Deployment of Analytics into the Healthcare Safety Net: Lessons Learned

David Hartzband[1] and Feygele Jacobs[2]

1. Director of Technology Research, RCHN Community Health Foundation

2. President and CEO, RCHN Community Health Foundation

Abstract

Background: As payment reforms shift healthcare reimbursement toward value-based payment programs, providers need the capability to work with data of greater complexity, scope and scale. This will in many instances necessitate a change in understanding of the value of data, and the types of data needed for analysis to support operations and clinical practice. It will also require the deployment of different infrastructure and analytic tools. Community health centers, which serve more than 25 million people and together form the nation's largest single source of primary care for medically underserved communities and populations, are expanding and will need to optimize their capacity to leverage data as new payer and organizational models emerge.

Methods: To better understand existing capacity and help organizations plan for the strategic and expanded uses of data, a project was initiated that deployed contemporary, Hadoop-based, analytic technology into several multi-site community health centers (CHCs) and a primary care association (PCA) with an affiliated data warehouse supporting health centers across the state. An initial data quality exercise was carried out after deployment, in which a number of analytic queries were executed using both the existing electronic health record (EHR) applications and in parallel, the analytic stack. Each organization carried out the EHR analysis using the definitions typically applied for routine reporting. The analysis deploying the analytic stack was carried out using those common definitions established for the Uniform Data System (UDS) by the Health Resources and Service Administration.[1] In addition, interviews with health center leadership and staff were completed to understand the context for the findings.

Results: The analysis uncovered many challenges and inconsistencies with respect to the definition of core terms (patient, encounter, etc.), data formatting, and missing, incorrect and unavailable data. At a population level, apparent underreporting of a number of diagnoses, specifically obesity and heart disease, was also evident in the results of the data quality exercise, for both the EHR-derived and stack analytic results.

Conclusion: Data awareness, that is, an appreciation of the importance of data integrity, data hygiene[2] and the potential uses of data, needs to be prioritized and developed by health centers and other healthcare organizations if analytics are to be used in an effective manner to support strategic objectives. While this analysis was conducted exclusively with community health center organizations, its conclusions and recommendations may be more broadly applicable.

Keywords: Community Health Centers, analytics, decision-making, data

Correspondence: dhartzband@rchnfoundation.org

DOI: 10.5210/ojphi.v5i3.4933

Introduction

Community health centers are the backbone of the health care safety net, providing comprehensive primary care for the nation's medically underserved communities and populations. In 2015, 1,429 community health centers operated in nearly 10,000 urban and rural sites across the country, serving over 25 million people. Buoyed by HRSA's long-standing focus on quality improvement and substantial investments in health center HIT systems, health center organizations have implemented electronic health record applications in record numbers. Ninety two percent (92%) of all federally qualified community health centers, and 85% of health center "look-alikes" - those entities that meet all requirements of the health center program but are supported by state and local funds rather than federal grants - report that an EHR was in use for all sites and all providers in 2015; only 2.4% have no EHR installed at any site and virtually all expect to adopt an EHR. In addition, 95.5% report using clinical decision support applications, and 64.1% exchange clinical information electronically with other key providers, health care settings or subspecialty clinicians.[3] In addition, 88.9% participate in the Centers for Medicare and Medicaid Services (CMS) EHR Incentive Program commonly known as "Meaningful Use." These statistics reflect a commitment to the adoption of new technologies to support the provision of high-quality clinical care and streamline operations. Yet as the movement to value-based payment accelerates and strategic planning becomes more complex, community health center organizations, along with all other providers, must be prepared for new and increasingly sophisticated analytics to support clinical care and operations.

As analytics are applied to ever-larger amounts of data and become both more important and more necessary, questions about their use become inevitable. How is data quality influenced by the use of health information technology (HIT) such as electronic health records (EHR), or acquisition through other means? On an operational level, how can analytic results best be understood and used to address and improve healthcare practice? Patient outcomes? Cost reduction? What are the implications of problematic data quality on operational capacity?[4]

To address these questions and help community health center organizations plan for future use and integration of contemporary analytics, several health center organizations were recruited to engage in a project to evaluate:

- Health center data accuracy: Do health center data systems ensure correct values and consistent formats for data?

- Health center data reliability: Do health center data systems collect and report results that are consistent and correspond to results from CDC data sources?
- Health center data completeness: Do health center data meet the criteria for all mandatory data items?

At each participating organization, which included several community health centers and one state primary care association, a Hadoop-based analytic stack was deployed alongside the organization's other data systems. Population-level statistics were compared for specific diagnoses and comorbidities calculated through the organization's normal means and through the analytic stack for comparability and utility.

Background and Literature

Documentation, reporting accuracy and data quality have been the focus of numerous studies. Yang and Colditz [4] recently undertook a review of NHANES survey data in an effort to benchmark the prevalence of obesity nationally. Al Kazzi *et al.* [5] examined the prevalence of obesity, overweight, tobacco and alcohol use comparing the data in a direct survey (the Behavioral Risk Factor Surveillance System - BRFSS) with that in the Nationwide Inpatient Sample administrative data base and found substantial differences between the two. O'Malley *et al.*[6] examined the ICD diagnostic coding process and potential sources of error in code accuracy. They found the principal sources of error to be related to both communication and documentation, citing lack of baseline information, communication errors, physician familiarity and experience with the presenting condition and insufficient attention to detail as well as training and experience of coders and discrepancies between electronic and paper record systems. Their prescription for improvement was the specification of clear coding processes and a focus on heightening the awareness of all staff engaged in documentation with respect to data quality.

Devoe, *et al.*[7] compared the entries in EHRs with the same data in the Medicaid claims data set for a group of 50 community health centers in Oregon. They found gaps in data congruence across the study group, with some services documented in the Medicaid data set but not the EHRs, and others documented in the EHRs but not in the Medicaid data set. For the latter group, nearly 50% of services documented in the EHR were not found in the Medicaid claims for HbA1c, cholesterol screening, retinopathy screening and influenza vaccination. They also evaluated demographic characteristics and found that Spanish speaking patients, as well as those who had gaps in insurance coverage, were more likely to have services documented in the EHR but not in the Medicaid claims data, a finding especially relevant to community health centers, which disproportionately serve poor, uninsured individuals and those best served in languages other than English.[5]

Outside of health care, other industries - including discrete manufacturing and financial services - have struggled with the overall issue of data quality [O'Connor, 8]. Over time, both of these industries have, for the most part, achieved very high levels of data quality and high levels of user confidence in their data, and the experience in these industries might provide some insight into data quality improvement in healthcare.[6] Two projects are especially instructive in this area. The C4 Project at General Motors, which began in 1986, [Bliss 9] was an attempt to develop an entirely paperless design and manufacturing specification system for automotive manufacturing. The data

quality effort associated with this project was immense. A staff of close to 50 people was assigned to the various parts of data acquisition, normalization, maintenance and life cycle management. The project emphasized the design of processes to ensure data quality and integrity. In particular, data governance was monitored at least as much as data entry, storage and usage. This went a long way toward ensuring a high level of data quality.

The same was true of a project at Goldman Sachs to develop an integrated trading system. The primary effort was the development of a set of data governance and data life cycle processes that focused on the awareness of data quality. [7]

The studies in other industries point to the processes deployed and potential for change relevant to the health care industry, while the studies of health services data frame some of the known challenges and complexities that remain in the health care industry. For many community health centers, which serve especially vulnerable populations, and tend to be less well-resourced than hospital providers, understanding the complexities related to data collection and use, and developing appropriate strategies to improve data collection and use information more effectively, remain challenging.

Methods

Two urban and one rural multi-site community health center, operating in three different states, were recruited along with one state primary care association to participate in the study in 2014-2015. The three health centers varied in size and in the aggregate, served approximately 124,000 patients. The participation of the primary care association, which administers a data warehouse for member health center organizations, resulted in a total data set representing 50 CHCs, operating more than 400 practice sites and serving approximately 1.3 million patients. The overall distribution of urban and rural sites was approximately equal. Each organization made available data for a period of either two years (2012-2014) or three years (2011-2014). It should be noted that the deployment was undertaken specific to each site; that is, each location was treated as a unique project site or case.

At each project site, a dual-path deployment and analytic process was used, wherein researchers worked with the IT group local to the organization to install and integrate a new software application to provide analytic capacity, alongside the centers' main systems.[8] The purpose of this was to help each health enter assess the reliability of its existing systems in deriving results consistent with the new application. This software, an open-source (OSS) Hadoop-based analytic stack, consisted of the Cloudera Express Hadoop distribution that includes the Hadoop Distributed File System (HDFS), Yarn (MapReduce2), Hbase (non-relational data management) and Impala (SQL-based query). Hadoop was selected as the analytic system for several reasons: it is a well-defined and well-understood technology; it is in current use in many sectors; it is relatively easy to install and test; it provides the opportunity to manage and analyze data from very heterogeneous sources; and it is easy to use because analytic queries may be produced in SQL. Most significant is the use of Hadoop for managing big data, and for predictive analytics. These are important considerations for health centers as they expand in capacity and require increasingly complex tools

for clinical and operational management. The local deployments provided an opportunity to test their use in the health center environment.

The OSS installation took between 1 week and 3 months, with the duration depending largely on the familiarity of the organization's IT staff with deploying open-source software. After deployment of this software, a read-only connection was made either to the database underlying the organization's electronic health record (EHR) system, or to a data extract or warehouse maintained by the primary care association and one of the urban community health centers.

The dual-path process provided for the comparison and separate evaluation of results and data quality for two different data collection and analysis approaches. The first method accessed and analyzed data through the normal processes in use at the health center. These processes included direct access to the EHR or EHR-based data warehouse and analysis through either the EHR's query facility or the business intelligence (BI) tool in use at the health center. The second method consisted of extracting data from the EHR (or warehouse), normalizing the data according to standard (UDS) definitions and conducting analysis using the Hadoop-based stack. This method allowed for greater transparency and permitted data quality issues such as differences in definitions or ambiguity due to EHR complexity to be identified and addressed.

After data were imported into the analytic stack and deployment was completed, an initial "level-up" exercise was performed. This exercise served both to test the analytic system and to facilitate the normalization of all core terms and data definitions between the CHC's operational systems and the analytic stack. The "level-up" exercise consisted of a number of defined queries performed through the organization's regular systems (EHR, SQL, BI tools) and compared with the same queries performed on the analytic stack with the data in the HDFS/HBase information store. The following queries were performed[9]:

- Number of patients served, per year
- Number of patients served presenting with specific diagnoses including hypertension, diabetes, obesity, heart disease, and behavioral health conditions
- Rank order of prevalent[10]comorbidities
- Cost[11] per patient, per year
- Cost, per comorbidity, per year.

This exercise, undertaken by the organizations in concert with the researchers, took between 2 weeks and 6 months to perform, and was largely dependent upon the organization's prior work on data normalization.

In parallel with this technical deployment, training was provided to each organization's IT personnel and other staff, including in all cases the CEO or Executive Director. The training focused on the uses of the stack-driven analytics, an exploration of its advantages and disadvantages, and addressed how it differed from extant business intelligence and other reporting.

Lastly, informal interviews were conducted with the chief medical officer or CEO at each organization to review results, discuss important findings, and consider potential challenges and approaches to continued analysis.

Study Limitations

The analysis was confined to community health center organizations, and included organizations operating in just three states. It may not be representative of a broader group of health centers nationally or of CHCs in different states. The data focused on specific years, and included centers that had undergone an EHR migration prior to the analysis period, which is an experience that may not be shared by health centers generally. Despite these limitations the data quality issues identified among the participants provide evidence of common data concerns and challenges.

Results

The exercise revealed common data quality issues across each of the organizations. These include missing or unusable data, as well as differences between the definition of core terms, such as patient or encounter, both within and across organizations even though these terms have standard definitions as required by HRSA. Finally, underreporting of certain diagnoses in comparison to the general population raises questions about the reliability of the data. While interesting in and of themselves, these initial results are important for what they may indicate about the bigger picture of data acquisition and use.

Table 1. below illustrates the values reported by the participating organizations, presented as ranges, for key diagnoses. As noted above, these results represent data from approximately 50 CHCs comprising over 400 clinical sites and a total of 1.3 million patients, for a period of two to three years. Reported population percentages (CDC FastStats) for the U.S. population as a whole are presented for comparison

Table 1. Population Percent Range Values for Selected Diagnoses

Diagnosis	Range from EHR Values	Range from Analytic Values	U.S. Population Percentage (CDC)[12]
Hypertension	17%-23%	4%*-22%	33.5%
Diabetes	6%-8%	2%*-8%	9.6%
Obesity	3%*-12%	3%*-12%	37.9%
Heart Disease	1%-4%	1%-3%	11.5%

Several issues are evident in the results reported above. The first is that the hypertension, diabetes and obesity results include a number of outliers (marked *). In each case, the outlier data are attributable to one (out of >40) health center organization. If these data are discarded, the comparative ranges for the results derived from EHR data are similar to the results derived from the data imported into the analytic stack. The table below shows the effect of removing the outlier organization from the analytic stack results. With the outlier removed, the EHR and stack-driven results are closer, but there is still some variation illustrated by the data ranges, reflecting inconsistencies in the data.

Table 2. Adjusted Population Percent Range Values for Selected Diagnoses

Diagnosis	Range from EHR Values	Range from Analytic Values	U.S. Population Percentage (CDC)
Hypertension	17%-23%	17%-22%	33.5%
Diabetes	6%-8%	5%-8%	9.6%
Obesity	9%-10%	7%-12%	37.9%
Heart Disease	1%-4%	1%-3%	11.5%

In addition, the percentage of the population with obesity and heart disease diagnoses, in both the full data table and second, adjusted, table are notably low in comparison with the CDC's reported figures for the U.S. population as a whole. We might expect the population percentages for these diagnoses in the community health center patient population to be at least the same as, if not higher than, those in the general population, given the documented level of disparities and characteristics of the population served. Possible reasons for these discrepancies will be discussed in the conclusions section of this paper.

More generally, the analysis revealed several types of potential data quality issues. Although the nature and extent of the problems varied across sites, the problems - including definitional conflicts, conversion issues and structural challenges - were not unique to any site and to some extent were evident at all sites. These include:

- o Errors resulting from deviation from standard definitions (for example, for patient, encounter, provider) even when guidelines for such definitions exist and are required for standard reporting (UDS, in the case of CHCs);
- o Errors caused by omission, that is data simply not recorded;
- o Errors resulting from incorrect entry
  - ▪ Values that are out-of-range & not caught by the EHR system, *e.g.* BMIs of >1000, BP values of 320/250, HbA1c values of >50;
  - ▪ Incorrect text entered for names, addresses, previous providers;
  - ▪ Values not entered into searchable fields;
  - ▪ Data recorded as text in clinical notes, but not into searchable fields;
  - ▪ Data imported from external sources (labs, registries, etc.) as text, but not into searchable fields;
- o Errors resulting from the structure and complexity of EHR systems
  - ▪ Several systems were found to be sensitive to the form that data was entered, specifically ICD-9 codes of 250., 250.0, 250.00 resulted in different query results as did 250.5, 250.50 etc.
  - ▪ Complexity of navigation and misalignment with provider workflows also appeared to be responsible for several types of errors.
  - ▪ Concentration on treatment of a single condition during an encounter, leading to low numbers of encounters with multiple diagnoses recorded.
- o Data corruption and/ or loss of data resulting from migration to a new EHR platform.

Conclusions and Recommendations

The length of time required to successfully complete the "level-up" exercise was substantially shorter in those organizations that had done extensive data normalization work prior to beginning the study. The organizations (PCA and one large, urban CHC) that took the least time (< 1 month) to deploy the analytic stack and perform the data quality exercise had previously undertaken substantial work to standardize definitions (semantic normalization) and to do format matching and format transformations (syntactic normalizations). This effort was not related to the study, and was in all cases done in conjunction with the creation and population of a data warehouse. In addition, these organizations had already begun an exploration of analytics that enabled them to quickly align with the deployment requirements and to think in terms of strategic analysis. Conversely, the organization requiring the longest period of time to complete the level-up exercise had the most widespread use of idiosyncratic, non-standard (i.e. non-UDS) definitions for core terms such as patient, encounter, and provider, as well as definitional mismatches between different clinical departments or between clinical & administrative departments within the organization.

The potential under-reporting of key diagnoses, as evidenced in the data for hypertension, diabetes, obesity and heart disease falling well below nationally reported figures, is of a different nature. The patient populations of community health centers are not generally thought to be healthier than the general U.S. population. Health centers patients are disproportionately poor, uninsured, and publicly insured, and disproportionately members of minority groups.[13] In addition, health centers are more likely to treat patients with chronic illnesses compared to other primary care physicians.[14] Yet in all cases, the reported percentages for key diagnoses were below the values reported for the population as a whole, and they are especially conspicuous for obesity and heart disease.

Al Kazzi et all [5] recently compared hospital discharge data reported in the U.S. Inpatient Reporting Sample (NIS, AHRQ) to interview data reported in the Behavioral Risk Factor Surveillance System (BRFSS, CDC) for 2011 data. Results for obesity showed a 9.6% population percentage in the NIS and 27.4% population percentage in the BRFSS. The population percentages reported in the BRFSS figures, which are based on direct participant surveys, are thus almost 3 times greater than the results from hospital discharge records [15], and more aligned with other recent results.[16] This suggests that the CHC-reported data are consistent with other provider-reported data, as demonstrated by the NIS results, but understated relative to other sources.

To better understand the anomaly with respect to obesity in the health center data sets, these results were reviewed with the Chief Medical Officers and other clinical staff at participating CHCs. Those interviewed estimated the obesity rate for the patients they served at forty percent. A recent paper in the Journal of the American Medical Association, Internal Medicine [4] estimated that in the United States, forty percent of adult men and thirty percent of adult women are overweight, while thirty-five percent of men and thirty-seven percent women are obese. The estimate provided by the participant CMOs is thus consistent with this data, and substantially higher than the data derived from the analysis. CMOs interviewed cited two possible explanations for this. First, it was noted that providers did not often diagnose obesity, and when they did, they did not use the full range of ICD-9 codes, which include three specific codes (278, unspecified obesity; 278.01,

morbid obesity, BMI>30; & 278.02, overweight, BMI>25). Further, while the UDS guidelines specify the use of the 22 V-codes for obesity, with a highly specific breakdown of BMI measurements, these apparently are also underutilized. It was conjectured that the data might reflect sensitivity to different cultural norms for defining obesity and overweight in the communities served.[17] While more investigation needs to be done to understand the data anomaly, the range of 3%-12% reported by the health center organizations in this study seems unlikely and could reflect both reporting and recording bias, as well as data quality issues.

Obesity might be subjective (although BMI values are a typically-used standard), but heart disease is a specific diagnosable occurrence. The apparent underreporting of heart disease in the study group – approximately seven to eight percent, as compared with eleven percent nationally per the CDC - is therefore harder to explain. Most CMOs thought that 20%-30% of their patients experienced some form of heart disease. Possible causes of underreporting are still under investigation, although it should be said that our analysis was not age-adjusted. It was also not adjusted for the fact that, particularly in 2014, many patients not previously known to the health centers were seen for the first time as coverage expanded, and the addition of new patients, may affect the distribution of diagnoses in ways that we do not yet understand.

Comparing the body of data quality work in aerospace and financial services industries with that in healthcare can be instructive. Each of the GM and Goldman Sachs projects referenced had several similarities besides emphasizing data governance. These included: 1) high level executive sponsorship – EVP and/or CEO who actually participated in introducing and reviewing the projects; 2) a long period of pre-work during which core terms were defined, data was normalized, and workflows and work processes were redesigned or newly created in order to provide an environment that promoted data quality; 3) broad participation from across the organization, not just IT; and 4) emphasis on standards where necessary or productive, but not as the primary or sole focus of effort. The most important characteristics in these industries' efforts were term definition, data normalization and process redesign as well as broad participation in the entire effort across the phases of planning, initiation, deployment and ongoing improvement. These are industries that spent decades reengineering their workflows and processes for operational and informational efficiency and effectiveness. [13,14].

In contrast, our experience suggests that: 1) governance and information life cycle are not at the core of how the healthcare industry approaches such projects; 2) while executive sponsorship is the norm, executive participation is rare; 3) many projects are designed, led and carried out by the IT group; and 4) standards are seen as a major part of the solution, including by Federal agencies and regulators (i.e. ONC, CMS, HRSA). These issues remain to be tackled in healthcare organizations.

This leaves us with the larger questions that were mentioned earlier. Our current findings provide some indications of the influences on data quality that might explain, at least in part, the variations and unexpected results. These influences include: quality degradation from system migration; inadequate or inappropriate data entry causing missing or incorrect data; inaccessible data in text entries from provider notes, text lists and external text imports; inadequate definition and format normalization resulting in unusable data; systemic errors due to current practice norms;

idiosyncrasies in how different EHRs process diagnosis codes; and complexity of navigation in EHRs. Many of these issues can at least be addressed by greater attention to detail at the data entry stage. Improving data quality directly in the EHR is more effective than trying to address it after the data are entered. The achievement of the Quadruple Aim - which encompasses improving the work life of health care providers, clinicians and staff as well as enhancing patient experience, improving population health, and reducing costs - clearly necessitates early preparation and consistent attention to data quality.

Several recommendations are suggested by these results:

- Definitions of core terms should be reviewed and consensus reached on their application and use. Moreover, data definitions and workflows should be aligned with standard practices.
- Workflows and other processes should be reviewed and redesigned as necessary to emphasize and promote data quality.
- Organizations should familiarize themselves with how their EHR processes data as it is entered (for example, how diagnosis codes are treated), and ensure that entered data is treated consistently by the EHR.
- Text data should be entered in a consistent manner that is retrievable for analysis as well as for use in diagnosis and patient care.
- Before migrating from one EHR platform to another, data should be cleaned and checked. Extensive data checking should also be done after EHR system migration. Care must be taken that the data from the retired system is backed up, potentially in a data extract, so that is available if any conversion loss occurs, and to vet the integrity of the migration process.

These recommendations are aimed at helping the health centers to answer two strategic questions. First, how good are your data? Clinical data from the EHR, data imported from labs and other providers, and financial data need to be carefully reviewed and vetted for accuracy, reliability and completeness. Second, how good are your systems? This includes infrastructure (servers, storage, network) and software systems and applications as well as processes and workflows. Finally, it includes staffing and staff training and engagement. Supporting health centers with the expertise and funds to undertake this work should be prioritized.

To date, the participating organizations have not moved substantially beyond the initial data quality exercise. The issue remains of how to integrate analytic results into the strategic and operational practice of a community health center, or a healthcare organization in general. The experience from this study indicates that while community health centers generally attain the highest standards of care and achieve good outcomes with respect to both quality and cost, considerable work may be needed to help all centers strengthen their awareness of data and information quality and move toward better integration of analytic results in practice. This awareness includes understanding how to perform complex analytic queries, and what applications might best be suited to particular types of analysis, but these are just two components of a forward–thinking data strategy. All staff at healthcare organizations, but particularly those engaged in using data for decision-making, must develop an awareness and appreciation of what data are available,

how it can best be analyzed and how these results relate to the clinical, operational and strategic needs of the organization.

Acknowledgements

The authors would like to acknowledge the assistance of Srini Rao, Ph.D., CEO of Datycs for deployment and analytic support.

Conflicts of Interest

The authors have no conflicts of interest to report.

References

1. Nambier R, Bhadwaj R, Sethi A, Vargheese R. 2013. A look at opportunities and challenges of Big Data analytics in healthcare. 2013 IEEE International Conference on Big Data. 10/2013.

2. Raghupathi, W., V. Raghupathi/ 2013. An Overview of Health Analytics. J. Health Med. Informat. 4:132. Dpi:10.4172/2157-7420.1000132.

3. Ward MJ, Marsolo KA, Froehle CM. 2014. Application of business analytics in healthcare. *Bus Horiz*. 57(5), 571-82. PubMed http://dx.doi.org/10.1016/j.bushor.2014.06.003

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

5. Al Kazzi E.S., B. Lau, T. Li, E.B.Schneider, M.A. Makary, S. Hutfless (2015) Differences in the Prevalence 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

6. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, et al. 2005. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 40(5 pt2), 1620-39. PubMed http://dx.doi.org/10.1111/j.1475-6773.2005.00444.x

7. Devoe JE, Gold R, McIntyre P, Puro J, Chauvie S, et al. 2011. Electronic health records vs Medicaid claims: Completeness of diabetes preventive care data in Community Health Centers. *Ann Fam Med*. 9(4), 351-58. PubMed http://dx.doi.org/10.1370/afm.1279

8. O'Connor L. 2007. Data Quality Management and Financial Services. Proceedings of the 2007 MIT Information Quality Industry Symposium. 5/2007.

9. Bliss FW. 1996. The C4 Program at General Motors. The CAD/CAM Handbook. Association for Computing Machinery. McGraw-Hill, NYC, NY. Pp. 309-320.

10. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. 2012. A Progmatic Framework for Single-site and Multi-site data quality Assessment in Electronic Health Record-based Clinical Research. *Med Care*. 50(0). doi:10.1097/MLR.0b013e318257dd67.

11. Wieskopf NG, Weng C. 2013. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *J Am Med Inform Assoc*. 20(1), 144-51. PubMed http://dx.doi.org/10.1136/amiajnl-2011-000681

12. Cai L, Zhu Y. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci J*. 14(2). doi:10.5334/dsj-2015-002.

13. Hammer M. 1990. Reengineering Work: Don't Automate, Obliterate. *Harv Bus Rev*. (Jul/Aug), 104-12.

14. Hammer M, Champy JA. 1993. Reengineering the Corporation: A Manifesto for Business Revolution. Harper Business Books. NYC. ISBN 0-06-662112-7

Footnotes

[1] As defined in Health Resources and Services Administration BUREAU OF PRIMARY HEALTH CARE, UDS Reporting Instructions for Health Centers: 2014 Edition (http://bphc.hrsa.gov/datareporting/reporting/2014udsmanual.pdf)

[2] http://whatis.techtarget.com/definition/data-hygiene. Data hygiene is the collective process conducted to ensure the cleanliness of data. Data is considered clean if it is relatively error-free

[3] HRSA 2015 Health Center Data. Table 5. Staffing & Utilization. http://bphc.hrsa.gov/uds/datacenter.aspx?q=tall&year=2015&state=

[4] *c.f.* [1] Nambier, *et al.* 2013; [2] Raghupathi, *et al.* 2013 and [3] Ward, *et al.* 2014.

[5] National Association of Community Health Centers A Sketch of Community health Centers [August 2016]

[6] Study author Dr. Hartzband has had extensive experience in these industries including as the external architect for the General Motors C4 project – an effort to develop a paperless design process for car manufacturing, and as a principal consultant to Ernst & Young for the Goldman Sachs integrated trading system effort

[7] Personal correspondence between Dr. Hartzband and Goldman Sachs team.

[8] Designed in accordance with various reviews of healthcare data quality assessment, especially: [10] Kahn, *et al.* 2012; [11] Weiskopf and Weng, 2013 and [12] Cai and Zhu, 2015.

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

[10] data quality and access issues prevented the accurate calculation of comorbidities

[11] actual cost (expenditure), not billed cost (revenue) – It is important to note that actual cost was not able to be calculated at any of the health centers and so these queries were not run.

[12] Figures from updated National Center for Health Statistics, CDC Fast Stats (2013-2014) for adults over 40 years of age. CDC definitions for diagnosis are identical to those used by HRSA for UDS except for heart disease, where the UDS definitions encompass more codes, & therefore more conditions.

[13] HRSA 2015 Health Center Data. Table 4. Selected Patient Characteristics. http://bphc.hrsa.gov/uds/datacenter.aspx?q=tall&year=2015&state

[14] NACHC Chartbook 2014. Figure 1.9. http://nachc.org/wpcontent/uploads/2015/11/Chartbook_December_2014.pdf

[15] NIS: discharge level data from approximately 8 million hospital stays (2011); BRFSS: 506,467 adult participants (2011)

[16] NIS: discharge level data from approximately 8 million hospital stays (2011); BRFSS: 506,467 adult participants (2011)

[17] Several Chief Medical Officers suggested that cultural norms for what is considered obesity very greatly among communities, and that providers might be unwilling to make a diagnosis not in line with such norms.