OJPHI

Data Lakes and Data Visualization: An Innovative Approach to Address the Challenges of Access to Health Care in Mississippi

Denise D. Krause, PhD[1]

1. Department of Preventive Medicine, University of Mississippi Medical Center, Jackson, MS USA

Abstract

Background: There are a variety of challenges to developing strategies to improve access to health care, but access to data is critical for effective evidence-based decision-making. Many agencies and organizations throughout Mississippi have been collecting quality health data for many years. However, those data have historically resided in data silos and have not been readily shared. A strategy was developed to build and coordinate infrastructure, capacity, tools, and resources to facilitate health workforce and population health planning throughout the state.

Objective: Realizing data as the foundation upon which to build, the primary objective was to develop the capacity to collect, store, maintain, visualize, and analyze data from a variety of disparate sources -- with the ultimate goal of improving access to health care. Specific aims were to: 1) build a centralized data repository and scalable informatics platform, 2) develop a data management solution for this platform and then, 3) derive value from this platform by facilitating data visualization and analysis.

Methods: A managed data lake was designed and constructed for health data from disparate sources throughout the state of Mississippi. A data management application was developed to log and track all data sources, maps and geographies, and data marts. With this informatics platform as a foundation, a variety of tools are used to visualize and analyze data. To illustrate, a web mapping application was developed to examine the health workforce geographically and attractive data visualizations and dynamic dashboards were created to facilitate health planning and research.

Results: Samples of data visualizations that aim to inform health planners and policymakers are presented. Many agencies and organizations throughout the state benefit from this platform.

Conclusion: The overarching goal is that by providing timely, reliable information to stakeholders, Mississippians in general will experience improved access to quality care.

Keywords: physician, workforce, data, geographic information systems, data warehouse, visualization

Correspondence: dkrause@umc.edu

DOI: 10.5210/ojphi.v7i3.6047

Introduction

Mississippi ranks 50th in the nation in overall health status [1]. Since it is common knowledge that access to care and provider availability are key factors contributing to overall health status, it is not surprising that Mississippi also ranks 50th in physicians per 100,000 residents with only 159 physicians per 100,000 population, compared to the national average of 220 per 100,000 [2]. As in many other states, unequal distribution of health care providers further exacerbates the problems related to access to care. In Mississippi, about 60% of primary care physician practices are located in urban areas; whereas, about 60% of the population resides in rural areas [3].

Of course, physicians are only one profession in the health care team and supply of health care providers is only one component of the supply/demand/utilization equation. Access to care is a multi-faceted and extremely complex issue that academic health sciences centers, other health professional training schools and programs, researchers, and health care planners continue to struggle with. How many providers and what types are needed to provide adequate quality health care to the population -- when and where? These remain extremely difficult questions to answer. What should the evolving health care team look like? How many slots are needed in training programs to prepare the workforce of the future? Again, there is not a "one size fits all" equation to address these complex questions. Models of estimation are continually being developed and tested, but estimates are constantly shifting. Not only is there is no right answer, but it's nearly impossible to get the same answer twice!

With that said, how can access to care be improved in Mississippi? It became quickly clear from strategic meetings with the Office of Mississippi Physician Workforce and other key policy, research, and other healthcare stakeholders across the state that they were grappling with similar issues. Each was responsible for a piece of the puzzle, but despite the fact that volumes of quality data relevant to these primary missions had been being collected for years, many were still at a loss because other data needed for informed decision-making were outside of their domains. Multiple mini siloes of data had been constructed throughout the state. Perhaps historically, some groups had protected their turfs. However, in this new age of data sharing and collaboration, walls seem to be crumbling naturally as stakeholders recognize the benefit of partnerships and become more interested in working together to move our state forward improving population health outcomes.

In response to this urgent need and opportunity, a strategy was developed to build and coordinate infrastructure, capacity, tools, and resources to facilitate health workforce planning throughout the state. Realizing data as the foundation upon which to build, the primary objective was to develop the *capacity* to collect, store, maintain, visualize, and analyze data from a variety of disparate sources with the ultimate goal of improving access to health care. Specific aims were to:

1) build a centralized data repository and scalable informatics platform,

2) develop a data management solution for this platform and then,

3) derive value from this platform by facilitating data visualization and analysis.

The stakeholders are many and varied – among them are the state's only academic health science center, the University of Mississippi Medical Center; William Carey College of Osteopathic Medicine; schools of nursing and other health professions throughout the state; the Office of Mississippi Physician Workforce; the Office of Nursing Workforce; the Mississippi State Health

Department; and a host of other state agencies and organizations. The overarching goal is that by providing timely, reliable information to groups such as these, Mississippians in general will experience improved access to quality care. While the focus is on Mississippi, this approach is highly generalizable and can be readily adopted by any other state.

Methods

Data Sources

We began collecting the most pertinent data needed to answer the most important initial questions. Publicly available datasets related to population statistics, socio-demographic data, and health services information, from sources such as the U.S. Census Bureau and the Health Resources and Services Administration, were obtained. Data contributors included health professional licensure boards, both medical schools in the state, rural physician and dentist scholarship programs, the Mississippi State Department of Health, the Mississippi Primary Health Care Association, training programs across the state, officials managing loan repayment programs, and many others. We were quickly inundated with valuable data from a variety of data contributors eager to share their data or, at least, benefit from having their data cleaned, organized, visualized, and analyzed. For the health professions data, current and historical licensure data on physicians were obtained initially, but dentists were soon added, and physician assistants and advanced practice nurses are currently being added. Other health professions will be included as they come forward and express interest. In addition to data on health professions, data on population health outcomes are being incorporated in order to establish baseline data to be able to chart the success, or lack thereof, of projects, initiatives, and interventions over time.

Data Lake

The usual approach to addressing the challenge of data stored in silos is to build a data warehouse [4]. Some states have well-developed data warehouse environments [5], but Mississippi does not have a mature data warehouse in production for health planning purposes. Efforts to coordinate health care workforce planning is still in its infancy in Mississippi, yet there is a heightened sense of urgency to address these issues and a realization that this work is time-sensitive and cannot wait for a full-fledged data warehouse solution to be implemented. In order to save time and bypass some of the limitations of traditional data warehouses, it was most expeditious to focus on the creation of a "data lake". This relatively new term refers to an easily scalable, potentially massive, but accessible, data repository built to store "big data" on (relatively) inexpensive computer hardware. Data lakes differ from data marts, which are optimized for data analysis by storing only some of the attributes. In contrast, data lakes are designed to retain all attributes of the raw data, which is especially important when the scope of the data or its uses may not yet be fully known [6]. The data lake solves the age-old problem presented by information silos while cutting costs, improving agility, and increasing efficiency by allowing for increased use of information. It becomes an ideal platform for visualizing and analyzing various sources of disparate data in their native forms.

Managed Data Lake

There is no perfect data solution and there are pitfalls to avoid in the use of data lakes [7]. Because data lakes inherently lack data governance, there is the danger of not being able to account for data quality or the lineage of findings. In order to proactively address these

limitations, the data lake concept was extended further to a "managed" data lake. To this end, a custom web-based data management tool for lifecycle management was developed (Figure 1), extending the data lake to include not only raw data, but also cleaned or processed files and revision histories. With this data management tool, data sources are logged as they come in, along with all pertinent information about their source and their refresh cycles, while also keeping track of their file processing and revision histories and who made what changes, when, where, and why. The administrative dashboard allows for easy monitoring of data sources that need to be refreshed within the next thirty days and if any data are overdue to be refreshed, thereby overcoming many potential pitfall of data quality issues.

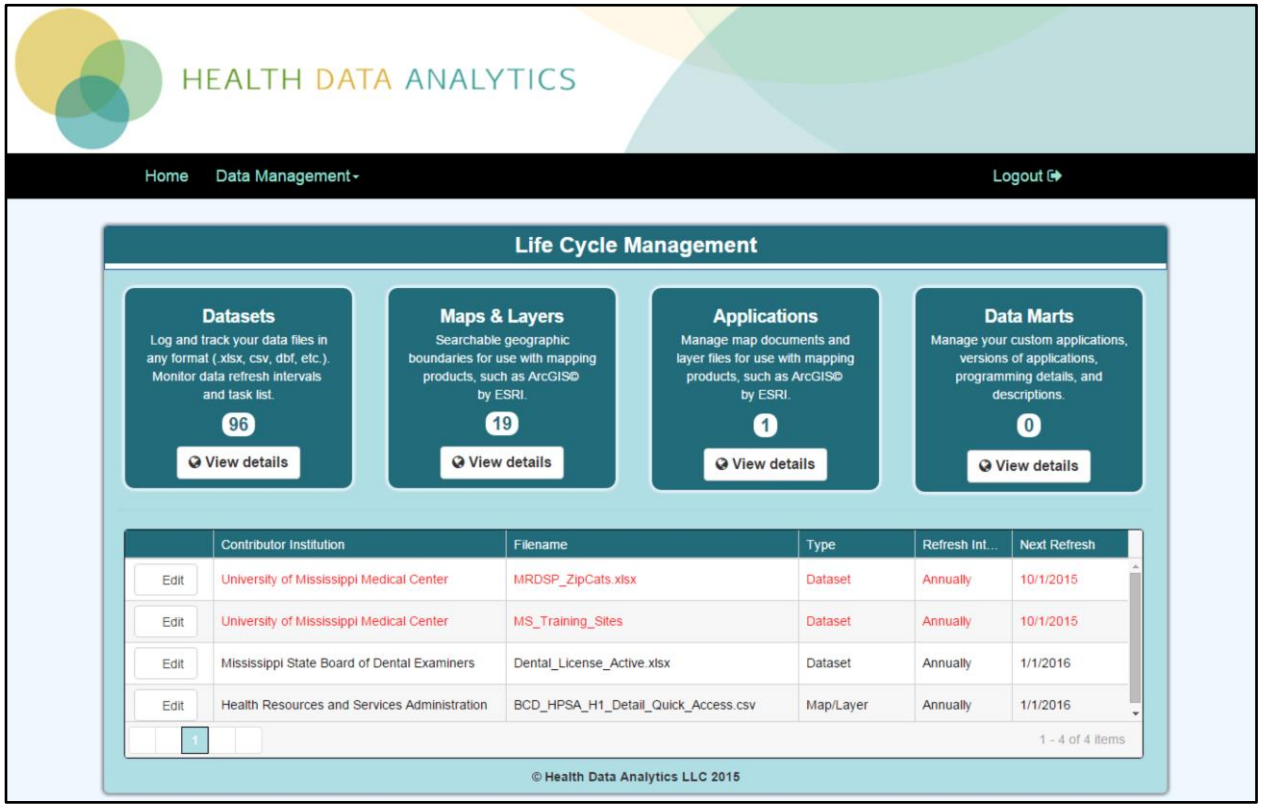

Figure 1. Custom web application developed specifically to manage data files, maps and geographies, custom applications, and data marts.

Additional risks that data lakes can introduce include a lack of security and access control [7]. In order to proactively address these critical issues, the data lake was not designed as an ungoverned data store, but as a centralized managed data repository. From this platform, a path was incorporated from which to extract, transform, and load (ETL) certain data into a traditional data warehouse as warranted to provide the single source of truth inherent in a traditional data warehouse environment [8,9]. When streamlined data marts are needed, the lifecycle management application also allows staff to monitor and track the development and delivery of data marts. So, in effect, the data lake is not replacing a traditional data warehouse, but is being used in addition to the data warehouse for a different purpose -- to provide a more agile platform for data analytics and the ability to quickly derive value from the data. This managed data lake allows development to be responsive, quick, and productive, without becoming bogged down by the typical lengthy delays associated with data warehousing.

Data Visualization and Analysis

Methods for collecting, storing, and maintaining data for health workforce planning have been discussed. Next we demonstrate how value is being derived from this highly scalable informatics platform consisting of data, hardware, software, and analytic tools by means of visualizing and analyzing data. A variety of tools are used for data visualization and data analysis.

First, data visualization software by Tableau Desktop Software® was used to examine data from different sources and run quality control checks. Because it was quickly apparent if data issues were present, feedback could be provided to data contributors along with attractive visualizations, giving them the opportunity to take appropriate action to improve data quality. Data contributors have been pleased to finally see answers present in their data and have been highly motivated to make improvements in data quality.

Next, using health professions licensure data and other supporting data sources, a geographic information system (GIS) web application was developed for mobile devices to explore the demographics and distribution of the physician workforce currently licensed in Mississippi and over time (Figure 2). Detailed information about the design and development of the health workforce mapping application can be found elsewhere [10]. The primary function of the GIS web application is mapping the health workforce and displaying results of custom queries geographically. To complement this application, we used Tableau® again to build visualizations and interactive dashboards of the various datasets, individually or in combination. These dashboards allow for the interactor to slice and dice the data on the fly, revealing answers that may not have been apparent without visualizations designed for data exploration.

Figure 2. Screenshot from the GIS web application developed to visualize and query physicians licensed in Mississippi. This visualization shows all physicians currently licensed in Mississippi with no query attributes defined.

Results

Key to success has been building strong working relationships with the Mississippi State Board of Medical Licensure (MSBML) and their governing board. The MSBML was given a presentation of the physician workforce mapping application and a series of visualizations and interactive, dynamic dashboards created using MSBML and related data. Board members were very interested in being able to visualize and understand the data that are being collected every year through new licensure applications and renewals. This relationship has been further strengthened by bringing back visualizations from their data and working with MSBML staff to improve data collection and data quality moving forward.

Sample dashboards created for the MSBML are shown in Figures 3 and 4. Each part of the dashboard shown in Figure 3 is interactive, meaning that the user can make selections as they choose and the data in all other parts of the dashboard redraw to meet the selected criteria. Figure 4 shows demographics of the physician workforce over the course of ten years and the distribution of the workforce. Work will continue with the MSBML, as well as other health licensure boards, to design and build visualizations that can be made available to the public on their websites or used for administrative purposes with security enabled. If organizations would like to build their own visualizations, assistance can be provided to develop this capability.

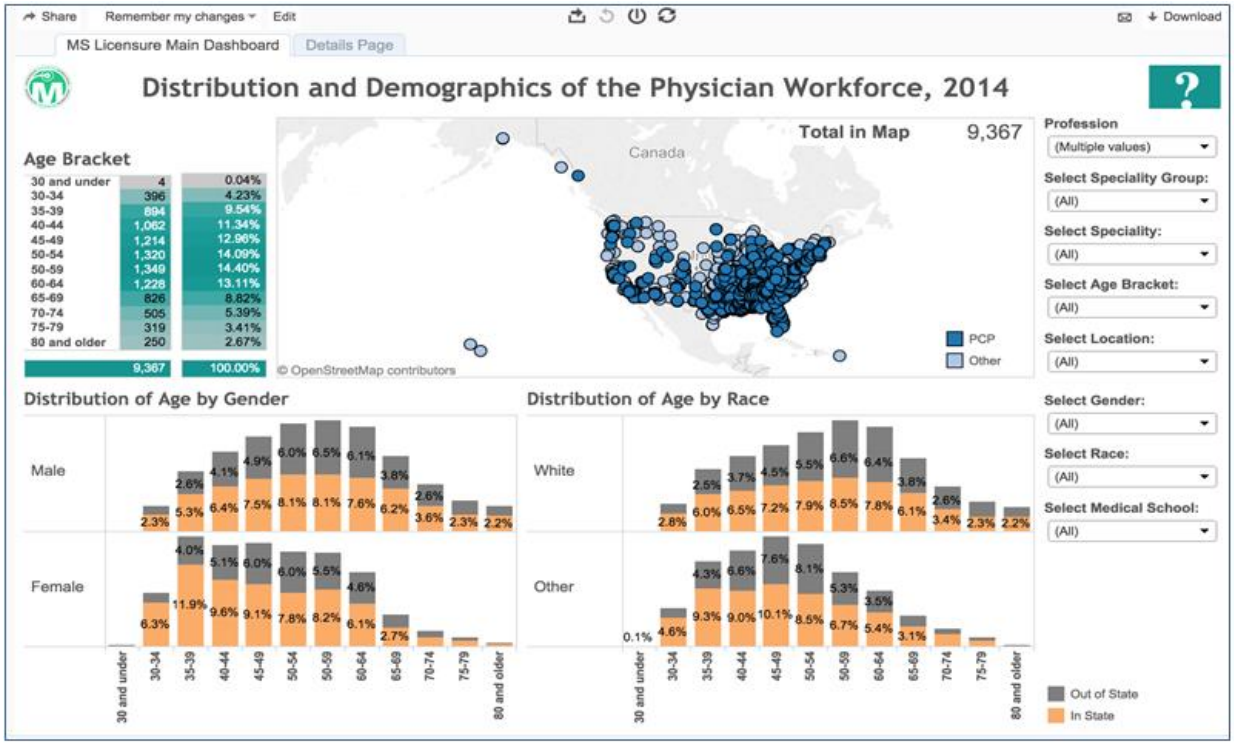

Figure 3. A dynamic dashboard created using current medical licensure data allowing the user to interact with the visualization to explore the demographics and distribution of the health workforce based on their selected criteria. Creatively designed dashboards can provide a wealth of information at one's fingertips.

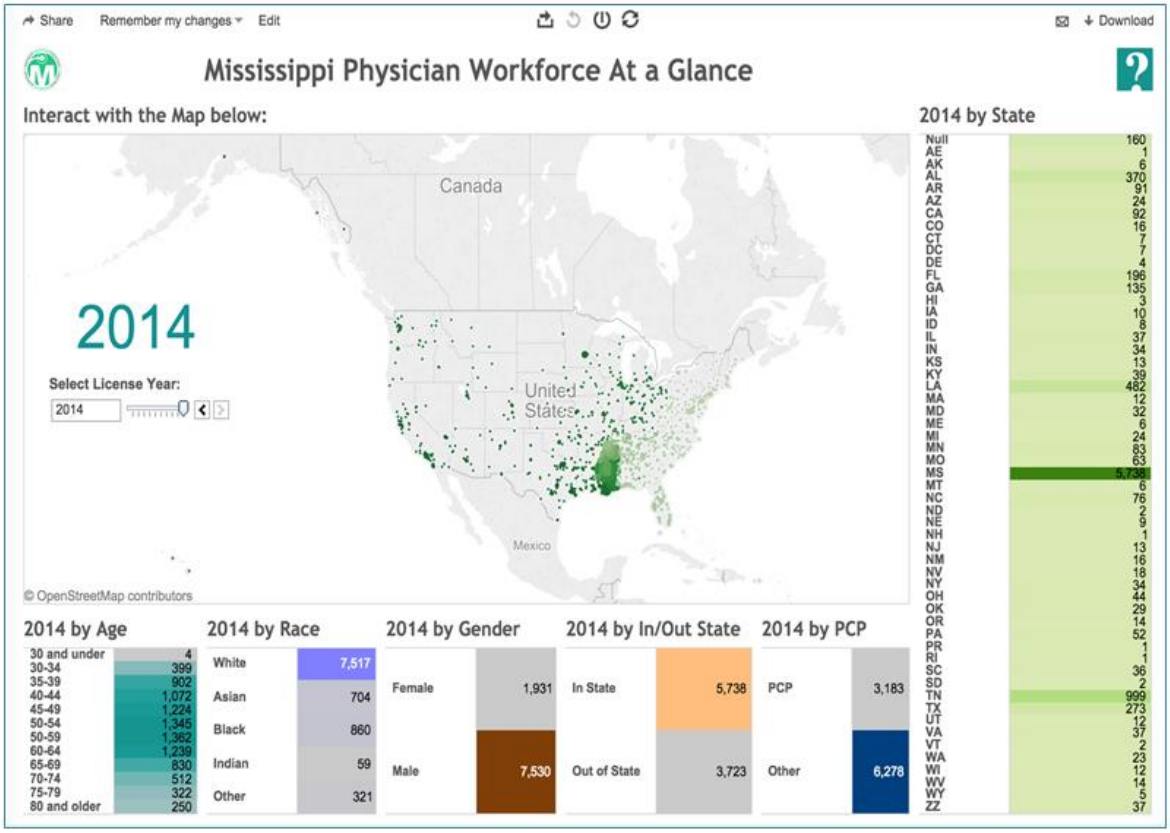

Figure 4. This dynamic dashboard displays pertinent information about the demographics and state of practice of physicians licensed in Mississippi a year at a glance. This visualization is built on ten years of medical licensure data.

This informatics platform also is being used to assist the University of Mississippi Medical Center (UMMC), the William Carey College of Osteopathic Medicine (WCCOM), and other training programs involved with health workforce planning. UMMC has been collecting huge volumes of data on a continual basis for many years and WCCOM, although new to medical training in Mississippi, is beginning to accumulate a wealth of data as well. One dataset that both medical schools collect includes information on each graduate and the program in which he/she matched for residency training. An example of a dashboard created for the medical schools to examine their match data is shown in Figure 5. Designing and deploying real-time visualizations of these types of data to administrators, health planners, researchers, and clinicians has great potential to boost advances in health education, research, and patient care.

Data Governance

Data governance is an essential component to a successful data lake. Our data lake is governed primarily by the data owners. Each contributing organization remains the data owner. Our team serves as data stewards -- not deciding how data will be used, but making recommendations, providing ideas and examples, and facilitating these exchanges. Memoranda of agreement are being put in place with data contributors with provisions regarding how data can or cannot be shared. Access is not provided to any data contributor's data without their express permission. This is key in building a *trusted* informatics platform [9]. Without trust and relationships, this initiative would fail before it got started.

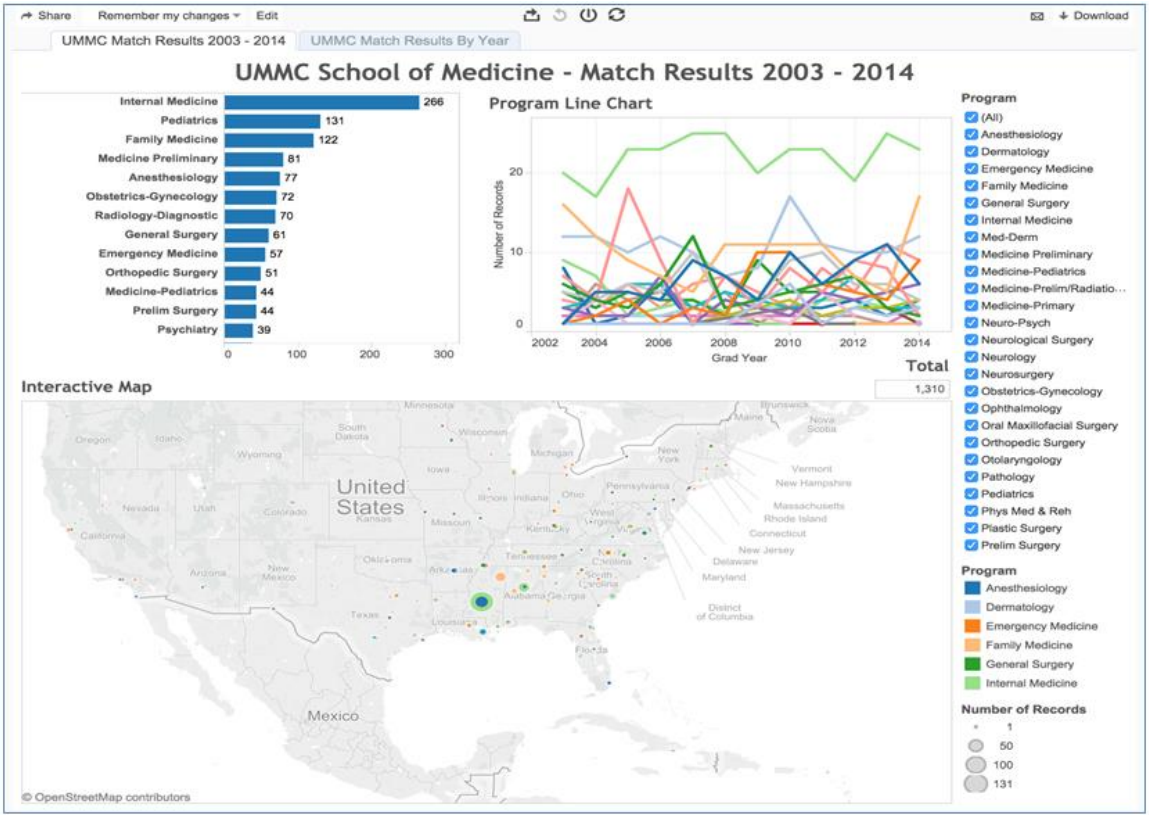

Figure 5. Dashboard designed to display results of medical school match data by program, volume, and location.

Discussion

These are but a few examples of how Mississippi is leveraging a data lake and data visualization tools and techniques to get information back to health planners so that they can make more informed decisions regarding how to best meet the health care needs of the state. To our knowledge, this is the first use of a data lake to be used for state health planning purposes. Many others in the field have been using data warehouses, which are valuable, but costly, slow to evolve, and often bogged down in bureaucracy related to data governance. Data lakes provide a centralized data repository, are much less expensive, and make analytics more readily available. However, they are not without limitations. It is important to plan proactively and address the lack of data governance inherent in data lakes, potential issues with security, and concerns about data quality. With proper forethought and planning, the advantages of implementing a data lake and data visualization and analysis tools can be realized without compromising integrity as described -- by extending the original data lake concept to a *managed* data lake with a data management solution and applying data visualization and analysis tools.

Conclusion

Healthcare has typically been much slower to realize the value of big data than other sectors. Getting data, information, and tools back to the business users who understand how the data should be used is a critical component of deriving value in a timely manner. Information silos

managed exclusively by IT along with heavy data governance structures have inherently been slow to realize such value.

Next steps are to continue to grow the data lake and to work closely with key stakeholders and data contributors to assist them in improving their data collection and quality and setting up strategies to address their most pressing questions. This may involve building visualizations, dashboards, or applications for them or providing the tools and skillsets needed to build their own. We will continue to push some select data to the data warehouse and assist in developing data marts where those are warranted. Assistance to other states interested in efforts to leverage such an approach in their states also can be provided.

The over-reaching goal was to bring data out of silos into a centralized accessible informatics platform in order to get information to those who need it when they need it and to provide the tools and techniques that enable health planners to do more – faster. An affordable, practical, and sustainable model was designed that transforms the field of health planning in Mississippi and could do so in other states. In addition to trying to provide the "perfect numbers" or "best guesses" of how many health care providers will be needed, information is being provided to decision-makers so they will have the knowledge they need when making those tough decisions. From an education and research perspective, the data lake and this platform is being used to support student research projects and further grow research programs in health workforce, health planning, and population health.

Acknowledgements

Thanks to Dr. John R. Mitchell, Director, Office of Mississippi Physician Workforce, and our many partners throughout the state, without whom this work would not be possible.

Conflicts of Interest

The author has no conflicts of interest to report.

References

1. America's Health Rankings. 2014. http://www.americashealthrankings.org/MS. Accessed March 25, 2015.
2. 2011 State Physician Workforce Data Book. Association of American Medical Colleges Center for Workforce Studies; 2011: https://www.aamc.org/download/263512/data/statedata2011.pdf. Accessed March 25, 2015.
3. King AN, Krause DD. Characteristics of Physicians Practicing in Urban, Rural, and Isolated Areas of Mississippi. April 30, 2015, 2015; 11th Annual AAMC Health Worforce Research Conference.
4. Shams K, Farishta M. 2001. Data warehousing: toward knowledge management. *Top Health Inf Manage*. 21(3), 24-32. PubMed
5. Chute CG, Beck SA, Fisk TB, Mohr DN. 2010. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. *Journal of the American Medical Informatics Association: JAMIA*. 17(2), 131-35. PubMed http://dx.doi.org/10.1136/jamia.2009.002691

6. Wiktionary. Definition of Data Lake. 2015. Accessed March 25, 2015. Retrieved from http://en.wiktionary.org/wiki/data_lake.

7. Gartner Inc. Gartner Says Beware of the Data Lake Fallacy. *Newsroom*2014. Accessed March 25, 2015. Retrieved from http://www.gartner.com/newsroom/id/2809117.

8. Schreiweis B, Schneider G, Eichner T, Bergh B, Heinze O. 2014. Health Information Research Platform (HIReP)--an architecture pattern. *Stud Health Technol Inform*. 205, 773-77. PubMed

9. Krause DD. 2013. Building a Trusted Healthcare Informatics Platform: Implemention of the Enterprise Data Warehouse at the University of Mississippi Medical Center. *Journal of the Mississippi Academy of Sciences.* 58(2), 169-75.

10. Krause DD. 2015. State Health Mapper: An Interactive, Web-Based Tool for Physician Workforce Planning, Recruitment, and Health Services Research. *South Med J*. 108(11), 650-56. doi:http://dx.doi.org/10.14423/SMJ.0000000000000369. PubMed