Massachusetts Institute of Technology
Engineering Systems Division

ESD Working Paper Series

Deployment of Analytics into the Healthcare Safety Net: Lessons Learned & Unlearned

David Hartzband
Research Affiliate
Sociotechnical Systems Research Center
Massachusetts Institute of Technology
Email: dhartz@mit.edu

ESD-WP-2015-12 December 2015 esd.mit.edu/wps

Deployment of Analytics into the Healthcare Safety Net: Lessons Learned & Unlearned

David Hartzband, D.Sc. Research Affiliate,
Sociotechnical Systems Research Center
Massachusetts Institute of Technology

Introduction

I have been working with the RCHN Community Health Foundation (rchnfoundation.org) for almost 10 years. The RCHN Community Health Foundation (RCHN CHF) is a not-for-profit operating foundation whose mission is to support and benefit the work of community health centers (CHCs) nationally. I am their Director of Technology Research., & in this role I spearhead the organization's continued evaluation, assessment and findings dissemination related to health information technology. I also do research[1] on emerging technologies, primarily in ultra large-scale software systems, machine learning & application of machine learning to current software systems, & I work with a number of technology start-ups, many of them in healthcare information technology.

In October of 2013, I made a proposal to the Foundation to start a project that would deploy a contemporary analytic software capability into community health centers that volunteered for the project & to work with their IT & executive staffs so that the capability could be productively used as part of how the health center made strategic decisions[2]. I wrote at the time: *"Everyone agrees that "analytics" are/will be important for health centers as they evolve to new organizational (participants HIEs, ACOs, HCCNs etc.) & sustainability (service providers, data providers) models. What this means & how to do it are hotly discussed topics, however, with no apparent tactic or strategy that seems feasible. There is no big bang in this effort. This capability will not spring forth complete & productive if health centers make the correct invocation or even spend a large amount of money. This memo specifies a program that would pilot an actual path for health centers (& other healthcare organizations with limited resources) to follow to begin to productively use analytics & to evolve a more & more effective capability in this area."*

I also wrote that: *"Complex analytics, multi-layered analytics and highly designed data warehouses are not necessary, and moreover, not appropriate if the questions that are asked aren't relevant or don't require them and the underlying data isn't complete and reliable. "*

That was just over two years ago. What happened with the project & what is going on with it now? What lessons have been learned? What lessons did we already know but needed to have reinforced by painful experience? Here is a project update.

---

[1] Research Fellow, Sociotechnical Systems Research Center, Massachusetts Institute of Technology

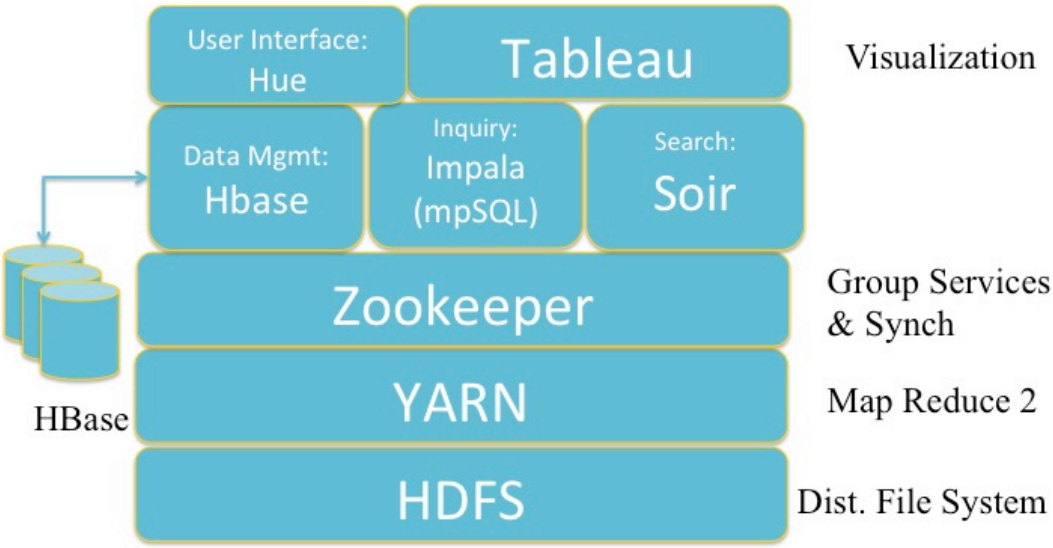
[2] The support of the RCHN Community Health Foundation is gratefully acknowledged, especially that of Feygele Jacobs, the Foundation's President & CEO

Project Initiation

By the early summer of 2014, two CHCs had volunteered to begin the project – one urban & one (very) rural. One more urban CHC & a Primary Care Association were added by late winter 2015. At that point, the project encompassed 33 health centers with over 400 sites & 1.3M patients/year for 2-3 full data years (depending on CHC). These participants will be referred to as PCA1, Urban CHC1, Urban CHC2 & Rural CHC1.

Each engagement began with a face-to-face meeting with the CHCs' IT & executive staffs. The goals of the project were discussed as was how to think about analytics as an enabler of strategic decision-making. Questions ranged from: "Isn't this just UDS[3] so why do we need it?" to "We already have more technology than we can use, so why do we need it?" to "Is this big data & if so, why do we need it?" In each case, the CEO &/or COO of the health center was interested & committed to the project so a good deal of time was spent (10 hours in a month for one CHC) working through how this was different than UDS, how analytics could be used to support decision-making, what the technology was & why it was needed (instead of what they already had) & many other topics/issues that came up. During this time, I was also meeting with the IT Staff to work out deployment of the new technology & arrange for data to be available. Deployment, data acquisition & testing was very dependent on both the amount of resources a CHC had to devote to it & the level of capability of the resource(s). At a minimum, the skills required to do the deployment, data acquisition & testing included: system management skills (program & application installation), database capability (schema awareness, SQL programming), network & connectivity troubleshooting, & testing skills.

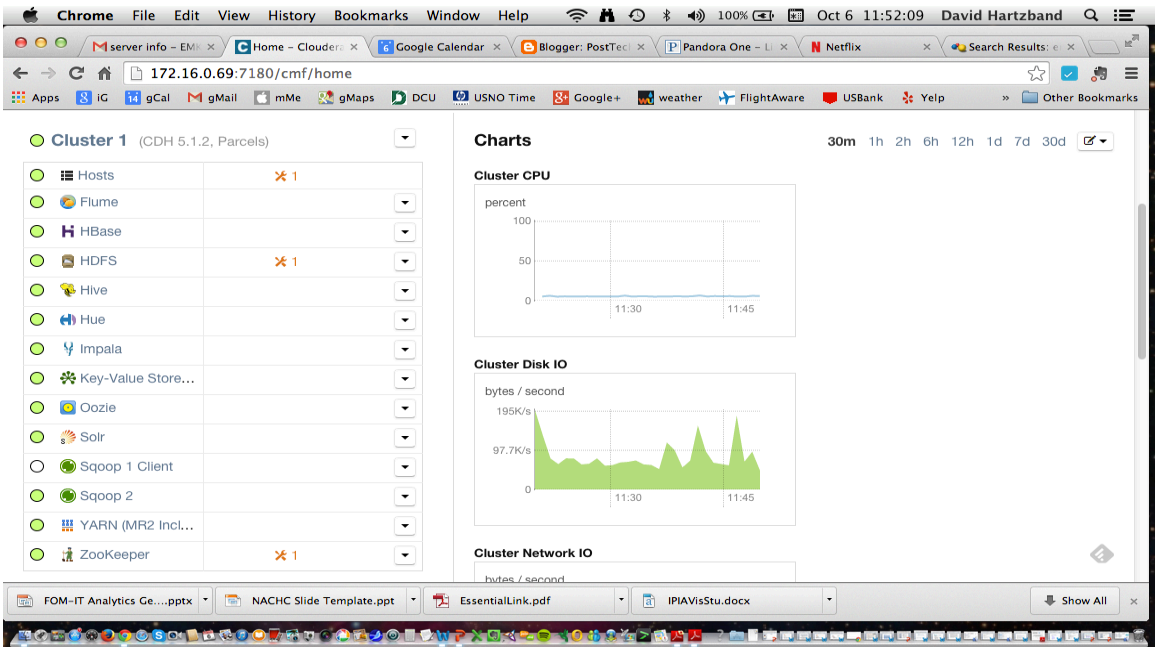
OK – so what technology did we deploy & what inquiry did the project start with?



[3] Uniform Data System yearly reporting required by HRSA (Health Resources & Services Administration, HHS)

The figure above shows the components that were deployed. They are all part of the Cloudera Express open source Hadoop deployment. Cloudera is a company that was started by people who came from Yahoo where Hadoop was originally developed. What's Hadoop & why did we choose it for the P2A project? Good questions…"Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware."[4] It was developed primarily by Doug Cutting (& Mike Cafarella) who was working at YAHOO at the time. He named it after his son's toy elephant. The framework consists of: 1) a massively parallel-distributed file system (Hadoop Distributed File System, HDFS), 2) MapReduce (currently Yarn MR2) which is a programming model & implementation for processing & creating very large data sets using parallel distributed algorithms, 3) Zookeeper, a centralized service for providing configuration, naming & synchronization services for distributed applications used with Hadoop systems, & 4) a large set of ecosystem applications, the most important of which for P2A are: 1) HBase, a non-relational data storage system optimized for very large data sets (billions of rows x millions of columns) optimized for use with HDFS & other Hadoop applications, 2) Impala, a query engine from Cloudera [5]that allows SQL queries to be run against HDF & HBase data, & 3) Hue, an web-base interface that supports Hadoop. All of these components are supported under an Apache open source license & available as part of the Cloudera Express deployment. The final piece in the figure is Tableau, an independent product (not open source or part of the Cloudera distribution) that provides visualization for large data sets. Visualization is often helpful when explaining complex analyses.

The next figure shows a screen shot of this system running at one of the project sites.



[4] https://en.wikipedia.org/wiki/Apache_Hadoop

[5] http://www.cloudera.com/content/www/en-us.html

Deployment took between 2 weeks & 3 months depending on the resources & personnel that each CHC had available. It was most effective at the PCA that did a centralized deployment for 30+ CHCs, had a IT staff that was experienced with provisioning Linux distributions & also had a staff member that had experience with Hadoop deployment. This group was able to go from the download of the Cloudera Express distribution to full deployment & provisioning of HDFS/HBase from their data warehouse in about 2 weeks.

Deployment was least effective at the Rural CHC1 that had a minimal IT Staff, only accessed their clinical data through an intermediate BI tool & did not (initially) have permission to access the database underlying their EHR. Deployment & data provisioning here took multiple months & was complicated by the design of the database underlying the EHR (that had >1000 tables most of which were redundant or unused).

Initial Inquiry

An initial inquiry or set of queries is done at each site after deployment is complete. This exercise acts as a final test of the analytic system & also is meant to allow the normalization of data, including data definitions, between the CHC's normal systems & the analytic stack. The exercise, called "level-up" consists of a number of queries performed both through the CHC's regular systems (EHR, SQL, BI tools) compared with the same queries performed on the analytic stack with the data in the HDFS/HBase information store. The following queries are performed[6]:

- # patients/year
- # patients/diagnosis/year (hypertension, diabetes, obesity, heart disease, behavioral)
- rank order of prevalent comorbidities
- (cost/patient/year)[7]
- (cost/comorbidity/year)

This exercise has been carried out at all sites except for rural CHC1.

Results for the level-up exercise have been,… instructive. Again, there was a large difference in the time that it took to run the queries among the sites. The PCA had this done within 2 weeks of completing deployment. Urban CHC1 required about a month, but Urban CHC2 required 5-6 months. There were several reasons for these discrepancies generally falling into 3 categories:

1. Differences in definitions used – The UDS definitions were strictly used for the P2A (analytic stack) queries, but each CHC deviated from the UDS defintions, often in major ways. Here are some examples:

[6] UDS definitions are used for all terms including: visits, patients & conditions, http://www.bphcdata.net/docs/uds_rep_instr.pdf

[7] actual cost (expenditure), not billed cost (revenue)

a. Each CHC had different, & in some cases quite idiosyncratic, definitions in use for visits. In most cases, these included visits that would have qualified under current (& recent) instructions (p. 17, *op. cit*).
b. Patients were defined in many different ways including one CHC that excluded dental patients from their counts (including what was reported to HRSA) if the patient had dental encounters, but not "medical" encounters (*i.e.* did not have a primary care provider assigned).
c. Although HRSA provides lists of ICD-9 codes to be reported for each diagnosis (condition), each CHC reported a different collection of codes as representing a particular condition. In addition, many CMOs reported that their providers did not generally report certain conditions (see below for obesity)
d. In at least one case, the CHC was reporting the number of patients by diagnosis if they: a) had a medical encounter during the data year, & b) had ever been diagnosed with the condition regardless of when. This is in contradiction to the instructions for Table 6A, Column B, Lines 1-20d (p. 76 *op. cit.*).

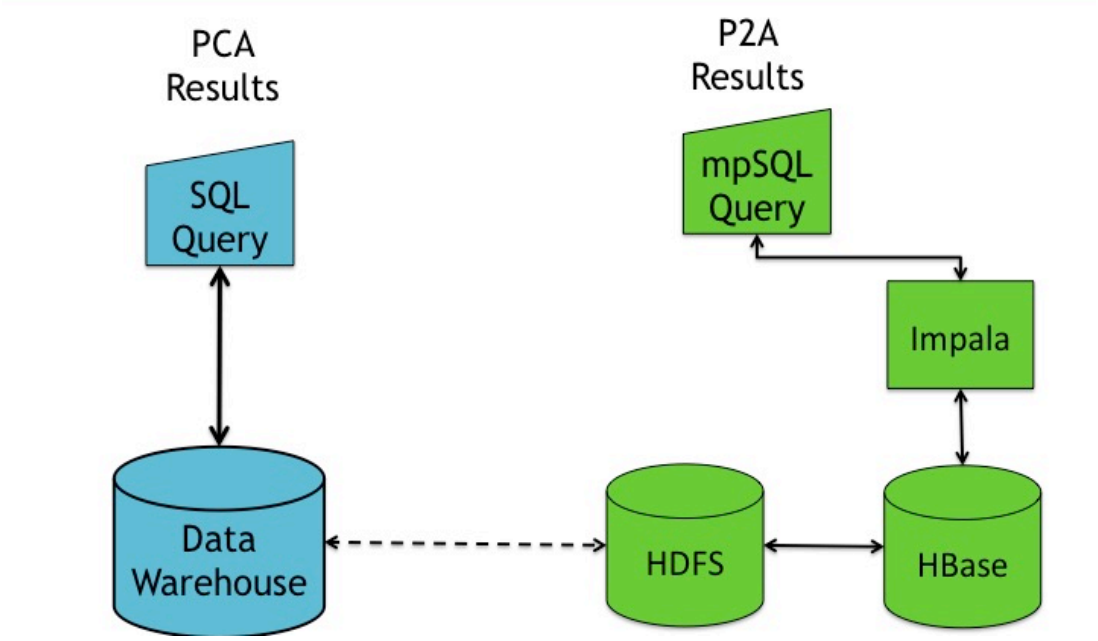
2. The PCA & Urban CHC1 had already done substantial normalization associated with the design & deployment of a data warehouse. Each of these organizations had both a data dictionary deployed & a written document with standard definitions. Even when the normalized definitions differed from the UDS definitions, they were able to be identified & modified much more easily than those centers that had not done this work & where the database schema or SQL code had to be examined to determine what definitions were used.
3. Issues with EHR structure & function – There 4 different EHRs in use as well as a variety of practice management, cost accounting & financial systems. EHRs included: NextGen, SuccessEHS, eClinicalWorks & GE Centricity.
    a. Each EHR dealt with ICD-9 differently, but each had some anomalies with queries based on the codes. Generally the following gave **different** results upon query: 250, 250.0, 250.00 & 250*.
    b. In addition 250* general did not return the same results as an enumeration of all the 5-digit codes (250.01, 250.02, 250.03, 250.10 ,…, 250.93)
    c. It was also found that navigation is complicated enough that it was very rare for more than one diagnosis to be recorded per encounter. This may also have to do with how encounters are reimbursed.
4. Lack of alignment between clinical (EHR) & financial (cost) data – While it is easy to associated a specific encounter with billing data (potential revenue) as this data is carried in the practice management & eBilling systems, it is quite difficult to associate actual costs with clinical encounter data. Cost accounting systems are separate from the PM & EHR systems used for clinical & billing data, & they are organized quite differently. Rather than keying on encounter, patient etc., they are usually organized around location, time period &/or payer. In no case did we find a direct key linking encounter data to actual cost data. This makes it difficult to analyze any clinical data in association with anything but billing (revenue) data, while what is actually required is the analysis with cost data.

Preliminary clinical results from the Level-Up exercise

Despite the difficulties described above, we have results for patients/condition/year for all but one of our participating CHCs. Results presented here are preliminary & should not be taken as anything other than an indication of trends. The project deployed all software within the security perimeter of the participating CHCs, the Foundation had BAAs with all relevant parties, all results are HIPAA de-identified & nothing but results ever left the security perimeter of the CHC. Results are presented for the Primary Care Association only as they are the most complete & are indicative of trends in the other result sets. The PCA's data set represents 30 CHCs with close to 300 sites.

	PCA	%		P2A	%		U.S.%
DX	2012	2013	2014	2012	2013	2014	
Hypertension	20.73	22.46	22.78	20.5	23.04	23.24	29%
Diabetes	6.59	6.56	6.58	6.49	7.26	6.70	9%
Obesity	4.38	10.00	11.68	4.37	10.23	11.78	35%
Heart Disease	1.43	1.59	1.63	1.39	1.61	1.63	11%
total Patients	350311	403286	440713	331968	384515	421159	

Results (table above) are presented for three data years (DX): 2012, 2013 & 2014. The PCA results were generated from SQL queries made directly to the PCA's data warehouse, The P2A results were generated by mpSQL queries made through Impala to the HBase representation of the data imported into HDFS. Results represent the percentages of each condition in the covered population. Total patients for the three data years are 1,194,310 in the PCA data & 1,137,642 in the P2A data (P2A figure = 95.3% of PCA figure). Percentages are compared with the CDC Fast Stats figures for the U.S. population as a whole (U.S.%).



Results for diabetes & hypertension are low compared to the CDC figures, but are generally within 75%-77% of the U.S. population. Results for obesity & heart disease are, however very low with obesity ranging from 14%-34% & heart disease about 14% of the CDC figures. These percentages are consistent across all CHC sites including the non-PCA CHCs.

As previously stated, results for the other CHCs in the study (3 CHCs, 45-50 sites, ~75,000 patients per year) were in line with the PCA results. All of them showed lower than expected figures for all conditions, but especially for obesity & heart disease.

In all cases, comorbidity results were not usable because of the low condition percentages. No calculations of cost vs. condition or comorbidity were made because in all cases: 1) it was not clear how to associate the actual cost data with clinical results, 2) comorbidity percentages were so low, generally <1%, that such calculations would not have been meaningful, even if possible.

Discussion of Preliminary Results

The Path2Analytics project is still in progress, but several results already stand out. The first is the large range in the ability of community health centers to deploy, provision & utilize Hadoop-based analytics. Next is the quality of the data available for use in analysis. Additionally, results of an initial analysis did not meet expectations for population health conditions, especially with respect to obesity & heart disease. Finally, lack of alignment with cost data makes analysis of such calculated parameters as cost/diagnosed condition/ patient difficult while the lack of credible comorbidity

data made it infeasible to calculate cost/comorbidity/patient, even if cost data had been available. Each of these issues will be discussed in more detail.

Condition Percentages

The real question is why are the percentages of chronic disease so low in these populations. We do not generally consider patient populations of community health centers to be as healthy or healthier than the general U.S. population, yet in all cases, condition percentages are below expectation & they are especially low for obesity & heart disease. I have explored this anomaly with respect to obesity in discussions with many of the Chief Medical Officers & other staff at participating CHCs. Many of them estimated their population at ~40% obese overall. A recent paper in the Journal of the American Medical Association, Internal Medicine estimated that 40% of men & 30% of women were overweight & that 35% of men & 37% of women were obese[8]. The estimate by CMOs of 40% would be in line with this study, but again the expectation might be that percentages for health center populations would be above this figure. CMOs & others thought there might be two explanations for the low percentages of obesity occurring as a diagnosis in their EHRs. The first is that most CMOs knew that their providers did not often diagnose obesity & when they did, they did not use the full range of ICD-9 codes. There are three specific codes for obesity (278, unspecified obesity; 278.01, morbid obesity, BMI>30; & 278.02, overweight, BMI>25). These are hardly ever used. The UDS guidelines specify the use of the 22 V-codes for obesity that give a highly specific breakdown of BMI measurements. Apparently these codes are used even less frequently than the 278 codes,… but why. Most people thought that there was a sociocultural bias against making this diagnosis & that in some demographics being overweight or obese was considered attractive or at least not unacceptable.[9] A recent paper in PLoS ONE[10] compared ICD-9 data reported in the U.S. Inpatient Reporting Sample (NIS, HRSA) to interview data reported in the Behavioral Risk Factor Surveillance System (BRFSS, CDC) for 2011 data & found that in the NIS data, that is hospital discharge data, the percentage of obesity reported was ~9%, & that in the BRFSS or interview data, the percentage of obesity was 27%. This is a significant difference between the recorded data & the observed data.

The underdiagnosis of heart disease is harder to attempt to explain. The overall U.S. percentage is ~11% but the CHC percentages were generally under 5%. Obesity might be subjective (although BMI values are supposed to be used), but heart disease is a diagnosable occurrence. You either have a myocardial infarction or systolic heart failure, or you do not. Most CMOs thought their populations were in the range of 20%-

---

[8] Yang , L. & G.A. Colditz. 2015. Prevelence of Overweight & Obese in the U.S., 2007-2012. JAMA Int. Med. Published online 22 June 2015.

[9] Please note that in ICD-10, now in use, there are 6 E66 codes for obesity & 6 O99 codes for obesity related to pregnancy

[10] Al Kazzi ES, Lau B, Li T, Schneider EB,Makary MA, Hutfless S (2015) Differences in thePrevalence of Obesity, Smoking and Alcohol in the United States Nationwide Inpatient Sample and the Behavioral Risk Factor Surveillance System. PLoS ONE 10(11): e0140165. doi:10.1371/journal. pone.0140165

30% for some form of heart disease. Possible causes of underdiagnosis are still under investigation.

Strategy & Normalization

CHCs differ greatly in their capabilities around both the deployment & maintenance of information technology & their ability to do strategic analysis, regardless of in what form (quantitative, qualitative, scenario-based, etc.). In two cases, the Primary Care Association & Urban CHC1, considerable work had already been done on strategy development including, in the case of Urban CHC1, hiring a Chief Strategy Officer (now COO) & in both cases, the development of a strategic analysis & development process. The other two CHCs had no real strategy development process, but appeared to react to both external & internal events on an *ad hoc* basis. Most health centers, in my experience, fall somewhere between these two extremes.

In addition, both the PCA & Urban CHC1 had put substantial effort into understanding the database underlying the EHR system in use & on creating unambiguous & standardized definitions for terms such as visits, patients etc. Even though in some cases, these terms differed from the definitions specified in the UDS guidelines, identification & correction of these differences (at least for the purpose of this analysis) was not difficult because in both cases a data dictionary had been deployed & in the case of Urban CHC1, a written data dictionary was also available.

In the case of Urban CHC2, a long time, 5-6 months, was needed to fully uncover all of the definitional differences between the health center analysis (SQL through intermediate BI tool to EHR reporting extract derived from EHR database) & the P2A analysis (mpSQL through Impala on HBase provisioned from data in HDFS imported from the EHR database). Partly this was due to starting from scratch on determining what the CHC's definitions were & needing ultimately to examine (& debug) their SQL coding, & partly it was due to multiple definitions being used for the same term.

The Rural CHC1 has still not completed the level-up exercise (after 14 months). There were a number of reasons for this. They do their queries for reporting through a dedicated BI tool that does not allow examination of the underlying generated SQL, the EHR in use has an extremely complicated schema design with over 1000 tables the majority of which are redundant or not used, they did not initially have a license that allowed them to examine or query the EHR database directly nor did they have any staff that could actually write SQL or other queries. This is not typical of rural health centers, but in this particular instance the impediments have proved to be difficult to overcome to date.

Data Quality

Data quality has also been a concern in the project. The PCA & Urban CHC1 had each gone through a process to design & deploy a data warehouse for their clinical (EHR)

data. In both cases, there were minor issues with data quality, mainly with missing data, but these issues could be resolved or adapted to so that analysis could be carried out. Urban CHC2 had potentially five years of EHR data (2010-2014). The center had done an EHR migration in 2011 that had been facilitated by their new vendor. As it turned out data years 2010-2012 were not usable do to corruption resulting from the conversion. The CHC had not looked at this data until the P2A project started, & so they were surprised that so much of their EHR data was not usable. This center started doing nightly extracts of the EHR data in 2013 & even though they did not have a standard data model or normalization procedures in place for the extract, the 2013-2014 data was much better in terms of missing or unusable data. The rural CHC1 had read-only access to the underlying database (Oracle) of their EHR. They had substantial problems with missing data, mostly data just not entered at the time of the patient visit according to anecdotal evidence & with unusable data, mainly data outside of normal (or in some cases even conceivable) ranges. An example would be many BMIs in the 400-500 range & several above 10,000. Their CMO told me that she knows they had several patients with BMIs in the 45-50 range, but clearly not in the 400-500 range. BMI is only one example of out of range & therefore unusable data. These problems of missing & unusable data were not limited to this rural health center but were found in every health center's data to a certain extent.

Clinical Results

Of course, the biggest data anomaly is the very low percentages of specific conditions in all of the centers' clinical data. Hypertension was in the 20% range where nationwide figures are closer to 30%[11]. Diabetes was generally in the 6%-7% range with nationwide figure in the 9%-10% range. More problematically, heart disease was generally <5% while the nationwide figure is more like 11%-12% & obesity was overall <10% while nationwide figures are more like 35%. Many of the CMOs informally interviewed about these results are not surprised that the percentage of diagnoses in the EHR data are low. Most of them admitted that the providers at their CHC do a poor job of coding many diagnoses without having an explanation for that apparent fact. It is also the case that these low percentages affect the accuracy & credibility of comorbidity percentages. I've already discussed the potential sociocultural & organizational biases that might be operating to affect the diagnosis of obesity. Similar factors may also be at work in the case of other diagnoses, but many of these are strictly empirical (as actually is obesity measured as BMI). You are either showing signs of cardiac distress at a specific visit is either high or you are not. The apparent inconsistent reporting of diagnoses during encounters, if true, is a real issue. If, as the recent PLOS One *(op. cit.*) points out, data in the National Inpatient reporting Sample (NIS) on obesity (overweight, tobacco use & alcohol abuse) are substantially different than data taken as part of a face-to-face interview (BRFSS), we have a real problem with the use of reported data such as EHR data for any type of population analysis. We have to assume that since providers are seeing patients in person, they are treating what they see & not just what they enter in the EHR.

[11] Nationwide percentages from CDC Fast Stats

This provides a segue to another possible cause for he low diagnosis percentages, especially with respect to comorbidities. In order for comorbidities to be identified, multiple diagnoses have to be entered for a patient, ideally for the same encounter. A patient that comes in with a severe foot infection might also have diabetes & potentially hypertension indicated as diagnoses. It is quite possible that the structure & function of current EHRs make this difficult. At the HIMSS 2015 conference, I proposed a use case to test this to 5 EHR vendors whose product is in use at CHCs. In each case, I sat with a company representative who was an expert in their EHR use & went through the use case that was designed to see if multiple diagnoses would be recorded for a specific type of visit[12]. In no case were multiple diagnoses recorded, & in all cases the EHR's user interface & fixed workflow made it very difficult to do so. To be fair, in two of the five cases, the EHR provided a mechanism to at least see other diagnoses for the specific patient (problem list navigation) & in one case, suggestions were made for additional diagnoses based on the diagnosis codes entered for the visit, but each representative said that in their experience, these mechanisms were rarely used in practice. It is also true that the reimbursement model (payment for care for primary diagnosis) does not create an environment where multiple diagnoses, & therefore the ability to identify comorbidities, are emphasized (or rewarded).

Analysis of Combined Financial & Clinical Results

Financial & cost accounting systems were evaluated at Urban CHC1 & Rural CHC1. In both cases, the financial systems were entirely separate from the EHR & other clinical information systems. This included different user level applications & different underlying databases. Also, in both cases, the data in the underlying database for the financial system was structured quite differently than the data in the database underlying the EHR system. Data in the EHR system was keyed on encounter (visit) date/time & patient identification. Data in the financial system was keyed on location where cost was accrued, cost date, payer type & other values not found or used as keys in the clinical systems. This made it difficult to impossible to calculate actual cost per patient per encounter per diagnosis. Patient identification is not usually present in the financial data. It is relatively to analyze revenue (or at least billed amount) per patient per encounter as the billing data is available in the practice management system that is keyed to the EHR data. It was not feasible at either CHC to assign actual cost from aggregated location costs (salaries, overhead etc.) to a specific patient visit, & so it is not possible to do analyses such as patient/encounter/diagnosis or yearly patient costs/diagnosis etc. These types of analyses however are essential if we are to understand & modify clinical & operational processes in order to both reduce overall costs & improve outcomes. CHCs are required to report annual cost/patient as part of their UDS reporting to HRSA, but this figure is calculated simply by dividing a "total cost per year" figure by total number of patients. There is no granularity in this

[12] See appendix for description of use case

number with respect to individual patients whose actual costs can vary between hundreds of dollars per year to hundreds of thousands of dollars per year.

Summary & Conclusions

Analytics have begun playing a larger & larger role in healthcare in the last 10-15 years. Healthcare organizations are using analytics for everything from facilitating & clarifying strategic planning to optimizing operational processes to providing data-based diagnosis & treatment planning to lowering the cost of care without compromising care quality. Many healthcare organizations have very large amounts of data in the petabyte range (Kaiser Permanente, Partners Healthcare, Geisinger Healthcare, Cleveland Clinic, etc.), but most have more modest amounts, in the gigabytes to low terabytes range. Regardless of the amount of data available for analysis, analytics are beginning to provide real leverage to healthcare organizations.

Healthcare safety-net organizations, such as community health centers (CHCs), often do not have the resources to explore new technological directions, especially if they require not only new hardware & software & analytic skills but also a new emphasis & understanding on the use of data. The Path2Analytics Project is deploying contemporary (Hadoop-based) analytics into CHCs & working with their staffs to make the necessary technological, organizational & cultural changes in order to leverage this new capability strategically. Now going into its third year, the P2A project is working with about 35 CHCs providing healthcare to about 1.3M people. Deployment capability varies widely with the Primary Care Association (PCA) & Urban CHC1 both deploying the analytic software & connecting to their data source quickly (1-2 weeks). Urban CHC2 & Rural CHC1 both took substantially longer (8-12 weeks) & required substantial assistance, especially to connect to their data sources. Similarly the PCA & Urban CHC1 both carried out the initial “level-up” exercise quickly while Urban CHC2 took 5-6 months & Rural CHC1 has still not completed it after 14 months.

Preliminary results, as presented & discussed above, were surprising. All population percentage measurements of patient/diagnosis/year were lower than expected & figures for obesity & heart disease were very much below nationwide figures as presented by the CDC. This is surprising as we generally think of the population served by CHCs as less healthy than the general U.S. population, & in fact their enhanced CMS reimbursement is based on this assumption & the fact that they provide a large amounts of enabling services to their patients (non-clinical services that enable the delivery of healthcare including at least: case management, benefit counseling, eligibility assistance, language interpretation, transportation & education services). Some possible technological, organizational & sociocultural causes of this underdiagnosis or data capture error have been discussed, but there are several larger issues including: 1) are these conditions being treated if they are not being diagnosed? 2) how good are the data for use in population health efforts & meta-analysis for clinical & operational process improvement? &, 3) can appropriate policy be developed based on our current understanding of individual & population health if

the data in EHRs is not actually representative of the health status of the population? These are serious questions that this study is neither designed nor prepared to answer,… but an attempt must be made to answer them.

The P2A project is continuing. We hope to include several new CHCs, especially rural CHCs, although it should be noted that the PCA involved includes many rural centers. A deeper examination of the issues raised by the preliminary results is also planned, & we expect that the health centers furthest along will begin including non-clinical data in the Hadoop analytic stack & use it to perform analysis to address specific strategic issues. Further reports will be made as the project progresses.

Appendix

Use Case for Multiple Diagnoses in EHRs:

1. Encounters:
    - 1.1. Patient who has not been seen before walks into CHC with severe foot infection (ICD-9: 730.97, unspecified infection of ankle & foot bone)
        - 1.1.1. Patient is treated & appointment made for follow-up in three days
    - 1.2. Patient returns to CHC for follow-up & reports severe headache (ICD-9: 339.10, tension type headache, unspecified)
        - 1.2.1. Foot infection inspected, disinfected & re-bandaged
        - 1.2.2. Headache discussed with patient, analgesic suggested
2. Results:
    - 2.1. Two encounters are recorded for the patient:
        - 2.1.1. Encounter 1 – diagnosis 730.97
        - 2.1.2. Encounter 2 – diagnoses 730.97, 339.10
    - 2.2. No other diagnoses recorded, no comorbidities (diabetes, hypertension, obesity) explored