable Semantics and Number: Data in the healthcare industry is amazingly inconsistent, both semantically and numerically. Semantically, electronic health record data encompasses diverse classes of variables and events. While a particular demographic feature (such as Date of birth) could be represented as a single attribute for a particular patient, events are more complex. An outcome of an event may be an associated measure and a timestamp. A medication treatment event might require the drug name, dosage, form and route of administration along with multiple timestamps representing each time it is taken. Administrative events might simply be a timestamp indication, for example, when the patient was admitted. Also, there are many different data types. Some variables are categorical, some ordinal, some continuous and some in date and time format [23]. Also, the number of these variables is not constant as there might be a few or dozens of drugs administered to a patient during a single admission. This leads to a relational database structure requiring multiple tables, instead of the traditional one-table structure consisting of observations and attributes.

Irregularity: This refers to inconsistency, noisiness and incompleteness in healthcare insurance data. The inconsistency here means the variation in the format that this data is presented. The data is sometimes structured, semi-structured and unstructured. Noisiness here means the high rate of errors in the data. This may occur as the result of improper logging of data into the EHR by healthcare personnel, either at wrong timestamps or with incorrect values [7].

Temporal Richness: Another obstacle that characterizes healthcare data is its rich temporal domain which prevents the breakdown of data into simple rows and features onto which traditional clustering, classification, and prediction tasks can be applied. Instead, a patient may have hundreds of tests performed and medications administered in interleaving sequences. Moreover, the sequences may be of different lengths, precluding a simple item-by-item comparison. Events may be associated with multiple timestamps: for example, a medication event has a timestamp for when it was ordered and another for each time it was administered. The data includes not only the relative sequence of events but also the actual time of each event, something that many current sequence mining algorithms do not consider. Finally, an additional problem is concurrent events [3][18]. In the clinical setting, the care team may submit multiple lab tests and medication orders at once, in various combinations. This makes it hard to form the true sequences.

2.3 Healthcare Insurance Data Analytics

Over the years, NHIS has developed some manual measures for identifying, investigating, and fishing out issues with data submitted by providers [24]. The experience, insight and intuition of these claims personnel have saved NHIS some monies in payments over the years. However, as good as this manual process is, the shortage of trained personnel to review every claim has been a significant problem. Thus, there are so many fraudulent claims that slip through the cracks. As a result, payments are sometimes made that should not be. This lack of many trained eyes to quickly identify fraud in claims has also culminated in the delays in the processing of claims, referrals, enrollments among others [25][26][27].

The primary purpose of data analytics in healthcare insurance is to checkmate fraud, waste and abuse of payments done for healthcare services [11][28][29][30]. Fraud is knowingly and wilfully carrying out or trying to implement, a scheme or artifice to trick any healthcare benefit program or to obtain (with the aid of false or fraudulent pretences, representations or promises) money or property owned or controlled by any healthcare benefit program. Waste refers to the overutilization of healthcare services or other practices that directly or indirectly, culminating in unnecessary costs to the healthcare insurance program. In most instances, waste is not considered as negligent actions by fraudsters but rather the ill use of resources. Abuse comprises actions that may directly or indirectly, bring about unnecessary costs to the healthcare insurance program, improper payment, payment for services that fail to meet professionally recognized standards of care, or medically unnecessary services [25], [28], [31], [32], [33]. Abuse consists of payment for items or services when there is no legal entitlement to that payment, and the provider has not knowingly and intentionally misrepresented facts to obtain payment [25]. The distinction between fraud, waste and abuse depends on specific facts and circumstances, intent and prior knowledge, and available evidence, among other factors. Thus, in this research, they are all treated as fraud.

Nigeria has a peculiar situation in combating fraud in health insurance data as the process is entirely manual [24][26][28][29][30]. This has given rise to most of the issues we have in the scheme today. The fear of financial leakages due to fraud is the key issue hampering the proper implementation of NHIS. Therefore a proper analytics tool must address fraud to curtail the other issues which are dependent on it. Most works done in the analytics of healthcare insurance data in other countries have been in the area of combating fraud. Though other aspects like the collection of data have been fully automated and hence, there are no issues with regards to delays in the processing of the other data [24][26][28][29][31][32][34]. Many works have been done in this area but none has come up with an Independent Mathematical Expressions for the analytics. In most works, a soft computing algorithm is customised to fit the problem of interest.

Sarraf and Ostadhashem [35] developed a new pipeline for medical imaging data (especially functional magnetic resonance imaging - fMRI) using Big Data Spark/PySpark platform on a single node which helps to read and load imaging data, convert them to RDDs in order to manipulate and perform in-memory data processing in parallel and convert final results to imaging format while the pipeline provides an option to store the results in other formats such as data frame. This was with the view of addressing big data analytics challenges in medical imaging which is a pillar in diagnostic healthcare deals with high volume of data collection and processing. The results revealed that Spark (PySpark) based solution improved the performance (in terms of processing time) around four (4) times on a single compared to the previous work developed in Python.

In a similar vein, [36] considered issues with big data analytics frameworks based on Hadoop/MapReduce. In this work, it was stated that these approaches cannot meet some vital requirements like scalability, security and large real-time streaming, because of some issues with I/O cost, algorithmic complexity, low-latency streaming jobs and fully disk-based operation. They proposed a scalable, secure and real-time healthcare analytics framework with apache spark. In this framework, Spark streaming was used to handle massive streaming data coming from streaming sources. On the other hand, SparkML was used to handle big static datasets coming from static data sources, and also for recommendation and diagnosis based on structured and unstructured knowledge processing.

Karthika and Porkodi [37] developed a fraud claim detection framework using Spark. Here, electronic health and medical data is used for detecting fault occurrence in healthcare insurance companies. Each patient is assigned with unique patient ID across the database. Apache spark is used for processing instantly on regular updates in medical records and finds the fraud occurrence by using map transformation, and reduce transformation is used to find records across the entire database. A rule-based model and Machine learning algorithm is used for automating the process, and result displays patient ID, city, time, hospital of patient in the claim. With this, fraudulent claims are reduced and it is most accurate when compared to the existing systems. In the existing system (manual system), it took 22hours to process the data which is received on a single day. But the proposed system takes only 20 minutes to process the real-time data which makes the detection of fraudulent claims very fast and highly accurate. Fraudulent claims were identified based on patientid details displays in dashboard. This also helps in reducing the treatments and readmissions.

Also, [38] developed a data analytics framework for Health Insurance data using Association Rule Mining. The researcher was able to detect fraudulent health insurance claims by identifying correlation or association between some of the attributes on the claim documents. With the application of a data mining techniques of evolving clustering method, association rule mining and support machine, this study was able to successfully determined correlated attributes to address the discrepancies of data in fraudulent claims and thus reduce fraud in health insurance. However, the study was used for structured data which made it unfit to be applied in big data which is highly unstructured. With the numerous data mining techniques implemented traditionally, it would consume more resources when applied to big data.

2.4 Mathematical Approach to Big Data Analytics

No matter how difficult a computational problem is, once reduced to some mathematical formulae or equations, it is better understood. In other words, the use of mathematics or logic to specify or verify the features of a system could further demystify the issues surrounding the design of such system. Big data analytics is not an exception to this. According to [4][39], there are no sound mathematical foundations for data analytics and this has presented serious challenges to the development of systems for big data analytics in many sectors. To curtail this problem, calculus and set theory are very pivotal towards the establishment of a firm mathematical foundation for big data analytics. The universality of the mathematics provides approaches to problem solving that developers understand with ease. According to [40][41], mathematical reasoning increases a software developer's competence. The lack of formal methods for specifying big data analytics solution is the reason responsible for the non-availability of sound data analytics tools in some areas such as the healthcare insurance. For instance, once the mathematical model for the analytics of healthcare insurance data is developed, every other issue concerning all the analytics of this will naturally be addressed [4][39][41].

In most of the works [35][36][37][38] reviewed above, none considered the development of a mathematical model for healthcare insurance data analytics. All they did was the implementation of some soft computing algorithms for the processing of data. But expressing a big data analytics problem using some systems of mathematical equations would better arm the data scientist with a tool similar to that of a hammer in the hands of the carpenter where every thing looks like a nail and wood; thus, every analytics problem becomes solvable.

inherent in decision making could be of help. The set theory is of great importance in this regards and it assures the subject of a place prominent in human culture which is the environment where mathematics takes place today. As such, it is expected to provide a firm foundation for the rest of mathematics. Because the fundamentals of Set Theory are known to all mathematicians, basic problems in the subject seem elementary [44][45][46].

3. Research Methods

This section discussed the collection of data and the formulation of the model.

3.1 Data Collection

The processing of NHIS data is a complete task which is manually carried out by a few personnel who have the responsibility of approving, modifying or rejecting these requests within a limited period from their reception. This has resulted in unnecessary delay in the process. These delays in processing the transactions of the scheme have been a discouraging factor in embracing the scheme. The implication is that the number embracing the scheme tends to be reducing or not encouraging. Collecting this data was one of the most difficulty in this work because most of the data were collected manually by NHIS through HMOs and stored in file cabinets. This was done via document examination and observation, which in either case, the data was collected from journals and NHIS databases. The categories and features of the data collected are presented in Table 1 and the sample data in different formats is shown in Fig. 1 – 4.

Table 1. Data Categories

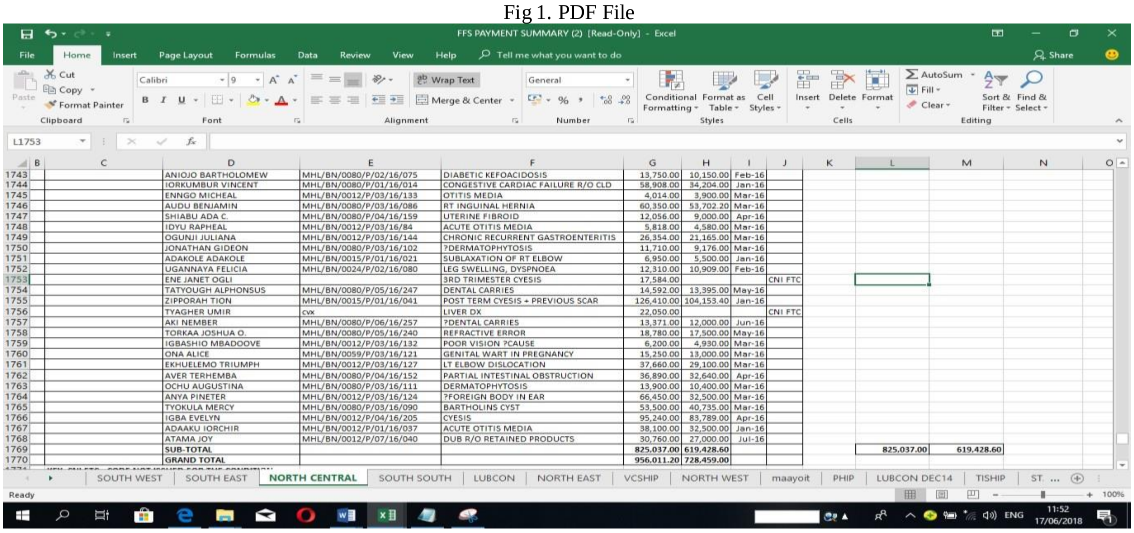
Data Category	Description
Enrolment	Provider: Name, Address, Telephone, Fax/Phone, Email, Type of facility, Category of registration, State registration no, Name of Director, Name of supervising Medical Director (If applicable), Affiliated HMOs, Affiliated Insurance companies, NHIS registration number, Incorporation/business registration.
	Beneficiary: Name, Address, Date of birth, Sex, Next of Kin, Email address, Mobile, Telephone no. fixed, National ID no, Employer NHIS no., Date of NHIS registration, Nationality, Location of Posting, Photograph, Blood group, Genotype, Allergies, Relationship (Principal, Spouse, Child, Extra-dependant), Expiry date, Primary provider
Payment	Claim: Name, NHIS No. of patient, Name and NHIS No. of patient's primary provider, Name and NHIS No. of Secondary Provider, drug prescription sheet, Diagnosis/disease code, Treatment given, Date of treatment, amount billed, Co-payment received (when applicable).
Update	Addition of dependent, change of facility, change of HMO
Referrals	Referral request, approvals, rejections

3.2 Mathematical Formulation

The formulation of the proposed model is based on the concept of searching which seeks to find patterns in a set of data; that is, it believes that patterns exist in every large collection of items. Much of big data analytics tasks involves searching for structures or items that are hidden within large and complex datasets. The concept of big data analytics involves the collection of large datasets similar to the haystack; there also, exist some small datasets within the given large datasets similar to the needle, existing together according to an unknown relationship. Usually, the data producers are quite different from data users, so that the cause-effect relation hidden in observation data is not clear to specific data users. These relationships are best described using set theory. The theory is about members and their relations in a set is general enough to deal with data analysis and problem-solving. In a sense, relations among set members corresponds to the association in big data.

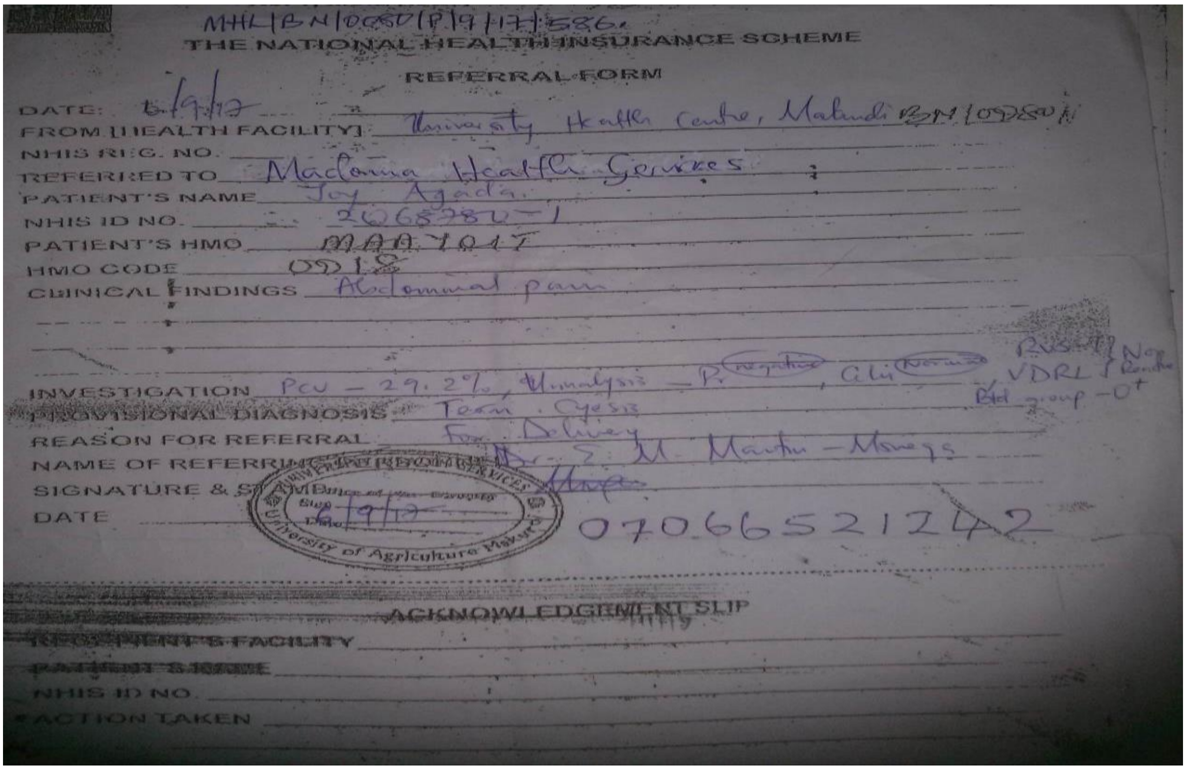
YB/0020	Ajiko Medical Center	9	+	20	+	0	=	29
YB/0022	Potiskum Medical Clinic	7	+	28	+	0	=	35
ZF/0001	General Hospital, Anka	1	+	0	+	0	=	1
ZF/0002	General Hospital, Bakura	1	+	5	+	0	=	6
ZF/0006	Daula Hospital & Mat. Home	14	+	16	+	0	=	30
ZF/0009	Gusau Medical Clinic	0	+	4	+	0	=	4
ZF/0019	General Hospital, Talata Mafara	1	+	1	+	0	=	2
ZF/0021	Federal Medical Centre Gusau	75	+	151	+	0	=	226
ZF/0027	Federal Polytechnic, Kaura Namoda	1	+	2	+	0	=	3
ZF/0029	Yariman Bakura Specialist Hospital	1	+	0	+	0	=	1

Fig 1. PDF File



	B	C	D	E	F	G	H	I	J	K	L	M
1743			ANIOJO BARTHOLOMEW	MHL/BN/0080/P/02/16/075	DIABETIC KEFOACIDOSIS	13,750.00	10,150.00	Feb-16				
1744			IORKUMBUR VINCENT	MHL/BN/0080/P/01/16/014	CONGESTIVE CARDIAC FAILURE R/O CLD	58,908.00	34,204.00	Jan-16				
1745			ENNGO MICHEAL	MHL/BN/0012/P/03/16/133	OTITIS MEDIA	4,014.00	3,900.00	Mar-16				
1746			AUDU BENJAMIN	MHL/BN/0080/P/03/16/086	RT INGUINAL HERNIA	60,350.00	53,702.20	Mar-16				
1747			SHIABU ADA C.	MHL/BN/0080/P/04/16/159	UTERINE FIBROID	12,056.00	9,000.00	Apr-16				
1748			IDYU RAPHEAL	MHL/BN/0012/P/03/16/84	ACUTE OTITIS MEDIA	5,818.00	4,580.00	Mar-16				
1749			OGUNJI JULIANA	MHL/BN/0012/P/03/16/144	CHRONIC RECURRENT GASTROENTERITIS	26,354.00	21,165.00	Mar-16				
1750			JONATHAN GIDEON	MHL/BN/0080/P/03/16/102	?DERMATOPHYTOSIS	11,710.00	9,176.00	Mar-16				
1751			ADAKOLE ADAKOLE	MHL/BN/0015/P/01/16/021	SUBLAXATION OF RT ELBOW	6,950.00	5,500.00	Jan-16				
1752			UGANNAYA FELICIA	MHL/BN/0024/P/02/16/080	LEG SWELLING, DYSPNOEA	12,310.00	10,909.00	Feb-16				
1753			ENE JANET OGLI		3RD TRIMESTER CYESIS	17,584.00			CNI FTC			
1754			TATYOUGH ALPHONSUS	MHL/BN/0080/P/05/16/247	DENTAL CARRIES	14,592.00	13,395.00	May-16				
1755			ZIPPORAH TION	MHL/BN/0015/P/01/16/041	POST TERM CYESIS + PREVIOUS SCAR	126,410.00	104,153.40	Jan-16				
1756			TYAGHER UMIR	cvx	LIVER DX	22,050.00			CNI FTC			
1757			AKI NEMBER	MHL/BN/0080/P/06/16/257	?DENTAL CARRIES	13,371.00	12,000.00	Jun-16				
1758			TORKAA JOSHUA O.	MHL/BN/0080/P/05/16/240	REFRACTIVE ERROR	18,780.00	17,500.00	May-16				
1759			IGBASHIO MBADOOVE	MHL/BN/0012/P/03/16/132	POOR VISION ?CAUSE	6,200.00	4,930.00	Mar-16				
1760			ONA ALICE	MHL/BN/0059/P/03/16/121	GENITAL WART IN PREGNANCY	15,250.00	13,000.00	Mar-16				
1761			EKHUELEMO TRIUMPH	MHL/BN/0012/P/03/16/127	LT ELBOW DISLOCATION	37,660.00	29,100.00	Mar-16				
1762			AVER TERHEMBA	MHL/BN/0080/P/04/16/152	PARTIAL INTESTINAL OBSTRUCTION	36,890.00	32,640.00	Apr-16				
1763			OCHU AUGUSTINA	MHL/BN/0080/P/03/16/111	DERMATOPHYTOSIS	13,900.00	10,400.00	Mar-16				
1764			ANYA PINETER	MHL/BN/0012/P/03/16/124	?FOREIGN BODY IN EAR	66,450.00	32,500.00	Mar-16				
1765			TYOKULA MERCY	MHL/BN/0080/P/03/16/090	BARTHOLINS CYST	53,500.00	40,735.00	Mar-16				
1766			IGBA EVELYN	MHL/BN/0012/P/04/16/205	CYESIS	95,240.00	83,789.00	Apr-16				
1767			ADAAKU IORCHIR	MHL/BN/0012/P/01/16/037	ACUTE OTITIS MEDIA	38,100.00	32,500.00	Jan-16				
1768			ATAMA JOY	MHL/BN/0012/P/07/16/040	DUB R/O RETAINED PRODUCTS	30,760.00	27,000.00	Jul-16				
1769			SUB-TOTAL			825,037.00	619,428.60				825,037.00	619,428.60
1770			GRAND TOTAL			956,011.20	728,459.00					

Fig 2. Excel File



MHL/BN/0080/P/9/17/586.
THE NATIONAL HEALTH INSURANCE SCHEME
REFERRAL FORM
DATE: 6/9/17
FROM [HEALTH FACILITY]: University Health Centre, Makurdi BN/0080/N
NHIS REG. NO.
REFERRED TO: Madonna Health Services
PATIENT'S NAME: Joy Agada
NHIS ID NO.: 2068280-1
PATIENT'S HMO: MAAYOIT
HMO CODE: 0018
CLINICAL FINDINGS: Abdominal pain
INVESTIGATION: PCV – 29.2%, Urinalysis – Pr (negative), Glu (normal), RVS (Non-reactive), VDRL (Non-reactive), Bld group – O+
PROVISIONAL DIAGNOSIS: Term Cyesis
REASON FOR REFERRAL: For Delivery
NAME OF REFERRING OFFICER: Dr. S. M. Mark – Manyi
SIGNATURE & STAMP
DATE: 6/9/17
07066521242
ACKNOWLEDGEMENT SLIP
RECEIVING FACILITY
PATIENT'S NAME
NHIS ID NO.
ACTION TAKEN

Fig 3. Hand Written document

```
MHL/FCT/0478/P/12/13/105,,,,,,4500.00,ATOPIL DERMATITION,2227.20,NOT AUDITED,NOT PAID,UATH,,,,,Dec-13,,,MAGAJI,HAFSAT
MHL/FCT/0478/P/12/13/106,,,,,,49050.00,MULTIPLE SCLEROSIS,30967.40,NOT AUDITED,NOT PAID,UATH,,,,,Dec-13,,,OGUNJUYIGBE,BOLANLE
MHL/FCT/0478/P/12/13/029,,,,,,21700.00,LOW BACK PAIN,6700.00,NOT AUDITED,NOT PAID,UATH,,,,,Dec-13,,,OLOWOOKERE,ORIMOLOYE
MHL/FCT/0478/P/12/13/116,,,,,,2600.00,CCF 2ND TO HTN,2972.00,NOT AUDITED,NOT PAID,UATH,,,,,Dec-13,,,SIMON,IKO
MHL/FCT/0478/P/12/13/108,,,,,,20770.00,RT ANKLE SPRAIN (DISLOCATION),20170.00,NOT AUDITED,NOT PAID,UATH,,,,,Dec-13,,,UBOM,AUGUSTINE
MHL/FCT/0568/P/11/13/010,,,,,,4700.00,SEVERE MENORRHAGIA,6050.00,NOT AUDITED,NOT PAID,UATH,,,,,Dec-13,,,NATHAN,MARY
MHL/FCT/0478/P/01/14/111,,,,,,1500.00,TRAUMATIC RED EYE,2000.00,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,DAUDA,ABDULRAHAMAN
MHL/FCT/0394/P/01/14/033,,,,,,10100.00,HEPATITIS B CARRIER,4709.50,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,ARAGI,TABITHA
,MHL/FCT/0336/P/01/14/011,,,,,,6150.00,DIABETES MELLITUS,6150.00,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,PHILIP,AKPE
,MHL/FCT/0478/P/02/14/134,,,,,,14850.00,COMPLICATED DM,6413.00,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,AKANDU,NGOZI
,MHL/FCT/0568/P/01/14/607,,,,,,9200.00,EXTRA-`DIGITAL (R) THUMB,4951.20,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,SAAVE,JOY
,MHL/FCT/0394/P/01/14/035,,,,,,7200.00,PARTIAL DEAFNESS,7000.00,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,OMOREGIE,BLESSING
,MHL/FCT/0478/P/02/14/125,,,,,,11500.00,TOOTH ACHE,8500.00,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,OTSE,BLESSING
,MHL/FCT/0478/P/02/14/119,,,,,,27400.00,HYPERTENSIVE HEART DISEASE,15175.00,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,FELIX,MATTHEW
,MHL/FCT/0478/P/02/14/129,,,,,,64100.00,FRACTURED MANDIBLE/ZYGEMA,58535.20,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,DANJUMA,GOYOLA
,MHL/FCT/0478/P/02/14/121,,,,,,22445.80,TRAUMATIC PERFORATION,22445.80,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,FREDRICK,ALEXANDER
,MHL/FCT/0568/P/02/14/004,,,,,,98750.00,GASTRIC OUTLET OBSTRUCTION,80320.40,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,KILLIAN,EKE
,MHL/FCT/0478/P/02/14/120,,,,,,18900.00,PYOGENIC GRANULOMA,18960.20,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,HENRY,UDEJI
,,,,,,,3650.00,,,NOT AUDITED,NOT PAID,UATH,,,,,,,,ONUBAIYE,YAKUBU
,MHL/FCT/0478/P/04/14/170,,,,,,15500.00,CEREBRAL PALSY,10107.40,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,AYEH,GOODNESS
,MHL/FCT/0478/P/12/14/139,,,,,,39850.00,HERNIA (INGUINOSCROTAL),27935.00,NOT AUDITED,NOT PAID,UATH,,,,,Jan-14,,,IBRAHIM,SANI
,MHL/FCT/0478/P/03/14,,,,,,15500.00,HYPERTENSIVE HEART DISEASE,11762.00,NOT AUDITED,NOT PAID,UATH,,,,,Mar-14,,,EGYE,AUGUSTINE
,MHL/FCT/0478/P/03/14/154,,,,,,33550.00,HYPERTENSIVE HEART DISEASE,17945.00,NOT AUDITED,NOT PAID,UATH,,,,,Mar-14,,,NELLYS,PATIENCE
,NIL,,,,,,5650.00,NIL,,NOT AUDITED,NOT PAID,UATH,,,,,,,,ABDULKAREEM,ABUBAKAR
,MHL/FCT/0478//P/03/14/155,,,,,,10200.00,DENTAL CARIES,7634.40,NOT AUDITED,NOT PAID,UATH,,,,,Mar-14,,,OGBOBI,ALEX
,NIL,,,,,,3700.00,NIL,,NOT AUDITED,NOT PAID,UATH,,,,,,,,SAAVE,RICHARD
,MHL/FCT/0568/P/03/14/005,,,,,,7500.00,HYPERTENSIVE HEART DISEASE,2040.00,NOT AUDITED,NOT PAID,UATH,,,,,Mar-14,,,OKORO,EDWIN
,MHL/FCT/0120/P/04/14/013,,,,,,7400.00,CYESIS,7050.00,NOT AUDITED,NOT PAID,LIMI HOSPITAL AND MATERNITY,,,,,Apr-14,,,ORIYOMI,ALADELOKUN
,MHL/FCT/0120/P/04/14/012,,,,,,1000.00,DATING OF PREGNANCY,1000.00,NOT AUDITED,NOT PAID,LIMI HOSPITAL AND MATERNITY,,,,,Apr-14,,,OLAJIDE,OBAFEMI
,MHL/FCT/0158/P/04/14/001,,,,,,2000.00,TENDON SPRAIN OF SHOULDER,3200.00,NOT AUDITED,NOT PAID,LIMI HOSPITAL AND MATERNITY,,,,,Apr-14,,,TINU,OYENIYI
,MHL/FCT/0120/P/05/14/015,,,,,,8250.00,CYESIS,7250.00,NOT AUDITED,NOT PAID,LIMI HOSPITAL AND MATERNITY,,,,,May-14,,,ALAUSA,HAPPINESS
,,,,,,,18650.00,,18500.00,NOT AUDITED,NOT PAID,,,,,,,,,SUBTOTAL,
,MHL/FCT/0546/P/05/14/018,,,,,,1000.00,CYESIS,1000.00,NOT AUDITED,NOT PAID,GARKI HOSPITAL,,,,,May-14,,,HOSEA,NAWUAL
,MHL/FCT/0546/P/05/14/017,,,,,,1000.00,UTI IN PREGNANCY,1000.00,NOT AUDITED,NOT PAID,GARKI HOSPITAL,,,,,May-14,,,AJAYI,KENECHI
```

Fig 4. CSV File

Let

D is the data stored in the repository

X is the data of interest extracted from D, i.e., $X \subseteq D$.

A is the set of features of X

K is the number of partitions of X

R is the relationship that exists between A and X.

G_D is a split function which implements the Divide and Conquers approach to split data. This function can be applied in three ways:

- It split D into smaller units, *i.e.*, $D = \{D_1, D_2, \ldots, D_n\}$, this is to ease the extraction of X.
- It is applied to split X into different categories: enrollment (X_e), fee-for-service (X_f), referrals (X_r) and Update (X_u).
- It is also applied to each category of X to split it into smaller units, e.g., U_r from Federal Medical Centre Owo, X_e from University College Hospital Ibadan etc.
- V_a is the value for an attribute, a, which is an element of A

Fa is an information mapping function or searching function that maps an object from D to a value in Va, Fa: D ➔Va and also uses A to search for insights from X. X^C is the data of non-interest which is the complement of X, i.e., $X^C \subseteq D$ n is the number of elements in D, i.e., n = |D|

BDA is the big data analytics function.

$$BDA = [D{:}X, A, R, \{Va|a \in A\}, \{Fa|a \in A\}, \{GD\,|1, 2, \ldots, k\}, XC] \qquad (1)$$

Where:

a) GD (n, k) = (Combinatorial Principle)

b) GD is used to partition a set of n objects into k nonempty subsets.

c) Fa = {x∩ | ∀X∈D} (Principle of Intersection of Set)

d) Fa is used to search for x from the partitions of the set D.

D which represents big data stored in MongoDB, according to Ramsey Theory, contains patterns (both regular and irregular), A, which are used to identify the insights in D. Also, there exist some relations, R, between D and A. This is represented mathematically as:

$$(D, A) \in R \qquad (2)$$

This is read, D has the attributes, A. To extract X from D, D is partitioned into smaller units following from the Principle of Divide-and-Conquer D = {D1, D2, …, Dn} Since $X \subseteq D$, this means, the features of X are common with some features of D, then X can be extracted by taking the intersection between X and D

∟ X = {(X ∩ D1) U (X ∩ D2) U … U (X ∩ Dn)}
∟ X = {X ∩ (D1 U D2 U … U Dn)}
∟ X = {X ∩ }

But X is complex and highly unstructured and extracting the attributes, A of X would be difficult; hence, according to Divide-and-Conquer Approach, X must be partitioned such that each partition would have attributes, A.

$$X = X1, X1, X2, \ldots, Xn \qquad (3)$$

Hence, for all X1, X1, X2… Xn ∈ X, there exist (X, A) ∈ R from equation (2).

Each partition could be further partitioned into n elements, i.e.,
X1 = {x1, x1, x2, x3, …, xn} n≠0
Xp = {x1, x1, x2, x3, …, xm} m≠0
According to Ramsey Theory, each of the 'Xs' contains A which when identified, the analytics task becomes easier. Again, the identification of A makes it easier to identify the relations R amongst the data which correspond to the associations in D (see equation 2). As stated earlier, analytics becomes the analysis of R, which in turn correspond to the analysis of the associations in the 'Xs'.
Using the categories of data listed in Table 1, we classify our 'Xs' as Enrollment (Xe), Fee-for-service (i.e., claims) (Xf), Referrals (Xr) and Update (Xu).

$$X = \{Xe, Xf, Xr, Xu\} \quad (4)$$

and each category has different features, i.e.,

$$A = \{Ae, Af, Ar, Au\} \quad (5)$$

$$Ae = \{Ae1, Ae1, Ae2, \ldots, Aen\} \quad (6)$$

These (the Aes) are enrollment features: first name, family name, gender, date of birth, etc.

$$Af = \{Af1, Af1, Af2, \ldots, Afn\} \quad (7)$$

These (the Afs) are fee-for-service features: amount billed, procedure codes, diagnoses, etc.

$$Ar = \{Ar1, Ar1, Ar2, \ldots, Arn\} \quad (8)$$

These (the Ars) are referral features: referral code, secondary/tertiary health facility, primary health facility, etc.

$$Au = \{Au1, Au1, Au2, \ldots, Aun\} \quad (9)$$

These (the Aus) are update features: new facility, new HMO, spouse, child, etc. From Association Rules Mining, we have X⊔Y, where X is the set of features of Y This also implies that

$$A ∟ X \quad (10)$$

Then, substituting the value of A from equation 5 and that of equation 4 in equation 10
(Ae, Af, Ar, Au) ∟ (Xe, Xf, Xr, Xu) i.e.,
Ae ⊔ Xe, Af ∟ Xf, Ar ⊔ Xr, Au ⊔ Xu
(Ae1, Ae1, Ae2, …, Aen) ∟ Xe
(Af1, Af1, Af2, …, Afn) ⊔ Xf
(Ar1, Ar1, Ar2, …, Arn) ⊔ Xr
(Au1, Au1, Au2, …, Aun) ⊔ Xu

Again, from Association Rule Mining, a context is a triple (X, A, R), where X and R are sets and R ⊆ X × A. The elements of G are called objects, and the elements of A are called attributes. For an arbitrary x ∈ X, and a ∈ A, we note xRa, when x is related to a, i.e., (x, a) ∈ R. Then, equation 4 can be rewritten as

$$X = \{(Xe, Ae), (Xf, Af), (Xr, Ar), (Xu, Au)\} \quad (11)$$

But equation 5 can also be expressed as

$$A = Ae \cup Af \cup Ar \cup Au \quad (12)$$

Using the values of equations 6 through 9 in 12

$$A = \quad (13)$$

But A ⊔X in equation 10, then

$$X ⊔ \quad (14)$$

Equation 1 can now be expressed as:
BDA = [D(k): X(k), A, R, {Va | a∈A}, {Fa | a∈A}, {GD | 1, 2,…, k}, XC]
D(k) ≡ {X(k): A, R, {Va | a∈A}, {Fa | a∈A}, {GD | 1, 2,…, k}} U XC
X(k) ≡ {XA: A, R, {Va | a∈A}, {Fa | a∈A}, {GD | 1, 2,…, k}}

Assumptions for the Model
The model is based on the following assumptions:

a) That the repository is not empty, i.e., D ≠ {ø}.
b) That the size of the data stored in the repository is not zero, i.e., |D| ≠ 0.
c) That the data of interest is a subset of what is stored in the repository, i.e., X ⊆ D.
d) That there exists a relationship between the data of interest and the respective attributes, i.e., (X, A) ∈ R.
e) That the set of attributes is not empty, i.e., A ≠ {ø}.
f) That there is no relationship between the data of non-interest and the attributes of the data of interest, i.e., (XC, A) ∉ R.
g) That there exists a relationship between that data in the repository and the attributes of the data of interest, i.e., (D, A) ∈ R.
h) That there exist the data of interest in the repository, i.e., X ∩ D ≠ {ø}.
i) That there is no intersection between the data of interest and the data of non-interest, i.e., X ∩ XC = {ø}.
j) That the combination of data of interest and the data of non-interest is not empty, i.e., X U XC ≠ {ø}.

4. Results and Discussion

We implemented our model using Java Programming Language, MapReduce and MongoDB. The system comprised of four modules: user management, enrollment processing, referral processing and claims processing. It has a standalone application that is used for the capturing of the manually submitted data. It preprocesses and exports the data for actual transformation by the web-based application. In order to demonstrate the concept of big data analytics in the health insurance domain, the National Health Insurance Scheme in Nigeria was considered. The NHIS in Nigeria, unlike other countries, processes its data manually and this has led to delay in all its processes due to the recent upsurge in healthcare data in terms of the volume, velocity, variety and even veracity that is key in the healthcare industry.

To ascertain the accuracy of the model, the accuracy of each category of data (see Table 1) was tested. This was carried out via two sets of experiments using different sizes of data (71543, 115427, 279950, 396428). In the first instance, the number of nodes of the MapReduce framework in the rules generation phase was kept constant; while the reverse was the case in the second sets of experiments. From the data depicted in Table 2 and Table 3, and the graphs in Fig 5, Fig 6, the accuracy of the model tends to decrease as the volume (number of records in this case) of the data increases in the first set of experiments. The accuracy was always at the peak (90.47%) during the processing of the first case and dropped to the least value in the last case (83.36%). In the second set of experiments, accuracy increases as the volume of data increases. The accuracy was always the lowest (93.23%) during the processing of the first case and increased to the highest value (97.14%) in the last case. The accuracy tends to vary among the four categories of data. One major cause of this was the different file formats in which the data exist. For instance, the processing of the updates and claims data yielded the highest accuracy while that enrollment and referral lagged in the first set of experiments; but this was not seen in the second case. Based on this evaluation, the model has an average accuracy of 87.08% for the first of experiments and 95.22% in the second case. This was obtained after carrying out a different experiment which in totality processed over ten million (10,000,000) records with the prototype system which implemented the model.

Abbreviations used in the Evaluation

- NR: Number of Records
- CAE: Accuracy for the processing of Enrollment data at Constant number of nodes
- CAC: Accuracy for the processing of Claims data at Constant number of nodes
- CAR: Accuracy for the processing of Referrals data at Constant number of nodes
- CAU: Accuracy for the processing of Update data at Constant number of nodes
- VAE: Accuracy for the processing of Enrollment data at Varying number of nodes
- VAC: Accuracy for the processing of Claims data at Varying number of nodes
- VAR: Accuracy for the processing of Referrals data at Varying number of nodes
- VAU: Accuracy for the processing of Update data at Varying number of nodes
- CAA: Average Accuracy for the processing of data at Constant number of nodes
- VAA: Average Accuracy for the processing of data at Varying number of nodes
- Table 2. Accuracy for the Data Processing

Table 2. Accuracy for the Data Processing

NR	CAE	CAC	CAR	CAU	VAE	VAC	VAR	VAU
71543.00	89.33	91.63	88.73	92.17	92.3	92.3	93.2	95.12
115427.00	86.37	90.97	87.96	89.58	93.42	93.73	94.49	95.58
279950.00	81.89	86.89	87.19	87.11	94.86	96.6	95.91	97.48
396428.00	77.90	85.89	85.12	84.52	96.18	97.55	96.85	97.99

Table 3. Average Accuracy for the Model

Data Category	CAA	VAA
Enrollment	83.87	94.19
Claims	88.85	95.05
Referrals	87.25	95.11
Updates	88.35	96.54
Average	87.08	95.22

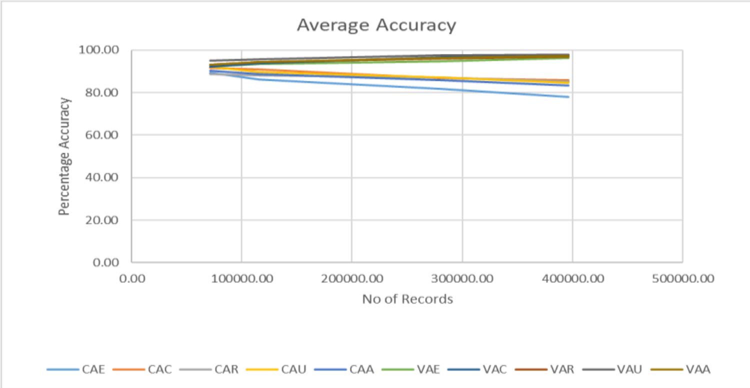


Fig 5. Graph for the accuracy of categories of NHIS data

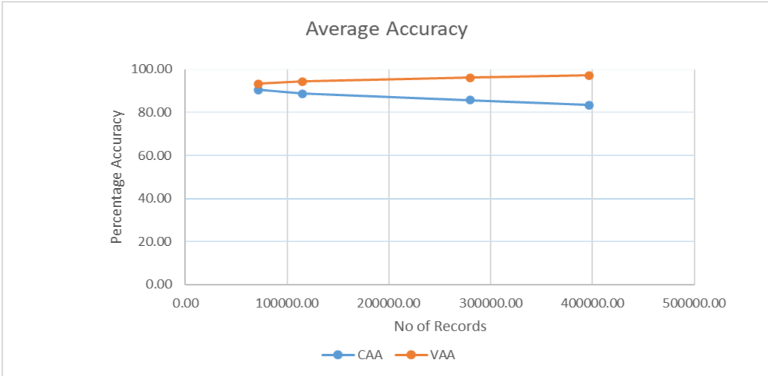


Fig 6. Graph for the average accuracy for the model

4.1 Discussion

The NHIS in Nigeria, unlike other countries, processes its data manually and this has led to delay in all its processes due to the recent upsurge in healthcare data in terms of the volume, velocity, variety and even veracity that is key in the healthcare industry. These delays in the NHIS have been a source of concern considering the prominent role the healthcare insurance play as a significant paymaster in the health sector. For instance, registration (new enrollment) and update (addition of a dependant, change of provider *etc.*) in NHIS takes ninety (90) to three hundred and sixty-five (365) days before it is completed [24], [26], [27]. Hence, many who are yet to register get discouraged by the harrowing and disheartening experiences of those who have started the process of documentation. The implication is that the number embracing the scheme tends to be diminishing as more exits are recorded. Service providers complain of excessive delay in processing their claims leading to delay in the payments of their bills which takes a period of six (6) to twelve (12) months [12], [24], [27].

Most importantly, delay in NHIS reduces access to healthcare, and the outcome of this has a severe negative effect on the health status of the populace due to delays in diagnosis and treatment. For NHIS to achieved its goal of providing easy access to healthcare to improve the quality of healthcare, the problem of delay must be tackled. Tackling delays would have a ripple effect as it would also tackle some of the challenges resulting from corruption. It is to this effect that this system was developed. Since delay is the major Achilles heel, the system was evaluated based on accuracy. The model was evaluated based on the four categories (see Table 1) of data: claims (fee-for-service), enrollment, referrals and updates. These data were collected in both image (jpg and png), pdf, excel, csv, docx and txt formats. Over ten million records were used in this whole process. The result of the evaluation is presented in the sections that follow.

From the results, it is evident that the accuracy of the model varies as the volume, velocity and variety of the data also varies. This variation affects the entire analytics process (*i.e.*, the extraction, preprocessing, analysis and visualization). This evaluation was done in two phases: keeping the nodes in the MapReduce framework to be constant during rules generation and varying the number of nodes in the MapReduce framework during rules generation. In the first instance, the evaluation shows that the average accuracy for the model was 87.08%. Much better and higher accuracy was obtained when the nodes in the MapReduce were increased as the data was increasing, and the value of accuracy at this point was 95.22%.

The accuracy decreases from 90.48% to 83.36% as the volume of the data increases from 71543 records to 396428 records in the first set of experiments. The reverse was the case in the second set of experiments as it increases from 93.23% to 97.14% as the volume of the data increases. The implication of this is that when the MapReduce was deployed with a constant number of nodes in the first set of experiments, the nodes are exposed to too much pressure which could lead to the generation of complex rules as the volume of the data increases. As the number of nodes was increased as a result of an increase in the volume of data, there was a significant improvement in the accuracy. This is because the processing algorithm is parallelized to allow the generation of rules for the processing of the different splits of the data to go on concurrently, thereby reducing the pressure that was exerted on few nodes as was the case with the first implementation.

The variety of data also affects the value of accuracy. As the volume and velocity of the data increase with varying formats of data from different sources, if the processing is maintained on the same number of nodes, the pressure on the nodes would be high, thereby decreasing the accuracy of the model. The accuracy decreases as the variety increases due to the varying number of attributes that are required for the processing and the different encoding systems employed by the different formats and sources of the data. This led to the loss of data which culminated into the generation of complex rules and the consequent reduction of accuracy of the model.

5. Conclusion

In conclusion, this paper developed a mathematical model for the analytics of healthcare insurance data using set theory for National Health Insurance Scheme data in Nigeria to address the delay in the existing manual data processing system. NHIS data is generated from different sources with varying formats and very high volume which otherwise is difficult to achieve manually. This has arm data scientists with a tool for processing both structured and unstructured data with high volume and velocity. Practically, the research has contributed to the development of a system that captures data in four (4) different format, generated at a higher rate with very large size so as to address the issues of data loss during data collection and storage. It has also, provided a platform for extracting meaningful insights from the data collected. The results show a highvalue accuracy. Thus, the processing time of NHIS data was reduced. This will enhance the flow of resources among stakeholders in the scheme and thus, a steady improvement in the structures, processes and outcomes leading to improvement in the quality of services rendered to beneficiaries by the facility would be attained.

Further researches should consider the issue of co-payment which requires a cover note such as drug prescription/dispense sheets, laboratory request/result sheet which analysing it would require handwriting recognition in order to extract the data before processing. Also, in this research we employed a non-invasive approach for our analytics. Further researches could employ invasive techniques such as deep learning to also address this problem. Again, evaluating stakeholders' satisfaction of the services of the service providers is another area that subsequent research effort could be geared to.

References

[1] Mirkin, B.: Core Concepts in Data Analysis: Summarization, Correlation, Visualization. Department of Computer Science and Information Systems, Birkbeck, University of London, Malet Street, London WC1E 7HX UK, 2010.

[2] Baldwin, D., Henderson, P.: The Importance of Mathematics to the Software Practitioner. IEEE Software, Vol. 19, No. 2, 112 – 111. (2002)

[3] Fang, R., Pouyanfar, S., Yang, Y., Chen, S. C. - Iyengar, S. S.: Computational Health Informatics in Big Data Age: A Survey. ACM Comput. Surv., Vol. 5, 1-35. (2016)

[4] Sun, Z., Wang, P. P.: A Mathematical Foundation of Big Data. Journal of New Mathematics and Natural Computation, Vol. 13, No. 2, 83-99. (2017)

[5] Pothuganti, A.: Big Data Analytics: Hadoop-Map Reduce & NoSQL Databases. International Journal of Computer Science and Information Technologies, Vol. 6, No. 1, 522-527. (2015)

[6] Ebenezer, G., Durga, S.: Big Data Analytics in Healthcare: A Survey. ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 8, 3645-3650. (2015)

[7] Famutimi, R. F.: Design and Implementation of In-Memory Technique for Managing Big Data Complexities. An Unpublished Ph.D. Thesis Submitted to the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria. (2018)

[8] Etobe, E. I., Etobe, U. E.: The National Health Insurance Scheme and Its Implication for Elderly Care in Nigeria. International Journal of Science and Research (IJSR), Vol. 4, No. 2, 128-132. (2015)

[9] Olaniyan, A. O.: Assessment of the Implementation of National Health Insurance Scheme (NHIS) in South-Western Nigeria. Unpublished PhD Thesis submitted to the Department of Public Administration, Obafemi Awolowo University, Ile-Ife, Nigeria. (2017)

[10] NHIS (National Health Insurance Scheme): National Health Insurance Scheme Operational Guidelines. 2013. Accessed on 01.06.2017 from http://www.nhis.gov.ng/images/stories/hmoregister/NHIS_OPERATIONAL_GUIDELINES.pdf.

[11] Moyano, L. G., Appel, A. P., de Santana, V. F., Ito, M., dos Santos, T. D.: GraPhys: Understanding Health Care Insurance Data through Graph Analytics. Proceeding of WWW'16 Companion, April 1115, 2016, Montréal, Québec, Canada, 227-230. (2016)

[12] Eteng, F. O., Ijim-Agbor, U.: Understanding the Challenges and Prospects of Administering the National Health Insurance Scheme in Nigeria. International Journal of Humanities and Social Science Research, Vol. 2, No. 8, 43-48. (2016)

[13] Egbunu, C. O., Oluoha, O., Malik, A. R., Yange, S. T., Atsanan, S. P.: Queue Management in Non-Tertiary Hospitals

for Improved Healthcare Service Delivery to Outpatients, International Journal of Applied Information Systems (IJAIS), Vol. 12, No. 31, 36-48, (2020).

[14] Das, S., Sismanis, Y., Beyer, K. S., Gemulla, R., Haas, P. J., McPherson, J.: Ricardo: Integrating R and Hadoop. SIGMOD'10, June 6–11, 2010, Indianapolis, Indiana, USA, 987-998. (2010)

[15] Ularu, E. U., Puican, F. C., Apostu, A., Velicanu, M.: Perspectives on Big Data and Big Data Analytics," Database Systems Journal, Vol. 3, No. 4, 3-14. (2012)

[16] Nalini, N., Suvithavani, P.: A Study on Data Analytics: Internet of Things & Healthcare. International Journal of Computer Science and Mobile Computing, Vol. 6, No. 3, 20-27. (2017)

[17] Chandola, V., Sukumar, S. R., Schryver, J.: Knowledge Discovery from Massive Healthcare Claims Data. KDD'13, August 11–14, 2013, Chicago, Illinois, USA, 1312-1320. (2013)

[18] Chen, C. L., Zhang, C-Y.: Data-Intensive Applications, Challenges, Techniques and Technologies: A survey on Big Data. Information Sciences, Vol. 275, 314–347. (2014)

[19] Fashoto, S. G., Owolabi, O., Sadiku, J., Gbadeyan, J. A.: Application of Data Mining Technique for Fraud Detection in Health Insurance Scheme Using Knee-Point K-Means Algorithm. Australian Journal of Basic and Applied Sciences, Vol. 7, No. 8, 140-144. (2013)

[20] LaBrec, P. A.: Linking Healthcare Claims and Electronic Health Records (EHR) for Patient Management – Diabetes Case Study. Treo Solutions-3M Health Information Systems, NY, USA, 1-17. (2014)

[21] Lin, C., Huang, L., Chou, S. T., Liu, C., Cheng, H., Chiang, I.: Temporal Event Tracing on Big Healthcare Data Analytics. In the Proceedings of the 2014 IEEE International Congress on Big Data, IEEE Computer Society Washington, DC, USA, 281-287. (2014)

[22] Li, J., Huang, K-Y., Jin, J., Shi, J.: A Survey on Statistical Methods for Healthcare Fraud Detection. Health Care Management Science, Vol. 11, 275-287. (2008)

[23] Bellazzi, R., Zupan, B.: Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines, International Journal of Medical Informatics, Vol. 77, No. 2, 81-97. (2008)

[24] Oyegoke, T. O.: Development of an Integrated Health Management System for National Health Insurance Scheme, An Unpublished M.Sc. Thesis Submitted to the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria. (2015)

[25] Borana, M., Giri, M., Kamble, S., Deshpande, K., Shubhangi, E. S.: Healthcare Data Analysis using Hadoop. International Research Journal of Engineering and Technology, Vol. 2, No. 7, 583-586. (2015)

[26] Alimi, O. M., Binuyo, O. G., Gambo, I. P., Jimoh, K.: Realtime National Health Insurance Scheme (RNHIS): Means to Achieve Health for All. International Journal of Computer Science, Engineering and Applications, Vol. 6, No. 2, 1-8. (2016)

[27] Oyegoke, T. O., Ikono, R. N., Soriyan, H. A.: An Integrated Health Management System for National Health Insurance Scheme in Nigeria. Journal of Emerging Trends in Computing and Information Sciences, Vol. 8, No. 1, 30-40. (2017)

[28] Yange, S. T., Soriyan, H. A., Oluoha, O.: A Fraud Detection System for Health Insurance in Nigeria, Journal of Health Informatics Africa, Vol. 6, No. 2, 64-73, (2019).

[29] Yange, S. T., Soriyan, H. A., Oluoha, O.: Design of a Data Analytics Model for National Health Insurance Scheme. Journal of Health Informatics Africa, Vol. 4, No. 1, 42-50, (2017).

[30] Yange, S. T., Soriyan, H. A., Oluoha, O.: A Schematic View of the Application of Big Data Analytics in Healthcare Crime Investigation. Journal of Health Informatics Africa, Vol. 4, No. 1, 32-41, (2017).

[31] Yange, S. T., Gambo, I. P., Ikono, R. N., Oluoha, O., Soriyan, H. A.: An Implementation of a Repository for Healthcare Insurance Using MongoDB, Journal of Computer Science and Its Applications, Vol. 27, No. 1, 33-51, (2020).

[32] Yange, S. T., Gambo, I. P., Ikono, R., Soriyan, H. A.: Multi-Nodal Implementation of Apriori Algorithm for Big Data Analytics using MapReduce. International Journal of Applied Information Systems (IJAIS), Vol. 12, No. 31, 8-28, (2020).

[33] Sheriff, C. I., Naqishbandi, T., Geetha, A.: Healthcare Informatics and Analytics Framework," In the Proceedings of the 2015 International Conference on Computer Communication and Informatics (ICCCI -2015), Jan. 08 – 10, 2015, Coimbatore, INDIA, 1-6. (2015)

[34] Agba, A. M., Ushie, E. M., Osuchukwu, N. C.: National Health Insurance Scheme (NHIS) and Employees' Access to Healthcare Services in Cross River State, Nigeria. Global Journal of Human Social Science, Vol. 10, No. 7, 9-16. (2010)

[35] Sarraf, S., Ostadhashem, M.: Big Data Spark Solution for Functional Magnetic Resonance Imaging. (2016)

[36] Karim, R., Sahay, R., Rebholz-Schuhmann, D.: A Scalable, Secure and Realtime Healthcare Analytics Framework with Apache Spark. In the Proceedings of 2nd Annual Insight Student Conference, NUIG, Ireland, Vol. 2, No. 1, 1-2. (2015)

[37] Karthika, I., Porkodi, K. P.: Fraud Claim Detection Using Spark. International Journal of Innovations in Engineering Research and Technology, Vol. 4, No. 2, 10-13. (2017)

[38] Kareem, S., Ahmad, R. B., Sarlan, A. B.: Framework for the Identification of Fraudulent Health Insurance Claims using Association Rule Mining. 2017 IEEE Conference on Big Data and Analytics (ICBDA), 99-104. (2017)

[39] Sun, Z., Wu, Z., Wang, P. P.: Calculus of Big Data.. BAIS No. 17005, Research Centre of Big Data Analytics and Intelligent Systems, PNG UoT, 2017. DOI: 10.13140/RG.2.2.26177.66407

[40] Duentsch, I., Guenter, G.: Rough Set Data Analysis: A Road to Non-Invasive Knowledge Discovery," Metoδos Publisher, Bangov, Bissendorf. (2000)

[41] Pawlak, Z.: Rough Set Theory and its Applications to Data Analysis. Cybernetics and Systems: An International Journal, Vol. 29, No. 7, 661-688. (1998)

[42] Zhang, H., Chen, G., Ooi, B. C., Tan, K. L., Zhang, M.: In-Memory Big Data Management and Processing: A survey. IEEE Transactions on Knowledge and Data Engineering, Vol. 27, 1920–1948. (2015)

[43] Riza, S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Śle, D., Benítez, J. M.: Implementing Algorithms of Rough Set Theory and Fuzzy Rough Set Theory in the R package "RoughSets"," Information Sciences, Vol. 287, 68-89. (2014)

[44] Zhang, J., Wong, S., Li, T., Pan, Y.: A Comparison of Parallel Large-Scale Knowledge Acquisition using Rough Set Theory on Different MapReduce Runtime Systems. International Journal of Approximate Reasoning, Vol. 55, No. 3, 896-907. (2014)

[45] Sarkar, S., Baidya, S., Maiti, J.: Application of Rough Set Theory in Accident Analysis at Work: A Case Study. In 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN- IEEE), 245-250. (2017)

[46] Das, A. K., Sengupta, S., Bhattacharyya, S.: A Group Incremental Feature Selection for Classification using Rough Set Theory Based Genetic Algorithm. Applied Soft Computing, Vol. 65, 400-411. (2018)