Designing a Robust Analytics Platform for the Analysis of Community Health Outcomes

Nicholas A. Davis
University of Oklahoma
School of Community Medicine
4502 E. 41st Street
Tulsa, OK 74135
nicholas-davis@ouhsc.edu

ABSTRACT

A key challenge for the healthcare industry in the Tulsa region lies in mitigating poor health outcomes and disparities. The OU School of Community Medicine has recently accepted the challenge and implemented a number of key programs designed to improve health outcomes along a number of dimensions. Two critical factors in this endeavor are the transition to electronic health records and the development and implementation of an analytics platform.

Multiple clinical and claims data feeds are integrated into the analytics environment and stored in an enterprise data warehouse. Extraction, transform, and load operations manipulate the raw source data into a form conducive for online analytics processing (OLAP) applications.

In this paper we describe the design of Pentaho, a Java-based analytics suite. The architecture, data flow, and implementation are discussed, along with practical applications. Results include automated care gap and performance email reports, NCQA patient-centered medical home certification, clinical team and patient health management, and meaningful use stage 2 metric evaluation.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Design, Management, Measurement, Performance

Keywords

ACM proceedings, data mining, health care, analytics, EHR, EMR, OLAP, electronic health record

1. INTRODUCTION

The School of Community Medicine at the University of Oklahoma (OUSCM) Tulsa was established to address the tremendous disparities in health outcomes in the greater Tulsa region and surrounding areas. A frequently-cited statistic that underscores the severity of poor health is the fourteen-year delta in life expectancy between zip codes in North Tulsa and those in South Tulsa[19]. As a state, Oklahoma ranks near the bottom in several national metrics for health rankings, including prevalence of smoking, obesity, and cardiovascular disease[21]. Thus, OUSCM has challenged the leaders in the local healthcare community to take bold steps to improve the health outcomes in Tulsa and indeed throughout the state.

OUSCM has closely tracked the shift in the industry from a "fee for service" basis to evidence-based medicine and patient-centered medical home[12], which employs a team of care providers (including clinicians, physician aids, nurses, and other staff) to maximize the efficacy of a patient's treatment while simultaneously reducing cost. A key strategy in the school's mission is the application of health information technology (HIT) in addressing these issues. HIT aids clinicians in the goal of providing timely, equitable, and affordable treatment options. Crucial in the deployment of HIT is the adoption of electronic health records (EHRs). EHRs are used to replace traditional paper charts and records. These have a number of benefits, including[20]:

- EHRs facilitate a more holistic diagnosis of patients
- Physicians receive lab results faster
- EHRs are more cost-effective compared with managing and storing paper
- Physicians are able to schedule more patient visits
- EHRs allow physicians to deliver better patient care

While EHRs provide compelling advantages on an individual patient basis, they also enable an aggregate view of population health and care team management not easily afforded by traditional paper charts. A parallel component in the deployment of EHR systems is the creation of a data warehouse and analytics software for storage and analysis of patient EHR data. OUSCM uses Pentaho, an open source

Java-based business intelligence (BI) platform, for its clinical analytics engine. Following is a discussion of the technical characteristics of OUSCM's clinical analytics implementation, as well as the processes involved in developing a transformative analysis tool for healthcare providers.

2. ARCHITECTURE AND DESIGN

Pentaho was chosen based on its support of a variety of databases, web-based analytics and reporting interface, and powerful data integration tools used for extract, transform, and load (ETL) operations. Being an open source project backed by a commercial entity conferred several advantages to Pentaho when compared with the competition. As with any open source application, the underlying Java source code can be readily evaluated for security vulnerabilities and software bugs. The software was also validated for conformance to OUSCM requirements prior to purchasing a support license. Additionally, Pentaho proved more cost effective than comparable analytics offerings that were considered.

OUSCM receives data from a handful of sources and uses Pentaho to provide analytics, reporting, and health performance dashboards. Data sources include clinical feeds as well as insurance claims data. These data are collected and processed using a sequence of steps described in detail below.

2.1 Data Sources and Formats

The OUSCM data warehouse stores clinical data generated from a number of sources, including the Schusterman Center Clinic, Bedlam Clinics[4], OU Family Medicine Center, specialty clinical services in neurology, urology, and minimally-invasive surgical procedures[5]. Source systems include the electronic medical record (EMR) system and the practice management system (PMS). Additionally, lab and test results are outsourced to regional laboratory services, so these external feeds are incorporated in the EHR data. Clinical data is complemented by an insurance claims data feed. Each data format, described in detail below, is commonly used in healthcare for conveying clinical details regarding patients or processing medical insurance claims for payment and clinical information.

These various formats contain a blend of both highly informative and less useful data elements for mining and knowledge extraction. Table 1 summarizes the data formats and sources used by OUSCM. These are described in detail in the sections below.

Table 1: OUSCM data formats

Format	Source	Usage
HL7	EMR systems/labs	Clinical support
CCD	EMR systems/labs	Clinical support
Delimited text	OHCA	Care mgmt
Delimited text	OHCA	HAN encounters
UCE 1500	OHCA	Institutional claims
UB 92	OHCA	Provider claims
NCPDP	OHCA	Pharmacy claims
XML	Doc2Doc	Referrals

2.1.1 Clinical Data Formats

Clinical data elements are stored using the native databases integrated in the EMR and PMS vendor products. Messages are transmitted to external systems using the Health-Level Seven (HL7) version 2 standard and format[7], as well as the Continuity of Care Document (CCD) format[18], based on a collaborative effort between HL7 and the American Society for Testing and Materials[3]. The CCD format is based on the HL7 Clinical Document Architecture (CDA) information model. Colloquially, the HL7 version 2 messages are referred to as "HL7", and the Continuity of Care Document format as "CCD".

There are a number of distinctions between the two clinical formats:

1. HL7 is ASCII character-delimited, while CCDs are XML-based
2. HL7 messages are relatively small in file size (tens of KB and smaller) compared with the more verbose XML-based CCD (hundreds of KB to several MB)
3. The HL7 ASCII format has existed since the late '80s, while the CCD is a more recent development (early 2000s)
4. As a result of 3, EHR vendor support of CCDs is nascent compared with the more established and older HL7 format, as of this writing

A wide array of data elements is captured from clinical events including provider encounters, diagnoses, procedures, lab orders and results, allergies, medications, immunizations, etc. Demographics data is also provided by most patients, as well as ancillary details such as insurance information, guarantor, and next of kin. Demographics data can include age, gender, ethnicity, place of residence, and contact information. When viewed in aggregate, this data provides a snapshot of population health from the Tulsa community.

2.1.2 Claims Data Formats

Complementary to the clinical data is the availability of insurance claims data from the Oklahoma Health Care Authority (OHCA)[17], which is Oklahoma's Medicaid Agency. OUSCM receives a claims feed for patients in its Sooner Health Access Network (HAN)[15], a network of SoonerCare (Oklahoma's Medicaid program) Choice providers that partners with OHCA to support care management in a Patient-Centered Medical Home (PCMH) model.

OHCA sends data in a number of formats: UCE 1500, UB 92, and NCPDP[9]. UCE 1500 is used by providers for insurance claims and is based on the standard HCFA 1500 used by CMS. UB 92 is similar in nature, but focuses on institutional (i.e. hospital and clinic) provided services, apart from those rendered by the patient's provider. Lastly, the NCPDP format is used to transmit prescription fill data from pharmacies.

Both the UCE 1500 and UB 92 are character-delimited files, while the NCPDP is an Electronic Data Interchange (EDI)-based format standardized by ANSI ASC X12[2]. In addition to the fee for the service(s) rendered, these formats

contain a wealth of clinically-relevant data on patients. Diagnosis codes are available on some claims using the ICD-9 coding system[14], an international standard code set for tracking diseases. Provider details are captured across all formats, as well as patient demographics data elements.

2.1.3 Other Formats

In addition to its claims/billing data, OHCA also sends encounter-based clinical data to OUSCM. There are two primary feeds: one for Sooner HAN patients and another for elevated status patient data. The former is derived from clinical encounters and includes ER visit details, in-patient visits, and screenings for conditions such as breast cancer, cervical cancer, and early and periodic screening, diagnostic, and treatment (EPSDT)[6] data for children. The latter represents patients who require an elevated level of care and treatment, and include high risk pregnancies, cancer patients, and individuals with hemophilia. Both feeds consist of character-delimited text files.

Doc2Doc[1] data is also available in Pentaho. Doc2Doc is an electronic referral management system used to refer a patient to a specialist. The service allows referring providers to keep track of their patients' statuses, as well as allow clinical office managers to track productivity. This feed arrives in an XML-based format.

2.2 Analytics Architecture

OUSCM's analytics environment is composed of a number of distinct architectural elements, including a data warehouse and application servers. Clinical data feeds are received from OUSCM EMR and practice management systems and insurance claims data from OHCA. Software used to process the data and perform analysis is provided via the Pentaho tool suite, accompanied by a number of data marts provided in the data warehouse. At a high level, the OUSCM analytics environment uses the following enterprise (commercially-supported) software components:

- Analysis Services, an online analytical processing (OLAP) engine based on the open source Mondrian[10] project
- BI Platform, a framework that provides core services such as logging, auditing, web services, and rules engines. It also integrates reporting, analysis, dashboards, and data mining components[13]
- Dashboard Designer, used to create an analytics dashboard that may include charts, data tables, external web sites
- Pentaho Data Integration (PDI) or Kettle, a suite of tools for performing ETL operations in data processing

Figure 2 illustrates the overall architecture currently in place. The top components represent distinct data sources, both internal and external. Each raw data feed is transformed using PDI and placed in the enterprise data warehouse (EDW). The EDW here is represented conceptually as the entire analytics environment, including the historical data store (HDS), custom data marts, and Pentaho web interface for reporting and user-driven analytics.

2.2.1 ETL Tools

The clinical and claims data feeds arrive in their native formats and must be processed in a sequence of steps. Via PDI, business intelligence developers construct intricate data ETL pipelines using a GUI-based visual design tool. Complex workflows include a number of steps to process raw data feeds. At OUSCM several of these workflows have been created to transform the source data into a state amenable for analytics operations.

PDI/Kettle provides a few components to perform ETL operations. The heart of its data processing capabilities lies in Pan, which is a data transformation engine that supports a variety of input and output formats for reading and writing data. Pan also incorporates a host of data manipulation tasks, such as the ability to extract columns or other subsets of data, sorting functions, and mathematical operations.

Kitchen is the job execution and scheduling engine. A job consists of a number of discrete data transformations, or steps in a sequence. Jobs can also be nested within each other. Kitchen executes these jobs, in either XML format or in a database repository. At OUSCM jobs are scheduled to run on a regular basis at specific times in batch mode. Both Kitchen and Pan are command line tools.

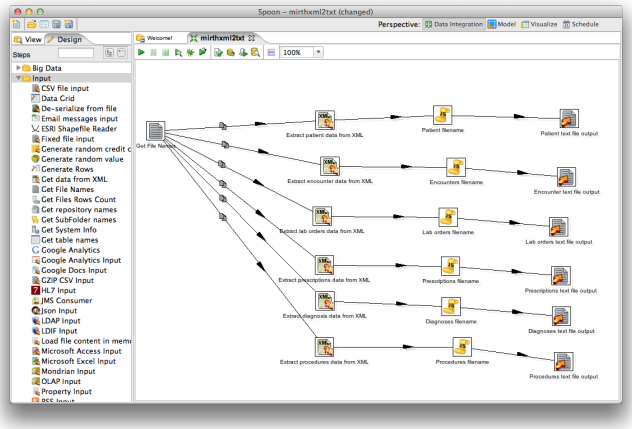

Figure 1: Spoon, a PDI GUI tool used to create ETL workflows

A visual design tool called Spoon is used to create a series of jobs containing several data transformation steps. Spoon, shown in Figure 1, facilitates the design of complex workflows whereby data is transformed in various stages. Each transformation is represented visually and multiple connections can be made between transformations. While some of these discrete steps represent input, manipulations, or output, others deal with error conditions and logging. Email notifications are typically a component of each job to notify BI developers of job failures and, in some cases, successful completion. Thus, error handling and logging are managed with each job.

2.2.2 Data Flow

There are a series of steps involved in a typical data workflow, for example, diabetes metrics. First, data is fed from

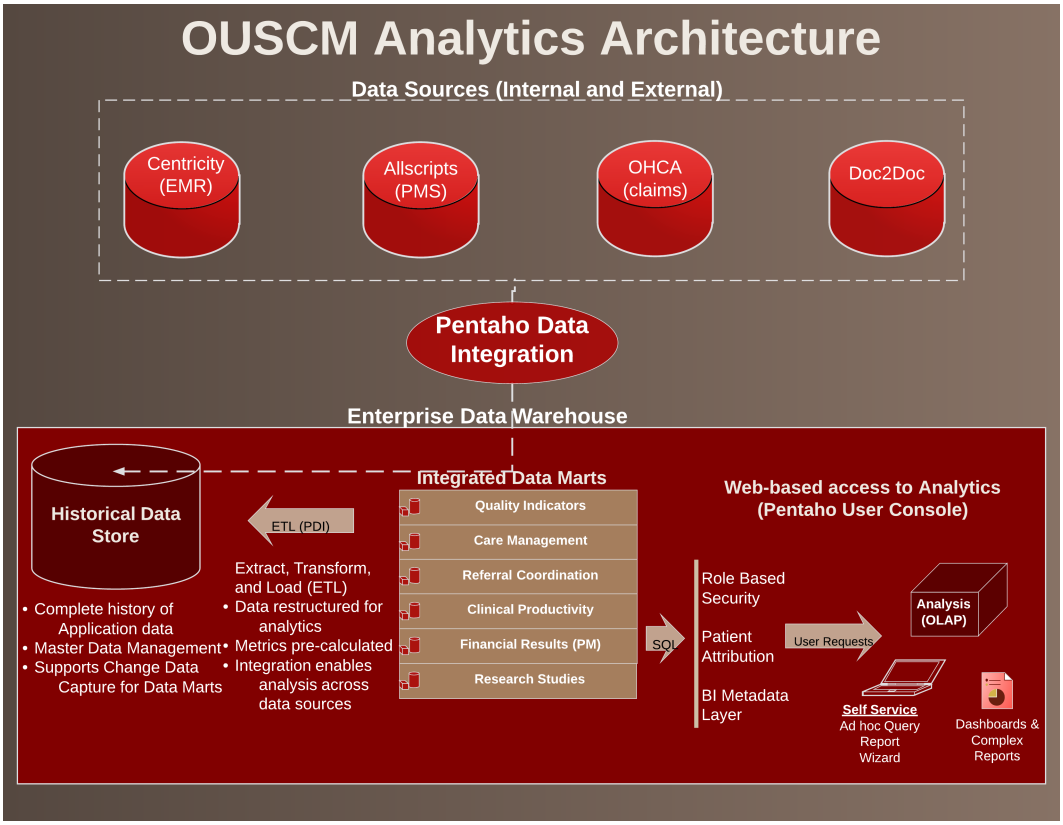

Figure 2: OU School of Community Medicine Analytics Architecture

the source systems (both internal and external) and brought into a staging area. This houses the raw data and includes observation header, values, location, person details, race, condition/problem codes (ICD-9).

Next, data is manipulated using ETL and stored in the HDS (see Figure 2). Data from the HDS is then used to develop dimensions in the data warehouse. For diabetes, these include a care team dimension, observation terms dimension, location dimension, assigned PCP, and patient dimension. These dimensions may incorporate data from multiple staging sources.

The final step in the process involves the construction of a fact table. This consists of measurements derived from the source data that has been massaged into an appropriate format in the surrounding dimension tables. In our diabetes example, each of the dimensions listed above is incorporated into a health value fact table specific to diabetes metrics for patients.

2.2.3 Web Interface

One of the key advantages of Pentaho compared with other analytics offerings, such as SAS or SPSS, is the inclusion of a robust web-based user interface, called Pentaho User Console. This allows a multitude of devices to interact with the application and perform data analysis functions. With version 4.5, the interface has been optimized for mobile devices such as tablets and smartphones.

The initial interface allows users to create a new report, analysis, or dashboard. A report is a static view of data that can be exported to a variety of formats such as PDF, Excel, CSV, and HTML. The purpose is to create a document that can be exported and shared with others. In contrast, an analysis is dynamic and encourages exploration of the data by including a number of rows, columns, and measures. Data in an analysis can be viewed in tabular format or as a variety of charts (bar, line, pie, scatter plot). Dashboards include a number of panels representing charts and tabular data arranged in a grid pattern. Additionally, each of these can be saved and stored on the server for subsequent usage.

Privacy and security are critical in the environment, and HIPAA compliance is a strict requirement. User administration and access control are incorporated in the User Console. Users and groups are created by the application administrator based on roles. Dashboards can be built that only show a portion of the data based on the user's role. For example, a clinic manager can only view the data from the clinics he or she manages. Similarly, a physician may only view details of patients under their care. Pentaho also has the ability to mask protected health information (PHI), such as names, contact information, and specific location. This allows a staff member to view data without revealing the identity of a patient. Thus, patient privacy is taken into account and enforced via these security features of the system.

2.3 Implementation

Being a Java-based platform, Pentaho is supported across several operating systems and hardware architectures, including Windows, Mac OS X, and Linux, in both 32- and 64-bit implementations. Currently, Pentaho 4.5 is deployed and there are plans to upgrade to 4.8 in the coming months.

OUSCM employs Windows Server 2008 as the operating system across all servers. There are three primary production servers, including a data warehouse and two application servers running Pentaho BI server and ancillary software (PDI tools, Tomcat Java app server). The data warehouse is an RDBMS running SQL Server 2008 R2.

Each production server is a Dell PowerEdge R910 with the following specifications:

- Intel Xeon 7500 CPU with 32 cores
- 256GB of DDR3 RAM
- 1 TB of 15000 RPM SAS storage in a RAID 10 configuration

Additional servers are used for development and testing purposes. The specs for these servers differ from production systems described above, but the software is nearly identical. However, when upgrading versions of any component (app server, OS, database), these versions may be temporarily mismatched until the migration is complete.

A load test was recently completed to gauge the performance of the analytics environment. HP LoadRunner was used to simulate a number of concurrent users, and run over a period of 2 hours. This test simulated a user running a set of saved Pentaho Analyzer reports in the Pentaho User Console. The OUSCM configuration was able to sustain 85 concurrent users with Pentaho 3.8. Additional load testing will be performed to determine the performance of Pentaho 4.5/4.8.

3. RESULTS

Pentaho and the analytics environment at OUSCM were instrumental in the school's achievement of PCMH certification through the National Committee for Quality Assurance (NCQA) in spring of 2013. OU Physicians teaching clinics have received Patient-Centered Medical Homes 2011 Recognition – Tier III[11]. As of this writing, the clinics are the only medical practices in Northeast Oklahoma to earn this distinction.

Related to PCMH, OUSCM clinicians use the analytics tools to manage teams of care providers as well as patients seen by the teams. Care gaps (areas where the patient may need additional treatment) and performance reports are generated and emailed to team leaders on a daily basis. These performance reports track on both an individual and team basis, and include population health metrics. Examples include the percentage of patients with elevated HbA1c (average level of glucose, used in diabetes diagnosis and treatment) values and wellness screenings. Screenings include asthma, diabetes, COPD, breast cancer, and cervical cancer. The reports also compare team population statistics and help physicians improve the screening rates and other health measurements. Temporal trends can be established to determine whether the average levels of HbA1c and other quantitative measurements are generally improving among a team's patients and the organization as a whole.

OUSCM was recently chosen as the test site for a new Meaningful Use Stage 2 metric[8], Closing the Referral Loop. This was a joint effort among the Office of the National Coordinator for Health IT (ONC), OUSCM, and NCQA. OUSCM was chosen primarily based on its implemented analytics architecture, specifically surrounding the reporting around electronic referrals. The data from Doc2Doc was used within Pentaho to create reports and dashboards used for the metric.

Academic research involving claims and clinical data are facilitated by the platform. A current project is medication adherence in various population segments using pharmacy claims records. Both the data warehouse and Pentaho are being leveraged to explore trends and identify areas where interventions are needed to improve adherence rates. Similar projects are planned that will utilize the healthcare data and scientific computing techniques to help researchers and clinicians understand community health trends.

The analytics platform provides services that will ultimately assist in the delivery of better, more affordable care to patients seen by University of Oklahoma Physicians Group (OUP) providers. Pentaho is being actively used in programs to analyze and treat patients with conditions including asthma, diabetes, and obesity. Physicians are using the software to help manage teams through the PCMH model, which improves access and patient engagement.

4. CONCLUSIONS

Pentaho provides a robust analytics engine and tool chain for OU's School of Community Medicine clinicians and staff. Disparate data sources are weaved together to support a variety of functions for analytics. Providers are offered a compelling tool suite to help manage their teams and patients, as well as explore trends in patient health. Research scientists have a platform with rich analytics tools to ask questions of the data and uncover patterns in community health. Public health outcomes can be tracked on an aggregate level, which can provide a useful rubric of organizational performance.

As is often stated, "just because you build it doesn't mean they will come." A formal program will be established at some point in the future for training and support purposes, to encourage broad adoption of the platform. Future automated reports may include a link to the analytics web portal to provide more details and encourage exploration. There are additional steps that will improve the quality of treatment and research initiatives. One area of improvement is the installation of additional components offered by Pentaho, such as the Weka[16] machine learning (ML) framework. Weka is Java-based, like Pentaho itself, and is available as a companion module in many of the existing Pentaho processes. Weka offers a rich suite of ML algorithms and will be used to mine patterns in health quality, treatment options, and healthcare costs.

Adoption of the analytics platform is still in the early stages, but is increasing as providers are shown the capabilities and benefits. More widespread adoption of Pentaho among OUSCM physicians and staff will come with expanded support and training options. Lastly, as additional data sources become available they will be integrated into the analytics architecture, enabling the improvement of patient health, disease management, and research opportunities.

5. ACKNOWLEDGMENTS

We'd like to acknowledge Jim Craddock and Kristian Foster for helpful discussions regarding the analytics workflows. Alan Gunderson provided technical details of the software and hardware used in the environment. Jeff Alderman and Renee Engleking provided a good overview related to healthcare quality initiatives at OUSCM. Other members of the OUSCM Medical Informatics Department also provided details that assisted in the development of this paper.

6. REFERENCES

[1] Oklahoma Doc2Doc Study. http://www.doc2docstudy.org, 2012.

[2] Accredited Standards Committee. http://www.x12.org, 2013.

[3] ASTM International - Standards Worldwide. http://www.astm.org, 2013.

[4] Bedlam - Tulsa - The University of Oklahoma. https://www.ou.edu/content/tulsa/community_medicine/bedlam.html, 2013.

[5] Clinics and Maps - Tulsa - The University of Oklahoma. http://www.ou.edu/content/tulsa/ou_physicians/clinics.html, 2013.

[6] EPSDT & Title V Collaboration to Improve Child Health, 2013.

[7] HL7 Standards Product Brief - HL7 Version 2 Product Suite. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=185, 2013.

[8] Meaningful Use Stage 2 Criteria | Policy Researchers & Implementers | HealthIT.gov. http://www.healthit.gov/policy-researchers-implementers/meaningful-use-stage-2, 2013.

[9] NCPDP. http://www.ncpdp.org, 2013.

[10] Open source analysis OLAP server written in Java. Enabling interactive analysis of very large datasets stored in SQL databases without writing SQL. | Mondrian: Pentaho Analysis. http://mondrian.pentaho.com, 2013.

[11] OU Physicians Receive National Committee for Quality Care Assurance Recognition for Patient Care. https://www.ou.edu/content/dam/Tulsa/scm/pdf/NCQA%20NR%200413%20(2).pdf, 2013.

[12] Patient-centered medical home. http://www.ncqa.org/Programs/Recognition/PatientCenteredMedicalHomePCMH.aspx, 2013.

[13] Pentaho BI Platform / Server. http://community.pentaho.com/projects/bi_platform, 2013.

[14] Resources - ICD 9 Medical Coding - AAPC. http://www.aapc.com/resources/medical-coding/icd9.aspx, 2013.

[15] Sooner Health Access Network: The University of Oklahoma. http://soonerhan.ouhsc.edu, 2013.

[16] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. http://www.cs.waikato.ac.nz/ml/weka, 2013.

[17] Welcome To The Oklahoma Health Care Authority. http://www.okhca.org, 2013.

[18] J. D. D'Amore, D. F. Sittig, A. Wright, M. S. Iyengar, and R. B. Ness. The promise of the CCD: challenges and opportunity for quality improvement and population health. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:285–94, Jan. 2011.

[19] F. D. Duffy. How Can MyHealth Access Network Help My Practice? http://greaterthan.securespsites.com/event/Provider%20Summit%20Presentations/Duffy-How%20Will%20an%20HIO%20Help%20Me%20Provide%20Better%20Quality%20Care%20pdf.pdf, 2011.

[20] E. Jamoom, V. Patel, J. King, and M. Furukawa. National perceptions of EHR adoption: Barriers, impacts, and federal policies. In *National Conference on Health Statistics*. National conference on health statistics., 2012.

[21] S. Muchmore. Oklahoma health ranking improves slightly | Tulsa World. http://www.tulsaworld.com/article.aspx/Oklahoma_health_ranking_improves_slightly/20121211_17_a14_oklaho390289, 2012.