BMC Medical Informatics and Decision Making

RESEARCH

Open Access

A statistical quality assessment method for longitudinal observations in electronic health record data with an application to the VA million veteran program

Hui Wang[1], Ilana Belitskaya-Levy[1], Fan Wu[1], Jennifer S. Lee[1,2,3], Mei-Chiung Shih[1,4], Philip S. Tsao[1,2] and Ying Lu[1,3,4*] on behalf of VA Million Veteran Program

Abstract

Background: To describe an automated method for assessment of the plausibility of continuous variables collected in the electronic health record (EHR) data for real world evidence research use.

Methods: The most widely used approach in quality assessment (QA) for continuous variables is to detect the implausible numbers using prespecified thresholds. In augmentation to the thresholding method, we developed a score-based method that leverages the longitudinal characteristics of EHR data for detection of the observations inconsistent with the history of a patient. The method was applied to the height and weight data in the EHR from the Million Veteran Program Data from the Veteran's Healthcare Administration (VHA). A validation study was also conducted.

Results: The receiver operating characteristic (ROC) metrics of the developed method outperforms the widely used thresholding method. It is also demonstrated that different quality assessment methods have a non-ignorable impact on the body mass index (BMI) classification calculated from height and weight data in the VHA's database.

Conclusions: The score-based method enables automated and scaled detection of the problematic data points in health care big data while allowing the investigators to select the high-quality data based on their need. Leveraging the longitudinal characteristics in EHR will significantly improve the QA performance.

Keywords: Data quality assessment (DQA), Electronic health record (EHR), Real world evidence, Clinical informatics, Health care big data, Vital signs

Background

The role of real-world evidence (RWE) is rapidly expanding over the last several years. It is now well recognized that RWE has potential for reshaping clinical research and clinical decision-making, even at regulatory level. As an example, the 21st Century Cures law passed in 2016 requires that FDA considers RWE for supporting regulatory decisions as a means of bringing new treatment to patients more quickly and efficiently. Electronic health records (EHR) data is a major source of RWE as well as a driving force behind RWE use, owing to its big size, rich dimensions, real-time update, and longitudinal characteristics. With the wide-spread adoption of EHRs in the US, the number of research studies based on the EHR data is rapidly increasing. Examples include disease

*Correspondence: ylu1@stanford.edu
[4] Department of Biomedical Data Science, Stanford University School of Medicine, 1265 Welch Road, X359, Stanford, CA 94305-5464, USA
Full list of author information is available at the end of the article

burden [1], post-marketing safety surveillance [2, 3], and comparative effectiveness including synthetic controls [4–6]. However, one major challenge for using the EHR data to support clinical decision making is whether EHR data is of satisfying quality for drawing any meaningful conclusions.

EHRs are routinely collected by providers at a patient care facility for administration use. They are typically not collected to the same standard of quality as those of research data, which are subject to routine monitoring, auditing, and verification. Therefore, before EHR data can be used to answer research questions, it must be assessed for its quality including conformance, completeness, and plausibility [7–11]. The conformance in data quality assessment (DQA) evaluates if data adheres to specified structural and formatting specifications of the database. The completeness examines if the presence or absence of data attributes are within expectation in a database. The plausibility determines the degree to which data values are believable. Plausibility can be further categorized into atemporal plausibility and temporal plausibility. Atemporal plausibility focuses on cross-sectional data features (e.g., height values must be non-negative), while temporal plausibility focuses on a sequence of values over time (e.g., adult height is stable over time). For data as big as the EHR, manual checking is infeasible, and algorithmic methods must be used. There is considerable effort for bringing forth automated quality assessment procedures to screen and clean EHR data [12, 13], but no standard is yet established. The data quality assessment of EHR depends on the EHR system, the protocol of how the data is collected (for example, provider-report vs. self-report), and the types of the data. As an example, methods for cleaning discrete data such as diagnosis codes and continuous data such as weight and height can be quite different. As a result, most DQAs implemented in major data sharing networks are rule-based methods [14, 15]. Rule-based methods are simple to implement but have limited power to detect data issues, especially for temporal plausibility. In this paper, we will introduce a score-based method that addresses the temporal plausibility for continuous measurements in EHR data.

The development of our method is based on the EHR data from the Veterans Health Care Administration (VHA). VHA has the largest integrated federal health care system and formally adopted an EHR system as early as in 1970s. VHA collects complete health-care history of veterans who use its care using VistA, an information technology infrastructure implemented in 1980s. The VistA data is extracted in SAS and SQL and stored in VHA's corporate data warehouse (CDW). The CDW data since year 2000 is made available to VHA researchers in a structured format. The CDW data is further standardized into the Observational Medical Outcomes Partnership (OMOP) Common Data Model for more efficient use in research [16].

We describe in this article an automated procedure for detection of implausible observations among continuous and autocorrelated variables from the EHR data such as height and weight, and body mass index (BMI) that is derived from these two variables. The algorithm can be applied to other types of continuous data and works best when data can reach a stationary distribution for a reasonable length of time.

Methods

A widely used method to identify problematic observations in continuous variables is simple thresholding: if a data point falls in an implausible range, it will be considered as erroneous. Thresholding method considers each measurement in isolation and falls into the atemporal plausibility DQA category. However, it ignores the longitudinal profile of a patient in EHR and can result in exclusion of good data or inclusion of erroneous data. For example, a patient weighs consistently of 380 pounds in ten visits in three years. With thresholding method, all the data of this patient is likely considered as error and excluded from analysis. On the other hand, a measurement of 120 pounds for an under-weight patient of 80 pounds is considered good data when it is indeed an error.

The proposed method in this paper addresses the drawbacks of thresholding method. A measurement of a continuous variable is considered of questionable quality if it experiences implausible changes over time. And our statistical procedure calculates a longitudinal plausibility score (Q_R) based on repeated measurements in that patient.

Longitudinal plausibility score (Q_R)

The number of repeated measurements and the time between measurements play important roles in the determination of whether a measurement is plausible or not. Using weight as an example, a patient has 6 weight measurements in 3 months: 5 of them are 200 pounds and one is 180 pounds. The measurement of 180 pounds is likely an error. Another patient has 2 weight measurement one year apart: the first is 200 pounds and the second is 180 pounds. There are likely no errors here.

Considering these factors, we have chosen the exponentially weighted moving average (EWMA) for calculation of the longitudinal plausibility score Q_R. The rational is two folds. Firstly, EWMA is a commonly used time-weighted method in quality control for manufacturing process due to its ability to model decaying dependencies among data over time. An alternative method also commonly used in quality control process is cumulative sum approach (CUMSUM). But CUMSUM assigns equal weights to every time point and cannot model decaying time dependency in autocorrelated data. Secondly and perhaps more importantly, EWMA is a simple method that can scale to the volume of the data linearly and, therefore, suitable for processing large amount of EHR data.

For an individual patient, suppose that we have a sequence of n measurements $y_1 ... y_n$ taken at time points $t_1 ... t_n$. The EWMA $\bar{y}_{i,EWMA}$ for a measurement y_i taken at time t_i, $t_1 \leq t_i \leq t_n$, is defined as a weighted average over the entire sequence:

$$\bar{y}_{i,EWMA} = \frac{\sum_{j=1}^{n} w_j y_j}{\sum_{j=1}^{n} w_j},$$

where the weight w_j is determined by the time interval between t_i and each t_j: $(1 \leq i, j \leq n)$:

$$w_j = \begin{cases} 1 & \text{if } j = i \\ e^{-\frac{|t_j - t_i|}{\tau}} & \text{if } j \neq i \end{cases}.$$

Parameter τ tunes for the dependency of $y_{i,EWMA}$ on its neighboring measurements and affects the smoothness of the EWMA estimates. Larger τ leads to more correlated and smoother EWMA estimates.

Let us define d_i as the absolute difference between the observed y_i and the EWMA estimate:

$$d_i = |y_i - \bar{y}_{i,EWMA}|$$

The variance estimates of d_i is, assuming y_is are independent of each other:

$$Var(d_i) = Var(y_{t_i,EWMA} - y_i) = var(\frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} - y_i) = \left(\frac{\sum_{i=1}^{n} w_i^2}{(\sum_{i=1}^{n} w_i)^2} + \frac{2w_i}{\sum_{i=1}^{n} w_i} + 1 \right) Var(y_i).$$

We can then derive a Z score for d_i and its corresponding two-sided p-value from a standard normal distribution:

$$Z_{R,i} = \frac{d_i}{SE(d_i)},\ \mathrm{Q}_{EWMA,i} = 2(1 - \Phi(Z_{R,i}))$$

The parameter τ tunes the smoothness of moving averages and plays an important role in identifying outliers. We provide a heuristic formula for setting *τ:*

$$\tau = -\frac{\xi/12}{log(\omega)},$$

In the formula, ω is the desired dependency in percentage a researcher wants to put on the neighboring observations ξ month away from *t*. For example, for height, we can assign 90% dependency on observations 1-year (12-month) away from the time point *t* because we expect the adult height to be stable over time. In this case, $\omega = 0.9$, $\xi = 12$, and it leads to $\tau = 9.49$. With this τ, the dependency on measurements 2-year and 5-year away is 81% and 59%. Meanwhile, the weight measurements have much greater variabilities over time, and it may be more reasonable to assign 90% weight on observations half a month away (i.e. $\xi = 0.5$ and $\tau = 0.4$). The calculation of $\mathrm{Q_R}$ also requires an estimate of the variance of measurements $\mathrm{y_i}$ –- $\mathrm{Var(y_i)}$. This variance can be estimated from data using a random effect model. When validation data is available, the tuning parameters τ and $\mathrm{Var(y_i)}$ can be searched for to achieve an optimal performance of the algorithm. For height, we used the heuristic formula to choose our tuning parameters. For weight, we optimized τ and $\mathrm{Var(y_i)}$ using a validation dataset on the false discovery rate and detection rate ($\tau = 0.5$, $\mathrm{Var(y_i)} = 210$) (Additional file 1: Methods).

Thresholding score ($\mathrm{Q_S}$)

For comparison purpose, we also used thresholding method and computed a QA score Qs for each measurement: the two-sided p-value of a simple Z-score that measures the distance between the observed value $\mathrm{y_i}$ and the population mean $\mathrm{y_u}$:

$$Z_{S,i} = \frac{y_i - y_u}{SD(y_i)},\ \mathrm{Q}_{S,i} = 2(1 - \Phi(Z_{S,i}))$$

The population mean y_u and standard deviations (SD) can be estimated from the reference data (Additional file 1: Methods).

Both $\mathrm{Q_R}$ and $\mathrm{Q_S}$ are invariant to the units of the measurements of the outcome. For time measurement in $\mathrm{Q_R}$, age was used for presentation simplicity. Other time measurements such as calendar years can also be used without changing the results if the time intervals among measurements are preserved. At least two measurements are required for calculation of $\mathrm{Q_R}$. The more measurements available, the smaller the standard error for EWMA, and hence $\mathrm{Q_R}$ is more precise.

Study ethics and participant consent

The Million Veteran Program received ethical and study protocol approval by the Veterans Affairs Central Institutional Review Board and informed consent was obtained for all participants. This methodology study protocol was approved by the Stanford University Institution Research Board. All analyses were based on deidentified data from VA CDW. All methods were carried out in accordance with relevant guidelines and regulations.

Table 1 Characteristics of the height and weight data in the MVP cohort

	Height **(N = 495,393)**	**Weight** **(N = 496,292)**
Total number of records	10,945,576	25,400,615
Number of subjects with multiple measurements, N (%)		
Single measurement	9,987 (2.0%)	3,206 (0.6%)
> =2 measurements	485,406 (98.0%)	493,086 (99.3%)
> =3 measurements	472,133 (95.3%)	488,684 (98.5%)
Number of measurements in a subject		
Mean (SD)	22.1 (20.4)	51.2 (59.8)
Median (IQR)	17.0 (8, 29)	38 (19, 68)
Number of years of follow-up for a subject		
Mean (SD)	11.5 (5.2)	12.2 (5.1)
Median	12.2 (7.2, 16.1)	13.1 (7.9, 16.8)
Number of measurements per calendar year in a subject[a]		
Mean (SD)	2.0 (1.2)	3.7 (3.9)
Median	2.0 (1.0, 2.0)	3.0 (2.0, 4.5)

[a] Median is used if subjects have multiple-year data

Results

Demographics and characteristics

We analyzed height and weight data in an MVP (Million Veteran Program) cohort of 496,311 patients using VA EHR data between year 2000 and 2016. A total of 10,960,056 height records and 25,548,357 weight records were analyzed. Most patients in this MVP cohort is male (91.4%), white race (71.6%), and non-Hispanic (89.3%). The median age is 64.4 years old at the enrollment (Additional file 1: Table S1).

For analysis, we required that a height measurement is within the range of 40 and 100 inches and a weight measurement within 40 and 1000 pounds. We also required that at the time a measurement was taken the patient was at least 17 years old in database. We removed height or weight records measured more than 3 times on the same day as these records are likely computer entry errors. These data preprocessing steps resulted in height data of 495,393 patients with 10,945,576 measurements. The weight data included 496,292 patients with 25,400,615 measurements. Among these patients, 485,406 had more than one measurement for height and 493,086 patients for weight. The median of the total number of years of follow-up is 12.2 in the height data and 13.1 in the weight data. The median frequency of measurements is 2 measurements per calendar year for height and 3 for weight (Table 1).

Observation-level QA

Our analysis focused on the patients with at least two measurements. Both the longitudinal plausibility score (Q_R) and the thresholding score (Q_S) were calculated for each measurement for height and weight.

Table 2 Proportion of positive and negative measurements identified by Q_R and Q_S

	Positive by Q_R (Q_R < =0.05)	**Negative by Q_R ($Q_R > 0.05$)**
Height (N = 10,945,576)		
Positive by Q_S ($Q_S \leq 0.05$)	1.0%	4.2%
Negative by Q_S ($Q_S > 0.05$)	5.3%	89.5%
Weight (N = 25,397,409)		
Positive by Q_S ($Q_S \leq 0.05$)	0.3%	8.8%
Negative by Q_S ($Q_S > 0.05$)	0.6%	90.2%

Table 2 compares the proportion of flagged measurements using a cutoff score of 0.05 for both Q_R and Q_S scores. For height, the two QA scores agreed on 90.5% of the records: 1.0% of data were positive findings and flagged as questionable by both methods, and 89.5% of data were classified as negative and not flagged. The discordance between the two scores counted for 9.5% of the data: 4.2% data were flagged by Q_S but considered negative by Q_R, and 5.3% of the data is vice versa. For weight, only 0.3% data were flagged by both scores; 8.8% data were flagged by Q_S while considered negative by Q_R, while 0.6% data were flagged by Q_R but not flagged by Q_S, and the remaining 90.2% data were in concordance as negatives. The histograms in Figs. 1 and 2 illustrate how different the Q_R and the Q_S scores can be for identifying problematic data. The observations flagged by Q_R (panel b) have a similar distribution as the normal values (panel a), which cannot be identified by Q_S scores. In the meantime, The Q_S score flagged out-of-range or infrequent values mostly considered negative by Q_R. Further examination of data suggests that the Q_S flags all observations for patients who are over- or under-weight even though these measurements are consistent over time and should be considered accurate (Additional file 1: Fig. S1).

Subject-level QA

A subject-level QA score can be calculated from the Q_R, defined as the proportion of flagged measurements in a subject. When the proportion of flagged measurements in a patient is too high, all the data of the patient can be considered unreliable and removed from analysis. In our MVP cohort, 9% subjects were identified for having more than 20% of their measurements flagged for height, and 0.47% subjects were identified for weight. These numbers dropped to 2.36% for height and 0.07% for weight when a 50% cutoff was used. More details can be found in Additional file 1: Table S2.

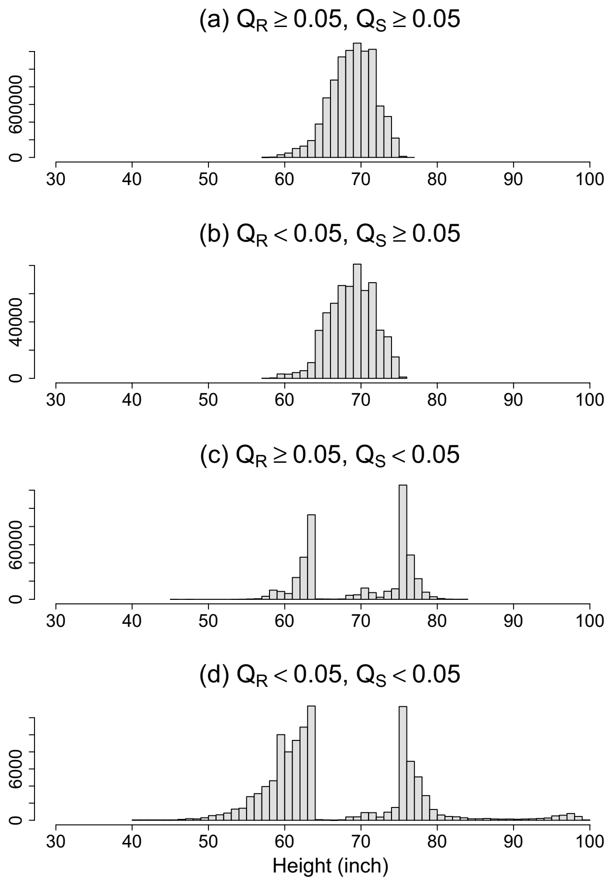

Fig. 1 Histograms of height measurements (inch) stratified by the agreements between Q_R and Q_S. **a** Height measurements where both Q_R and $Q_S > 0.05$ cutoff value; **b** Height measurements where $Q_R \leq 0.05$ and $Q_S > 0.05$; **c** Height measurements where $Q_R > 0.05$ and $Q_S \leq 0.05$; **d** Height measurements where both Q_R and $Q_S \leq 0.05$ cutoff value. X-axis represent the values of height, and y-axis is frequency counts

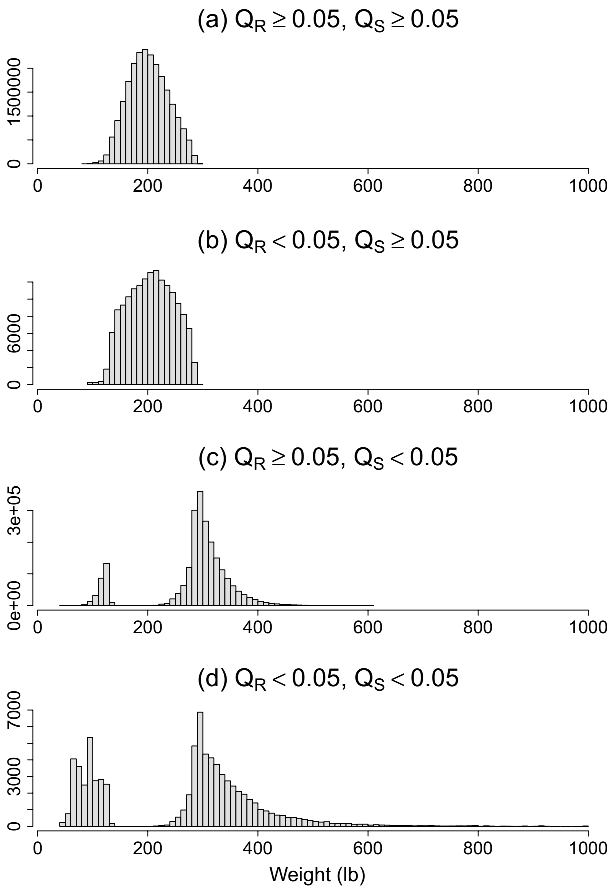

Fig. 2 Histograms of weight measurements (lb) stratified by the agreements between Q_R and Q_S. **a** Weight measurements where both Q_R and $Q_S \geq 0.05$ cutoff value; **b** Weight measurements where $Q_R \leq 0.05$ and $Q_S \geq 0.05$; **c** Weight measurements where $Q_R \geq 0.05$ and $Q_S \leq 0.05$; **d** Weight measurements where both Q_R and $Q_S \leq 0.05$ cutoff value. X-axis represent the values of height, and y-axis is frequency counts

Validation

To evaluate the performance of the EWMA algorithm, we randomly selected 100 patients whose proportion of flagged measurements were between 0 and 20% using a p-value cutoff of 0.05. The data of these patients were manually reviewed independently by two biostatisticians for identification of problematic measurements. All discrepancies between the two reviewers were reviewed and called independently by a third biostatistician. All reviews were blinded to the results of the algorithm. The incidence of positives is low: for height, 343 out of a total of 6,652 (5%) unique measurements were identified to be problematic in manual review; and for weight, 175 out of 16,825 (1%) unique measurements were considered problematic.

Both the EWMA QA algorithm and the simple thresholding method were evaluated against the manual-review results. False positive rate (FPR), power, and positive and negative predictive values (PPV/NPV) were calculated. The EWMA QA method has an FPR of 1.6% for height and 0.3% for weight. The power of detecting a problematic measurement is 88.9% for height and 75.4% for weight. The PPV (proportion of true positives among called positives) is 75.9% for height and 71.4% for weight. The NPV (proportion of true negatives among called negatives) is 99.4% for height and 99.7% for weight. In contrast, the thresholding method has a much lower power with a higher FPR. The positive predictive values are also much lower than the EWMA method, while the negative predictive values are similar (Table 3). The results are consistent when a cutoff of 0.01 was used for Q_R and Q_S. (Additional file 1: Table S3).

Use case: BMI

BMI is an indicator of obesity and often serves as prognostic factors for a variety of diseases such as diabetes and cardiovascular conditions. The accurate assessment of BMI is required in many studies. With EHR data, we have a way to assess BMI longitudinally. But we want to be mindful of the quality of BMI calculated from height and weight of EHR data. To demonstrate how the quality of height and weight data impacts BMI calculation, we computed BMI in our data using all the data without any QA, data QAed with the thresholding method (Q_S), and data QAed with the longitudinal method (Q_R). We then grouped BMIs into four categories of underweight (< 18.5), normal to overweight (≥ 18.5 and < 30), obese class I/II/III (≥ 30), and obese class III (> 40) according to WHO classifications. The proportion of incidence of a patient who ever falls into each BMI class were then calculated. Table 4 compares these proportions among the three methods.

BMI was calculated when height and weight were measured on the same day. There are 10,377,511 such records in 485,316 subjects. Among these records, 12.32% (1,278,801) were flagged by thresholding QA, and 7.42% (770,578) were flagged by longitudinal QA. The proportion of BMI measures that were obese class III was impacted most by QA methods: 14.54% without QA, 6.69% with thresholding QA, and 11.25% with longitudinal QA. The big differences between thresholding QA and longitudinal QA for extreme obesity illustrates that thresholding QA probably has called many large values as errors while they were not. The proportions in other BMI groups were also affected by QA method but not as greatly.

Table 3 Validation results

	Longitudinal QA (Q_R)		Thresholding QA (Q_S)	
	Height (%)	Weight (%)	Height (%)	Weight (%)
False positive rate (FPR)	1.6	0.3	6.6	12.9
Power	88.9	75.4	17.8	28.0
Positive predictive value (PPV)	75.9	71.4	12.8	2.2
Negative predictive value	99.4	99.7	95.4	99.1

Q_R and Q_s cutoff = 0.05

Table 4 Proportion of subjects who ever had any BMI in the listed BMI classes

BMI class	All data, No QA[a] (N = 10,377,511) (%)	Thresholding QA[b] (N = 9,098,710) (%)	Longitudinal QA[c] (N = 9,606,933) (%)
Underweight (BMI < 18.5)	4.54	1.47	2.24
Normal to overweight (BMI ≥ 18.5 and < 30)	75.56	76.46	73.07
Obese Class I/II/III (BMI ≥ 30 and < 40)	64.36	61.68	61.06
Obese Class III (BMI ≥ 40)	14.54	6.69	11.39

[a] Excluded 0% data

[b] QA excluded 12.3% data

[c] QA excluded 7.4% data

QA methods excluded all records with a Q_S or Q_R score ≤ 0.05

N is the number of usable records included in the BMI classification

Conclusion and discussion

The quality of the EHR data impacts the validity of a study. Therefore, the assessment and the control of the data quality is of the utmost importance in any EHR based research. We have adopted a score-based approach for assessment of the plausibility of the values of continuous variables in EHR data. A quality score is calculated for each observation, and users can select 'good' data using user-defined cutoffs based on the need of the study. This is different from the rule-based method where binary calls are made for data quality and users have no control over the QA process. There is always a trade-off between using a strict cutoff and having less data and using a relaxed cutoff but having more errors in data. Researchers also should be mindful about any bias that could be introduced by leaving out any data from an analysis. When necessary, sensitivity analysis may need to be performed using different cutoffs.

In our validation study, the Q_R score had 89% power with 14% false discovery rate for height, while for weight, only 75% power was achieved with a 28% false discovery rate. The performance of Q_R for weight seems suboptimal although the Q_S performs much worse. The reason can be two folds. One is that there is still space to improve the algorithm. For example, the choice of the tuning parameters does impact the performance of the algorithm, and our algorithm also did not model random measurement errors. The other is that many false positives/negatives are borderline cases that are challenging to identify by either an automated algorithm or manual reviews.

Meanwhile, an "implausible" value flagged by the QA method may not necessarily be untrue. For example, a sudden drop of body weight may be related to the change of health condition that itself is an important signal for EHR data mining, while a sudden drop in height may be less plausible. By controlling the number and proportions of flagged records for weight (or BMI), this algorithm may also serve as an option for data science to search clinical events around these records to further understand if they are indeed related to changes of health conditions or no clinical explanations. Still, the unflagged data can reflect the stable status of subjects in the system.

Furthermore, the proposed method is most applicable to detect outliers in stationary data such as weight and height. For non-stationary data on frequent fluctuations (for example, inpatient blood pressure measurements), more advanced methods are needed to derive useful signals for analysis. The longitudinal QA method can also be applied to routine laboratory data such as lipid panels. However, the laboratory data typically have gone through internal QA from the lab, and the added value of using the longitudinal QA is limited. Simple thresholding method is adequate for detecting outliers in the VA lab data. In summary, the effective QA of the EHR data will require a multitude of methods and approaches that can be adapted to specific database and study need. The longitudinal QA method serves as one of the tools in the method toolbox for producing EHR data of research quality.

An alternative approach for handling outliers is to use analytic models less sensitive to outliers and errors in data such as non-parametric statistics and robust regressions. This approach "embeds" outlier handling in the statistical modeling of clinically interesting parameters and hence requires high customization to individual studies. The DQA method, in contrast, identifies problematic data that can be used for many studies and can provide systematic solutions to quality improvements for EHR data. For example, implausible height and weight values may derive a plausible BMI value that is not recognized by roust statistical methods. Nevertheless, robust methods can always be used in addition to DQA methods to improve the quality of the analysis.

As a final remark, we would like to note that the proposed methods detect data errors ad-hoc. Although EHR data are currently being used retrospectively, a better practice is to implement quality control measures in the data collection stage in EHR systems. Simple rule-based checks and prompts such as confirmation of an unexpected value at the input of the data can be much more effective than ad-hoc remedies and save an enormous amount of downstream cleaning work.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-021-01643-2.

Additional file 1: Supplement Information.

Acknowledgements

This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration. This publication does not represent the views of the Department of Veteran Affairs or the United States Government.

Authors' contributions

HW contributed to study design, data analysis, and manuscript writing. IBL contributed to the validation of the data and results. FW contributed to the data extraction and cleaning of this study. JSL, MCS, and PST provided funding and data support for this manuscript. YL contributed study design, key methods to this research, the validation of the data and results, and the manuscript writing. All authors contributed in critical review and revision of the manuscript.

Funding

This research was supported by award # I01-BX003362 (KM Chang and PS Tsao) and #5I01- HX002487 (JS Lee).

Availability of data and materials

The datasets analyzed in this study are currently available to VHA investigators and other approved partners by the Million Veteran Program. Data may become available to non-VA investigators in the future. Please inquire with the VA Million Veteran Program (https://www.research.va.gov/mvp/) for details and updates on access to data.

Declarations

Ethics approval and consent to participate

This research was conducted under the approved IRB of Palo Alto Cooperative Studies Program Coordinating Center (CSPCC) (Umbrella) (SQL 78831) with Federal-wide Assurance (FWA) number FWA00000929 (VA).

Consent for publication

Not Applicable.

Competing interests

There is no competing interest for this study.

Author details

[1]Department of Veterans Affairs, Cooperative Studies Program Palo Alto Coordinating Center, 701B North Shoreline Blvd, Mountain View, CA 94043, USA. [2]Department of Medicine, Stanford University School of Medicine, 1265 Welch Road, Stanford, CA 94305-5464, USA. [3]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA 94305, USA. [4]Department of Biomedical Data Science, Stanford University School of Medicine, 1265 Welch Road, X359, Stanford, CA 94305-5464, USA.

Received: 21 March 2021 Accepted: 21 September 2021
Published online: 20 October 2021

References

1. Mathur R, Bhaskaran K, Edwards E, et al. Population trends in the 10-year incidence and prevalence of diabetic retinopathy in the UK: a cohort study in the Clinical Practice Research Datalink 2004–2014. *BMJ Open* 2017;7(2):e014444.
2. Liu M, Hinz ERM, Matheny ME, et al. Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. J Am Med Inform Assoc. 2013;20(3):420–6.
3. Moore TJ, Furberg CD. Electronic health data for postmarket surveillance: a vision not realized. Drug Saf. 2015;38(7):601–10.
4. Stapff M, Hilderbrand S. First-line treatment of essential hypertension: a real-world analysis across four antihypertensive treatment classes. J Clin Hypertens. 2019;21(5):627–34.
5. 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. 2016;4(1):1244.
6. Chau I, Le DT, Ott PA, et al. Developing real-world comparators for clinical trials in chemotherapy-refractory patients with gastric cancer or gastroesophageal junction cancer. Gastric Cancer. 2020;23(1):133–41.
7. Suchard MA, Schuemie MJ, Krumholz HM, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet. 2019;394(10211):1816–26.
8. Weiskopf NG, Bakken S, Hripcsak G, et al. A data quality assessment guideline for electronic health record data reuse. EGEMS. 2017;5(1):14.
9. Weiskopf NG, Hripcsak G, Swaminathan S, et al. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. 2013;46(5):830–6.
10. 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. 2016;4(1):1244.
11. Liaw ST, Guo JGN, Ansari S, et al. Quality assessment of real-world data repositories across the data life cycle: a literature review. J Am Med Inform Assoc. 2021;28(7):1591–9.
12. 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(1):144–51.
13. von Lucadou M, Ganslandt T, Prokosch HU, et al. Feasibility analysis of conducting observational studies with the electronic health record. BMC Med Inform Decis Mak. 2019;19(1):202.
14. Callahan TJ, Bauck AE, Bertoch D, et al. A comparison of data quality assessment checks in six data sharing networks. EGEMS. 2017;5(1):8.
15. Wang Z, Talburt JR, Wu N, Dagtas S, Zozus MN. A rule-based data quality assessment system for electronic health record data. Appl Clin Inform. 2020;11(4):622–34.
16. Lynch KE, Deppen SA, DuVall SL, et al. Incrementally transforming electronic medical records into the observational medical outcomes partnership common data model: a multidimensional quality assurance approach. Appl Clin Inform. 2019;10:794–803.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.