Barriers to Achieving Economies of Scale in Analysis of EHR Data

A Cautionary Tale

Mark P. Sendak[1]; Suresh Balu[1]; Kevin A. Schulman[2,3]

[1]Duke Institute for Health Innovation; [2]Duke Clinical Research Institute; [3]Department of Medicine, Duke University School of Medicine, Durham, North Carolina

Keywords
Chronic Kidney Diseases, Health Services Research, Informatics; Primary Health Care

Summary
Signed in 2009, the Health Information Technology for Economic and Clinical Health Act infused $28 billion of federal funds to accelerate adoption of electronic health records (EHRs). Yet, EHRs have produced mixed results and have even raised concern that the current technology ecosystem stifles innovation. We describe the development process and report initial outcomes of a chronic kidney disease analytics application that identifies high-risk patients for nephrology referral. The cost to validate and integrate the analytics application into clinical workflow was $217,138. Despite the success of the program, redundant development and validation efforts will require $38.8 million to scale the application across all multihospital systems in the nation. We address the shortcomings of current technology investments and distill insights from the technology industry. To yield a return on technology investments, we propose policy changes that address the underlying issues now being imposed on the system by an ineffective technology business model.

Correspondence to:
Kevin A. Schulman, MD,
Duke Clinical Research Institute, PO Box 17969
Durham, NC 27715
Phone: 919–668–8101
Email: kevin.schulman@duke.edu

Appl Clin Inform 2017; 8: 826–831
https://doi.org/10.4338/ACI-2017-03-CR-0046
received: 22. March 2017
accepted in revised form: 15. June 2017
published: August 9, 2017
Citation: Sendak MP, Balu S, Schulman KH. Barriers to Achieving Economies of Scale in Analysis of EHR Data. Appl Clin Inform 2017; 8: 826–831
https://doi.org/10.4338/ACI-2017-03-CR-0046
Funding/Support
This work was supported internally by the Duke Clinical Research Institute.
Additional Contributions
Damon M. Seils, MA, Duke University, assisted with manuscript preparation. Mr Seils did not receive compensation for his assistance apart from his employment at Duke University.

1. Background and Significance

In 2009, President Obama signed the Health Information Technology for Economic and Clinical Health Act to accelerate adoption of electronic health records (EHRs). Seven years and $28 billion in federal funds later, 84% of hospitals collect data in EHRs [1]. Yet, EHRs have produced mixed results. Information technology (IT) systems were envisioned to spawn "software applications that can retrieve data, organize them, apply decision algorithms, and provide the results to clinicians and managers when and where they need it" [2]. The value of such applications has been well articulated [3], but the current health IT ecosystem makes this vision challenging to implement. Health care systems are locked into EHRs that lack data model standardization [4], and the business model persists whereby EHRs control all data rather than liberate data for use in innovative applications [5]. To illustrate the opportunities and limitations of current health IT solutions, we documented the costs of validating a chronic kidney disease (CKD) analytics application.

In 2013, a preliminary analysis in an adult primary care clinic at Duke Health—a health system in North Carolina with 3 hospitals, 61,000 inpatient admissions, and 1,800,000 outpatient visits annually – revealed that only 11% of patients with laboratory values indicating stage 4 CKD were coded to reflect the diagnosis. Diagnosis codes used for billing failed to capture the burden of CKD, suggesting that clinicians were either unaware of or not documenting the condition. Clinical practice guidelines recommend nephrology referral for patients with greater than 10% to 20% risk of 1-year progression to end-stage renal disease (ESRD) to delay progression to dialysis and prevent emergency starts. Adapting lessons from Kaiser Permanente Hawaii [6, 7], Duke's Medicare Shared Savings Program (MSSP) secured funding to build an analytics application to identify high-risk patients for referral.

2. Handoff to Technology Team

Our technology team had to extract data from Epic and raw Medicare claims, transform the data for analysis, and load the data into an application, a process known as "extract, transform, and load" (ETL).

Our analytics application sourced data from Clarity [8] (a reporting database behind Epic), Medicare Part A and Part B claims, and an MSSP master beneficiary file. Clinicians specified variables of interest, which a database architect mapped to source data elements, and a data scientist performed exploratory data analysis for each variable.

Clinicians validated every variable from each data source at the aggregate and individual levels, because data from EHRs and claims is inaccurate and inconsistent [9, 10]. For example, we found 14 names for serum creatinine as a descriptor for over 4.4 million laboratory values stored in the database. Minor spelling variations were associated with major differences: "Creatinine – LabCorp" with 1 space after the dash was excluded for having a significantly different distribution of values than "Creatinine – Labcorp" with 2 spaces. Furthermore, Epic implementation did not remedy these differences. In Clarity, 3 laboratory variables with different identifiers are named "CREATININE" and have identical external names, base names, and abbreviations. Although Clarity does have a field for standard LOINC lab identifiers, the field is optional and not used in any internal functionality [11]. In our Epic system, only 2 serum creatinine lab names have the correct standard identifier populated. The validation process is even more complex for nonnumeric values that lack standard ranges and units.

Data extracts were used to build 22 variables, including 6 demographic features (age, race, gender, address, payer), 4 diagnoses (hypertension, diabetes, metastatic cancer, ESRD), 7 lab values (creatinine, urine protein, urine albumin-to-creatinine ratio, albumin, phosphorous, bicarbonate, calcium), 3 vital signs (systolic and diastolic blood pressure, weight), and 3 events (death, dialysis, nephrology visit). Age, labs, and vitals were structured as continuous numeric variables, diagnoses were structured as binary indicator variables, and race, gender, and payer were structured as categorical variables. All continuous variables were normalized, and missing urine albumin-to-creatinine ratio values were imputed from urinalysis protein.

Variables were then transformed using 2 validated models, including a regression model with 13 variables to predict 1-year risk of ESRD progression [12] and a regression of estimated glomerular filtration (eGFR) rate over time [13]. Clinicians requested displaying results of both models to convey risk and rate of disease progression. These models were developed and validated on large population data sets by investigators at University of Manitoba and Johns Hopkins, respectively, but are not standard in EHR systems. Finally, our team developed a scalable application to automate the analysis and present clinicians with relevant information about high-risk patients.

3. Results

Using the Kidney Failure Risk Equation [12], 1875 of 46,143 (4.1%) of patients enrolled in the MSSP had an ESRD risk greater than 15%. Risks scores were calculated for over 42,000 patients who had age, gender, and eGFR data. Patients who were seeing a nephrologist, deceased, or on dialysis were excluded from the pilot. Patients at high risk who were not excluded were reviewed by an interdisciplinary team during "population rounding" sessions to identify the best course of action. During the first 9 months, 438 patients at high risk of ESRD were reviewed and 84 patients (19.2%) were referred to a nephrologist. The primary outcome of the pilot was the number of patients identified by the analytics technology who required changes in management. After the pilot, the MSSP invested in deploying components of the analytics technology, hired a nephrologist to review high-risk patients, and launched a similar initiative to improve management of diabetes.

The total cost of our pilot in 2015 dollars was $217,138. The 4 categories of cost included $130,000 for design and development of the analytics application, $14,900 for query development to extract variables, $24,450 for exploratory data analysis and data pipeline development, and $47,788 for clinical validation of variables and workflow. The steep pilot costs were supported by 2 internal innovation grants, and maintenance of the system is being supported by the MSSP and the health system IT support group.

Given that EHR data models vary, other providers that wish to adopt our approach to screening patients with CKD will have to invest about $90,000 to validate the ETL data pipeline and clinical workflow. That assumes providers interface with a prebuilt application capable of ingesting standardized data, obviating the need for local application design and development. Based on our experience, the cost of scaling this application to all 432 multihospital health systems in the United States would total $38.8 million. If each hospital required custom validation of data and workflow, the cost of analyzing CKD data for the 4474 hospitals participating in the federal government's EHR incentive program [14] surpasses $400 million.

The cost of maintaining valid data extractions and transformations is also substantial. Purchase of a new automated blood chemical analyzer in the laboratory requires an effort to ensure that the ETL process is still valid. The laboratory does not know to notify data scientists of this change, so users of the analytics application are left to discover gaps in application performance on their own. In fact, it's not uncommon for participants in distributed research networks, such as the FDA's Sentinel, to learn of local data changes from validation efforts that ensure consistency of data within and across sites [15]. If the queries, data pipeline, and clinical workflow are refreshed annually, the cost to maintain the algorithm would be $5766 per site and $2.5 million for 432 multihospital health systems nationwide.

Despite these costs, there remains tremendous enthusiasm and opportunity to use EHRs to improve CKD management [16]. Since completing the pilot, our team developed the business case for using health IT to manage CKD for the National Institutes of Health [16]. We now help organizations across the country assess the opportunity to invest in health IT, but we believe more must be done at a policy level to reduce redundant costs.

4. Lessons

Our project revealed that analytics software can improve identification and management of progressing CKD. However, our efforts offer a cautionary tale about efficiencies of scale in health IT.

Other industries have faced poor data quality, variation in data sources, and massive legacy databases. Their solutions to these challenges may be instructive.

Automated quality-assurance systems identify data values that fail to meet specifications (ie, a required data type in a specified range) or are outside a set number of standard deviations from the mean. These systems then present data visualizations to subject matter experts to verify the data [18, 19]. Process management teams set dynamic dashboards to run at specified time intervals to ensure that processes generating or transforming data perform as expected. Using these techniques, it recently took 4 engineers 2 months for Inflection, a Silicon Valley company that uses public data to build trust, to integrate a new data source with over 1 billion records for only $60,000 to $80,000 (Sendak personal communication, 6/28/2016). Their system monitors dynamic data streams at a fraction of the cost it would take to build a similar system using current health IT approaches.

Beyond cleaning existing data, organizations must improve how new data are created and stored. Chevron launched a program to clean data associated with existing wells but quickly realized that cleanup would take as long as 5 years [20]. To avoid having to clean the newly generated data, a senior manager changed internal processes so that 100% of data on new wells was generated correctly. In health care, much of the data stored in EHRs without standardized metadata is generated by pharmacies, laboratories, or other external partners. Payment for a lab should be contingent on the transmission of complete metadata into the EHR along with the lab result. The entity that generates a piece of data should be responsible for curating metadata. In the early 1990s, AT&T reengineered its bill-verification process by applying data tracking internally until it became evident that the quality of externally sourced data needed to be improved [19]. AT&T specified exactly what data it needed from suppliers, required suppliers to report data tracking results, and performed periodic audits of supplier data. Following these steps, suppliers improved bill quality measures by a factor of 10, cutting AT&T's billing system expenses by two-thirds.

Beyond these practices, an entire industry of open-source technology has emerged to improve the efficiency of observational research [21]. Groups such as ours would be able to publicly share code to promote collaboration around critical steps in the ETL process. Other groups would use our results to accelerate their own efforts and reduce the massive redundancy of developing their own technology. Open-source approaches improve the economics of application development. However, health care has embarked on a different trajectory under most vendor contracts. Generally, the underlying structure of data is considered proprietary by the vendor, as are queries and transformations that reveal the underlying data model.

5. Conclusion and Recommendations

Despite valiant efforts to develop analytics applications that integrate with EHRs, the current health IT ecosystem imposes significant barriers to achieving economies of scale and providing safe, effective care. We must focus on the underlying issues of the massive and redundant costs being imposed on the system by an ineffective technology business model. We must also learn from the successes of other industries in grappling with similar legacy systems problems. We recommend the following:

- Health IT vendors supplement current products offerings with quality assurance systems that monitor the processes that generate and transform health care data
- Entities that generate and transmit data into EHRs are held financially responsible for curating thorough metadata
- EHR vendor contracts allow public sharing of ETL processes required to map proprietary databases to standard data models for analytics applications

If we fail to address these problems, significant cost barriers will prevent the efficient scaling of analytics software to improve CKD management across the nation. Organizations such as ours will continue to rely on innovation or pilot grants to build redundant technologies. The solution for CKD and more generally requires a transformation of how we treat our technology infrastructure rather than as local system or practice issues, and instead requires us to adopt a more systematic approach to making data interpretable and accessible.

Our description of our application development can help to engender a debate around these questions, one that needs to extend well beyond our colleagues in the health IT community. Our hope is that the transparency of this analysis will support that discussion.

Multiple Choice Questions

1.When scaling an analytics application that ingests electronic health record data to different health systems, significant, redundant costs are required to...?
A. Gather feedback from providers on the user interface
B. Compute risk scores for validated input variables
C. Validating and normalizing data across sources
D. Deploy the application to the cloud

2.Data companies in other industries have taken which of the following approaches to rapidly integrate new data sources into analytics applications?
A. Deploy quality-assurance systems with dashboards for human review
B. Ensure that data is generated in a clean and standardized manner
C. Foster an open source environment to share code and programs
D. All of the above

Answers to Multiple Choice Questions

1. C. Validate and normalize data across sources

Every site that wishes to adopt our chronic kidney disease analytics application must rely on local experts to map data source elements to the input variables. Current health information technology infrastructure prevents our team from being able to ensure that our local data validators will function properly in a new environment. Most health systems do not have data quality and assurance systems and are often unaware of the cleanliness of data being captured from laboratories or at the point of care.

The other answer choices referencing the user interface and computation of risk scores are components of the application that do scale. Of course, there may be users who want to modify the interface, but the front end developed at our system will function in a new environment. However, to ensure that the risk scores and user interface are displaying correctly, the redundant work must be done to validate transformations from the data source. The last answer choice is irrelevant, because an application can be deployed either on premise or on the cloud. The effort required to install applications in either setting should be minimal.

2. D. All of the above

Private industries have adopted all of these approaches in enhancing how rapidly new data sources can be integrated into analytics applications. Specific examples are provided from a Silicon Valley data company, Inflection, Chevron, and existing open source movements in clinical informatics.

Clinical Relevance Statement

Health information technology can be used to improve the detection and management of chronic kidney disease at the population level, but requires significant investment. Unfortunately, existing electronic health record systems do not enable rapid and efficient use of data to drive population health management programs. Health care systems must transform their technology infrastructure to achieve efficiencies of scale and advance population health.

Conflict of Interest

None reported.

Human Subjects Protections

No human subjects were involved in this work. The study was approved by the institutional review board of the Duke University Health System.

References

1. Henry J, Pylypchuk Y, Searcy T, Patel V. Data Brief 35: Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008–2015. Office of the National Coordinator for Health Information Technology, May 2016.
2. Blumenthal D, Glaser JP. Information technology comes to medicine. N Engl J Med 2007; 356(24): 2527–2534.
3. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big Data In health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014; 33(7): 1123–1131.
4. Koppel R, Lehmann CU. Implications of an emerging EHR monoculture for hospitals and healthcare systems. J Am Med Inform Assoc 2015; 22(2): 465–471.
5. Mandl KD, Kohane IS. Escaping the EHR trap - the future of health IT. N Engl J Med 2012; 366(24): 2240–2242.
6. Lee BJ, Forbes K. The role of specialists in managing the health of populations with chronic illness: the example of chronic kidney disease. BMJ 2009; 339: b2395.
7. Lee B, Turley M, Meng D, Zhou Y, Garrido T, Lau A, Radler L. Effects of proactive population-based nephrologist oversight on progression of chronic kidney disease: a retrospective control analysis. BMC Health Serv Res 2012; 12(1): 1.
8. Data Analytics Center, Perelman School of Medicine at the University of Pennsylvania. Epic Clarity. http://www.med.upenn.edu/dac/epic-clarity-data-warehousing.html. Accessed June 11, 2017.
9. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, Lehmann HP, Hripcsak G, Hartzog TH, Cimino JJ, Saltz JH. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013; 51(8 Suppl 3): S30-S37.
10. Fung V, Brand RJ, Newhouse JP, Hsu J. Using Medicare data for comparative effectiveness research: opportunities and challenges. Am J Manag Care 2011; 17(7): 488–496.
11. Adamusiak T, Shimoyama N, Shimoyama M. Next generation phenotyping using the unified medical language system. JMIR Med Inform 2014; 2(1): e5.
12. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, Levin A, Levey AS. A predictive model for progression of chronic kidney disease to kidney failure. JAMA 2011; 305(15): 1553–1559.
13. Coresh J, Turin TC, Matsushita K, Sang Y, Ballew SH, Appel LJ, Arima H, Chadban SJ, Cirillo M, Djurdjev O, Green JA, Heine GH, Inker LA, Irie F, Ishani A, Ix JH, Kovesdy CP, Marks A, Ohkubo T, Shalev V, Shankar A, Wen CP, de Jong PE, Iseki K, Stengel B, Gansevoort RT, Levey AS; CKD Prognosis Consortium. Decline in estimated glomerular filtration rate and subsequent risk of end-stage renal disease and mortality. JAMA 2014; 311(24): 2518–2531.
14. Office of the National Coordinator for Health Information Technology. Certified Health IT Vendors and Editions Reported by Hospitals Participating in the Medicare EHR Incentive Program. Health IT QuickStat 29. September 2016. dashboard.healthit.gov/quickstats/pages/FIG Vendors of EHRs to Participating Hospitals.php.
15. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care 2013; 51(8 Suppl 3): S22–S29.
16. Drawz PE, Archdeacon P, McDonald CJ, Powe NR, Smith KA, Norton J, Williams DE, Patel UD, Narva A. CKD as a model for improving chronic disease care through electronic health records. Clin J Am Soc Nephrol 2015; 10(8): 1–12.
17. Brajer N. CKD Population Health Cost Model. http://www.dihi.org/news/ckd-population-health-cost-model. Accessed May 6, 2017.
18. Rothenberg J. A Discussion of Data Quality for Verification, Validation, and Certification (VV&C) of Data to be used in Modeling. Rand Project Memorandum. 1997.
19. Redman TC. Improve data quality for competitive advantage. Sloan Manag Rev 1995; 36(2):99–107.
20. Redman TC. Data's credibility problem. Harv Bus Rev 2013 December; 84–88.
21. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015; 216: 574–578.