5-10-2017

DQe-v: A Database-Agnostic Framework for Exploring Variability in Electronic Health Record Data Across Time and Site Location

Hossein Estiri
Harvard Medical School, Massachusetts General Hospital, hestiri@mgh.harvard.edu

Kari Stephens
University of Washington, kstephen@uw.edu

Recommended Citation
Estiri, Hossein and Stephens, Kari (2017) "DQe-v: A Database-Agnostic Framework for Exploring Variability in Electronic Health Record Data Across Time and Site Location," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 5: Iss. 1, Article 3.
DOI: https://doi.org/10.13063/2327-9214.1277
Available at: http://repository.edm-forum.org/egems/vol5/iss1/3

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

DQe-v: A Database-Agnostic Framework for Exploring Variability in Electronic Health Record Data Across Time and Site Location

Abstract
Data variability is a commonly observed phenomenon in Electronic Health Records (EHR) data networks. A common question asked in scientific investigations of EHR data is whether the cross-site and -time variability reflects an underlying data quality error at one or more contributing sites versus actual differences driven by various idiosyncrasies in the healthcare settings. Although research analysts and data scientists have commonly used various statistical methods to detect and account for variability in analytic datasets, self service tools to facilitate exploring cross-organizational variability in EHR data warehouses are lacking and could benefit from meaningful data visualizations. DQ^e-v, an interactive, database-agnostic tool for visually exploring variability in EHR data provides such a solution. DQ^e-v is built on an open source platform, R statistical software, with annotated scripts and a readme document that makes it fully reproducible. To illustrate and describe functionality of DQ^e-v, we describe the DQ^e-v's readme document which includes a complete guide to installation, running the program, and interpretation of the outputs. We also provide annotated R scripts and an example dataset as supplemental materials. DQ^e-v offers a self service tool to visually explore data variability within EHR datasets irrespective of the data model. GitHub and CIELO offer hosting and distribution of the tool and can facilitate collaboration across any interested community of users as we target improving usability, efficiency, and interoperability.

Acknowledgements
This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000423. The authors would like to thank Roy Pardee from Group Health Research Institute and Alison Kosel for their constructive feedback on the tool and paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Keywords
Electronic Health Records, Data Quality, Data Variability, Data Warehouse

eGEMs
Generating Evidence & Methods to improve patient outcomes

DQ^{e}-v: A Database-Agnostic Framework for Exploring Variability in Electronic Health Record Data Across Time and Site Location

Hossein Estiri, PhD, MS;[i,ii] Kari Stephens, PhD[iii]

ABSTRACT

Data variability is a commonly observed phenomenon in Electronic Health Records (EHR) data networks. A common question asked in scientific investigations of EHR data is whether the cross-site and -time variability reflects an underlying data quality error at one or more contributing sites versus actual differences driven by various idiosyncrasies in the healthcare settings. Although research analysts and data scientists have commonly used various statistical methods to detect and account for variability in analytic datasets, self service tools to facilitate exploring cross-organizational variability in EHR data warehouses are lacking and could benefit from meaningful data visualizations. DQ^{e}-v, an interactive, database-agnostic tool for visually exploring variability in EHR data provides such a solution. DQ^{e}-v is built on an open source platform, R statistical software, with annotated scripts and a readme document that makes it fully reproducible. To illustrate and describe functionality of DQ^{e}-v, we describe the DQ^{e}-v's readme document which includes a complete guide to installation, running the program, and interpretation of the outputs. We also provide annotated R scripts and an example dataset as supplemental materials. DQ^{e}-v offers a self service tool to visually explore data variability within EHR datasets irrespective of the data model. GitHub and CIELO offer hosting and distribution of the tool and can facilitate collaboration across any interested community of users as we target improving usability, efficiency, and interoperability.

[i]Harvard Medical School, [ii]Massachusetts General Hospital, [iii]University of Washington

Introduction

The Health Information Technology for Economic and Clinical Health Act of 2009 (HITECH) and the Patient Protection and Affordable Care Act of 2010 have contributed to the widespread adoption of Electronic Health Record (EHR) systems in the United States.[1–5] EHRs provide valuable information about determinants of health and treatment effectiveness,[1] and their proliferation offers a huge potential for secondary use of EHR data in health care research and decision-making.[6–8] However, despite this potential, examples of actual use of EHR data for improving the efficiency of the health care system are surprisingly scarce.[9]

Data quality and variability are two major concerns hindering utility of EHR data for health care research and policy.[7,10–12] EHR data quality issues can jeopardize scientific inference and obstruct policy evaluations.[13] Yet, data quality has often been defined on a case-by-case basis with no unifying standard approach,[7,14–17] and defining its indicators in health care systems has been full of ambiguities and inconsistencies.[2,17]

Variability has often been characterized as an indicator of EHR data quality. For example, the harmonized data quality assessment framework uses variability to define temporal and atemporal plausibility in EHR data quality assessment.[18] Given that variability is not a well-defined term in the context of EHR data, we define "variability" as "the extent of data dispersion (in value and meaning), relative to a similar group of observations." Significant variability in EHR data has been observed across time,[19] geography,[20] and data sources (e.g., EHR versus claims- and payer sources),[21] and between clinical site locations of practice.[22,23] Potential causes of variability include differences in operational data structures, formats, and standards;[24] clinical data collection and analytic extraction methodologies and workflows;[16,24–26] extraction criteria;[8,27] and patient populations.[28]

Whether or not variability in EHR data reflects an underlying data quality issue or reasonable differences within patient populations and practice patterns, EHR researchers need to understand and properly address variability in data.[22,29] Variability can complicate data use and undermine comparability when data are combined from multiple sources.[11,24] Data variability has important implications, particularly for the conduct and interpretation of comparative effectiveness research[16,23,30] as it can bias comparison results.[26,27]

Different methods are used to identify data anomalies on a case-by-case basis.[17,31] At the research study level, researchers use a variety of statistical methods to evaluate variability in research data sets. Nevertheless, the lack of standard, applied approaches to monitor and detect variability in an EHR data warehouse hinders a harmonized effort to extract high-quality EHR data for research.

Variability across time and site location are common axes that must be accounted for when utilizing EHR data. Combining data from multiple site locations—through a network of EHR systems or one large EHR system with multiple location sites—can strengthen external validity of health care research[16,26] and provide new opportunities to conduct comparative research on a variety of topics. One opportunity that the combined data offer is the identifying of potential data anomalies by monitoring variability between data sources and across time. That is, similar to the leave-one-out cross validation approach, data characteristics (e.g., distribution, frequency, deviation) in a multi-EHR and multisite combined data set can be used to validate characteristics of data from an individual site location within a given EHR. This prospect highlights the role that EHR data warehouses, and

Volume 5 (2017) Issue Number 1

organizations such as the Distributed Ambulatory Research in Therapeutics Network (DARTNet) Institute[32]—which acts as a support entity for practice-based research networks across the United States—can play in providing high quality EHR data to researchers.

To our knowledge, however, no reproducible database-agnostic tools are available for in-depth exploring of EHR data variability across site location and time—the key axes and drivers of data variability. We present an interactive reproducible tool, DQ^e-v, which provides a framework to help data warehouse administrators, analysts, researchers, and other EHR data users with basic understanding of R statistical programming to explore and track site location and time-driven variability in EHR data networks. DQ^e-v provides an operational tool and an expandable platform to explore variability in multisite EHR data. We describe below the tool architecture that encompasses the workflow, data model, and execution. To demonstrate DQ^e-v's functionality, we use data from the WWAMI region Practice and Research Network's Data QUEST EHR data warehouse[33] to present examples of the tool's outputs.

Tool Architecture

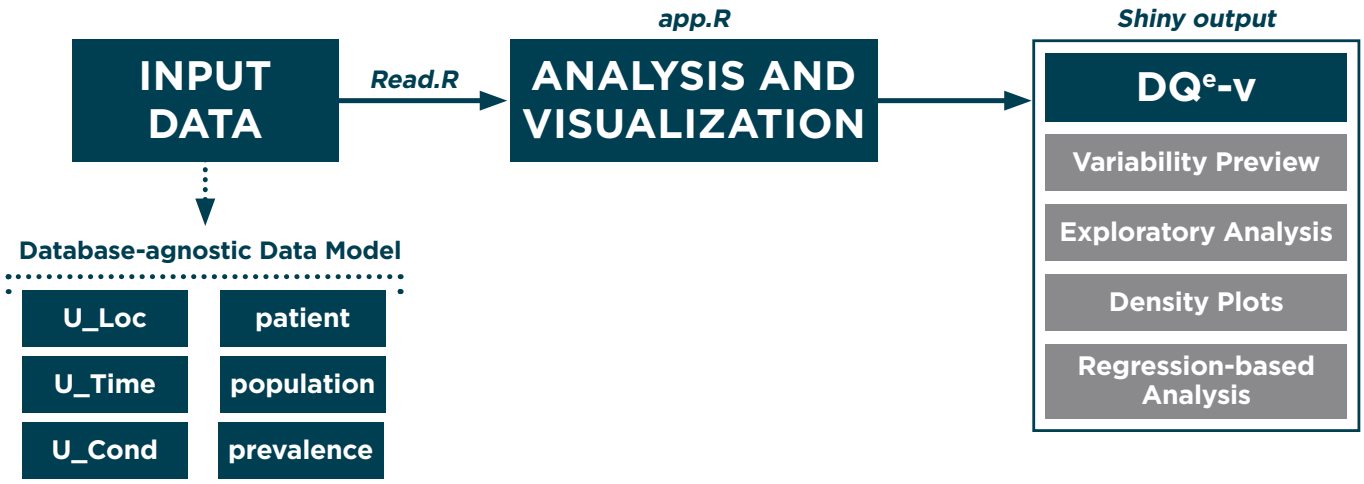
DQ^e-v builds upon the approach to profiling data variability used in the Variability Explorer Tool (VET).[19] VET is a Web-based tool for FindIT, a data profiling tool,[34] that visualizes variability in International Classification of Diseases-Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis codes over time and across site locations. DQ^e-v offers several enhancements over VET, expanding its usability and application. First, it provides more visualizations of variability than the VET, which will allow the users to obtain more knowledge about variability in the EHR data. Second, DQ^e-v is interactive, which will put the power of specification and interpretations into the users' hands. Third, it is open source, using a flexible architecture that can be customized and expanded by users according to the EHR data warehouse needs. Lastly, the input data model for DQ^e-v can be easily generated from any EHR data model, making it database agnostic and applicable across EHR systems. DQ^e-v is a component of a growing toolset for examining data quality, Data Quality Explorer (DQ^e), which aims to provide a suite of interactive database-agnostic solutions to explore EHR data quality.

Multiple statistical procedures exist for detecting anomalies in data. Most of these procedures have been designed for time series data, which by definition do not completely match characteristics of patient records stored daily in EHR systems—even though time series data can be extracted from EHRs. Other methods for detecting variability in data might be considered too technical for clinical researchers without help from statisticians, and are expected to be conducted on a study-by-study basis—not as an overall view of variability when engaging in initial examinations of EHR data. DQ^e-v provides an interactive visual overview of variability in data through exploratory and predictive lenses, to assist with exploration of data in service of developing future studies and designing appropriate methods for analyses.

DQ^e-v's workflow is illustrated in Figure 1. It uses data aggregated by units of location, time, and medical condition; performs data preparation and analyses; produces visualizations; and presents the outputs in a web browser. This work is achieved via an interactive interface powered by the shiny package version 0.13.0 in R.[35]

Interactivity is a key attribute in DQ^e-v's design, through which the tool does not impose a specific point of variability to the users. Instead, the tool is designed to allow the users to explore relevant

Figure 1. Workflow for DQ^e-v



slices of data sets as they examine variability. Once variability is discovered, users must decide how statistical analyses will be conducted to account for the variability.

Preparing and Reading the Data

The R script, Read.R, reads the data and feeds it to the tool. It uses the data.table package to read the input data. All packages that are being used to run the tool are also loaded in Read.R. The script is annotated with instructions for preparing the data for DQ^e-v and configuring its location. The user needs to extract the input data out of the EHR system according to the data model described below.

The Data Model

The data model that the tool feeds from consists of three aggregation units (u_Loc, u_Time, and u_Cond), two counts (population, patient), and a percentage (prevalence) derived from the two counts. The data model is flexible in that the three aggregation units can be defined at any spatial level, time scale, and condition of interest. The input data columns are described in the following:

- Column u_Loc stores the location/spatial unit of analysis as character. Examples of spatial unit are clinic, organization, census tract, and county.
- u_Time, of type integer (currently), stores the time unit, which can be at any interval (i.e., hourly, daily, monthly, annual, etc.) as long as the unit is consistent across the data set.
- The phenotype- and cohort of interest can be stored in the u_Cond column as type character. Data in this column include an extract from one or multiple EHR data sources that aggregate counts by the unit of location, and unit of time for a user defined phenotype and cohort (e.g., a particular patient cohort, a set of medications, or clinical conditions). The tool's interface is designed to automatically read all unique values stored in u_Cond, so more than one condition can be stored in this column.
- Column population stores the total size of the patient population at the indicated location and spatial unit and time unit.
- Column patient is the size of the subset of population who have the indicated condition.
- Prevalence is patient divided by population for a condition at a given time unit and location/spatial unit.

Volume 5 (2017) Issue Number 1

Table 1 presents sample rows from the provided example data. The Read.R adds a new column to this data model where it copies the u_Time variable as a factor that will be used later by the app to generate two of the plots. For illustration purposes, an example data set database.csv is provided. The output of the Read.R is the source data set, srcdt, which feeds all tool functions in app.R.

User Interface (UI) and Server Data

DQ^{e}-v currently uses navbarPage format for shiny. app.R processes source data srcdt that has been read into R by Read.R to produce data for the user interface (UI) (datUI) and reactive data (dat and datREG) for the server. The tool utilizes datUI to read in the unique condition units, u_Cond, from the source data and automatically fill in the "Select Data" option on the left. Since "old" EHR data are often not reliable, but are often present due to various artifacts in EHR systems, the data on the UI panel is currently limited to years after 1980 for illustration purposes—this can be changed from the R script in app.R, if the user would like to choose a different time range.

Through data processing, the code adds two new columns to the source data prepared for the server (dat), where it stores two ratio indices for Interquartile Range (IQR) and Standard Deviation (STD), which will then be used for the "Exploratory Analysis" tab.

- IQR Ratio. Column iqr stores the ratio of interquartile range for each time unit (u_Time) to the mean interquartile range over all time units.
- STD Ratio. Column std stores the ratio of standard deviation for each time unit (u_Time) to the mean standard deviation over all time units.

Using the two indices, "iqr" and "std," users can interactively highlight time units where the two values are significantly above the overall patterns.

Tool Outputs

DQ^{e}-v produces a Web interface with four outputs; we refer to the outputs as "tabs" (Table 2). Together, the first three tabs (Variability Preview, Exploratory Analysis, and Density Plots) provide different visual representations of the data for the users to explore variability from their own point of view by targeting display of overall data distribution, high and low variability, and density functions. A final tab was added, the Regression-based Analysis tab, to demonstrate a predictive analytic visualization, which uses polynomial regression modeling to identify and recommend possible anomalies. All graphics in these outputs were produced using ggplot2 package version 2.0.0, and all dynamic graphics are enhanced with plotly package version 2.0.16.

We explain, below, the four tabs provided by DQ^{e}-v with snapshots of the output for each tab that were generated using the example data that comes

Table 1. Excerpt of Input Data from the Provided Example Data

U_LOC	U_TIME	U_COND	POPULATION	PATIENT	PREVALENCE
LOC_P	2010	Condition_F	807	298	0.369
LOC_V	2009	Condition_F	5456	1411	0.259
LOC_Y	1903	Condition_C	21514	46	0.002
LOC_Y	1950	Condition_C	21514	46	0.002

Table 2. DQ^e-v's Four Tabs and Their Functionality

TAB NAME	OUTPUT DESCRIPTION
Variability Preview	An overall visualization of data distribution and variability over time
Exploratory Analysis	Interactive visualizations to explore high-low variability
Density Plots	Visualization of probability density functions
Regression-based Analysis	Predictive analytics to recommend anomalous site locations and times

with the tool, and we demonstrate how one may interpret the outputs. The outputs we use here are for illustration purposes and were generated for different conditions and time units using data from Data QUEST.[33] Our interpretations of the output may not be exactly extended to the outputs and interpretations of other tabs because they are case specific to different queries from the Data QUEST data.

1. The Variability Preview Tab

The "Variability Preview" tab provides a dynamic interactive overview of the data through a box plot, which is a conventional plot for looking at variability, and a scatter plot (Figure 2). The goal of this tab is to show an overall view of data distribution across time and clinic location for a user-specified phenotype and cohort. The Y axes in both plots represent the prevalence of the selected condition, and the X axes show the selected time unit. The numeric values of a box's hinges are displayed when the users hover their mouse pointers over the box, as shown in Figure 2.

Because box plots give an excellent sense of the spread of a statistic but no indication of the size of the population depicted, the scatter plot is included to address this deficiency. The dots plotted are median prevalence values from each source data warehouse, and are sized proportional to population (on a log scale). Thus, the user can view prevalence value in light of how big or small the population at each location unit is.

For example, in Figure 2, the box plot shows that within-site-location variability in the number of diagnoses per visit was relatively higher between years 2006 and 2009 than the variability after year 2010. The scatter plot reveals that there were only two to three data points (data from two to three site locations) available between years 2006 and 2009. Altogether, in this example, the two plots in this tab raised concerns about reliability of data from before 2010 as there were few data points and high observable variability in that time frame.

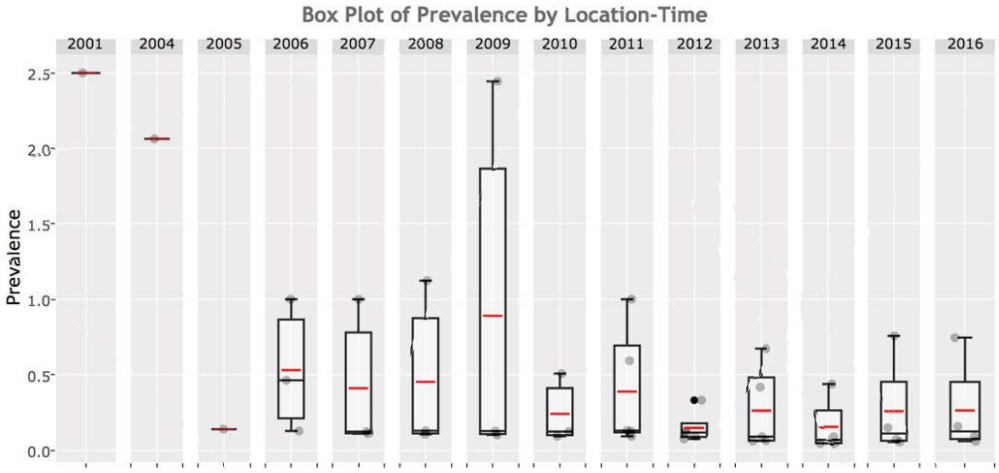
This tab provides a general view of the data, to raise questions to be investigated further using the subsequent tabs. The users can begin their data exploration in more detail through the next two tabs.

2. The Exploratory Analysis Tab

The "Exploratory Analysis" tab provides four visualizations for the user to explore potential variability (or lack thereof) in the data. On the left-hand side menu, the user can select one or more conditions to view. The UI automatically recognizes and lists unique values in u_Cond. The user can also

Volume 5 (2017) Issue Number 1

Figure 2. The Variability Preview Tab Previewing Diagnoses Per Visit Per Year



Hover over the Box plot to see the actual values for Max, 3rd quantile, mean (in red), 1 standard error over mean (in blue), median, 1st quantile, and Min.

Size of the poits on Scatter plot represent log of population at each location unit and time unit. Hover over to see values for the prevalence and log of population size for each location unit.

select a time unit range to zoom in on a particular period or zoom out to see the overall trend. These settings feed the necessary data to generate the four visualizations for exploring variability.

Figure 3 illustrates the plots generated in this tab, with data from the provided example data set. Here again, the X axes on all four graphics are the time unit and the Y axes are prevalence.

Prevalence Box Plots

The first two graphics on the Prevalence by Location-Time tab are ordinary box plots of the prevalence variable, which show distribution of prevalence (i.e., patient divided by population) of the selected condition among units of locations across the selected period. A scatter plot with jittered data points is also overlaid in the background. Box

plots at any specific time unit illustrate maximum, minimum, first and third quartiles, and median for the prevalence of the selected unit condition. In the example data set, the unit of time is year, but the time interval can be set at any uniform time interval by the user. The two graphics show variations in the annual prevalence of conditions G, H, and B between years 1998 and 2013. Box plots are conventionally used to visualize variability. However, to add more to these plots, users can highlight time units in which IQR (first plot) and STD (second plot) are within a certain range.

The interactivity allows users to highlight particular time units that have higher IQR or STD than the average over the selected period. A continuous color spectrum from gold to red is dynamically assigned to the minimum and maximum of selected ranges for better visual presentation of variability.

In Figure 3, for example, years are highlighted in which IQR and STD for diagnoses per visit data were between one to four times more than the average IQR and the average STD within 2001 and 2016. With these IQR and STD settings, in this example, the user can see in the box plots that six (in IQR plot—first plot) and five (in STD plot—second plot) time units were respectively highlighted as time units with high variability. As the figure shows, it is possible that some time units with low IQR ratio have high STD ratio, and vice versa, due to the spread of data points and outliers. For this reason, this tab provides users with the ability to consider both indices in their exploration of data variability. Also the users can speculate about the actual values for each time unit by changing the ranges for IQR and STD.

Scatter Plots

Two scatter plots complement information provided by box plots. Plot number three (Prevalence Over Time) adds a smoothed regression line with confidence intervals to the scatter plot of the Prevalence Over Time unit. The regression line illustrates the overall variability pattern. The confidence interval visually represents data points across time units, helping the users make more informed decisions about variability conveyed through the box plots.

The last plot, Overall Patient Population Over Time, is a scatter plot of the overall population within location units (e.g., primary care clinics), regardless of condition unit. This plot illustrates the changing pattern in population seen at location units within the selected time units. For example, in Figure 3, high variability between years 2005 and 2010 (highlighted with a yellow box) can be related to the relatively smaller population size across location units. The two scatter plots in this example show that as the total number of patients recorded in the EHRs increased over time, the number of diagnoses per visit slightly decreased before 2010 and stabilized thereafter.

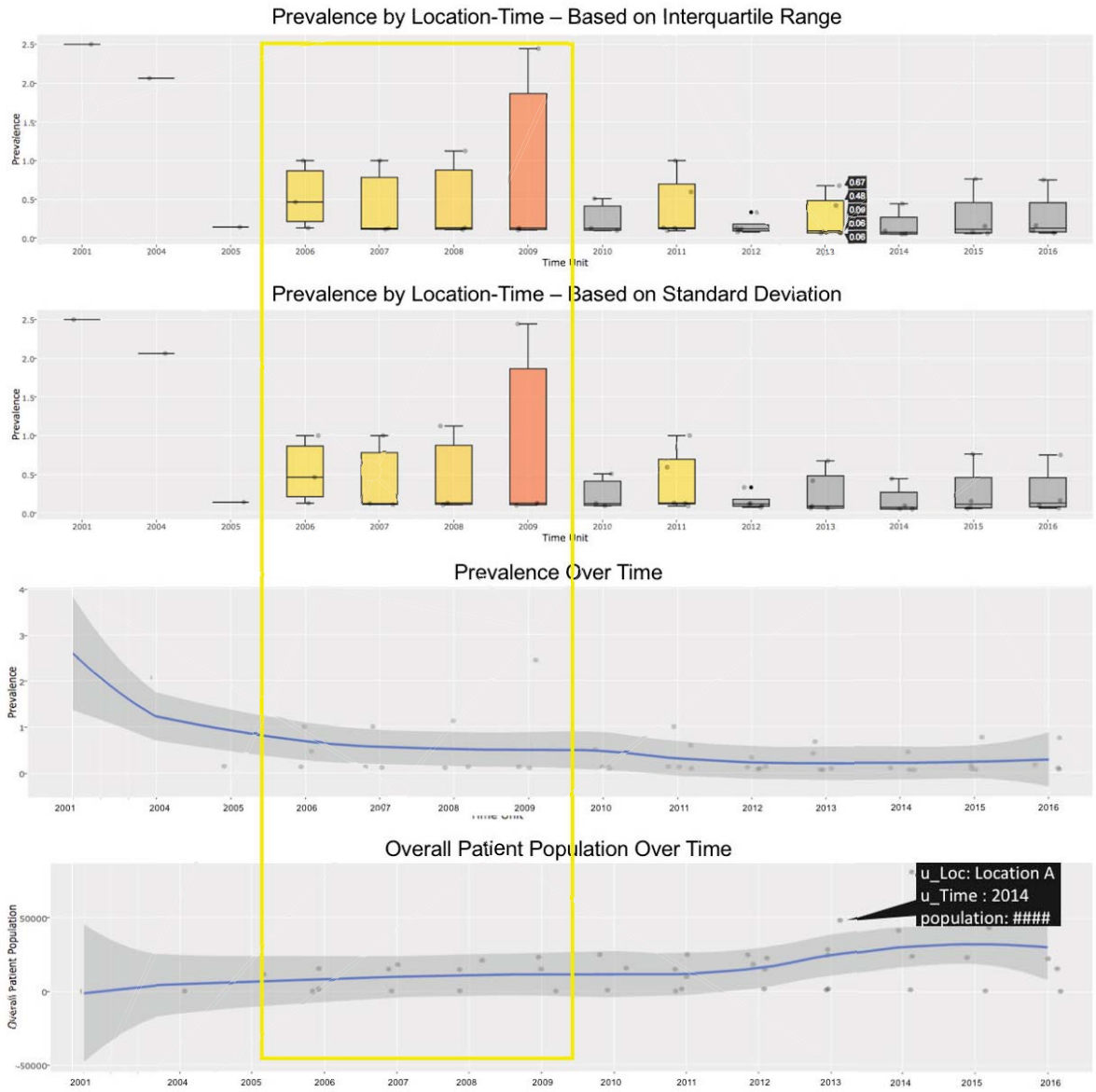
3. The Density Plots Tab

The "Density Plots" tab displays smoothed probability density function of a user-selected variable for a selected condition and time range. Visualizations in this tab are intended to complement the exploratory information obtained from the two preceding tabs. The menu bar in this tab allows the users to look at changes in distribution of their variable of interest, be it prevalence, patient, or population. Allowing the user to change the variable of interest provides a more holistic preview of data distribution. Figure 4 illustrates annual probability distribution functions for the number of creatinine labs per patients with a chronic kidney disease (CKD) diagnosis and 1+ visits between years 2005 and 2016.

The X axis in each plot represents the selected variable from the UI, and the Y axis represents density. This tab is especially useful for learning

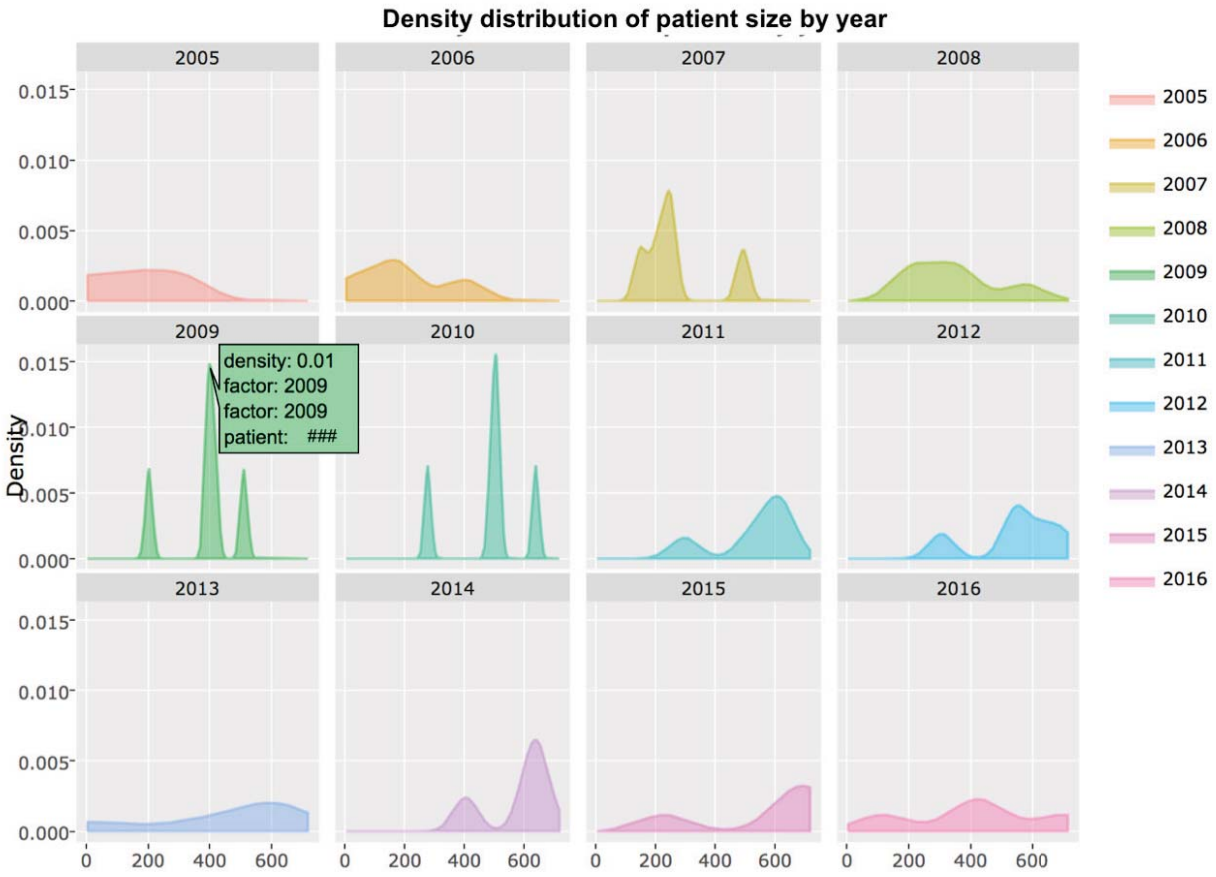
Volume 5 (2017) Issue Number 1

Figure 3. The Exploratory Analysis Tab's Plots Help Identify Years and Site Locations with High Variability in the Number of Diagnoses Per Visit



Note: The yellow box highlights a period with relative high variability.

Figure 4. Visualization of Probability Density Functions in Density Plots Tab for Number of Creatinine Labs Per Patients with a CKD Diagnosis and 1+ Visits



Note: Colors signify time units.

about distribution of data over time and to identify time units with abnormal distribution patterns. In Figure 4, we selected patient as the variable of interest (selection choices are prevalence, patient, and population) between years 2005 and 2016. According to the figure, density distribution of number of creatinine labs per patients with CKD diagnosis and 1+ visits was skewed to the left before 2008 and to the right after 2011. Also, in 2007, 2009, 2010, and 2014, the plot shows that there were two peaks in the observed density values.

4. The Regression-based Analysis Tab

The "Regression-based Analysis" tab uses predictive analytics to detect data points (i.e., single observations within a given unit of condition, time, and location) that are statistically aberrant given the data observed in preceding and ensuing time units. This tab takes a prescriptive approach that recommends the user site locations and times with anomalous data. These data points are marked as anomalous and displayed on two visualizations. Calculation of expected values are based on a series

Volume 5 (2017) Issue Number 1

of raw polynomial regression models that predict prevalence based on time unit and population for each location unit, using the function poly() in R. The regression function internally sets up the model matrix with the raw coding χ, χ^2, ..., χ^p, when χ is the variable of interest and p is the degree of the polynomial regression, which can be defined by the user.

Accordingly, the two regression models are estimated using the following function:

$$\theta = \beta_1 \sum_{i=1}^{p} \text{u\_Time}^i + \beta_2 \sum_{j=1}^{p} \text{population}^j$$

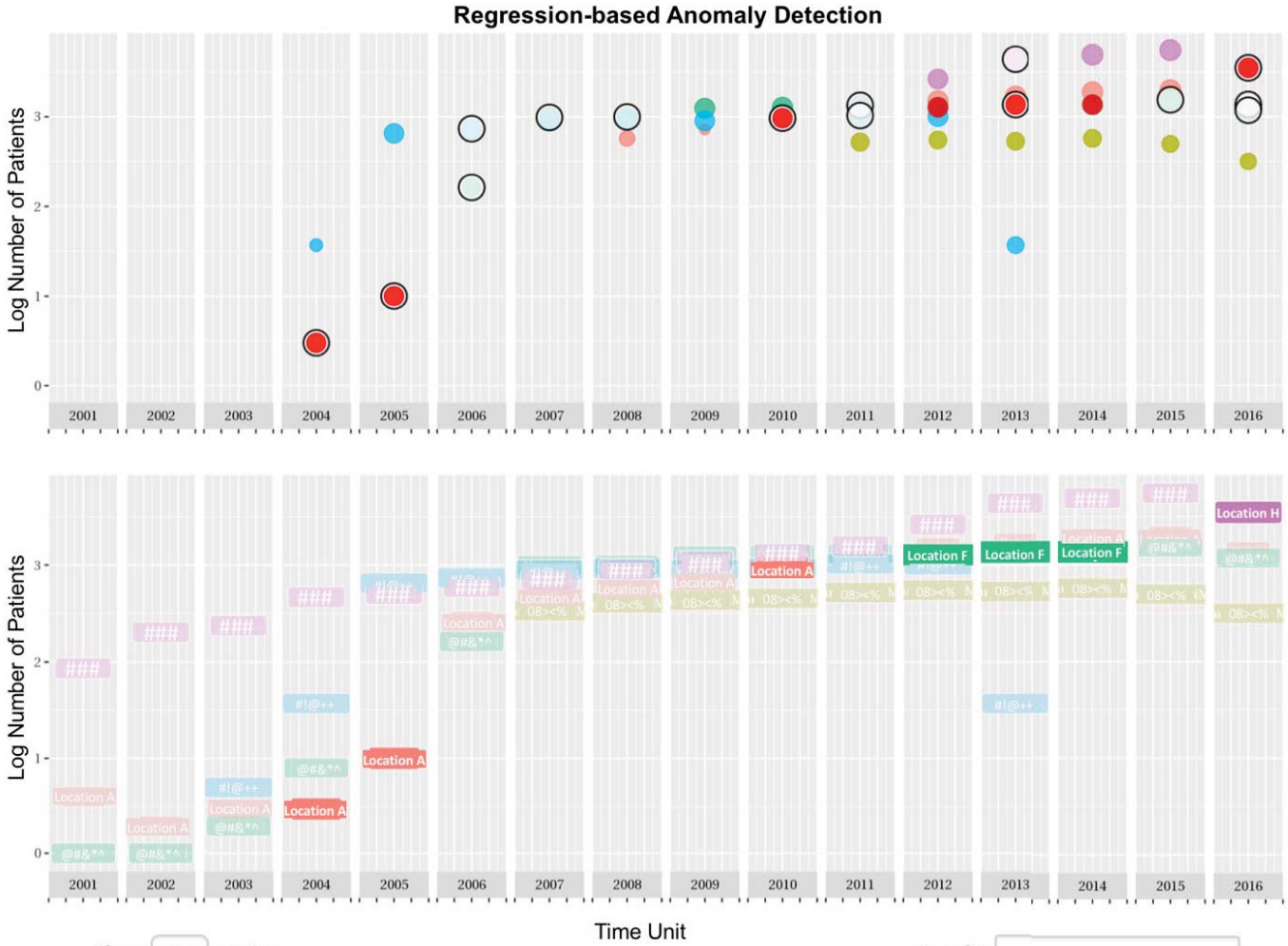
where p is the degree of polynomial regression and is the estimated value for the prevalence given all other data points provided for a unit of condition.

The tool uses polynomial regression as a smoothing mechanism. That is, the users can set up the smoothing degree in the UI (maximum has been set to five) and compare the difference in anomalies detected via two user-defined smoother polynomial lines. As illustrated in Figure 5, the regression based analysis produces three outputs: two plots and a consensus table.

The top plot is a scatter plot of log of number of hemoglobin A1c labs per patients with a diabetes diagnosis and 1+ visits at each site location and year. Dots are color coded by unit location and, similar to the scatter plot in Figure 2, are sized proportionally to the log of total population at each clinic and time unit. On top of the prevalence scatter plot, location units that have been detected as anomalous using polynomial regression estimates are highlighted. White dots with a black stroke line are anomalies detected with the first set of polynomial regression models (where the actual prevalence value at a given site location and time is statistically different from the estimated values given all data points for location x within the selected time unit range). Red dots on top of white dots show anomalies detected with the second set of polynomial regression models. The default is set to one, so that red dots overlap the white dots. Once the user increases the polynomial degrees for the first and second model (i.e., moves toward smoother regression lines), fewer red dots are expected. The second plot, which has the same x and y axes as the first plot, highlights only the location names detected as anomalous with the second polynomial models, as a reference. The data table shows the time and location units where the outcomes of the two models are in consensus. This interactivity enables the users to apply different smoothing scenarios and to see how anomaly identification may differ based on the defined models—model selection can vary from Ordinary Least Square (OLS) to a five-degree polynomial.

In the Figure 5 example, the first plot has identified 15 data points (site locations and years) as having potentially anomalous numbers of hemoglobin A1c labs per diabetes patients with more than one visit, according to model number one—using an OLS algorithm (polynomial with degree set to one). Five of these data points were also identified as anomalies by the second model—a degree-three polynomial regression algorithm. The second plot highlighted the location ID (u_Loc) for the three data points—i.e., it shows that site location A in years 2004, 2005, and 2010, site location in years 2012 to 2014, and site location H in year 2016 were identified in both models as having anomalous numbers of hemoglobin A1c labs per patients with a diabetes diagnosis and 1+ visits. We obstructed the site-location names for this illustration, but it is useful to identify sites directly in the visualizations to help facilitate hypotheses for the noted variability. As the two models agree on these seven data points as anomalies, the table underneath the plots illustrated three of the seven anomalies in this example—the

Figure 5. Outputs of the Regression-Based Analysis on Hemoglobin A1c Labs Per Patient with a Diabetes Diagnosis and 1+ Visits. (Site-location Names Are Obstructed.)



Show 3 entries Search:

u_Loc	u_Time	POLY1	POLY2
Location A	2005	Anomaly	Anomaly
Location F	2013	Anomaly	Anomaly
Location H	2016	Anomaly	Anomaly
u_Loc	u_Time	POLY1	POLY2

Volume 5 (2017) Issue Number 1

number of rows to show can be defined by the user. The output indicates that the 2010 numbers of hemoglobin A1c labs per patients with a diabetes diagnosis and 1+ visits in site-location A were significantly different from the trend lines predicted from data from all other years in the same location. The users can then go back to the previous tabs to explore the recommended anomalous site location and time combinations with more details.

Discussion

Variability in EHR data can reflect both noise (i.e., data quality errors) and signal (i.e., real data characteristics such as demographic and practice pattern differences). When observing variability, we must take necessary steps to understand and address noise within data sets to extrapolate unbiased information. Data analysts apply various statistical methods to detect and account for variability in analytic data sets. Such methods—often designed on a case-by-case basis—could be too technical, or could lack appropriate outputs to be used by database administrators or researchers at the data warehouse level where EHR data are stored and maintained from multiple organizations. Agglomeration of data from multiple observation units (e.g., clinical settings) at EHR data warehouses presents an opportunity to conduct cross comparisons and evaluate variability in individual data sets. However, due to the lack of tools that are designed for high-level variability exploration, data variability is not being actively monitored before analytic data sets are extracted, when there is more potential to capture variability.

We introduced an interactive, open source, database-agnostic tool that provides a framework to explore variability in EHR data. DQ^e-v provides a suite of views to examine variability from both predictive and exploratory approaches. DQ^e-v's operationalization of the term "variability" in EHR settings maps the temporal and atemporal subcategories of plausibility as defined in the harmonized data-quality assessment framework.[18] The tool is primarily intended to work at the data warehouse level to help inform database administrators of the variability in EHR data as they examine data quality during extraction, translation, and loading (ETL) processes. Yet, given its flexible user-autonomous design (users can define the dimensions to look at) and agnostic nature to the backend data set, DQ^e v can also be used on a specific analytic data set by a research team. We are currently implementing DQ^e-v on a regional practice-based research network and a national EHR data registry to explore variability in a set of expert-defined aggregate counts in primary care settings. We highlighted examples of DQ^e-v's outputs using data from the WWAMI region Practice and Research Network's Data QUEST EHR data warehouse, using preliminary data loads to illustrate variability likely driven by errors in ETL processes and clinic anomalies. It is important to mention that our brief interpretations of the outputs are not necessarily indications of data quality in Data QUEST, but rather are presented here for demonstration.

The Observational Health Data Sciences and Informatics Collaborative (OHDSI) consortium has an open source tool, Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES), which provides visualizations of Observational Medical Outcomes Partnership (OMOP) data. DQ^e-v has several notable differences that address functionality that ACHILLES does not provide for OMOP repositories, as well as scalability to other data models. DQ^e-v is database agnostic, meaning that its input data can be extracted from any data model. Also, DQ^e-v provides visual analytics (adds analytics to visualization) and has a modular architecture, which provides flexibility to the user for adding to and customizing its

analytics tabs. Finally (and perhaps most notably), beyond being a tool, DQ^e-v provides a framework for exploring variability in clinical databases. The interactive functionalities and data visualizations that DQ^e-v offers in a tabular modular format provide the users with a framework to zoom in or zoom out into specific data points in order to investigate variability in data.

Limitations and Future Work

The next steps for DQ^e-v involve testing and improving its usability, efficiency, and interoperability. Usability research is needed to understand whether the features and outputs of the tool are intuitive and comprehensive enough for the targeted users (e.g., researchers, analysts, and database administrators). In addition, needs assessments are required to understand whether the specific visualizations and analytics are what users need in order to explore variability in clinical data, and to incorporate unmet user needs into future versions of the tool. Although we intend to keep the tool database agnostic, we may need to provide more guidelines to the users for extracting aggregate counts out of their EHR repository to produce the input data. In addition, efficiency of the R code for analysis and visualization has not been tested on a large amount of data, which may require programming methods to scale functionality. Namely, depending on the response time to perform the analyses and produce the outputs, we might need to improve the R code to make the tool faster. It is also possible to use faster versions of R (e.g., Microsoft R Server), apply parallel computing techniques, or seek out-of-memory solutions to improve computing efficiency when running DQ^e-v on a large amount of data. Also, evolving the tool to distributed networks to eliminate physical centralizing of the data across data repositories would expand scalability.

Conclusion

DQ^e-v is database agnostic, open source, and reproducible—providing a framework for users to access a user friendly, dynamic graphical user interface (GUI) to explore various levels of data variability across time and site location. From the data preprocessing, to analyses and visualizations, all procedures are written in R statistical language, a widely used programming language and software environment for statistical computing and data visualizations. To make the tool fully reproducible, annotated R codes, a test data set, and an in-depth readme document are provided on GitHub and CIELO. Database administrators and researchers can download the codes from one of these repositories, read the detailed instructions (also provided in this paper), and run DQ^e-v with minimal R knowledge. The test data set is provided for the setup and initial implementation of the tool in the users' environment. Once the tool is set up, to run DQ^e v against actual data from the users, the input data model needs to be extracted from the database in which the users aim to explore variability. Further expansion upon the available tool is also possible by users, and we welcome Github pull requests to improve the tools functionality in the future.

Acknowledgements

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000423. The authors would like to thank Roy Pardee from Group Health Research Institute and Alison Kosel for their constructive feedback on the tool and paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Volume 5 (2017) Issue Number 1

References

1. Institute of Medicine. *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. Washington (DC): The National Academies Press; 2014.
2. Arts DGT, De Keizer NF, Scheffer G-J. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J *Am Med Inform Assoc*. 2002;9:600-611. doi:10.1197/jamia.M1087.
3. Hsiao C-J, Hing E, Socey TC, Cai B. Electronic health record systems and intent to apply for meaningful use incentives among office-based physician practices: United States, 2001-2011. *NCHS Data Brief*. 2011:1-8. http://www.ncbi.nlm.nih.gov/pubmed/22617322.
4. 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.
5. Grinspan ZM, Banerjee S, Kaushal R, Kern LM. Physician Specialty and Variations in Adoption of Electronic Health Records. *Appl Clin Inform*. 2013;4:225-240. doi:10.4338/ACI-2013-02-RA-0015.
6. Hersh WR. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *Am J Manag Care*. 2007;13:277-278.
7. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20:144-151. doi:10.1136/amiajnl-2011-000681.
8. Bae CJ, Griffith S, Fan Y, et al. The Challenges of Data Quality Evaluation in a Joint Data Warehouse. *eGEMs*. 2015;3(1):1125. doi:10.13063/2327-9214.1125.
9. Murdoch T, Detsky A. The inevitable application of big data to health care. *J Am Med Inform Assoc*. 2013;309(13):1351-1352. http://dx.doi.org/10.1001/jama.2013.393.
10. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46:830-836. doi:10.1016/j.jbi.2013.06.010.
11. 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.
12. Hogan WR, Wagner MM. Accuracy of Data in Computer-based Patient Records. *J Am Med Informatics Assoc*. 1997;4:342-355. doi:10.1136/jamia.1997.0040342.
13. Hennessy S, Leonard CE, Palumbo CM, Newcomb C, Bilker WB. Quality of Medicaid and Medicare data obtained through Centers for Medicare and Medicaid Services (CMS). *Med Care*. 2007;45:1216-1220. doi:10.1097/MLR.0b013e318148435a.
14. Liaw ST, Rahimi A, Ray P, et al. Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. *Int J Med Inform*. 2013;82:10-24. doi:10.1016/j.ijmedinf.2012.10.001.
15. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health*. 2014;11:5170-5207. doi:10.3390/ijerph110505170.
16. 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.
17. Kahn MG, Brown JS, Chun AT, et al. Transparent reporting of data quality in distributed data networks. *eGEMs*. 2015;3(1):1052. doi:10.13063/2327-9214.1052.
18. Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *eGEMs (Generating Evid Methods to Improv patient outcomes)*. 2016;4(1). doi:10.13063/2327-9214.1244.
19. Estiri H, Chan Y-F, Baldwin L-M, Jung H, Cole A, Stephens KA. Visualizing Anomalies in Electronic Health Record Data: The Variability Explorer Tool. *AMIA Jt Summits Transl Sci* Proc AMIA Summit Transl Sci. 2015;2015:56-60. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4525227&tool=pmcentrez&rendertype=abstract.
20. 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.
21. 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.
22. Jordan K, Clarke AM, Symmons DPM, et al. Measuring disease prevalence: A comparison of musculoskeletal disease using four general practice consultation databases. *Br J Gen Pract*. 2007;57:7-14.
23. 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.
24. 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.
25. 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. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243268&tool=pmcentrez&rendertype=abstract%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243268/.
26. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care*. 2013;51:S22-9. doi:10.1097/MLR.0b013e31829b1e2c.
27. Li R, Abela L, Moore J, et al. Control of data quality for population-based cancer survival analysis. *Cancer Epidemiol*. 2014;38:314-320. doi:10.1016/j.canep.2014.02.013.
28. 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.

29. Walker AM. Matching on provider is risky. In: *Journal of Clinical Epidemiology*. Vol 66. ; 2013. doi:10.1016/j.jclinepi.2013.02.012.
30. 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.
31. Hartzema AG, Reich CG, Ryan PB, et al. Managing data quality for a drug safety surveillance system. *Drug Saf*. 2013;36 Suppl 1:49-58. doi:10.1007/s40264-013-0098-7.
32. Pace WD, Fox CH, White T, Graham D, Schilling LM, West DR. The DARTNet Institute: Seeking a Sustainable Support Mechanism for Electronic Data Enabled Research Networks. *eGEMs*. 2014;2(2):1063. doi:10.13063/2327-9214.1063.
33. Stephens KA, Anderson N, Lin C-P, Estiri H. Implementing partnership-driven clinical federated electronic health record data sharing networks. *Int J Med Inform*. 2016;93:26-33. doi:10.1016/j.ijmedinf.2016.05.008.
34. Stephens KA, Lin C-P, Baldwin L-M, Echo-Hawk A, Keppel GA. A web-based tool for cataloging primary care electronic medical record federated data: FInDiT. In: *CTSA 2011 Informatics Annual Meeting*. Bethesda, MD; 2011.
35. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R. 2016. https://cran.r-project.org/package=shiny.