algorithms

MDPI

Article

A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management

Fernando López-Martínez [1,*,†,‡], Edward Rolando Núñez-Valdez [1,‡], Vicente García-Díaz [1,‡] and Zoran Bursac [2,‡]

[1] Department of Computer Science, Oviedo University, 33003 Oviedo, Spain; nunezedward@uniovi.es (E.R.N.-V.); garciavicente@uniovi.es (V.G.-D.)
[2] Department of Biostatistics, Florida International University, Miami, FL 33199, USA; zbursac@fiu.edu
* Correspondence: uo259897@uniovi.es; Tel.: +1-551-587-0112
† Current address: Sanitas Medical Center Corporate Offices, 8400 NW 33rd St, Doral, FL 33122, USA.
‡ These authors contributed equally to this work.

Received: 28 February 2020; Accepted: 21 April 2020; Published: 23 April 2020

Abstract: Big data and artificial intelligence are currently two of the most important and trending pieces for innovation and predictive analytics in healthcare, leading the digital healthcare transformation. Keralty organization is already working on developing an intelligent big data analytic platform based on machine learning and data integration principles. We discuss how this platform is the new pillar for the organization to improve population health management, value-based care, and new upcoming challenges in healthcare. The benefits of using this new data platform for community and population health include better healthcare outcomes, improvement of clinical operations, reducing costs of care, and generation of accurate medical information. Several machine learning algorithms implemented by the authors can use the large standardized datasets integrated into the platform to improve the effectiveness of public health interventions, improving diagnosis, and clinical decision support. The data integrated into the platform come from Electronic Health Records (EHR), Hospital Information Systems (HIS), Radiology Information Systems (RIS), and Laboratory Information Systems (LIS), as well as data generated by public health platforms, mobile data, social media, and clinical web portals. This massive volume of data is integrated using big data techniques for storage, retrieval, processing, and transformation. This paper presents the design of a digital health platform in a healthcare organization in Colombia to integrate operational, clinical, and business data repositories with advanced analytics to improve the decision-making process for population health management.

Keywords: decision support systems; population health management; big data; machine learning; deep learning; personalized patient care

1. Introduction

Colombia's health system is formed by the public sector and the private sector. The general social security system has two plans, contributory and subsidized. The contributory regime covers salaried workers, pensioners, and independent workers, with the subsidized plan covering anyone who cannot pay. Enrollment coverage increased from 96.6% in 2014 to 97.6% in 2015 [1].

The National Health Authority's primary purpose in Colombia is to improve the quality of healthcare and strengthening supervision, surveillance, and control of the health system. The 2015 Statutory Health Law No. 1751 places the responsibility for guaranteeing the right to health with the

health system and recognizes health as a fundamental social right and makes it the state's responsibility to pursue an approach in health promotion and disease prevention [2].

The health sector in Colombia supports all initiatives for implementing new technologies to prevent cardiovascular diseases, disabilities, and high-cost hospitalization cases [3]. There is a remarkable need to improve the prediction of the risk of conditions for the population through the integration and unification of massive volumes of data and the implementation of effective advance analytic solutions to improve the decision-making process and population health management in Colombia's population [4].

Keralty organization is formed by a group of insurance and health services companies with a global presence, which together develops an integral health model, whose purpose is to produce health and well-being to people throughout their lives. The organization is committed to keeping its users healthy and autonomous, focusing on prevention, identification, and management of health risks, control, and care of disease and dependency [5]. The organization is a leader in Colombia by providing integrated health services and is recognized for their human, scientific, technical, and ethical approach [6].

This paper presents how we can obtain value from a large volume of heterogeneous data generated by different data sources in healthcare, and the architecture implemented. The development of proper advanced data analytics methods such as machine learning and big data analytics to perform meaningful real-time analysis on the data to predict clinical complications before it happens and to support the decision-making process are challenging but much needed to handle the complexity of the data-driven problems we are currently facing.

1.1. Related Work

Several initiatives in Europe, Asia, and North America aim to develop healthcare digital platforms with collaborative access tools to allow the exchange and sharing of information and knowledge wherever and whenever needed throughout the attention process. This type of frameworks and architectures will allow maximum quality and efficiency for patient's care, and to provide appropriate attention to the patient's condition and risks.

Castilla and Leon, for example, implemented a digitalization of health services as a tool to increase the efficiency of the services and increase the security in the attention to patient [7]. A healthcare cyber-physical system assisted by cloud and big data is being developed in the department of computer science at Pace University in New York [8]. This system consists of a data integration layer, a data management layer, and a data-analytics service layer to improve the functioning of the healthcare system. In France, a group of researchers implemented a wearable knowledge as a service platform to cleverly manage heterogeneous data coming from wearable devices to assist the physicians in supervising the patient health [9]. Another interesting work was presented at the International Conference on Computational Intelligence and Data Science (ICCIDS 2018). The authors proposed a hybrid four-layer healthcare model to improve disease diagnostic [10]. In India, a centralized architecture for an end to end integration of healthcare systems deployed in the cloud environment was developed using fog computing [11].

Medical organizations are investing more and more in developing a healthcare platform that integrates data, applications, business processes, and user interfaces to gain knowledge and useful insights for clinical decisions, drug recommendation systems, and better disease diagnoses. Some other examples of big data applications in healthcare can be found in healthcare monitoring, where data captured from wearable devices can assist providers in managing symptoms of patients online and adjust their prescriptions [12]. An analytical platform called "MedAware" has been developed to detect errors in medical prescriptions and clinical errors, reducing the hospital admission and readmission in real-time [13]. In the healthcare prediction field, a healthcare system called "Gemini (Generalizable Medical Information analysis and Integration system)" was developed to collect, process, and analyze large volumes of clinical data and apply machine learning algorithms for performing predictive

analytics [14]. Other platforms have been implemented for genomics data analytics to generate predictions based on DNA molecular changes and mutations [15]. Another type of healthcare platform is related to the healthcare knowledge system, defined as the combination of clinical data and physician expertise to support clinical decision-making and diagnosis [16].

1.2. Why Big Data and Machine Learning?

Big data and machine learning are redefining healthcare goals for the future. Healthcare data are impacting the way disease research is performed, and the level of complexity in population health management is increasing as the traditional fee for service approach is transformed into the value-based care model [17,18].

Population health management is basically the aggregation of patient health data from multiple data sources, and the analysis and transformation into actionable insights to generate informed decisions to improve clinical and financial outcomes [19].

Big data technologies will allow us to bring large volumes of structured and unstructured data from disparate data sources into a data repository to be examined and analyzed. Machine learning models will assist in discovering insights from complex datasets with capabilities such as finding unseen patterns, making new predictions, and analyzing trends on health data. Machine learning is being used in a variety of clinical domains with the analysis of hundreds of clinical parameters resulting in effective and efficient models to improve the outcomes and quality of medical care models [20].

The implementation of this platform shows the enormous potential in using big data to individualize medical treatment, the opportunity for improving the lives of the patients, delivering better medical care, and reduced waste at an operational level [21]. Other chances for big data in healthcare for Keralty organization are:

- A physician would know before prescribing whether the patient is at high-risk to become dependent and different treatment plans can be selected based on this information.
- Psychosocial and clinical medical data could inform about the development of a chronic illness that can be properly diagnosed.
- The organization can use big data to understand how they are performing, the opportunities to improve clinical care, and their capacity to redesign care delivery to their patients.
- Using the platform's analytics component to improve the quality of care and patient experience at the lowest possible cost is core to the organization.
- Capturing streaming data and wearable data can provide to healthcare providers real-time insights about a patient's health that will allow them to improve their decision-making process for treatment and medication.
- Big data analysis can help the organization to deliver information that is evidence-based and can improve the efficiency, understanding, and implementation of the best practices associated with any disease.

In addition to the big data technologies used to build the platform, another essential component is the advanced analytic module of the platform. This module contains several machine learning algorithms to support clinical diagnosis. However, the organization should feel confident in these models and how they can be applied to specific use cases. These first models will alert providers to changes in high-risk conditions such as sepsis and hypertensive patients.

The main objective of this paper is to present the developed platform and its components to allow Keralty organization to derive better and more actionable insights from their data, i.e., to derive meaningful information from all these data in a way that allows them to improve care and lower costs needed for value-based reimbursement and business objectives while providing the highest quality care for population health management [22]. The goal is to be aligned with the triple aim framework developed by the Institute for Healthcare Improvement that describes an approach to optimizing healthcare system performance.The implementation of this platform intends to resolve

several problems in health services to assist patients and their families in managing their health by providing better access to healthcare services [23].

2. Proposed Digital Health Platform

Keralty organization currently have several information systems such as Health Information Systems (HIS), Lab Information Systems (LIS), Radiology Information Systems (RIS), Enterprise Resource Planning (ERP), and Customer Relationship Management (CRM), among others, in their ambulatory care centers, hospitals, and home care, which support their integrated health model. The information from these systems was not consolidated on a single platform, and its access and availability generated an operative load, which obstructs all health management processes and the support of clinical decisions for physicians. Consequently, we proposed the design and implementation of a healthcare, clinical, and business data repository with advanced analytic capabilities to consume machine learning prediction models to improve the decision-making process and population health management at the organization. The digital health platform conceptual framework is shown in Figure 1.

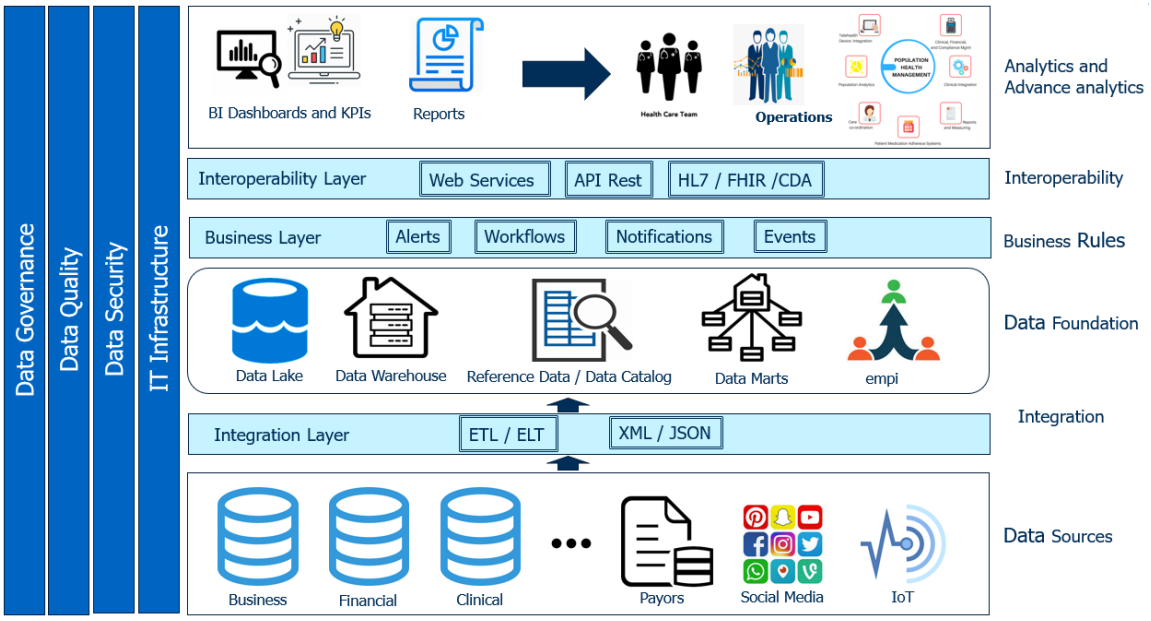

Figure 1. Conceptual Framework—Keralty Health Portal.

The implementation of the platform was an ambitious project that required integrating health information from disparate sources, building numerous technological and functional components, and the definition of IT management processes robust enough to support interoperability with other systems. The digital health information platform included patient-related data, Electronic Health Records (EHR), diagnostic reports, prescriptions, medical images, pharmacy records, research data, operational data, financial data, and human resources data.

This project was innovative and pioneered the designing and building of a comprehensive health digital platform for a healthcare organization in Colombia, with the patient being at the center of it and all of its information aggregated and summarized based on the standardized enterprise data repository. This information can be accessed quickly and intuitively when and where it is needed, hiding all technical complexity and providing longitudinal process management tools, as well as tools for decision support for professionals. The difference of this platform with other implementations was the development of a medical portal with a patient 360 view that uses data from the enterprise data repository to generate real-time early warning scores, patient surveillance, open API for hospitals integration, prediction of health risk patterns, high-risk markers, co-morbidity

detection to predict critical diseases, early diagnosis of diseases, treatment comparison with medical guidelines, and measurement of efficiency of specific drugs to provide the best quality of care.

The Digital Healthcare platform architecture can ingest data from over 50 different source systems at the granular level, including claims, clinical, financial, administrative, wearables, genomics, and socioeconomic data. Few platforms today can integrate that many heterogeneous data sources successfully. The platform can consume machine learning models on-demand without the need for further development. The data logic models are on top of the raw data and can be accessed, reused, and updated through open APIs without the need for clinical and business logic changes. The platform was able to integrate successfully structured and unstructured data. It is commonly seen that this type of platforms in the market is built to either integrate structured data or unstructured but few cases successfully integrate both. Open microservices APIs were created for operations such as authorization, identity management, interoperability, and data pipeline management. These microservices enable the development of third-party applications to interoperate with the platform.

2.1. Platform Architecture

The initial approach was to build a big data processing pipeline with a Microsoft Azure lambda architecture to support real-time and batch analytics. This approach is shown in Figure 2. This architecture has different mechanisms to consume data depending on the source and timing needed to generate insights. In addition, with this approach, we can have professionals with different skills working in parallel to build the platform.

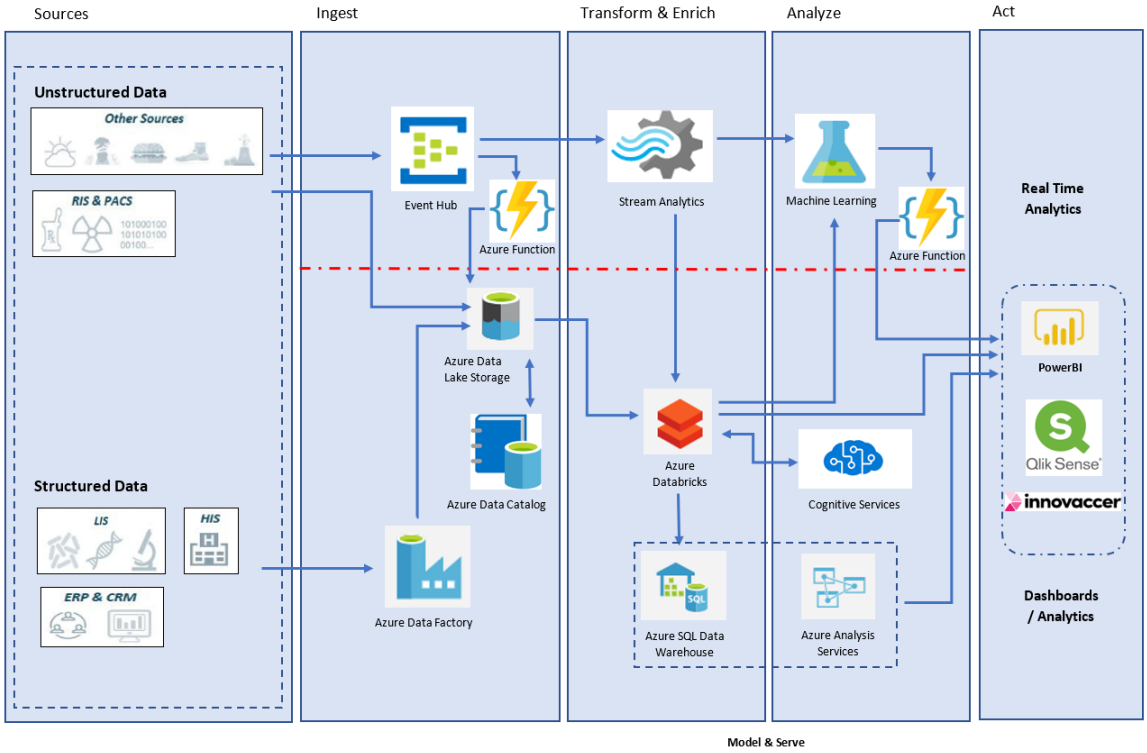

Figure 2. Azure Big Data and Machine Learning Lambda Architecture.

The architecture contains a batch layer, a real-time layer, and a serving layer. The batch layer is in charge of persistent storage and is able to scale horizontally. The real-time layer process streaming data and performs dynamic computation. The serving layer query data on the repositories and consume the prediction models.

From the infrastructure point of view, the platform offers the flexibility of being implemented in a hybrid environment, namely the cloud and the local data processing center, through the use of virtualization techniques, containers, and load balancing systems. The design of the infrastructure was

prepared to provide a flexible set of resources that can be used on-demand and based on the specific workload requirements. The infrastructure deployment relied heavily on automation to provide fluid operations.

2.2. Data Repository

An enterprise-wide staging repository for the big data analytics platform was considered. The data lake allows capturing data of any volume, type, and ingestion speed in one single place for storing heterogeneous data. This staging area included capabilities such as security, scalability, reliability, and availability. The data can be passed, processed directly from the staging area, or can be ingested to an enterprise data warehouse for historical load, preparation, and serve for BI and machine learning needs. This data warehouse repository has a scale-out architecture and massively parallel processing (MPP) engine.

Data models were developed to cover clinical, social, and healthcare program domains. Each model performs validations and processing on the data received, decoupling the processing and administration of the data from the source. These data models can also be extended to store additional attributes specific to the implementation, allowing these models to subscribe to certain types of messages, using the mapping and filtering options provided by the data processing pipelines. Once these subscriptions are created, the model will be loaded with all relevant messages to those who are subscribed and stored in the data lake.

For data storage, the data are loaded into a data warehouse with a daily refresh. This healthcare data repository contains a highly normalized data model for fast and efficient querying and analysis. This repository is read-only.

2.3. Integration and Interoperability

The platform provides a mechanism to integrate data from heterogeneous sources, define workflows to ingest data from different data stores, and transform and process data to data stores to be consumed by BI applications. A cloud-based data integration service is used to create these data-driven workflows and orchestrate all automation, transformation, and data movement in the platform. The main tasks this integration service should perform are: creation and scheduling of data pipelines to ingest data from different data sources, processing and transformation of the data, and store data in data stores such as data lakes or data warehouses.

Azure Data Factory automates and orchestrates the entire data integration process from end to end in the platform. We built the ETL (extract, transform, and load) pipelines with this Azure component. The data are extracted from the source locations, transformed from its source format to the target Azure data lake's schema, and loaded into Azure data lake and the data warehouse, where they can be used for analytics and reporting. Azure Data Factory defines control flows that execute various tasks in the transform and load process.

We used the mechanism called mapping data flows, combining control flows and data flows to build the data transformations with an easy-to-use visual user interface. These data flows are then executed as activities within Azure Data Factory pipelines. Data Factory is certified by HIPAA (Health Insurance Portability and Accountability Act), which protects the data while they are in use with Azure. In the data flow, we created transformation streams where we define the source data and create the graph with the transformations, schema operations such as derived column, aggregate, surrogate keys and selects, and the output settings.

2.4. Data Security and Privacy Model

In terms of security, the platform guarantees authentication, access control, and encryption capabilities. The security mechanisms of the platform can provide protection, alert monitoring, and support the OAuth 2.0 protocol for authentication with REST interfaces. ACLs are enabled on folders, subfolders, and files. The platform also provides encryption mechanisms to protect the data.

All these capabilities are accompanied by the implementation of enterprise security policies and regulatory compliance requirements.

2.5. Stream Analytics

The platform can handle mission-critical real-time data and offer end to end streaming pipelines with continuous integration and continuous delivery (CI-CD) services. Other capabilities such as in-memory processing, data encryption, and support of international security standards including HIPAA (Health Insurance Portability and Accountability Act), HITRUST (Health Information Trust Alliance), and GDPR (General Data Protection Regulation).

2.6. Advanced Analytics

The analytic data component consists of two areas: The first area is the BI models we develop for tactical, operational, and strategic decisions. The second area comprehends several prediction models that need to be developed. Currently, there are two prediction models developed by the authors of this paper to support population health management, specifically the diagnosis of sepsis and hypertension prediction [24,25]. These insights assist providers in the detection and tracking of chronic diseases. The machine learning component is used to build, test, consume, and deploy predictive analytic models on-demand and as requested for the organization. The platform provides self-service dashboards and visualizations that use data from the repositories to drive the decision-making process. The machine learning application layer is one of the essential layers of this platform.

Once the data are integrated, aggregated, and normalized in the system, the platform offers a tool to provide knowledge management through the business intelligence interface providing data analysis, design, and training of machine learning models, as well as development and management of results-based care indicators or population health management. The platform provides a tool where clinicians, researchers, and scientists can mine the data and get valuable information.

Machine learning models can be trained and customized in preconfigured data domains, allowing the storage of the results for future use. Data researchers and scientists can develop advanced tools to obtain information and value of the data stored in the solution, taking advantage of the model design, training, and validation component. We briefly present the predictive models implemented in the platforms.

- **Machine Learning Classification for a Hypertensive Population:** This prediction model evaluates the association between gender, race, BMI (Body Mass Index), age, smoking, kidney disease, and diabetes using logistic regression. Data were collected from NHANES datasets from 2007 to 2016 to train and test the model, a dataset of 19,709 samples with (83%) non-hypertensive individuals and (17%) hypertensive individuals. The results show a sensitivity of 77%, a specificity of 68%, precision on the positive predicted value of 32% in the test sample, and a calculated AUC of 0.73 (95% CI [0.70–0.76]). The model used to estimate the probability that a person will belong to the hypertensive or non-hypertensive class is:

$$p = \frac{e^{(\beta_0+\beta_1 gender+\beta_2 age+\beta_3 race+\beta_4 bmi+\beta_5 kidney+\beta_6 smoke+\beta_7 diabetes)}}{1+e^{(\beta_0+\beta_1 gender+\beta_2 age+\beta_3 race+\beta_4 bmi+\beta_5 kidney+\beta_6 smoke+\beta_7 diabetes)}}$$

We used the logistic regression classification model in this experiment to evaluate the importance of the risk factor variables and their relationship with the prevalence of hypertension among a nationally representative sample of adults ≥20 years in the United States (n = 19,759). The distribution of the samples by hypertensive patients, gender, and race is shown in Table 1.

Table 1. Number of samples by hypertensive class, gender, and race.

Hypertension, Adults 20 and over—2007–2016			
Class	**Gender**	**Race**	**n**
Non Hypertensive	**Female**	**Mexican American**	1269
		Non-Hispanic Black	1674
		Non-Hispanic White	**3674**
		Other Hispanic	951
		Other Race—Including Multi-Racial	864
	Male	**Mexican American**	1255
		Non-Hispanic Black	1599
		Non-Hispanic White	**3714**
		Other Hispanic	774
		Other Race—Including Multi-Racial	843
Hypertensive	**Female**	**Mexican American**	205
		Non-Hispanic Black	420
		Non-Hispanic White	**662**
		Other Hispanic	149
		Other Race—Including Multi-Racial	114
	Male	**Mexican American**	214
		Non-Hispanic Black	478
		Non-Hispanic White	**670**
		Other Hispanic	138
		Other Race—Including Multi-Racial	132
		Total	**19,799**

We computed chi-square test between each independent variable and the dependent variable to indicate the strength of evidence that there is some association between the variables. Chi-square was selected due to the categorical form of the data used in the model, and it is considered one of the best methods to estimate the dependency between the class and the features when the feature can take a fixed number of possible values that belong to a group or nominal category.

Table 2 shows the p-value for each variable; the null hypothesis is reject for any $p \leq 0.05$, while the null hypothesis is not rejected when $p > 0.05$. p-values for the variables GENDER, BMIRANGE_1, BMIRANGE_3, and KIDNEY_2 are not statistically significant at 0.05 alpha level; the clinical importance of these variables in the model for interpretation allows us to include them. We ran the model with and without the variables, and there were no significant changes in the accuracy score, positive predicted value rate, and true positive rate.

The training dataset was derived from a random sampling of 70% (13,831) of the extracted study population and the test sampling the remaining 30% (5928) to evaluate the model on the ground-truth that was never used for training. We ran the logistic regression model on the entire dataset to verify the accuracy score of the model.

Table 2. Chi2 test and *p*-value for the independent variables.

Chi-Squared between Each Indicator Variable and the Baseline for the Model				
Feature	**Description**	**Dummy**	***p*-Value**	**Score**
GENDER	**Male**	**GENDER_1**	0.1416446	2.160001
	Female	**GENDER_2**	0.1450268	2.123795
AGERANGE	**20–30**	**AGERANGE_1**	0.0000001	560.890568
	31–40	**AGERANGE_2**	0.0000001	299.675698
	41–50	**AGERANGE_3**	0.0000001	98.221463
	51–60	**AGERANGE_4**	0.0000035	21.520345
	61–70	**AGERANGE_5**	0.0000001	342.879412
	71–80	**AGERANGE_6**	0.0000001	1037.137074
RACE	**Mexican American**	**RACE_1**	0.0067797	7.330429
	Other Hispanic	**RACE_2**	0.0275756	4.854409
	Non-Hispanic White	**RACE_3**	0.0455912	3.996636
	Non-Hispanic Black	**RACE_4**	0.0000001	91.264812
	Other Race	**RACE_5**	0.0000278	17.562718
BMIRANGE	**Underweight = <18.5**	**BMIRANGE_1**	0.6730361	0.178071
	Normal weight = 18.5–24.9	**BMIRANGE_2**	0.000033	17.234712
	Overweight = 25–29.9	**BMIRANGE_3**	0.9174572	0.010741
	Obesity = BMI of 30 or greater	**BMIRANGE_4**	0.0006362	11.666854
KIDNEY	**Yes**	**KIDNEY_1**	0.0000001	58.963059
	No	**KIDNEY_2**	0.1872889	1.738816
SMOKE	**Yes**	**SMOKE_1**	0.0021759	9.394891
	No	**SMOKE_2**	0.0053461	7.758468
DIABETES	**Yes**	**DIABETES_1**	0.0000001	217.214128
	No	**DIABETES_2**	0.0000001	39.351672
	Borderline	**DIABETES_3**	0.0000051	20.798905

The Logistic Regression model uses the logit function to express the relationship of the risk factors as:

$$logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i$$

The probability of success can be expressed as:

$$p = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i)}}$$

where p is the predicted probability of having hypertension, X_i are the risk factors or independent variables, and β_i are the coefficients that are estimated by using the method of maximum likelihood and allow us to calculate the odds that, for every unit increase in X_i, the odds of having hypertension changes by e^{β}.

- **A neural network approach to predict early neonatal sepsis:** We developed a non-invasive neural network classification model for early neonatal sepsis detection. The data used in this study are from Crecer's Hospital center in Cartagena-Colombia. A dataset of 555 neonates with (66%) of negative cases and (34%) of positive cases was used to train and test the model. The study results show a sensitivity of 80.32%, a specificity of 90.4%, precision on the positive predicted value of 83.1% in the test, sample and a calculated area under the curve of 0.925 (95% Confidence Interval [91.4–93.06]). The neural network architecture can be seen in Figure 3.

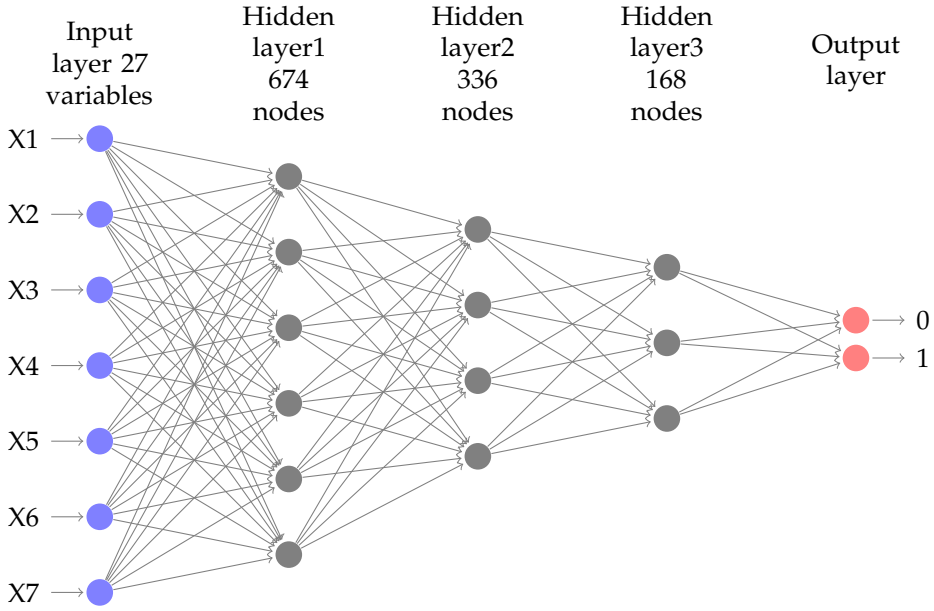

Figure 3. Multilayer Perceptron Architecture.

Table 3 shows the parameters of the architecture. Labels X1–X7 are informative only, and the input size is 27 variables.

Table 3. Model architecture parameters.

Model Architecture Parameters	
Parameter	**Value**
Input Dimension	27
Num Output classes	2
Num Hidden Layers	3
Hidden Layer1 Dimension	674
Activation Func Layer1	Relu
Hidden Layer2 Dimension	336
Activation Func Layer2	Relu
Hidden Layer3 Dimension	168
Activation Func Layer3	Relu
Minibatch size	8
Num samples to train	388
Num minibatches to train	48
Loss Function	cross entropy with softmax
Eval Error	Classification error
Learner for parameters	momentum sgd
Eval Metrics	Confusion Matrix, AUC

The model used an anonymous dataset from a private medical institution in Cartagena, Colombia, from 2016 to 2017. Demographic, laboratory data, blood pressure, and body measures data were part of the dataset. This dataset includes cases of live newborns of ages inferior to 72 h with a diagnosis of early neonatal sepsis by clinical criteria and laboratory blood cultures. Control cases were part of the dataset including all newborns healthy by clinical diagnosis and who returned healthy for a follow up at 72 h.

This retrospective study includes 186 cases and 368 controls based on a case-control relationship of 1:2 with a 95% trust factor and power of 80%. Bivariate analysis and logistic regression were performed to detect the variables associated with early sepsis, and the statistical significance was considered at the alpha level of 0.05.

This model considered nine sociodemographic, fourteen obstetric, nine neonatal, and four maternal infectious related pathology variables. Table 4 shows the quantitative sociodemographic variables, Table 5 shows the qualitative sociodemographic variables, Table 6 shows the quantitative neonatal variables, Table 7 shows the qualitative neonatal variables, Table 8 shows the quantitative obstetric variables, Table 9 shows the qualitative obstetric variables, and Table 10 shows the qualitative maternal infections of the cases and controls.

A bivariate chi-square test with correction was performed to the qualitative variables to find a statistical association between the independent variable and the possibility to develop early neonatal sepsis. For continuous variables, the Mann–Whitney U test was performed. From this statistical analysis, it is essential to show that we did not find significant statistical evidence for the variables age, start of marital status at younger than 18 years old, gender, APGAR (Appearance, Pulse, Grimace, Activity, and Respiration) value less than 7 after 1 and 5 min, the number of pregnancies, and the type of birth. Prenatal control is not associated with the case of sepsis; however, assisting to five prenatal controls are associated with the protection to avoid the appearance of early neonatal sepsis. There was no evidence with the variables IUGR (Intrauterine Growth Restriction) background and multiple pregnancies. Twenty-seven (27) variables were selected as input variables for our artificial neural network architecture.

Table 4. Quantitative sociodemographic variables in cases (186) and controls (369).

Quantitative Socio Demographic Variable	Cases				Controls				
	Mean	Median	SD	RIQ	Mean	Median	SD	RIQ	*p*-Value
Age	23.93	23.5	4.99	20–26	24.22	23	6.19	19–28	0.793
Onset of sexual activity	16.06	16	0.945	15–17	15.6	16	0.971	15–16	0.0001

Table 5. Qualitative sociodemographic variables in cases (186) and controls (369).

Qualitative Socio Demographic Variable	Categories	Cases		Controls		X2	*p*-Value
		N	%	N	%		
Teen Mother	Yes	15	8.1	69	18.7	10.88	0.001
	No	171	91.9	300	81.3		
Health Regimen	Government	183	98.4	349	94.6	4.51	0.041
	Commercial	3	1.6	20	5.4		
Origin	Rural	42	22.6	5	1.4	71.87	0.00001
	Urban	144	77.4	364	98.6		
Marital Status	Married or in common law married	128	68.8	101	27.4	87.64	0.00001
	Single, divorced or widow	58	31.2	268	72.6		
Level of education	Elementary School	86	46.2	80	21.7	35.57	0.00001
	High School	100	53.8	289	78.3		
Start of Marital status life younger than 18 yo	Yes	178	95.7	357	96.7	0.39	0.531
	No	8	4.3	12	3.3		
Start of Marital status life younger than 16 yo	Yes	47	25.3	147	39.8	11.54	0.001
	No	139	74.7	222	60.2		

Table 6. Quantitative Neonatal variables in cases (186) and controls (369).

Quantitative Neonatal Variable	Cases				Controls				
	Mean	Median	SD	RIQ	Mean	Median	SD	RIQ	*p*-Value
New born weight in grams	2639.9	2768.5	546.5	2500–3020	3202.4	3224	412.1	2950–3500	0.0001
APGAR after 1 min of birth	7.73	8.0	0.611	8.0	8.09	8.0	0.598	8.0	0.0001

Table 7. Qualitative Neonatal variables in cases (186) and controls (369).

Qualitative Neonatal Variable	Categories	Cases		Controls		X2	*p*-Value
		N	%	N	%		
Premature	Yes	100	53.8	25	6.8	156.4	0.0001
	No	86	46.2	344	93.2		
Gender	Male	109	58.6	202	54.7	0.748	1.672
	Female	77	41.4	167	45.3		
Less than 1500 grams	Yes	11	5.9	2	0.5	15.6	0.00001
	No	175	94.1	367	99.5		
Less than 2500 grams	Yes	44	23.7	9	2.4	64.44	0.00001
	No	142	76.3	360	97.6		
APGAR less than 7 after 1 min of birth	Yes	2	1.1	3	0.8	0.095	0.999
	No	184	98.9	366	99.2		
APGAR less than 7 after 5 min	Yes	4	2.2	9	2.4	0.045	0.999
	No	182	97.8	360	97.6		
Respiratory distress	Yes	89	47.8	27	7.3	122.8	0.0001
	No	97	52.2	342	92.7		

Table 8. Quantitative Obstetric variables in cases (186) and controls (369).

Quantitative Obstetric Variable	Cases				Controls				
	Mean	Median	SD	RIQ	Mean	Median	SD	RIQ	*p*-Value
Gestational age at the time of birth	35.6	36.0	3.47	34–39	38.4	39.0	1.62	38–39	0.0001
Number of prenatal controls	4.08	5.0	1.83	3.75–5.0	4.32	5.0	1.83	4–5.0	0.002
Number of pregnacies	1.77	1.0	1.15	1.0–2.0	1.6	1.0	1.15	1–2.0	0.076
Number of births	1.04	1.0	1.03	0–1	0.7	1.0	1.03	0–1	0.0001
Numbers of C-sections	0.65	1.0	0.68	0–1	0.76	1.0	0.68	0–1	0.029

Table 9. Qualitative Obstetric variables in cases (186) and controls (369).

Qualitative Obstetric Variable	Categories	Cases		Controls		X2	*p*-Value
		N	%	N	%		
Type of birth	Vaginal	98	52.7	162	43.9	3.833	0.05
	C-Section	88	47.3	207	56.1		
IUGR Background	Yes	5	2.7	13	3.5	0.275	0.6
	No	181	97.3	356	96.5		
Assistance for prenatal control	Yes	165	88.7	318	86.2	0.702	0.402
	No	21	11.3	51	13.8		
Assistance for at least 4 prenatal control	Yes	140	75.3	301	81.6	3.01	0.083
	No	46	24.7	68	18.4		
Assistance for at least 5 prenatal control	Yes	105	56.5	254	68.8	8.301	0.004
	No	81	43.5	115	31.2		
Premature rupture of membrane with more than 18 hours	Yes	95	51.1	17	4.6	165.7	0.00001
	No	91	48.9	352	95.4		
Chorioamnionitis	Yes	23	12.4	3	0.8	36.96	0.00001
	No	163	87.6	366	99.2		
Premature membrane rupture with more than 6 hours	Yes	161	86.6	194	52.6	61.96	0.0001
	No	25	13.4	175	47.4		
Multiple Pregnacies	Yes	2	1.1	10	2.7	0.39	0.353
	No	184	98.9	359	97.3		

Table 10. Qualitative maternal infections variables in cases (186) and controls (369).

Qualitative Maternal Infections Variables	Categories	Cases		Controls		X2	*p*-Value
		N	%	N	%		
Maternal Fever	Yes	67	36.0	40	10.8	50.38	0.0001
	No	119	64.0	329	89.2		
Yeast Infections	Yes	31	16.7	15	4.1	25.83	0.0001
	No	155	83.3	354	95.9		
Sexualy transmitted disease history	Yes	27	14.5	7	1.9	34.24	0.0001
	No	159	85.5	362	98.1		
Urinary Tract Infections	Yes	11	5.9	9	2.4	4.29	0.0381
	No	175	94.1	360	97.6		

In terms of computational timing, It is difficult to evaluate the complexity and timing of a machine learning algorithm. However, based on the algorithmic complexity, we can measure the time performance in terms of its training time complexity using big O notation because the classification time of the models can vary depending on the stress in the computational performance and power. In terms of timing, the classification prediction with the trained models is less than 1 s. The time complexity of the logistic regression could be expressed as $O((f+1)csE)$, where f is the number of features (+1 because of bias), c is the number of possible outputs, s is the number of samples, and E is the number of epochs to run. For the neural network approach, $O(pnl_1 + nl_1nl_2 + \dots)$, where p is the number of features and nl_i is the number of neurons at layer i in a neural network [26].

3. Actual Platform Benefits

The implementation of the platform became the digital healthcare ecosystem for the organization. The organization can populate workflow information systems with critical decision-making insights, accurate and reliable healthcare data that significantly increased the value of the healthcare outcome to patients and care providers. This platform delivers significant benefits to the organization, such as physicians having an intelligent application that can be configured to their preferences and optimized to their disciplines, patients receiving more personalized care, an improvement in healthcare workflow and patient care, and personalized care for physicians and patients.

We describe in the following subsections several use cases that effectively present the change and digital transformation of the organization with the implementation of the platform.

3.1. Reduce Total Cost of Care for Care Coordination

With a robust data analytic component, the organization was able to prioritize opportunities for improvement and to improve the way care is coordinated and delivered throughout its network of hospitals and medical facilities. The results include a considerable increase in financial results in just six months.

The organization uses the platform to generate timely, meaningful, and actionable data to drive change and improve the quality of care for patients. The organization uses the data for risk-stratification of the network's population, prioritization of the care coordination activities, and prevention activity's interventions. Risk stratification was completed for all patients, enabling care managers to identify individuals at various risk levels for unnecessary services and high-cost utilization, improving patient outcomes and experience. The analytical component also reduces unnecessary visits, facilitates access to specialty care and community-based services, and achieves healthcare outcomes. Other benefits include 3% increase in the detection of high-risk patients with primary care, 20% increase in the number of patients with ongoing care managed, and 10% percent reduction in emergency department utilization per member among care managed patients.

3.2. Self-Service Analytic

As described in this paper, the healthcare platform combines and standardizes data across different source systems to provide actionable insights in a single platform. The platform integrates data from different sources, such as claims data, cost data, financial data, clinical data, and other patient data. With self-service analytics, the organization increases the number of users accessing the analytic component, improving data visibility and providing actionable insights to improve patient outcomes.

3.3. Reduce Deaths from Sepsis

The organization improved sepsis mortality rates and improving care outcomes by using the advanced analytic component of the platform. Sepsis impacts almost 1.7 million adults in the U.S. and is responsible for nearly 270,000 annual deaths. One-third of all hospital deaths are patients with sepsis [27]. The machine learning prediction model used in the platform was developed by one of the authors of this paper, as described before. It is still too early to mention the results of the utilization of this feature. However, the goal of the organization is to reduce its sepsis mortality rate, the costs of the creation of its sepsis care transformation team, and the implementation of an evidence-based sepsis care practice.

3.4. Discussion and Limitations

The digital health platform helps Keralty organization with closing the gaps between multiple datasets, improving clinical benefits, improving patient's lives, supporting better decision-making to manage larger populations, and improving overall health outcomes. However, the need for algorithms with high accuracy in medical diagnosis is still a challenge that needs to be improved precisely and efficiently [28]. The increasing complexity of building end-to-end platforms to integrate disparate systems and to apply machine learning techniques in specific areas such as computer vision, natural language processing, reinforcement learning, and other generalized methods present many challenges when forming the interdisciplinary team needed and the set of technological components used for the implementation.

Some challenges should be considered in the design and implementation of machine learning projects for healthcare. One of the most critical challenges requires algorithms that can answer causal questions. These questions are beyond classical machine learning algorithms because they require a formal model of interventions [29]. To address this type of question from the analytical component of the platform, we need to learn from data differently and to gain knowledge in causal models to understand how machine learning algorithms need to be trained. Another challenge is to create reliable outcomes from heterogeneous data sources with the participation of SME (Subject Matter Experts) who understand the disease; the machine learning predictive accuracy and correct clinical interpretation depend on the criteria and context of the disease. Providers and machine learning engineers should work together on model interpretability and applicability. Machine learning implementation is not an easy task; the selection of predictive features and optimization of hyperparameters is another challenge that needs to be mastered to implement models that provide useful insights [30]. The success and meaningful use of these algorithms, and their integration into the platform depends on the accuracy of the models and their interpretability.

4. Results of Advanced Analytics

After training and testing the logistic regression model for predicting hypertension, we generated some evaluation metrics to evaluate the classifier. Table 11 shows the confusion matrix with the classification results, include the true positive value (730), true negative value (3407), false negative (216), and false positive value (1575). The classification report in Table 12 shows the calculated precision and sensitivity.

Table 11. Confusion matrix.

		Predicted	
		Non-Hypertensive	**Hypertensive**
True	**Non-Hypertensive**	3407	1575
	Hypertensive	216	730

Table 12. Classification report.

Classification Report				
	Precision	**Recall**	**f1-Score**	**Support**
Non-Hypertensive	0.94	0.68	0.79	4982
Hypertensive	0.32	0.77	0.45	946
avg/total	0.84	0.7	0.74	5928

The test sampling of 5928 contains 4982 (84%) non-hypertensive and 946 (16%) hypertensive patients. The model shows a sensitivity of 730/946 = 77% and a specificity of 3407/4982 = 68%. The precision of the model was 730/2305 = 32% and the negative predicted value 3407/3623 = 94%. The false negative rate of the model was 216/946 = 22%. The model was better at identifying individuals who will not develop hypertension than those who will develop hypertension.

For the neural network approach to predict early neonatal sepsis, Table 13 shows the confusion matrix with the classification results of actual class label vs. the predicted ones, including the true positive value (49), true negative value (95), false negative (12), and false positive value (10).

Table 13. Confusion matrix.

		Predicted	
		Non-Sepsis	**Sepsis**
True	**Non-Sepsis**	95	10
	Sepsis	12	49

The classification report in Table 14 shows the precision and sensitivity. The sensitivity of the model is moderately acceptable due to the imbalanced testing dataset, and there is still a high number of false negatives.

Table 14. Classification report.

Classification Report			
True Positive	**False Negative**	**Precision**	**Accuracy**
49	12	0.83	0.867
False Positive	**True Negative**	**Recall**	**f1-score**
10	95	0.803	0.817
Positive Label: 1		**Negative Label: 0**	

A sensitivity of 80.3% and a specificity of 90.4% show that the model might be useful for detecting positive cases, and the true negative rate shows that the model is also efficient at identifying negative cases. The high precision value of 83.1% and the AUC of 0.925 confirm the adequacy of the model as a preliminary screening tool. The percentage of positive cases shows that the model works better than random guessing and the conditional probability of negative test results is considerably low.

The accuracy of 86.74% shows that the model correctly identifies negative cases and positive cases based on the characteristics of the dataset and the small number of cases examined.

5. Comparison with Other Platforms

A review of several healthcare platforms shows that the architecture presented in this paper covers all the categories from integration, interoperability, security care, and advanced analytics. Generally, other implementations only focused on one specific area, as shown in Table 15 and taken from the International Conference on Computational Intelligence and Data Science (ICCIDS 2018) and a healthcare frameworks review proposed in the Journal of King Saud University [31].

Table 15. Comparison of healthcare big data platforms.

Author and Year	Patient Centric	Predictive Analysis	Real Time Monitoring	Improve Treatment	Interoperability	Workflow and Rules	Pop Health	Patient 360
Our Health Platform	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Raghupathi et al. (2014) [32]	Yes	No	Yes	Yes	Partial	No	Partial	No
Patel et al. (2016) [33]	Yes	Yes	Yes	Yes	Partial	Partial	Yes	No
Chawla et al. (2013) [34]	Yes	Yes	No	Yes	Partial	Partial	Yes	No
Abinaya et al. (2015) [35]	Partial	Yes	Partial	Yes	Yes	Partial	No	No
Balladini et al. (2015) [36]	Yes	No	Yes	Yes	Partial	Yes	No	No
Belle et al. (2015) [37]	Partial	No	Yes	Yes	Partial	Yes	No	No
Mezghani et al. (2015)	Partial	No	Yes	Yes	Yes	Partial	No	No
Chen et al. (2017) [38]	Yes	Partial	Yes	Yes	Yes	Partial	Yes	No

We designed and implemented a healthcare platform using big data technologies with actionable insights to augment human decision-making at the organization impacting the population's health, public health, and to capture social determinants of health. This platform comprehends all the features we use in the comparison. Raghupathi et al. reported a conceptual architecture to present big data analytic outlines in healthcare with no predictive analytic capabilities and no patient 360 view. Patel et al. designed a big data architecture platform to improve data aggregation in the healthcare industry and to provide a reduction in healthcare cost, predicting analytic, preventive care, and drug discovery capabilities but without patient 360 view capabilities. Chawla et al. presented a patient-centric healthcare framework—Collaborative Assessment and Recommendation Engine (CARE)—to improve patient-centric treatment and diagnosis without real-time monitoring and 360 view capabilities. Abinaya et al. implemented a fascinating e-Health service application for diagnosing heart diseases. Balladini et al. designed a real-time architecture of big data for Francisco Lopez Lima Hospital in Argentina to process physiological data. This platform did not include predictive analytic and patient 360 view. Belle et al. implemented a genomic data processing platform that provides image analytic and signal processing of psychological data. Mezghani et al. designed a big data platform for integrating heterogeneous wearable data in healthcare for real-time monitoring and diagnosis. Lastly, Chen et al. presented a real-time big data platform to improve communication and collaboration between patients and providers, increasing the quality of care that clinical teams can provide.

6. Conclusions and Future Work

This paper provides details of an optimized and secure healthcare platform that revolutionizes the healthcare industry in Colombia by providing better information to patients and care teams. The use of this technology reduces the costs associated with healthcare.

The proposed digital health platform allows us to address population health challenges, to understand better patient's health, and to find hidden patterns that traditional data analytics fail to

find. The organization can use unified patient-generated data, financial data, and socioeconomic data to detect patterns and to discover a group of patients who share similar health behavior. The analysis of clinical and non-clinical data allows predicting patient's health with better accuracy. The platform also allows better health discoveries and actions based on treatment history for individuals and groups of patients.

Keralty organization recognized that better care coordination was required for patients receiving care. The organization wanted to improve quality outcomes, provider engagement and recruitment, and its own economic health. To meet these objectives, the organization focuses on clinician engagement and organizational alignment, ensuring widespread access to meaningful, actionable data, and the use of the healthcare analytics platform to inform decisions and drive improvement. Keralty believes the use of machine learning will be one of the most important, life-saving technologies ever introduced to the organization. We believe the opportunities are virtually limitless for the platform to improve and accelerate clinical, workflow, and financial outcomes.

More future work needs to be done on the platform to continue improving all the benefits for the entire organization. Tools for performing knowledge discovery process will be added to the ecosystem. The organization is planning to start the implementation of prescriptive analytics models to assist the organization in making smarter decisions in population health management. The architecture team will look at the possibility of implementing Map/Reduce-based computations for processing data with high scalability and to execute low latency and high concurrency analytical queries on top of Hadoop clusters.

Author Contributions: Conceptualization, F.L.-M., V.G.-D. and E.R.N.-V.; Methodology, F.L.-M., V.G.-D. and E.R.N.-V.; Software, F.L.-M.; Validation, F.L.-M., V.G.-D., Z.B. and E.R.N.-V.; Formal Analysis, F.L.-M.; Investigation, F.L.-M., V.G.-D., Z.B. and E.R.N.-V.; Resources, F.L.-M.; Data Curation, F.L.-M. and Z.B.; Writing—Original Draft Preparation, F.L.-M., V.G.-D. and E.R.N.-V.; Writing—Review and Editing, F.L.-M., V.G.-D., Z.B. and E.R.N.-V.; Visualization, F.L.-M., Z.B. and E.R.N.-V.; Supervision, F.L.-M. and V.G.-D.; Project Administration, V.G.-D. and E.R.N.-V.; Funding Acquisition, F.L.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This document presents an independent study supported by the company Sanitas USA. The points of view expressed are those of the authors and not necessarily those of Sanitas USA. We thank Ivan Murcia VP of Healthcare Services at Sanitas USA and Santiago Thovar, CIO at Keralty who provided insight and expertise that greatly assisted the study.

Conflicts of Interest: The authors declare no conflict of interest. The founders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ACL	Access Control List
BI	Business Intelligence
CRM	Customer Relationship Management
EHR	Electronic Health Record
ERP	Enterprise Resource Planning
GDPR	General Data Protection Regulation
HIS	Hospital Information System
HIPAA	Health Insurance Portability and Accountability Act
HITRUST	Health Information Trust Alliance
LIS	Lab Information System
MPP	Massive Parallel Computing
RIS	Radiology Information System
REST	Representational State Transfer

References

1. Glassman, A.; Giuffrida, A.; Escobar, M.L.; Giedion, U. Chapter 1 Colombia: After a Decade of Health System Reform. In *From Few to Many*; Inter-American Development Bank: Washington, DC, USA, 2009; Volume 1, pp. 1–13.
2. Ruíz, F.; Gaviria, A.; Norman, J. Plan Decenal de Salud Pública. *Bogotá* **2020**, in press.
3. Legido, H.; Lopez, P.A.; Balabanova, D.; Perel, P.; Lopez-Jaramillo, P.; Nieuwlaat, R.; Schwalm, J.D.; McCready, T.; Yusuf, S.; McKee, M. Patients' knowledge, attitudes, behaviour and health care experiences on the prevention, detection, management and control of hypertension in Colombia: A qualitative study. *PLoS ONE* **2015**, *10*, e122112. [CrossRef]
4. Lopez, F.E.; Bonfante, M.C.; Arteta, I.G.; Baldiris, R.E. IoT and big data in public health: A case study in Colombia. In *Protocols and Applications for the Industrial Internet of Things*; IGI Global: Hershey, PA, USA, 2018; pp. 309–321, ISBN 978-1-5225-3806-6.
5. Dennis, R.J.; Caraballo, L.; García, E.; Rojas, M.X.; Rondon, M.A.; Pérez, A.; Aristizabal, G.; Peñaranda, A.; Barragan, A.M.; Ahumada, V. Prevalence of asthma and other allergic conditions in Colombia 2009–2010: A cross-sectional study. *BMC Pulm. Med.* **2012**, *12*, 12. [CrossRef]
6. About Keralty. Available online: https://www.keralty.com/en/about-keralty (accessed on 27 January 2020).
7. León, G.R. Digitalización de Historia Clínica. Available online: https://contrataciondelestado.es/wps/wcm/connect/3236c434-7ce1-484f-bb50-b8942bdc7d66/DOC20190314132936Estandar_digitalizacion_SACYL-+9.pdf?MOD=AJPERES (accessed on 27 January 2020).
8. Zhang, Y.; Qiu, M.; Tsai, C.W.; Hassan, M.M.; Alamri, A. Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. *IEEE Syst. J.* **2017**, *11*, 88–95. [CrossRef]
9. Mezghani, E.; Exposito, E.; Drira, K.; Da Silveira, M.; Pruski, C. A Semantic Big Data Platform for Integrating Heterogeneous Wearable Data in Healthcare. *J. Med. Syst.* **2015**, *39*. [CrossRef]
10. Kaur, P.; Sharma, M.; Mittal, M. Big Data and Machine Learning Based Secure Healthcare Framework. *Proc. Procedia Comput. Sci.* **2018**, *132*, 1049–1059. [CrossRef]
11. Thota, C.; Sundarasekar, R.; Manogaran, G.; Varatharajan, R.; Priyan, M.K. Centralized Fog Computing security platform for IoT and cloud in healthcare system. In *Fog Computing: Breakthroughs in Research and Practice*; IGI Global: Hershey, PA, USA, 2018; pp. 365–378, ISBN 978-1-5225-5650-3.
12. Edet, R.; Afolabi, B. Prospects and Challenges of Population Health with Online and other Big Data in Africa. *Adv. J. Soc. Sci.* **2019**, *6*, 57–63. [CrossRef]
13. MedAware—Using AI to Eliminate Prescription Errors—Digital Innovation and Transformation. Available online: https://digital.hbs.edu/platform-digit/submission/medaware-using-ai-to-eliminate-prescription-errors/ (accessed on 8 March 2020).
14. Ling, Z.J.; Tran, Q.T.; Fan, J.; Koh, G.C.H.; Nguyen, T.; Tan, C.S.; Yip, J.W.L.; Zhang, M. GEMINI: An integrative healthcare analytics system. *Proc. VLDB Endow.* **2014**, *7*, 1766–1771. [CrossRef]
15. Manogaran, G.; Thota, C.; Lopez, D.; Vijayakumar, V.; Abbas, K.M.; Sundarsekar, R. *Big Data Knowledge System in Healthcare*; Springer: Cham, Switzerland, 2017; pp. 133–157. [CrossRef]
16. Bates, D.W.; Saria, S.; Ohno-Machado, L.; Shah, A.; Escobar, G. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs* **2014**, *33*, 1123–1131. [CrossRef]
17. Farooqi, M.M.; Shah, M.A.; Wahid, A.; Akhunzada, A.; Khan, F.; ul Amin, N.; Ali, I. Big Data in Healthcare: A Survey. *Appl. Intell. Technol. Healthc.* **2019**, 143–152. [CrossRef]
18. Hulsen, T.; Jamuar, S.S.; Moody, A.R.; Karnes, J.H.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafler, D.A.; McKinney, E.F. From big data to precision medicine. *Front. Media* **2019**, *6*, 34. [CrossRef]
19. Hatzigeorgiou, M.N.; Joshi, M.S. Population Health Systems: The Intersection of Care Delivery and Health Delivery. *Popul. Health Manag.* **2019**, *22*, 467–469. [CrossRef]
20. Koti, M.S.; Alamma, B.H. Predictive analytics techniques using big data for healthcare databases. In *Proceedings of the Smart Innovation, Systems and Technologies*; Springer Science and Business Media: Singapore, 2019; Volume 105, pp. 679–686. [CrossRef]
21. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, Analysis and Future prospects. *J. Big Data* **2019**, *6*, 54. [CrossRef]
22. Puaschunder, J.M. Big Data, Algorithms and Health Data. *SSRN Electron. J.* **2019**. [CrossRef]

23. Moreira, M.W.; Rodrigues, J.J.; Korotaev, V.; Al-Muhtadi, J.; Kumar, N. A Comprehensive Review on Smart Decision Support Systems for Health Care. *Inst. Electr. Electron. Eng.* **2019**, *13*, 3536–3545. [CrossRef]
24. López-Martínez, F.; Núñez-Valdez, E.R.; Lorduy Gomez, J.; García-Díaz, V. A neural network approach to predict early neonatal sepsis. *Comput. Electr. Eng.* **2019**, *76*, 379–388. [CrossRef]
25. López-Martínez, F.; Schwarcz.MD, A.; Núñez-Valdez, E.R.; García-Díaz, V. Machine Learning Classification Analysis for a Hypertensive Population as a Function of Several Risk Factors. *Expert Syst. Appl.* **2018**, *110*, 206–215. [CrossRef]
26. Singh, A. Foundations of Machine Learning. *SSRN Electron. J.* **2019**, *486*. [CrossRef]
27. Rhee, C.; Dantes, R.; Epstein, L.; Murphy, D.J.; Seymour, C.W.; Iwashyna, T.J.; Kadri, S.S.; Angus, D.C.; Danner, R.L.; Fiore, A.E.; et al. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014. *JAMA J. Am. Med. Assoc.* **2017**, *318*, 1241–1249. [CrossRef]
28. Mahindrakar, P.; Hanumanthappa, M. Data Mining In Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges. *Int. J. Eng. Res. Appl.* **2013**, *3*, 937–941.
29. Ghassemi, M.; Naumann, T.; Schulam, P.; Beam, A.L.; Chen, I.Y.; Ranganath, R. A Review of Challenges and Opportunities in Machine Learning for Health 2018. *arXiv* **2018**, arXiv:1806.00388.
30. Waring, J.; Lindvall, C.; Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artif. Intell. Med.* **2020**, *104*, 101822. [CrossRef]
31. Palanisamy, V.; Thirunavukarasu, R. Implications of big data analytics in developing healthcare frameworks—A review. *J. King Saud Univ. Comput. Inf. Sci.* **2019**, *31*, 415–425. [CrossRef]
32. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health Inf. Sci. Syst.* **2014**, *2*. [CrossRef] [PubMed]
33. Patel, S.; Patel, A. A Big Data Revolution in Health Care Sector: Opportunities, Challenges and Technological Advancements. *Int. J. Inf. Sci. Tech.* **2016**, *6*, 155–162. [CrossRef]
34. Chawla, N.V.; Davis, D.A. Bringing big data to personalized healthcare: A patient-centered framework. *J. Gen. Intern. Med.* **2013**, *28*. [CrossRef] [PubMed]
35. Abinaya, K. Data Mining with Big Data e-Health Service Using Map Reduce. *IJARCCE* **2015**, *4*, 123–127. [CrossRef]
36. Balladini, J.; Rozas, C.; Frati, F.; Vicente, N.; Orlandi, C. Big Data Analytics in Intensive Care Units: Challenges and applicability in an Argentinian Hospital. *J. Comput. Sci. Technol.* **2015**, *15*, 61–67.
37. Belle, A.; Thiagarajan, R.; Soroushmehr, S.M.R.; Navidi, F.; Beard, D.A.; Najarian, K. Big data analytics in healthcare. *BioMed Res. Int.* **2015**, *2015*. [CrossRef]
38. Chen, D.; Chen, Y.; Brownlow, B.N.; Kanjamala, P.P.; Arredondo, C.A.G.; Radspinner, B.L.; Raveling, M.A. Real-time or near real-time persisting daily healthcare data into HDFS and elasticsearch index inside a big data platform. *IEEE Trans. Ind. Inform.* **2017**, *13*, 595–606. [CrossRef]