Visualizing Anomalies in Electronic Health Record Data: The Variability Explorer Tool

Hossein Estiri, PhD[1], Ya-Fen Chan, PhD[2], Laura-Mae Baldwin, MD, MPH[1,4], Hyunggu Jung, MS[3], Allison Cole, MD, MPH[4], Kari A. Stephens, PhD[1,2,3]
[1]Institute of Translational Health Sciences; [2]Department of Psychiatry & Behavioral Sciences; [3]Department of Biomedical Informatics and Medical Education; [4]Department of Family Medicine; University of Washington, Seattle, WA

Abstract

As Electronic Health Record (EHR) systems are becoming more prevalent in the U.S. health care domain, the utility of EHR data in translational research and clinical decision-making gains prominence. Leveraging primary care-based, multi-clinic EHR data, this paper introduces a web-based visualization tool, the Variability Explorer Tool (VET), to assist researchers with profiling variability among diagnosis codes. VET applies a simple statistical method to approximate probability distribution functions for the prevalence of any given diagnosis codes to visualize between-clinic and across-year variability. In a depression diagnoses use case, VET outputs demonstrated substantial variability in code use. Even though data quality research often characterizes variability as an indicator for data quality, variability can also reflect real characteristics of data, such as practice-level, and patient-level issues. Researchers benefit from recognizing variability in early stages of research to improve their research design and ensure validity and generalizability of research findings.

Introduction

To promote meaningful use and adoption of health information technology[1], the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 has accelerated the increasing trend in Electronic Health Record (EHR) adoption among health care practices[2]. By 2012, EHR systems were adopted by three-quarters of office-based physicians in the U.S., a more than 100% increase since 2007[3]. Information collected in EHRs provides clinical and administrative stakeholders, policy-makers, and researchers with the opportunity to evaluate delivery of health services, and quality and effectiveness of care. The Institute of Medicine has cited the enormous potential of EHR data in facilitating the creation of the learning health system, in which clinical decision-making is guided through the iterative real-time process of capturing and using/transforming knowledge from the care experience[4].

Given that significant resources are at stake for implementation of EHR systems, strong expectations exist for EHR systems to improve quality of care in the US. Meaningful use and health care reform are pushing the use of EHR systems with financial incentives[4]. Toward this goal, use of tools that foster improvements in clinical decision-making need to be coupled with basic EHR functions[5]. In addition, effectiveness of an EHR system is also a function of the quality of its data, as low quality data can obstruct clinical decision-making[6,7]. Therefore, tools that can profile data quality issues can play an important role in clinical decision-making and improving quality of care.

Data variability is commonly observed in health research[8,9,10]. In EHR-driven research, data variability has often been characterized as a data quality issue. However, many reasons can contribute to variability in EHR data. For example, a diversity of data models are often used on standard EHR forms[11]. Also, clinical coding behaviors can vary–for example, due to introduction of incentives to tackle certain conditions[12]. Both of these reasons can result in a substantial degree of variation in EHR data aggregated from multiple sites.

Variability in EHR data can influence comparability of the data[13], and complicate data extraction from multiple sites[14], and therefore, can represent a threat to generalizability of clinical trials by introducing bias across treatment effects[15]. In comparative effectiveness research, it is critical to account for data variability[10]; otherwise the results are 'subject to validity concerns.'[16] Variability in treatment effects between different trials (also known as heterogeneity) has important implications for the research design and interpretation of results in comparative effectiveness research[15]. Properly addressing the conditions that may influence the prevalence of disease in an EHR will empower researchers in their research design choices[17]. Comparative effectiveness studies become meaningful only when variability in study population, treatment exposure, and clinical outcomes reflects real differences in clinical practices[18].

Because data quality and provenance issues are often not being comprehensively assessed in EHR data, evaluating data variability in EHR-based research is of special importance[18]. To improve the quality of research, we need to be able to recognize variability in EHR data quickly and at early stages of a research project's lifecycle. Variability in EHR data could be observed among multiple EHR systems or within a single EHR system over time. We need data tools to translate variability of complex datasets into actionable knowledge. This paper introduces a new web-based

visualization tool, the Variability Explorer Tool (VET), to help researchers explore variability in EHR data. VET provides a visualized demonstration of variability across time and between clinical sites for selectable EHR data variables and values. The goal of VET is to help researchers at initial stages of their research to: (1) generate research questions and hypotheses about possible reasons for observed variations (or lack of variation) in prevalence of specific clinical phenomena/observations (e.g., diagnoses, procedures, medications), (2) to inform choice of analytical methods in order to increase a study's external validity, and (3) to help with cohort selection / data extraction.

Methods

The Variability Explorer Tool (VET) is a web-based tool that produces custom visualizations of data variability across time and between clinical sites. VET visualizes data variability based on a simple approximation of the probability distribution functions in each year for the prevalence of specific data values that the researcher defines. The current version of VET profiles variability within EHR diagnoses by allowing dynamic entry of any combination of International Classifications of Diseases codes (ICD-9). First, the researcher identifies one or a cluster of ICD-9 code(s) of interest, using the search box on VET webpage. This will initiate a SQL query that calculates the prevalence of the requested ICD-9 code(s) in each year and in each clinic based on the following formula: $\frac{n_{ij|k}}{N_{ij}}$, where $n_{ij|k}$ is the number of patients in clinic i and year j that were associated with the requested ICD-9 code (or cluster) k, and N_{ij} is the total number of patients in clinic i and year j.

The query results table directly feeds into the Candlestick Chart[19] layout from Google Charts to generate visualized approximations of the annual probability distribution functions. VET plots use the Candlestick Chart template to visualize the distribution of data points based on means and standard deviations. By using mean and standard deviation, VET approximates where 95.4% of the data points are, as well as illustrates the range of values. To approximate the annual probability distribution functions, the SQL query returns maximum, minimum, mean, and two standard deviations below and above the mean of the prevalence of the requested ICD-9 code(s) in each year that the requested data values exist in the data table. Therefore, in addition to visualizing the variability, VET's output shows the time period in which the requested proportion of data is available.

Data and Web Platform

For this investigation, the Variability Explorer Tool used anonymized aggregated counts of data from the Data QUEST electronic data-sharing architecture, hosted by the University of Washington Institute of Translational Health Sciences (ITHS). Data QUEST is an infrastructure that facilitates sharing of EHR data across diverse primary care organizations. Data QUEST includes EHR data from 15 primary care clinics in the Washington, Wyoming, Alaska, Montana, and Idaho (WWAMI) region[20]. These clinics use a diverse set of electronic health record systems, including Allscripts and Centricity, semantically aligned within Data QUEST.

As a web-based tool, VET functions on Data QUEST's federated information dictionary web platform, called FindIT. Designed to help researchers understand the depth and breadth of the data, FindIT profiles the data shared across the Data QUEST network. FindIT's interface is built with Drupal, a PHP framework with a relational database, in this case PostgreSQL, using Microsoft SQL Server. The SQL Server database is a centralized collection of aggregated, anonymized counts from the Data QUEST federated architecture.

Results

To illustrate VET's variability plot, case examples of depression are presented. Depression is commonly tracked on patients seen in primary care and therefore offers good natural examples of variability in EHR data. ICD-9 codes for depression used in this study include 296.2x (Major Depressive Disorder, single episode), 296.3x (Major Depressive Disorder, recurrent), 300.4 (Dysthymic disorder), and 311 (Depressive Disorder, Not Otherwise Specified). Figure 1 shows VET's visualization of the variability in the proportion of patients with any of the selected ICD-9 depression codes in each of the years for which data are available. The horizontal axis represents the time period in which depression codes are available in the database, 1990 to 2013. Blue boxes in a given year represent where approximately 95.4% of data points are distributed across clinics. The number of clinics providing data to the tool can vary from year to year. Therefore, a taller box represents more variation in prevalence of these depression diagnoses between clinics in the given year. Variability across years can be inferred from comparing the height of boxes over time.
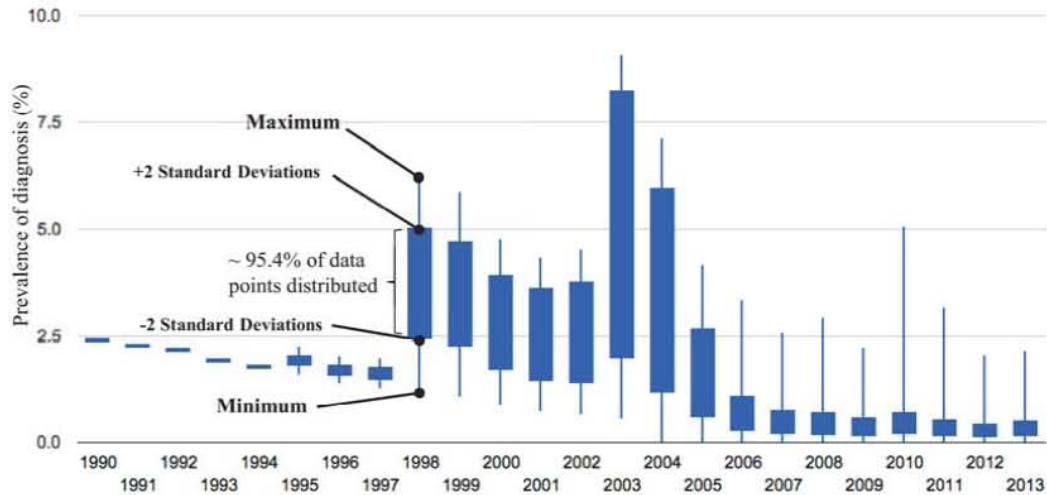
Figure 1: Outcome of the Variability Explorer Tool on the full cluster of depression ICD-9 codes

Based on the VET's output plot for depression ICD-9 codes, there appears to be substantial between-clinic variation in the prevalence of these depression diagnoses across time and between clinical sites. The prevalence of these depression diagnoses has changed noticeably – from about 9% (maximum in 2003) to less than 0.1% (minimums since 2004) over the 23-year period.

The VET plot shows that data are available since 1990. Within this time frame, a clear concave (increasing and then decreasing) trend is distinguishable in between-clinic variability. What seems to be very small variability in the early 1990s intensifies over the following few years until maximum variability occurs in 2003 and 2004. There is no between-clinic variability in depression diagnosis between 1990 and 1994. The VET plot shows that both the between-clinic variations and the prevalence of depression diagnosis are higher between 1998 and 2005, especially in 2003 and 2004, than in earlier and later years. In contrast to the prior period, as the figure shows, both between-clinic variability and prevalence of depression diagnosis has reduced substantially and stabilized since 2006.

In addition to visualizing variability between clinics and across years using a cluster of diagnoses (as shown in Figure 1), VET can be used to explore variability at the single diagnosis level. For example, the researcher can use VET plots to break down the cluster of diagnoses into a single diagnosis VET plot to compare variability across clinics, years, and the individual diagnoses. Figures 2 and 3 are VET plots using the 296.3x and 311 ICD-9 codes, respectively. Both between-clinic and across-year variability differ when data are pulled for these two different ICD-9 codes. It also appears that ICD-9 code 311 was a more prevalent depression diagnosis than ICD-9 code 296.3 in the dataset.
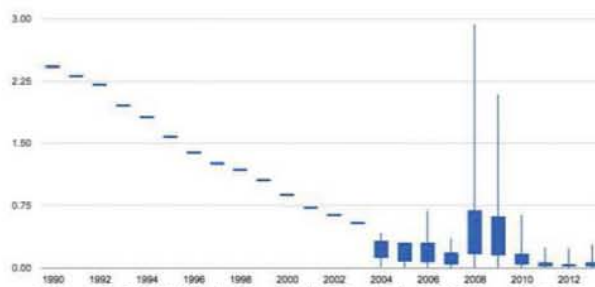

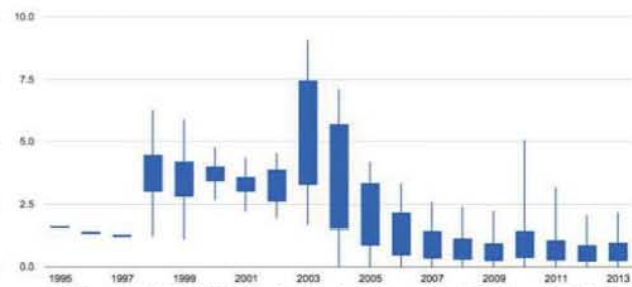Figure 2: Variability in depression data using ICD-9 code 296.3x

Figure 3: Variability in depression data using ICD-9 code 311

Discussion

Output plots from VET for depression ICD-9 codes visualized a substantial degree of variability between clinics and across years. The observed variability notifies the researchers about two issues at early stages of research design: (1) there are complexities to defining a cohort with depression, and (2) simple analytical methods may not account for the substantial variability present in these data. Variability across units of analysis (e.g., patients) across time and space (e.g., clinics) is an essential characteristic of any human-related phenomenon, making comparative research

meaningful. Variability between clinical sites can be due to many factors, from demographic differences in patient populations,[21] data capture and terminologies, to local practice patterns[16].

Even though data variability is typically categorized as a data quality issue, variability in data can also represent real characteristics of the population under study, caused by exogenous factors influencing prevalence of a certain condition. For example, under treatment of patients with depression between primary care clinics in the early 1990s versus after the passage of Mental Health Parity Act in 1996[22] may have led to changes in diagnostic coding for patients with depression. It is crucial for researchers to question, examine, and understand the underlying causes for variability in EHR data and distinguish between 'real' and 'spurious' data variability[18].

VET visualized 23 years worth of data for the selected depression ICD-9 codes. Given that none of the clinics had an EHR system in place in the early 1990s, it is likely that EHR data reflecting these years are either in error, or represent attempts to document historical data. The lack of variability between 1990 and 1994 may reflect that data are from only one clinic or that all clinics had the same depression prevalence in this period. In the case of depression ICD-9 codes from Data QUEST used in this paper, the lack of variability stemmed from data from only one clinic. Low variability in the distribution of values also can represent the presence of 'fabricated' data, due to data imputation or interpolation[21]. Researchers should be careful about inclusion of patients from time periods with extremely low variability in the study cohort without understanding the etiology of this finding.

Higher prevalence of depression diagnoses observed in 2003 and 2004 may be due to data quality issues, occurrence of a periodic event in a certain location, such as a targeted effort at screening for depression in these years within one or more of the clinics, or smaller number of total patients (denominator, N_{ij}) within those years relative to the recording of depression diagnosis in the EHR. Simultaneous high prevalence and high between-clinic variability, however, is less intuitive. Large variability in distribution of values can reflect systematic data errors, such as error in a measurement instrument[21]. Temporal trends related to clinical use of ICD-9 codes could affect variability over time. ICD-9 codes in primary care are primarily assigned by the provider, but are sometimes assigned by coding personnel. New trainings of providers and/or coding personnel in diagnosis coding and the introduction of new diagnosis codes could both affect variability in prevalence of diagnoses over time.

Conclusion

Variability in EHR data has important implications for the validity and generalizability of translational research that uses it. With a solid understanding of both how EHR data were collected and the variability in the dataset, clinical and administrative stakeholders and policy makers can make better decisions as they evaluate health services delivery and quality of care. When using EHR data, researchers need tools to allow for quick evaluation of data variability early in the research process to improve their research design and ensure validity and generalizability of research findings. This paper has introduced a web-based tool, the Variability Explorer Tool, which provides researchers a quick way of examining EHR data variability using a visualization approach on a scalable platform supporting replicability across data domains. Existence of anomalies in data can generate important questions and hypotheses about the data and phenomena under study. As demonstrated by the depression use case, the VET allows researchers to identify data variability, a key element of EHR data quality, and use this information to refine research questions and procedures to maximize validity and generalizability of research findings generated from EHR data.

The development of VET is an ongoing project with content experts from diverse fields, including health informaticists, computer scientists, clinicians, biostatisticians and health services researchers. User tests are needed to improve the tool's usability for clinical investigators and other potential users. Further investigation is necessary to determine root causes of variation in data before the data can be interpreted. Critical factors to interpreting VET's illustrations of variability in data include denominators for the boxes (number of patients) and counts of clinics in each year. Future enhancements to VET could explore ways to incorporate number of patients and clinics, as well as other data characteristics, into the visualizations. Further, VET's methodology provides replicability to other dimensions of variability in EHR data (e.g., variability other data domains beyond diagnoses), and scalability to features allowing deeper exploration of variability.

References

1. U.S. Department of Health & Human Services. HITECH Act Enforcement Interim Final Rule. 2009. http://www.hhs.gov/ocr/privacy/hipaa/administrative/enforcementrule/hitechenforcementifr.html.

2. Patel V, Jamoom E, Hsiao CJ, Furukawa MF, Buntin M. Variation in electronic health record adoption and readiness for meaningful use: 2008-2011. *J Gen Intern Med*. 2013;28:957-964. doi:10.1007/s11606-012-2324-x.

3. Hsiao C-J, Hing E, Ashman J. Trends in electronic health record system use among office-based physicians: United States, 2007-2012. *Natl Health Stat Report*. 2014:1-18.

4. (IOM) Medicine I of. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care. In: *The Foundation for Continuous Improvement in Health and Health Care Workshop Serier Summary*.; 2011:1-311.

5. Zhou L, Soran CS, Jenter CA, et al. The Relationship between Electronic Health Record Use and Quality of Care over Time. *J Am Med Inform Assoc*. 2009;16:457-464. doi:10.1197/jamia.M3128.

6. Bowman S. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspect Health Inf Manag*. 2013;10:1c.

7. Dentler K, Cornet R, ten Teije A, et al. Influence of data quality on computed Dutch hospital quality indicators: a case study in colorectal cancer surgery. *BMC Med Inform Decis Mak*. 2014;14:32. doi:10.1186/1472-6947-14-32.

8. Angier H, Gold R, Gallia C, et al. Variation in Outcomes of Quality Measurement by Data Source. *Pediatrics*. 2014. doi:10.1542/peds.2013-4277.

9. Vanasse A, Niyonsenga T, Courteau J, et al. Spatial variation in the management and outcomes of acute coronary syndrome. *BMC Cardiovasc Disord*. 2005;5:21. doi:10.1186/1471-2261-5-21.

10. Cooperberg MR, Broering JM, Carroll PR. Time trends and local variation in primary treatment of localized prostate cancer. *J Clin Oncol*. 2010;28:1117-1123. doi:10.1200/JCO.2009.26.0133.

11. Abernethy N, DeRimer K, Small P. Methods to identify standard data elements in clinical and public health forms. *AMIA Annu Symp Proc*. 2011;2011:19-27.

12. Buchan I, Winn J, Bishop C. A Unified Modeling Approach to Data-Intensive Healthcare. In: *The Fourth Paradigm: Data-Intensive Scientific Discovery*.; 2009:91-97. http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx.

13. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev*. 2010;67:503-527. doi:10.1177/1077558709359007.

14. Roth CP, Lim Y-W, Pevnick JM, Asch SM, McGlynn E a. The challenge of measuring quality of care from the electronic health record. *Am J Med Qual*. 2009;24:385-394. doi:10.1177/1062860609336627.

15. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: heterogeneity--subgroups, meta-regression, bias, and bias-adjustment. *Med Decis Making*. 2013;33:618-640. doi:10.1177/0272989X13485157.

16. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care*. 2013;51:S22-S29. doi:10.1097/MLR.0b013e31829b1e2c.

17. Walker AM. Matching on provider is risky. In: *Journal of Clinical Epidemiology*.Vol 66.; 2013. doi:10.1016/j.jclinepi.2013.02.012.

18. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research. *Med Care*. 2012;50:S21-S29. doi:10.1097/MLR.0b013e318257dd67.

19. Google. Visualization: Candlestick Chart. *Google Charts*. 2014. https://google-developers.appspot.com/chart/interactive/docs/gallery/candlestickchart.

20. Stephens KA, Lin C-P, Baldwin L-M, et al. LC Data QUEST: A Technical Architecture for Community Federated Clinical Data Sharing. *AMIA Summits Transl Sci Proc*. 2012;2012:57.

21. Venet D, Doffagne E, Burzykowski T, et al. A statistical approach to central monitoring of data quality in clinical trials. *Clin Trials*. 2012;9:705-713. doi:10.1177/1740774512447898.

22. Hennessy KD, Goldman HH. Full parity: Steps toward treatment equity for mental and addictive disorders. *Health Aff*. 2001;20:58-67. doi:10.1377/hlthaff.20.4.58.