Big Data Research 4 (2016) 13–24

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

# Assessing the Quality of Service Using Big Data Analytics With Application to Healthcare

Feras A. Batarseh [a,*,1], Eyad Abdel Latif [b,2]

[a] College of Science, George Mason University (GMU), Fairfax, VA, United States
[b] MedStar Georgetown University Hospital (MGUH), Washington, DC, United States

ARTICLE INFO

*Article history:*
Received 28 May 2015
Received in revised form 5 September 2015
Accepted 24 October 2015
Available online 10 November 2015

*Keywords:*
Big data analytics
Quality of Service (QoS)
Healthcare
Forecasting

ABSTRACT

Many industries are riding the wave of big data as the new era of data-driven decision making is unveiling. The field of big data analytics is gaining fast traction in industry, academia and the government; the healthcare arena is no different. In this paper, big data analytics are applied to healthcare data that is collected from multiple sources to gain quality insights and apprehend best practices of the field (using new healthcare specific big data tools). The US states are unceasingly pursuing potential improvements to their healthcare's Quality of Service (QoS). Recent changes in data sharing provisions, such as the disposition of the recent Affordable Health Care Act (ACA), changed the rules of the game, and provided the US states with a new set of measurable health quality variables that they couldn't overlook anymore. Individuals in those states without health insurance tend to ignore visiting the clinic even if they feel symptoms of a disease; healthy young individuals with insurance can also have the same behavior. Health experts constantly recommend closer immersion in one's health and more engagement with preventive healthcare. In three experimental studies, this multidisciplinary paper examines historical health data from all over the country, assesses the medical QoS for multiple US states using a new healthcare-specific analytical infrastructure, delivers forecasts and correlations for future healthcare activities, and provides data-driven

## 1. Introduction and background

Due to the exponential demand for healthcare-specific analytical tools, many existing software vendors are repositioning their tools to allow for their usage in the healthcare realm. This section discusses the existing data analytics tools, introduces the data analytics lifecycle, the processes and the factors that are studied in this paper to analyze healthcare's QoS in multiple US states.

### 1.1. Big data analytics in healthcare

With the recent mandated adoption of electronic health records (EHRs) by the US Department of Health and Human Services (HHS) [1], healthcare professionals are getting access to abundant amounts of data that can provide more insights and better takeaways that were not possible before. EHRs are not the only ample and rich source of healthcare data, for instance, wireless health monitoring devices and behavioral social media sources also provide more opportunities and could be serious game-changers. Multiple data analytics software vendors are building tools that are connected to these sources and are specifically tailored to healthcare; for example, IBM provides a tool for health paperwork content management. The tool helps healthcare providers with recording their patients' health data, and provides access to tools for data analysis and visualizations [2]. One of the main users of this tool is the State of North Dakota's (ND) Department of Human Services (DHS). The department along with the tool provides services that help citizens of all ages with maintaining and enhancing their health. On their website, IBM claims the following: "ND's DHS replaced paper-based processes with a central electronic content management system that makes information more accessible and streamlines workflows, helping staff work together more effectively" [2]. SAS [3] on the other hand, is another data-analytics provider that created a dedicated *Center for Health Analytics and Insights* to address the increasing demand from hospitals, clinics, and health professionals across the world. SAS is a global provider of analytics that covers multiple industries, such as financial institutions, education, govern-

[*] Corresponding author.
*E-mail addresses:* fbatarse@gmu.edu (F.A. Batarseh), eyad.abdellatif@gunet.georgetown.edu (E.A. Latif).
[1] Research Assistant Professor.
[2] Clinical Administrator, Nursing Admin.

http://dx.doi.org/10.1016/j.bdr.2015.10.001
2214-5796/© 2015 Elsevier Inc. All rights reserved.

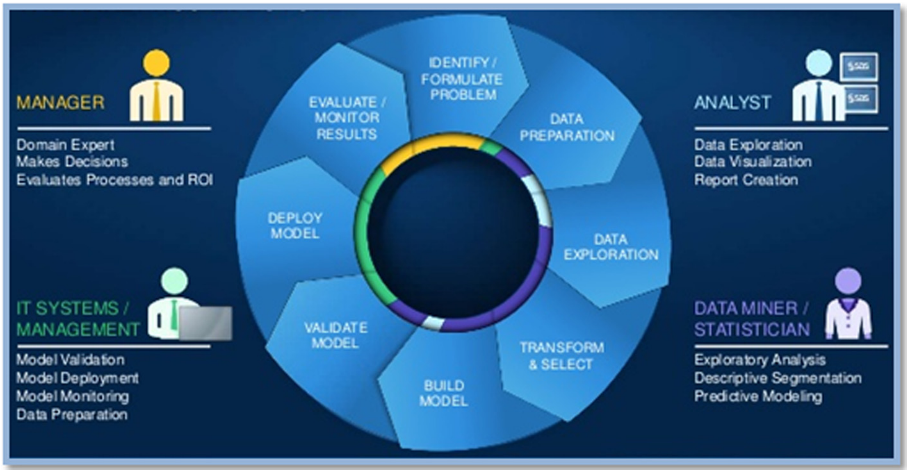


**Fig. 1.** Industry's data analytics lifecycle [3].

ment, and healthcare. Other vendors however chose to focus their business and their tools solely on healthcare. Mede Analytics [4] for instance provides a number of solutions that are tailored to different areas of healthcare (such as: health insurance, health records management, provider engagement, patient engagement, and many others). Mede, SAS, IBM and many other tools (such as SPSS [5], MicroStrategy [6], Tableau [7], Cognos [8], and Qlik [9]) provide state-of-the-art software for health management, nonetheless, *traditional* healthcare software providers such as Epic [10] are now focusing on creating and collecting EHRs. Using Epic's EHR database, VMware launched the *Care Systems Analytics* [11], which is a set of VMware virtual machines (VMs) that gather EHRs, and uses them for analysis through a set of tools on the VMware cloud.

One common aspect that all of these tools deploy however, is a well-defined lifecycle process for healthcare data analytics. The next section discusses the *de-facto* steps of any data analytical process.

### 1.2. The data analytics lifecycle

Although data mining/analytics research has been of interest to many researchers around the world for a long time, data analytics didn't see much light until it was adopted by the industry. Many software vendors (examples listed earlier in the paper) shifted the focus of their conventional software development to include a form of data analytics, big data, data mining, statistical modeling and data visualization.

Based on multiple long and challenging deployments in many areas, trials and errors, and multiple consulting exchanges with many customers from many fields, those vendors coined a lifecycle for data analytics. SAS (based on Gartner's research [12] is one of the pioneer vendors in this field), provided a lifecycle (illustrated in Fig. 1) which includes the following phases:

1. Identifying and formulating the problem
2. Preparing the data (pivoting and data cleansing)
3. Data exploration (summary statistics, bar charts and other means of exploration)
4. Data transformation and selection (select ranges, and data subsets)
5. Statistical modeling development, validation and deployment
6. Evaluating and monitoring the results of models, delivering and refine the analytical models

These phases are applied to any data analytical project, and they require four main human roles, the data manager, the system manager, the analyst, and the data miner (see Fig. 1 for more details on the tasks). Although this is a generic process, it is highly applicable to healthcare analytics. In the study presented in this paper, this lifecycle model is used to analyze the health data. In software, multiple lifecycles have been used throughout the years, examples of the most commonplace lifecycles include: the waterfall model, the spiral model, agile and the scrum development models [13].

Most importantly however, lifecycles for developing medical systems is key to the scope of this paper. Such systems were one of the first drivers of the Artificial Intelligence (AI) research in the sixties (1960s). Medical systems were the first deployments of Knowledge-Based Systems (KBS), or what most publications refer to as Expert Systems, which are introduced and discussed in the next section.

### 1.3. Traditional medical expert systems

Knowledge-based systems (expert systems) are intelligent systems that reflect the knowledge of a skillful person. Expert systems are a special kind of intelligent system that make extensive use of knowledge. Those systems were first introduced in the sixties during processes of capturing medical knowledge from healthcare practitioners. Expert systems are different from conventional software systems and data analytical systems because they use heuristic rather than algorithmic approaches for decision making. The idea of a general problem solver (GPS), that later turned into the idea of building a medical expert system used generic search techniques aided by heuristic knowledge to solve problems [14,15]. The GPS notion was instrumental in the development of MYCIN, a system that diagnosed blood disorders. MYCIN is a landmark medical rule-based system developed at Stanford University (known to be the first expert system). More importantly, some claim that MYCIN even influenced the creation of the field of expert systems. Afterwards, by learning from MYCIN, PROFORMA [16] was developed as a generic model for building clinical and medical expert systems [14]. It was developed to address the high demand of requests to build intelligent healthcare systems. Nonetheless, the recent flashy data analytical tools reviewed in this paper incorporate all these "expert" ideas and more. One thing that both recent and traditional tools share is the need for data and knowledge modeling, the most common ways of modeling include [14]: 1. Rule-based systems – where knowledge is represented in term of rules. 2. Case-based systems – this kind of system depend on the gathering of cases. 3. Logic-based systems – in this form of system, knowledge is a set of logical operators. 4. Frame-based systems –

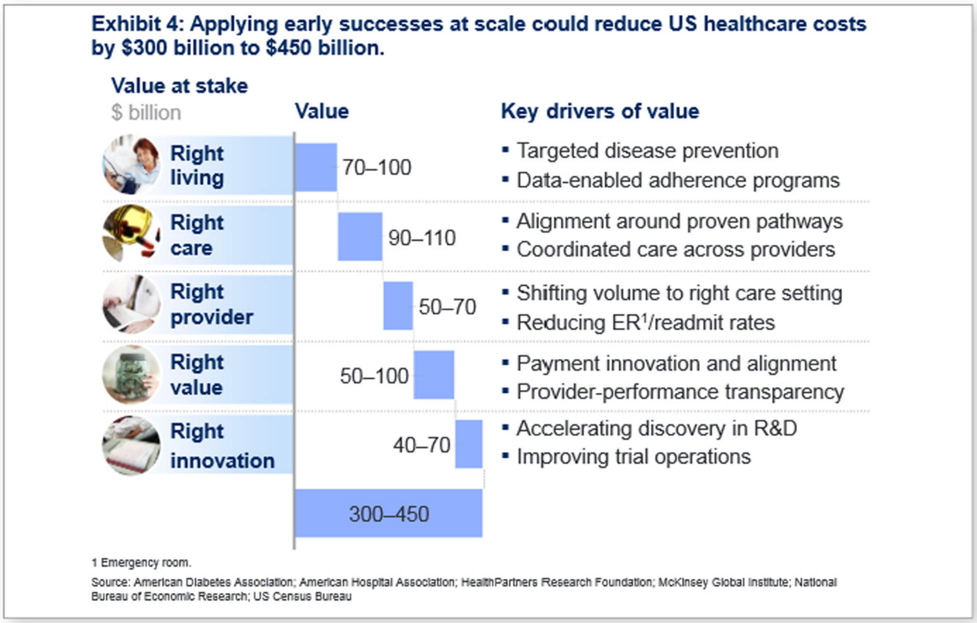


**Exhibit 4: Applying early successes at scale could reduce US healthcare costs by $300 billion to $450 billion.**

| Value at stake<br>$ billion | Value | Key drivers of value |
|---|---|---|
| Right living | 70–100 | ▪ Targeted disease prevention<br>▪ Data-enabled adherence programs |
| Right care | 90–110 | ▪ Alignment around proven pathways<br>▪ Coordinated care across providers |
| Right provider | 50–70 | ▪ Shifting volume to right care setting<br>▪ Reducing ER[1]/readmit rates |
| Right value | 50–100 | ▪ Payment innovation and alignment<br>▪ Provider-performance transparency |
| Right innovation | 40–70 | ▪ Accelerating discovery in R&D<br>▪ Improving trial operations |
|  | 300–450 |  |

1 Emergency room.

Source: American Diabetes Association; American Hospital Association; HealthPartners Research Foundation; McKinsey Global Institute; National Bureau of Economic Research; US Census Bureau

**Fig. 2.** US healthcare potential improvements [18].

knowledge is represented in templates or "frames". Frames have a number of slots to be filled by the engineers. 5. Object-based systems – in this type, an object is a collection of information that represents an "entity" from the real world and describes its functions. Expert systems have many advantages, most importantly, they include efficient replacements of the human experts, as they can act as a repository for human knowledge in case of loss of human expertise. With data analytics, knowledge is saved in form of data, that data is transformed into knowledge on the fly using data mining methods, and advanced statistical models. The study in this paper introduces a number of statistical measures that analyze big data of healthcare for all US states. The context of the healthcare data used in the experiment of this paper is introduced in the next section.

### 1.4. Healthcare's study motivation

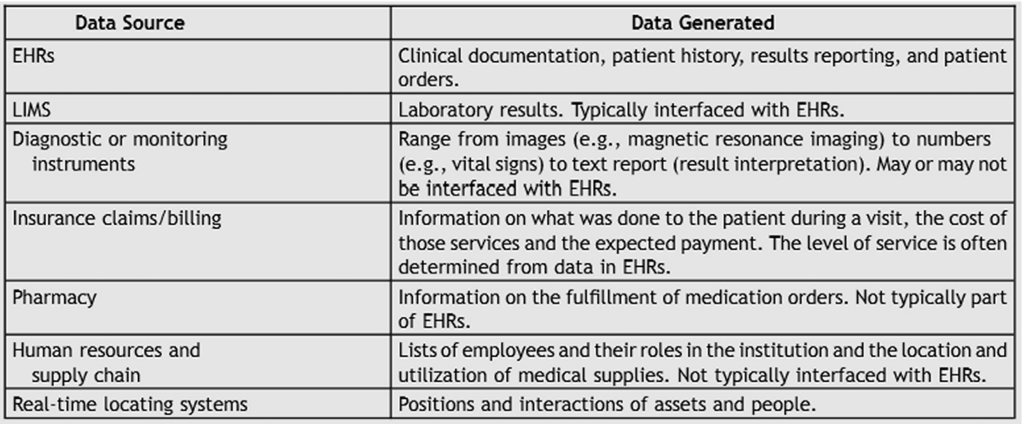
According to a recent McKinsey & Company report on big data [17,18], in 2010, an estimated 30 million patients in the US were chronically diagnosed with diseases most relevant to **diabetes and hypertension**. The diagnosed patients accounted for more than **80%** of health system costs. These two disease are the main focus of the analytical experiment presented in this paper. More data on these disease is presented in Sections 2 and 3. Engaging and enlightening patients to make informed decisions about preventive healthcare can improve overall health and reduce demand and most importantly waste in the world of healthcare. This waste and unnecessary costs are one of the main drivers for the need of studies that evaluate the healthcare system in the US. As it is well established in research [14], applying incremental process improvements is a key driver for cost reduction (refer to Fig. 2). **The study in [18] illustrates how healthcare costs could be reduced by $300 billion to $450 billion merely by applying lessons from success stories across different practices. This paper aims to locate such success stories on a state-level, so other *not-so-successful* states can learn from these best practices**. Such studies can accelerate the improvement of QoS, and accelerate the development of better practices for healthcare in the US.

However, the healthcare industry is constantly changing, especially in terms of data. For example, in 2005, 30% of office-based physicians started using EHRs; that number rose to more than 50% for physicians and 75% for hospitals by the year 2012. More interestingly, the government has been a huge advocate for big data, specifically for health. In the 2009, the Open Government Directive, as well as consequent actions from the Department of Health and Human Services (HHS) under the Health Data Initiative (HDI), are starting to liberate data from agencies like the Centers for Medicare and Medicaid Services (CMS), the Food and Drug Administration (FDA), and the Centers for Disease Control (CDC) [18]. More visibly, the Affordable Care Act (ACA), enacted in March 2010, included a provision that let HHS release their data to research institutions and the public. HHS releases a major report every year about the overall health of the US [19]. The executive office of the US president published a report discussing the effects of the ACA. The report claims the following: "The evidence is clear that recent trends in health care spending and price growth reflect, at least in part, ongoing structural changes in the health care sector. The slowdown may be raising employment today, and, if continued, will substantially raise living standards in the years ahead. The evidence also suggests that the ACA is already contributing to lower spending and price growth and that these effects will grow in the years ahead, bringing lower cost, higher quality care to Medicare and Medicaid beneficiaries and to the health system as a whole" [20]. As this quote confirms, the ACA, among other novel healthcare trend shifts, changed the field and had impacts that are discussed in this paper. One of the main challenges of finding data was finding *complete* data sets. We decided to only include sources that have comprehensive data; all the health dimensions in the paper come from trusted, complete, and validated data sources.

This paper is organized as follows: the next section introduces the analytical health study motif (among other existing studies), as well as the new big data healthcare platform – CHESS (main technological contribution), Section 3 presents the three analytical experiments and their results (main analytical contribution), and lastly, Section 4 presents conclusions, future work and data-driven healthcare lessons learned for the US states.

## 2. The analytical healthcare context

This section reviews closely related studies published in literature, introduces the healthcare-specific big data platform and tool



| Data Source | Data Generated |
|---|---|
| EHRs | Clinical documentation, patient history, results reporting, and patient orders. |
| LIMS | Laboratory results. Typically interfaced with EHRs. |
| Diagnostic or monitoring instruments | Range from images (e.g., magnetic resonance imaging) to numbers (e.g., vital signs) to text report (result interpretation). May or may not be interfaced with EHRs. |
| Insurance claims/billing | Information on what was done to the patient during a visit, the cost of those services and the expected payment. The level of service is often determined from data in EHRs. |
| Pharmacy | Information on the fulfillment of medication orders. Not typically part of EHRs. |
| Human resources and supply chain | Lists of employees and their roles in the institution and the location and utilization of medical supplies. Not typically interfaced with EHRs. |
| Real-time locating systems | Positions and interactions of assets and people. |

**Fig. 3.** Most data sources used in literature [23].

(CHESS), presents the void in research, and the problem that this paper tackles with the three experimental studies introduced in Section 3.

### 2.1. Related work in healthcare analytics

Medical documentation and predictions have always been prominent areas of study. The famous study in [21] presents a detailed historical trace of analytical studies for healthcare that date back to ancient Mesopotamia, Greece, ancient Rome and Arabia [21]. Since these ancient researchers worked on this area, there has always been an obvious need for analytics to improve the quality of human's health–gladly, data analytics now provide the tools to address this historical challenge. Multiple recent publications in the field covered use cases that address problems such as monitoring the hospitals quality, improving the treatment methods, and providing patient centric services [22]. However, **no study was found in literature that covers state-level healthcare analytics.** For example, in [23], data sources that are considered are EHRs, diagnostic instruments, insurance claims and so on, all sources are on the "patient" level, no state-level data was used for the study. Fig. 3 shows general data sources that most relevant publications used. Studies like that are interested in the human health, and are mostly trying to solve a specific "health" issue using data. It is important to note that this paper is *multidisciplinary* and is focused on the overlap between the following 3 areas: 1. *Geographical Data Analytics*, 2. *Nation-wide Healthcare*, and 3. *QoS Evaluation*.

To solve problems relevant to health, multiple databases were created that host international human health data; that includes initiatives such as the 1000 Genomes Project, Database of Genotypes and Phenotypes, among others. These databases help answer questions relevant to genetics, their effect on humans' health, and overall health trends in the world with association to genes [23, 24]; again, addressing a specific "health" issue. In [24,25], multiple aspects of improving nursing and medical activities are evaluated, and recommendations on nurses' activities are discussed; such activities include: patient monitoring, home nursing, preventing inpatient morbidity, patient comfort and telemedicine. All these areas are key for healthcare, and studies like the one in [24] help nurses improve their practices, however, it can't be stressed enough that this paper aims to introduce an overall study that covers healthcare practices across the country and not only specific to a certain nursing or medical aspect. After introducing the most relevant work in Sections 1.1, 1.2, 1.4 and 2.1, it is evident that the aforementioned *state-level* analytical gap in literature is critical, and is what this paper aims to fill. The context, experiments and overview of our study are presented in Section 3. However, before that, the healthcare-specific big data system needs to be introduced, CHESS is presented next.

### 2.2. The big data healthcare-specific software system

The big data system that was used for this analytical experiment in this section is presented in Fig. 4, we call it **CHESS**: The Comprehensive Healthcare Electronic Software System. Due to the recent rise of analytical forecasting importance in literature [26], and using the analytical lifecycle presented previously in the paper, data from the Michigan Quality Improvement Consortium (MQIC) is used as the main data source, that data is then used for analysis and is contrasted with data from Centers for Medicaid and Medicare Services (CMS), NIH and HHS. All the data files reside in Hadoop [27], a big data engine described by Apache Software (their open source "owner") as: "The Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage". A state-level data aggregation is then migrated to a SQL server for analysis (much less number of data rows are saved in SQL), afterwards, that data is used for advanced analytics using Tableau [7], and Excel's Pivot tables.

The tool shown in Fig. 4 allows researchers to upload their EHRs, or many other types of data sets into CHESS (such as excel and comma separated files). The CHESS tool then moves the data set(s) to Hadoop (without any data massaging or structuring). Users then can access their data through a variety of tools, listed in the GUI above. Note that the data opens with all the mentioned tools, however, without any structuring. The user will have to rely on Hadoop for handling big data issues, and only query smaller amounts of data to the analytical tools.

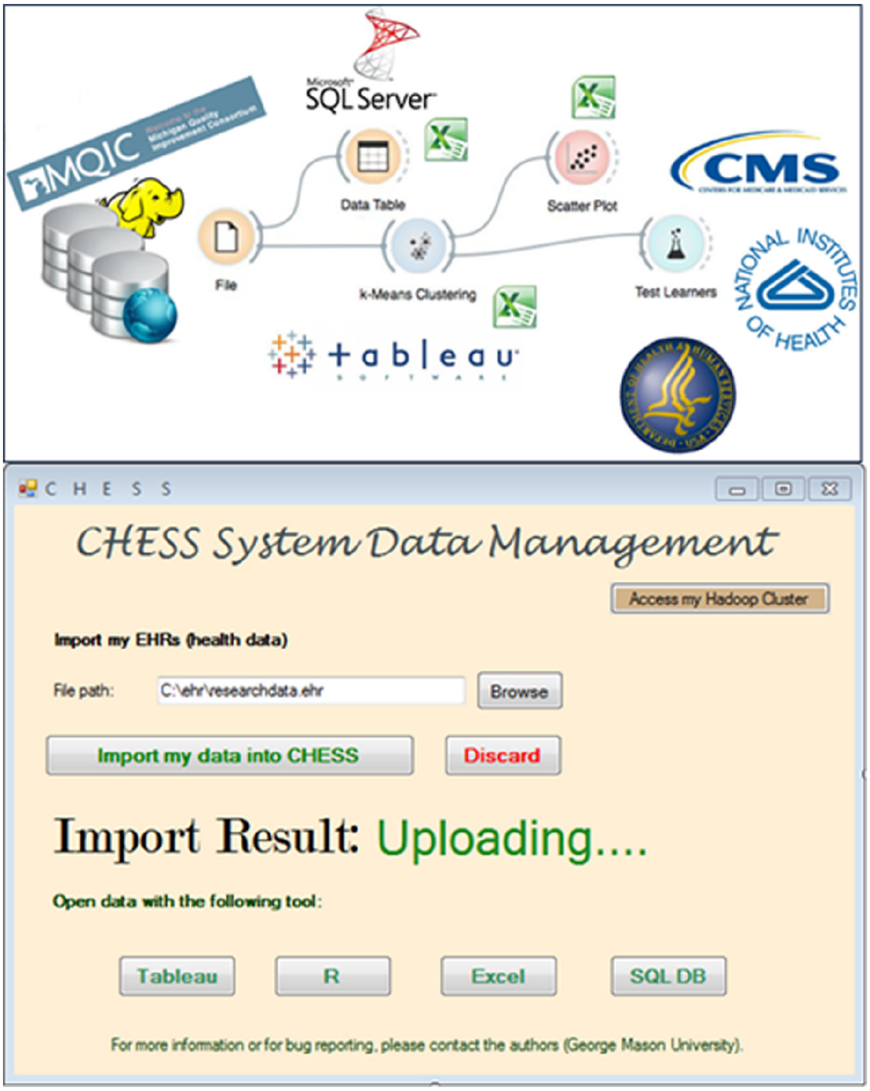
Additionally, the user has to pivot, format, and reorganize their data based on what they aim to accomplish with their studies. The tool is developed using Microsoft Visual Studio, in C#.

The CHESS workflow consists of the following major steps:

The next section uses data analytics and the CHESS system to describe existing and previous trends in healthcare across the country, focuses on some states that present a unique story, forecasts the average number of patient visits for these states, and reflects on the effects of recent regulations on the states' health.

## 3. The analytical healthcare experiments

This section represents the main analytical contribution of this paper. As it is noted in the titles of the 3 sub-sections, the first study is ***descriptive*** (it describes historical health data from 2005


MQIC
Welcome to the Michigan Quality Improvement Consortium
Microsoft SQL Server
Data Table
Scatter Plot
CMS
CENTERS FOR MEDICARE & MEDICAID SERVICES
File
k-Means Clustering
Test Learners
NATIONAL INSTITUTES OF HEALTH
tableau SOFTWARE
DEPARTMENT OF HEALTH & HUMAN SERVICES USA
CHESS
CHESS System Data Management
Access my Hadoop Cluster
Import my EHRs (health data)
File path:
C:\ehr\researchdata.ehr
Browse
Import my data into CHESS
Discard
Open data with the following tool:
Tableau
R
Excel
SQL DB

Import Result: Uploading....

For more information or for bug reporting, please contact the authors (George Mason University).

**Fig. 4.** The comprehensive healthcare electronic software system (CHESS) and its user tool.

1. Select a complete and verified two dimensional data set.
2. Enter the full name and address of your Hadoop cluster (You can get Hadoop for free from Apache's website: hadoop.apache.org).
3. Put your data in excel, or if it exceeds excel's limits, place it in a comma separated files (unless it's an EHR file).
4. Use the CHESS tool interface to upload the data by browsing, choosing the file, and selecting "import my data into CHESS".
5. Once uploading is done, choose your preferred tool for advanced analytics. Note: if the tool is not available in the selection, the user can connect the tool directly to the same Hadoop cluster, and use their tool of choice for statistical analysis.

and after), the second study is ***predictive*** (it looks at data for the end of 2015 and after), and the third is ***prescriptive*** (it describes how the recent health trends are effecting on the states position in the healthcare realm). The three categories are based on Gartner definition of data analytics types [12]. Main contributions and outcomes that are deemed important in our analytical studies are underlined in the 3 sub-sections below.

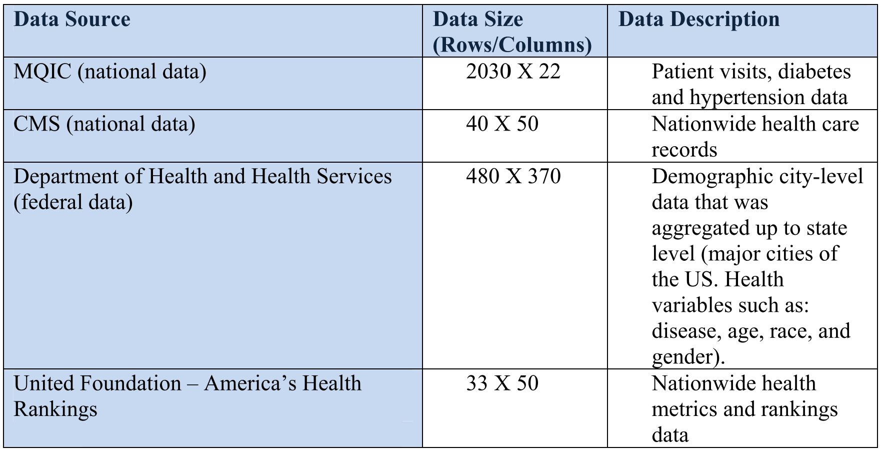
It is appropriate to introduce the dimensions of the dataset before discussing the experiments. As previously mentioned, the data used in this paper came from multiple sources, however, only sources with valid and complete was used. The raw data is described in Table 1. The data is available for download at (for any technical or data download difficulties, please contact the authors): https://www.dropbox.com/sh/zfroxhaf825lrw6/AABk7o5kttdHpXmozhaBu02da?dl=0.

Using the process from the previous section defined above, the CHESS system combined all these data sources, and we were able to build the models presented in the three subsections below.

### 3.1. The first analytical study: health of the US states (descriptive)

Americans lead different lifestyles, and have very different health habits. They especially react differently to their health requirements. Some are very anxious and paranoid about being sick that they keep seeking professional advice and undertaking all kinds of tests to find something. Although high patient engagement is suggested to be a good practice, these patients might end

**Table 1**
Raw datasets information.



| Data Source | Data Size (Rows/Columns) | Data Description |
|---|---|---|
| MQIC (national data) | 2030 X 22 | Patient visits, diabetes and hypertension data |
| CMS (national data) | 40 X 50 | Nationwide health care records |
| Department of Health and Health Services (federal data) | 480 X 370 | Demographic city-level data that was aggregated up to state level (major cities of the US. Health variables such as: disease, age, race, and gender). |
| United Foundation – America's Health Rankings | 33 X 50 | Nationwide health metrics and rankings data |

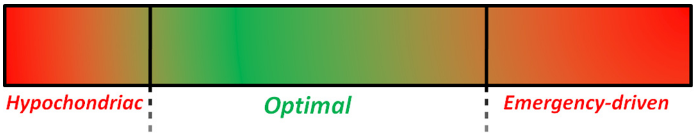


**Fig. 5.** Levels of patient's engagement with their health.

up overcrowding the system – they are referred to as hypochondriacs [28]. Another category of Americans are young, "careless", and only seek healthcare when they are severely sick (which is a bigger category than hypochondriacs). Health experts however, refer to patients' engagement levels that are in between as *Optimal*, refer to Fig. 5.

Optimal patients' engagement is referred to as *preventive care*. The number of visits of a patient to a clinic is a major indicator [29]. As it is previously established, diabetes and hypertension are the most diagnosed diseases across all states of the US. This study evaluates the involvement of patients of these two diseases with their health. Data for the experiments was collected from the United Health Foundation [30], the Michigan Quality Improvement Consortium (MQIC) [31], the Department of Health and Human Services [1], and the Centers for Medicaid and Medicare Services (CMS) [32]. The data collected is used at different stages of this study, and entered into CHESS. First, the average ***number of visits per patient*** at each state is measured by aggregating the data up from patient to state level, and by deriving that from the number of office visits and the number of diabetes and hypertension patients. That is then contrasted with the ***health*** of these states. Health is derived from weight values, Body Mass Index (BMI), and number of patients with diabetes and hypertension. These two factors (***number of visits per patient*** and ***health***) represent the adopted definition for QoS in this paper: *QoS is defined as a combination of states' health info and measures of patients' engagement with their own health.*

This descriptive part of the data showed that the overall trend for patient visits in the US is on the rise (refer to Fig. 6). Since 2005, almost all states witnessed an increase in patients' engagement. Some states however, are on the extreme sides of the spectrum, Connecticut for example is the highest hypochondriac state, and Kentucky is the most emergency-driven state (refer to Fig. 7), data shown in Fig. 6 is for five example states (time range from: 2005 to 2010) – same time range is for Fig. 7 as well. The $X$-axis in Fig. 6 shows the average number of visits per patient, and the $Y$-axis shows yearly trend in the five states. Nonetheless, the probability of a high number of visits to occur is very high in all US states. Visits more than 15 times per patient of hypertension or diabetes are almost guaranteed, the probability of that gets close to 99% from the MQIC data. Nonetheless, the *average overall* number of visits in the US is 11.50. The scattered plot in Fig. 7 shows how the US states are dispersed in terms of number of visits. In the chart, four states are highlighted with a red square: Connecticut and Kentucky (for reasons mentioned previously), Utah and Alabama (although both are close to the *optimal* area, the next section highlights radical changes with the healthcare QoS in these 2 states). Refer to Fig. 7 for all the states, $X$-axis is the state names, and the $Y$-axis is the average number of patient visits per year. Health variables (such as: diabetes, smoking, obesity, hypertension, and blood pressure) were also considered, and for these 4 highlighted states, all fall within the intermediate area as well.

The 4 states are not extremely healthy nor are they extremely unhealthy (based on data from between years 2005–2010). Additionally, in terms of weight, Body Mass Index (BMI), and number of patients with diabetes and hypertension, Maine and South Carolina were the unhealthiest states (mostly due to weight), Oregon, Washington and California were the healthiest/leanest states.

Although this study was able to point to health levels at different states, and possibly states like Kentucky and Connecticut can look at states that are in the optimal area such as Illinois, New York, and Pennsylvania and learn from their practices. So far, the analytics introduced are ***descriptive***, as they analyze historical data, and describe the trends and provide insights, however the **power** of data analytics is in the ***predictive*** side. In the next section, predictions for the mentioned states are introduced. These predictions are then contrasted with the recent (2014 and 2015 data) to address the previously discussed healthcare trends in the US.

The goal is to assess the QoS at some states that recently changed their legislations based on the ACA and other discussed factors. This analytical study is especially interested in diabetes and hypertension (due to the high number of Americans that have these two diseases). These two diseases are highly effected by the patients' behavioral activities [33,34], for example: Diabetes is exacerbating all over the US; the hypertension rise is not as pronounced as diabetes, but both diseases are already very high and need to be controlled (see Figs. 8 and 9).

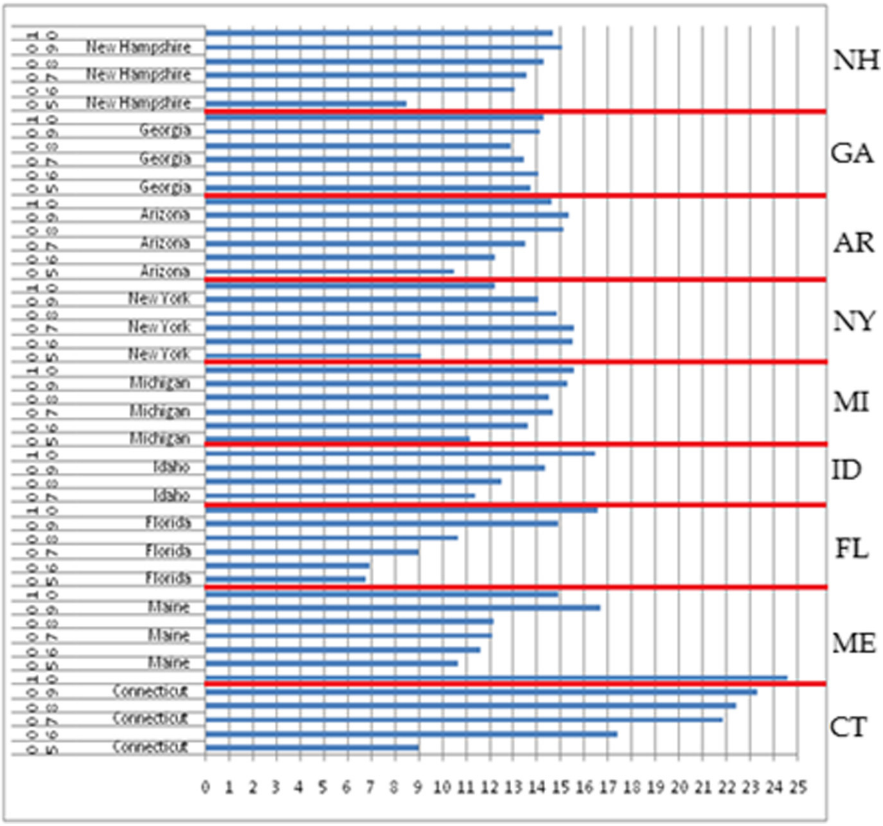


**Fig. 6.** An obvious increase in visits per patient in most US states (average number of visits on $X$-axis and States on $Y$-axis).

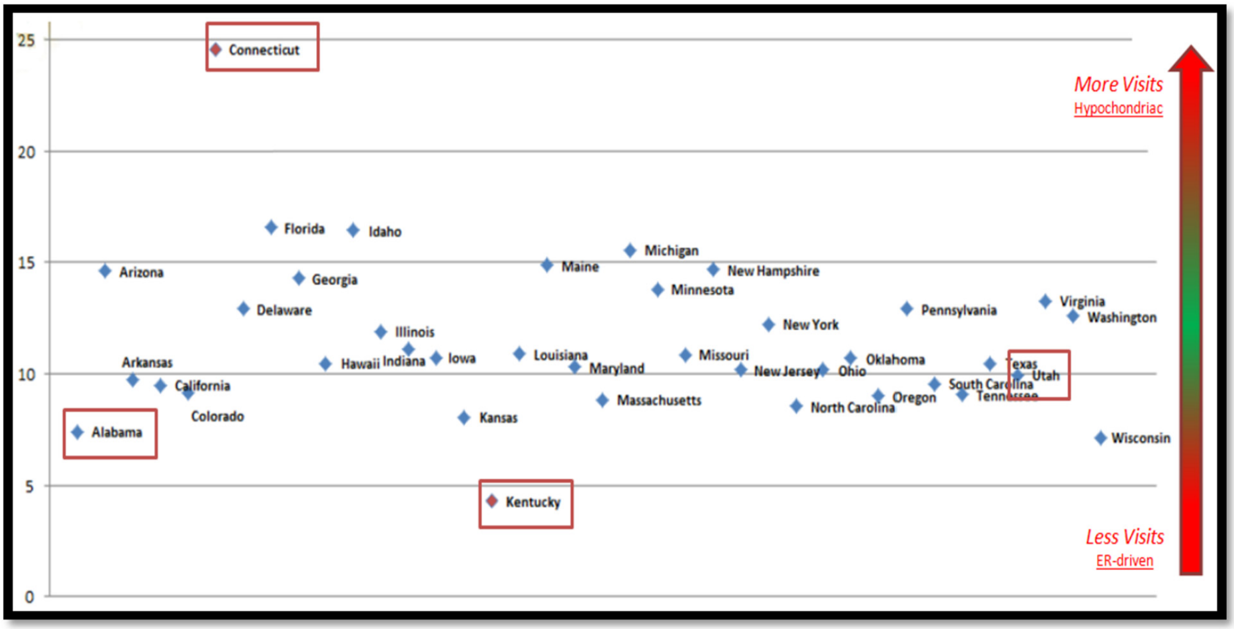


**Fig. 7.** Patients engagement per US state (states on $X$-axis and visits per state on $Y$-axis).

Uncontrolled diabetes can lead to hypertension: If a patient doesn't go to a primary care physician, uncontrolled diabetes will lead to vascular changes which will result in hypertension.

Both diseases are highly relevant (high prevalence) to each other and to two major activities (obesity and smoking) [33,34].

*Health and its relevant research* can have multiple factors and dimensions. Based on the World Health Organization (WHO) [35], health dimensions fall under the five categories, shown in Fig. 10. In this paper, the focus is on two types of factors only, health system factors (HSF), and patient related factors (PRF) – highlighted below.

This subsection discussed how patients interact with their health (i.e. PRF), the next two sections introduce HSF studies. For more information on health dimensions, please refer to [35]. The

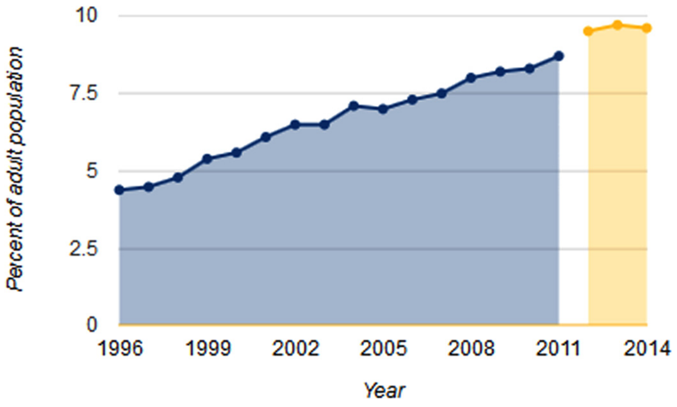


**Fig. 8.** Diabetes is on the rise [30] (year on $X$-axis and percent of population on $Y$-axis).

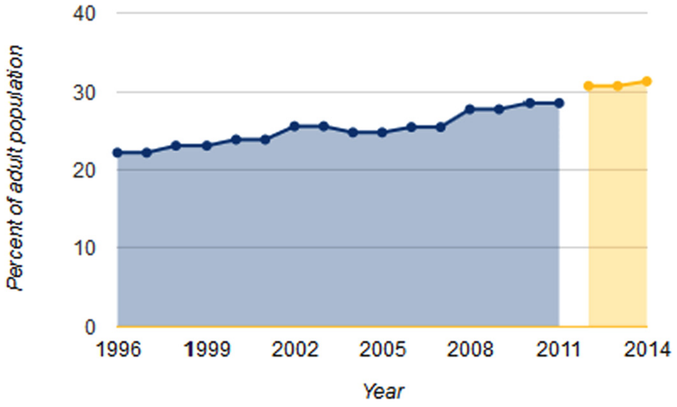


**Fig. 9.** High blood pressure is witnessing a slower linear increase in the US [30] (year on $X$-axis and percent of population on $Y$-axis).

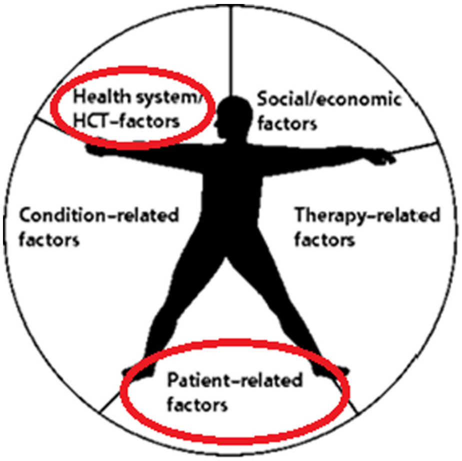


**Fig. 10.** WHO's five dimensions of health [35].

WHO defines the PRF as: "Patient-related factors represent the resources, knowledge, attitudes, beliefs, perceptions and expectations of the patient", and defines HSF as a health service that falls under an insurance plan, and involves a healthcare provider. More information on both PRF and HSF in the following sections.

### 3.2. The second analytical study: trends in patients' engagement (predictive)

In this section, CHESS is used to create the predictions for the number of visits for the US states based on historical data. The most interesting ones are presented and discussed. The two extreme states on the hypochondriac/ER-driven scale are Connecticut and Kentucky. Regression forecasting for these states is shown in Figs. 11 and 12. Top chart in Figs. 11 and 12 is historical data, and bottom chart includes the forecasts. Each data point represents the number of visits in different counties and data points from the MQIC data source for that state. Multiple forecasts were tried (linear, exponential, and logarithmic); illustrated results in Figs. 11 and 12 are the ones with the highest statistical confidence.

**Table 2**
The 2014–2015 smoking and obesity ranking in the four states. (For interpretation of the references to color in this table, the reader is referred to the web version of this article.)

| State | Smoking Rank | Obesity Rank |
|---|---|---|
| Utah | 1 (best) | 4 |
| Connecticut | 4 | 8 |
| Kentucky | 49 | 46 |
| Alabama | 38 | 43 |

The forecasts in this paper are linear regression forecasts. It is apparent to note that Kentucky has been fairly consistent with patients' visits, and therefore the statistical regression model predicted (with high statistical confidence) that there will ***not*** be a serious rise in future years. However, the state of Connecticut has a predicted exponential rise in the average number of patients' visits. The same forecast was applied to four states using CHESS (Utah, Alabama, Connecticut and Kentucky). Utah and Alabama showed similar forecasts as Connecticut and Kentucky respectively.

It is already established why Kentucky and Connecticut were chosen for this further predictive analysis. However, why were Utah and Alabama chosen for our study as well?

Also, are these forecasted values accurate when ignoring the recent changes in healthcare policies and trends across the country? Both these questions are answered in the section below.

### 3.3. The third analytical study: post-ACA era, analysis and correlations (prescriptive)

This section introduces a 3rd analysis study that is prescriptive in nature; study's major takeaways, and the overall lessons learnt for the US states. As it is noted in the titles of the sections, the first study is descriptive, the second is predictive and the third is prescriptive (based on Gartner definition of data analytics types [12]).

As it was previously indicated (in Fig. 7), the states of Utah and Alabama were in the *optimal* area in previous years, however, after the introduction of multiple new healthcare policies, such as the ACA (discussed in previous sections), both states moved in two very different directions in terms of health, and patients engagement. Based on data from [30], Alabama, Mississippi, and West Virginia are states with the worst health, and least patient involvement with their own health. Utah, however, and Connecticut are doing much better in 2014. Both states moved closer to the center of the *optimal* area in patients' involvement and health measures. The state of Alabama on the contrary, moved away from the *optimal* area into the unhealthy and the emergency-driven area. Thus, the forecasts presented in the previous section are not very representative of what will happen without considering the new trends of health in the US and incorporating them in the regression model. Besides these 4 states, no obvious shifts were noticed in the data. Using recent data from [30], and due to the established notion about the high influence of obesity and smoking on the two diseases under study (diabetes and hypertension). In Table 2, healthier states in terms of smoking and obesity in **2014/2015** are shown in green (Utah and Connecticut), and unhealthy states are Kentucky and Alabama. Smoking in the US is drastically decreasing, but obesity is on the rise [30].

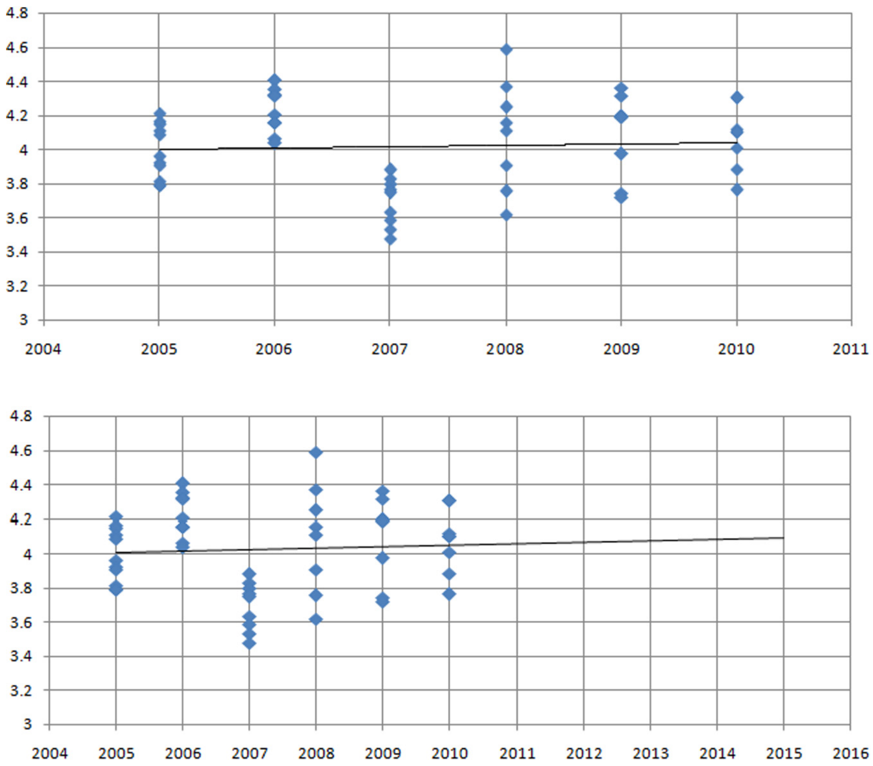


**Fig. 11.** Kentucky's forecasted number of patient visits (up to 2015) (year on $X$-axis and number of visits on $Y$-axis).

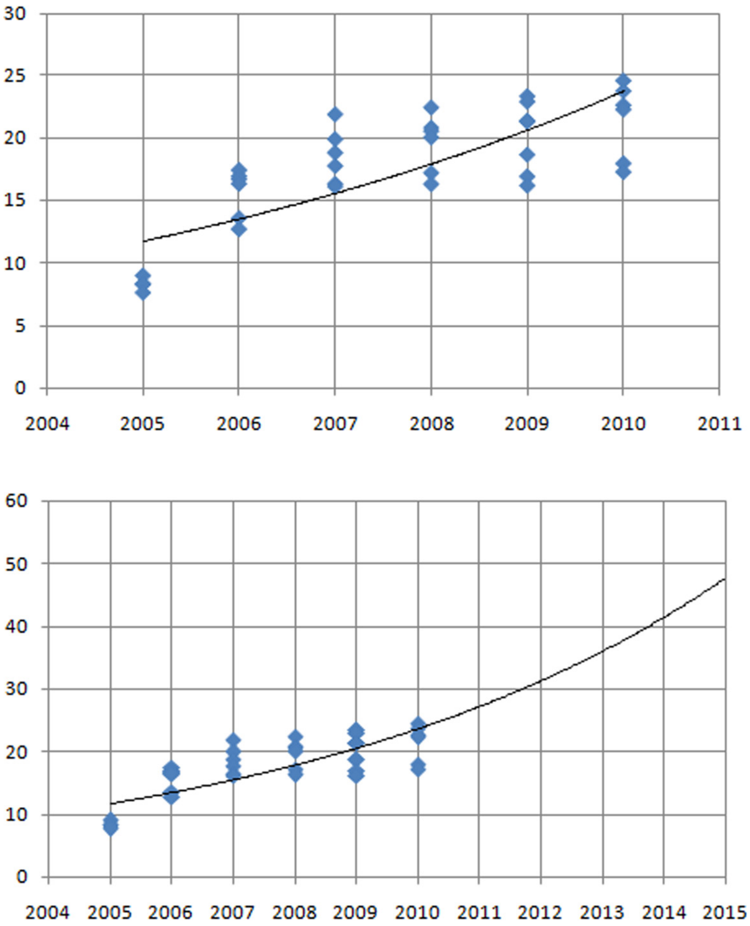


**Fig. 12.** Connecticut's forecasted number of patient visits (up to the end of 2015) (year on $X$-axis and number of visits on $Y$-axis).

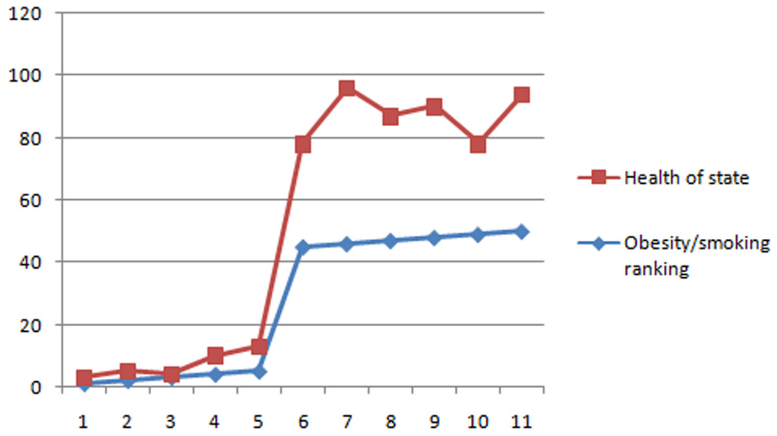


**Fig. 13.** Correlation between the overall health of a state ($Y$-axis) and it's obesity and smoking rankings ($X$-axis).

We think that the ACA event influenced these outcomes (this doesn't reflect a political opinion whatsoever, rather, it is based on the data and the analysis of this paper), some states deployed the law with high approval rates, but the law is not popular in other states. Therefore, we recommend for states that aim to improve their QoS to consider lessons learned from the states in this study and implement best practices locally. Based on the data in [30], the following are the main healthcare QoS drivers of the states in 2014 (only considering factors mentioned in this paper):

**Utah**: Low prevalence of smoking, and Low rates of preventable hospitalizations (ER-driven practices). **Connecticut**: Low prevalence of smoking, and Low obesity rates.

**Alabama:** although smoking decreased in the past 2 years by 12%, there is still a high prevalence of diabetes, and abnormal child weights. **Kentucky:** high prevalence of smoking, and still a high rate of ER-driven practices by its citizens.

The statistical correlation between a health of a state and an average of its ranking in obesity and smoking was eye opening, it was equal to **0.956**. Fig. 13 shows that high correlation for the top 5 states and bottom 6 states. We expect this correlation number to reduce (but still be closer to 1) once the 50 states are included. We included the main states that we focused on in the previous parts of this paper.

For more details and information on this data, please contact the authors. The next section introduces the main summaries, conclusions and future work of this research.

## 4. Conclusions

The research and experiments performed using CHESS resulted in 3 main data-driven analytical studies. These studies aim to evaluate the QoS of healthcare in the US on the state level. This section summarizes the work that was done, introduces the major takeaways and the future work.

### 4.1. Summary of contributions

The story told in this paper is a deep dive into the trends of QoS in relevance with health habits in all the US states. The focus shifts to a certain number of states that represent unique cases, and have a different narrative to tell. In the studies presented, QoS is defined as a combination of states' health info and measures of patients' engagement with their own health. The work was performed using CHESS, a big data system that is dedicated to healthcare. The data that was used in this paper was extracted from multiple government and private sources, all the data was saved into a repository and used for the analytics. Tools like Tableau and Pivot tables are used to perform the advanced statistical methods. Multiple summary statistics are presented in this paper, furthermore, regression forecasts are also deployed. The three studies aim to identify trends in the US especially with the recent changes in healthcare policies, regulations and the changing public's health habits. Most studies that were found in literature focus on a specific health aspect, and aim to solve to a certain medical issue (even the traditional ones that historically used expert systems [36]), however, in the three descriptive, predictive and prescriptive [12] data analytical sections of this paper, data is aggregated up to the state level, and the analysis is done per state.

The Health Financial Management Association (HFMA) [37] summarized how better quality could be accomplished in healthcare through the recent changes in the domain. Fig. 14 summarizes the main pillars of the work presented in this paper through CHESS, and compares it with what HFMA presented as major pillars for increased health care QoS/value [37].

In HFMA's "temple" (left side), they list healthcare practices that should improve the overall value of service for US states. In the "temple" on the right side, CHESS puts these practices to the test of analytics and statistics, and provides insight by implementing studies and analyzing big amounts of data (using Hadoop, Pivot tables, Forecasting, and Aggregations). The studies presented in this paper are examples on that.

Fig. 14 is a good high-level summary of the goal that this paper aims to address. In our studies, the states of Utah, Connecticut, Kentucky and Alabama presented exceptional cases worth digging deeper into. By looking at the practices of these states, best practices can be extracted, and health practices relevant to QoS (hypochondriac activities, smoking, and obesity) could be identified. The main **contributions** of this paper are:

### 4.2. Future work

The work done in this paper lends itself to further investigations into causes and effects of healthcare QoS in the US. The following data analytics ideas are the strongest candidates for next steps:

1. Perform similar study on a lower level of aggregation, for example look at specifically healthy US cities and counties and identify health habits in these areas.
2. Build advanced clustering models (such as k-means) within CHESS, and try to group the states in other different ways. That would help in identifying the statistical significance of certain healthcare factors. This will also answer the question of which factors have the most effect on the state's position on

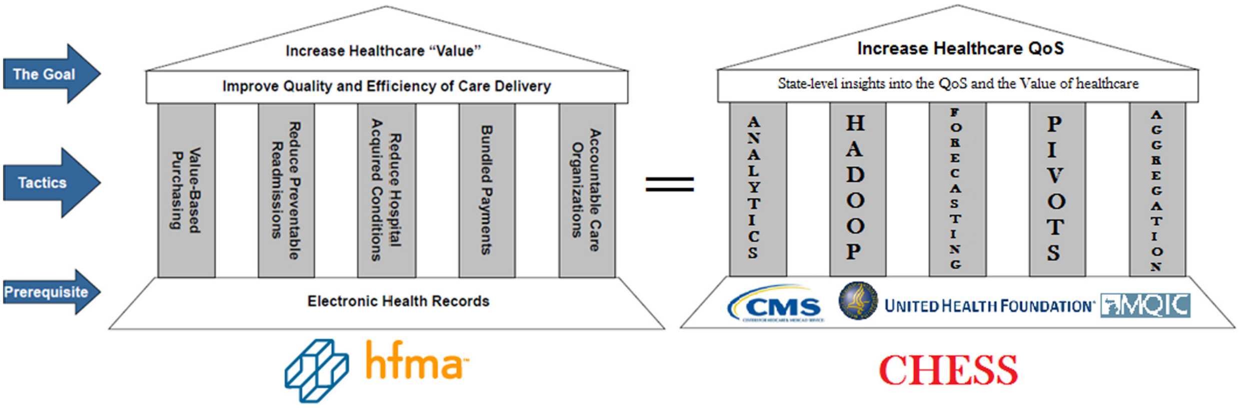

The Goal
Tactics
Prerequisite
Increase Healthcare "Value"
Improve Quality and Efficiency of Care Delivery
Value-Based Purchasing
Reduce Preventable Readmissions
Reduce Hospital Acquired Conditions
Bundled Payments
Accountable Care Organizations
Electronic Health Records
hfma
=
Increase Healthcare QoS
State-level insights into the QoS and the Value of healthcare
ANALYTICS
HADOOP
FORECASTING
PIVOTS
AGGREGATION
CMS
UNITED HEALTH FOUNDATION
MQIC
CHESS

**Fig. 14.** CHESS analytical accomplishments vs. HFMA's aspirations.

1. A comprehensive big data system that can read health data and provide means for analysts to run experiments, and analyze health data using their preferred tool.
2. A user-interface that allows the users to upload data to Hadoop for analysis.
3. An overview of the current state of the art in healthcare data analytics.
4. Three analytical studies on how different states are performing in terms of health and healthcare.
5. Ranking states based on multiple different health variables – among other insightful analytical discussions (all presented in the paper).
6. An experimented novel method that is reusable for further data analytics of healthcare data.
7. Established a high statistical correlation between smoking/obesity ranking and state's health and Predicted (using linear regression) the health of some states in the future.

the healthcare/patient-engagement QoS scale (similar to what is shown in Fig. 7).

3. Use the CHESS system to study the correlation and covariance between patients' health activities (ER-driven vs. hypochondriac) and the states' healthcare QoS.
4. Tell the stories of more states in more detail, similar to what was done in this paper for states like Utah and Alabama.
5. The CHESS system is already in place for more data analysis, it could be used to build further models if the right data is available. Models that are part of future work include: clustering of states, classification of patients, association rules for patients' visits to the clinic/physician, t-tests to study the relations between multiple health dimensions to the health status and quality of service at a state.

As it is mentioned in previous sections of this paper, healthcare analytics are on the rise, and there is a serious need for insightful takeaways for US states to improve their QoS. This paper presented answers to some of the QoS questions and recent healthcare trends.

## References

[1] The US Department of Health and Human Services: http://www.hhs.gov/.
[2] A. Weintraub, C. Le Clair, C. Mckinnon, The forrester wave: enterprise content management, Published by the Forester Research, Inc., 2013, Q3.
[3] Using analytics to navigate health care reform, A white paper, Published by the SAS Institute, Inc., 2015.
[4] Mede Analytics' website: http://medeanalytics.com/.
[5] SPSS website: http://www-01.ibm.com/software/analytics/spss/.
[6] MicroStrategy's website: http://www.microstrategy.com/us/.
[7] Tableau's website: http://www.tableau.com/.
[8] Cognos' website: http://www-01.ibm.com/software/analytics/cognos/.
[9] Qlik's website: http://www.qlik.com/us/company.
[10] Epic's website: http://www.epic.com/software-index.php.
[11] VMware solutions for an epic environment, A technical white paper published by VMware, 2013.
[12] Data analytics yearly report, Gartner magic quadrant 2014, a yearly report published by Gartner, 2014, http://www.gartner.com/technology/research/methodologies/research_mq.jsp.
[13] I. Sommerville, Software Engineering, 8th edition, Addison-Wesley, 2007, Chapter 4.
[14] F. Batarseh, Incremental lifecycle validation of knowledge-based systems through CommonKADS, Doctoral dissertation published at the Florida State University Library Services and the Library of Congress, 2011.
[15] A.J. Gonzalez, D. Dankel, The Engineering of Knowledge-Based Systems, Theory and Practice, Prentice Hall, 1993.
[16] S. Smith, A. Kandel, Validation of expert systems, in: Proceedings of the Third Florida Artificial Intelligence Research Symposium (FLAIRS), 1990, pp. 197–201.
[17] J. Manyika, M. Chui, B. Rbown, et al., Big data: the next frontier for innovation, competition, and productivity, Report by the McKinsey & Company, May 2011.
[18] P. Groves, B. Kayyali, D. Knott, S. Van Kuiken, The big data revolution in healthcare, Report by the McKinsey & Company's Center for US Health Reform, January 2013.
[19] US Department of Health and Human Services, Health, United States, 2013 – with special feature on prescription drugs, Government report, US Department of Health, 2013.
[20] Trends in health care cost growth and the role of the affordable care act, A Paper published by the Executive Office of the President of the United States, November 2013.
[21] L. Miner-Winters, P. Bolding, J. Hilbe, et al., in: Practical Predictive Analytics and Decisioning Systems for Medicine, Science Direct/Academic Press, 2015, pp. 5–22, Chapters 1–3.
[22] J. Archenaa, M. Anita, A survey of big data analytics in healthcare and government, Proc. Comput. Sci. 50 (2015) 408–413.
[23] M. Ward, K. Marsolo, C. Froehle, Applications of business analytics in healthcare, Bus. Horiz. 57 (5) (October 2014) 571–582.
[24] L. Miner-Winters, P. Bolding, J. Hilbe, et al., Predictive analytics in nursing informatics, in: Practical Predictive Analytics and Decisioning Systems for Medicine, Academic Press, 2015, pp. 969–974, Chapter 16.
[25] T. Strome, Healthcare Analytics for Quality and Performance Improvement, John Wiley and Sons, ISBN 978-1-118-51969-1, October 2013.
[26] F. Batarseh, A.J. Gonzalez, Predicting failures in contextual software development through data analytics, Softw. Qual. J. (2015) 1–18, http://link.springer.com/article/10.1007/s11219-015-9285-3. First online: 09 August 2015.
[27] The Hadoop website: https://hadoop.apache.org/.
[28] The hypochondriac, Dublin Penny J. 46 (May 1833) 366–368.
[29] C. Preston, M. Alexander, Medicare coverage of preventive care services, J. Am. Med. Assoc. (2010), https://www.medicare.gov/Pubs/pdf/10110.pdf.
[30] The American Heath Rankings – by the United Health Foundation: http://www.americashealthrankings.org/.
[31] The Michigan Quality Improvement Consortium: http://mqic.org/.
[32] The Centers for Medicare and Medicaid Services: http://www.cms.gov/.
[33] J. Hum. Hypertens. (Jun. 2015) 339–397;
J. Hum. Hypertens. (Jun. 2015) 281–337;

J. Hum. Hypertens. (Jun. 2015) 141–210.

[34] The diabetes care – predicting type 1 diabetes using biomarkers, A report published by the American Diabetes Association, 2015.

[35] World Health Organization (WHO), Adherence to long term therapies study – evidence for action, http://apps.who.int.

[36] F. Batarseh, A.J. Gonzalez, Validation of knowledge-based systems – a reassessment of the field, Artif. Intell. Rev. (Feb 2013), http://dx.doi.org/10.1007/s10462-013-9396-9.

[37] The Healthcare Financial Management Association, HFMA's Leadership Publications from: https://www.hfma.org/.