Is Data Quality Enough for a Clinical Decision?: Apply Machine Learning and Avoid Bias

Kim Hee
Frankfurt Big Data Laboratory
Goethe University Frankfurt
Frankfurt, Germany
hkim@dbis.cs.uni-frankfurt.de

Abstract—**This paper provides a practical guideline for the assurance and (re-)usage of clinical data. It proposes a process which aims to provide a systematic data quality assurance even without involving a medical domain expert. Especially when (re-)using clinical data, data quality is an important topic because clinical data are not purposely collected. Therefore, data driven conclusions might be false, because a given dataset is not representative. These false data driven conclusions could even harm the life of patients. Thus, all researchers should adhere to some basic principles that can prevent false conclusions. Twelve empirical experiments were conducted in order to prove that my process is able to assure data quality with respect to the descriptive and predictive analysis. Descriptive results obtained by applying stratified sampling are conflicting in four out of nine population inputs. Sampling is carried based on the top ranked feature drawn by the Contextual Data Quality Assurance (CDQA). Between datasets these features are confirmed by the Mutual Data Quality Assurance (MDQA). Stratified sampled inputs improve predictive results compared to raw data. Both Area Under the Curve (AUC) scores and accuracy scores increase by three percent.**

Keywords-**data quality; decision quality; healthcare; clinical data; EHRs; data quality assurance; stratified sampling; bias**

I. Introduction

Data quality refers to the degree of fulfillment of all those requirements defined for data, which is needed for a specific purpose [1]. This definition is aligned to the widely adopted definition of quality: ”Quality is defined through fitness for use” [2]. Those definitions imply two interesting traits of data quality. First, a comprehensible assurance of data quality can be made only after setting up a data mining goal. Second, an appropriate degree of quality is defined based on the data mining goal. Another definition of data quality defined by Jeff Saltz: ”Data quality is being an integrated aspect of an end-to-end process” [3]. It means that data quality is not limited to a single step in a data mining process. For instance, ”If a data product can serve a data mining goal, then data quality is high.” is a true statement, but the converse of that statement ”If data quality is high, then a data product can serve a data mining goal” is not necessarily true. In other words, high data quality is *not sufficient* to meet the data mining goal, but multiple steps in the process need to be addressed. For instance, data cleaning (e.g., improve data completeness, uniqueness, correctness and more), data selection (e.g., selection bias and confounding factors) and modeling (e.g., appropriate algorithms) are the relevant steps.

In the medical domain, data quality is a corner stone for ensuring robust healthcare services. Without data quality, it is hard to make a high quality decision with evidences at the right timing [4]. In fact, a minor defect of data quality can lead to various negative impacts. It is particularly true in the medical domain because bad data quality could harm the life of a patient. For instance, consider there is a clinical application which diagnoses a type of tumor whether it is benign or malignant. A wrong decision, particularly the false negative error (also known as a Type II error), derived by the application could risk the life of a patient, because patient would miss a right timing to get a proper treatment.

With respect to the type of medical data, data quality plays an important role in a clinical data type. Two types of data are available in the medical domain: research data and clinical data. The research data are collected from Randomized Controlled Trials (RCTs), whereas the clinical data are collected for a billing purpose. Clinical data are a history of patients' records also known as Electronic Health Record (EHR) data or Electronic Medical Record (EMR) data. RCTs are the gold standard data in clinical research, but it is resource intensive to conduct RCTs for every disease case and for a certain patient cohort. Furthermore, it is often not ethical to implement RCTs because it requires to discriminate two patient groups receiving a medical treatment. As a consequence, clinical decisions driven by purposefully collected research data are not yet popular among the healthcare providers. Only 10 - 20 % of clinical decisions are made based on the RCTs' results, according to the 2012 Institute of Medicine Committee Report [5]. Recently, observational studies have gained more attention among researchers due to its ease of patient recruitment, while not charging additional costs. Many researcher are trying to reuse existing EHR data to discover a clinical evidence.

However, despite the numerous advantages of reusing EHR data, there are some concerns against reusing clinical data. The main reason is that EHR data are just historical data, not purposely collected to find a clinical evidence. For instance, some researchers argue that one should only use data for the purpose it was collected for [6]. Moreover, numerous observational studies suffer from biases and therefore results may not be the ground truth. Often even two observational studies using the same data sets produce paradoxical results. For example, two independent observational studies [7], [8] produced conflicting conclusions, despite the fact that the two studies analyzed the same database over approximately the same period [9]. Therefore, considering data quality is not an option, but a must before reusing observational data.

In fact, it is ideal to take advice from a domain expert on data quality, as well as intermediate and final results and also on whether the research objective is achieved or not. Unfortunately, it is not easy to motivate a committed domain expert to participate in a recursive and therefore a tedious data mining process. Thus, there is a growing demand for a new approach in data quality assurance that is more objective but and less reliant on input from a domain expert. This paper meets such demand by combining numerous conventional data quality assurance methods and machine learning algorithms. Section 2 provides an overview of a conventional approach and proposes two innovative approaches for data quality assurance. Section 3 is a case study on two EHR datasets on which three data quality assurances are applied. The case is continued in Section 4 which focuses on achieving the better data mining objective through the three steps of quality assurance. In Section 5, a conclusion is drawn and some proposals for future research are proposed.

II. Data Quality Assurance

If something cannot be measured, then it is impossible to be managed. This simple principle also applies toward data quality management. Data quality cannot be managed without data quality assurance. Data quality assurance is a set of activities in the data preparation phase along the data mining process. Toward the systematic data quality management, data quality assurance needs to be quantified and conceptualized as a set of dimensions. Many researchers have investigated and identified various dimensions [10], [11], however there is no clear consensus among researchers. BT Hazen et al. [12] identify four dimensions in common among the following eight studies [13], [14], [15], [16], [17], [18], [19], [20]: Completeness, Accuracy, Consistency and Recency.

Table I. Five Data Quality Dimensions in Three Categories

Category	Dimension	Description
Intrinsic	Completeness	Are all data recorded? No missing value is the ideal condition
	Uniqueness	How is the cardinality in a column? Are there many duplicated records?
	Recency	Are data up-to-date? Can the data mining goal be achieve within the time frame?
Contextual	Relevance	How relevant is each feature to the target feature (data mining goal)?
Mutual	Consistency	How similar is the given dataset to another dataset in terms of the relevance of features?

This paper uses only three primary dimensions except Accuracy, which indicates a degree of the errors in a variable. Accuracy is an essential dimension to measure data quality. However the correct set of values or correct value range, needs to exist or to be predefined by the domain expert. Thus, this dimension is excluded to minimize the need to rely on domain experts. On top of these three dimensions, two additional dimensions are added as shown in Table I. Five dimensions are classified into three groups with respect to the consideration of assurance objective. *Intrinsic Data Quality Assurance (IDQA)* attempts to avoid a source of errors. The objective of IDQA is to ensure the quality of data its own. *Contextual Data Quality Assurance (CDQA)* attempts to avoid selection bias also known as selection effect. It aims to ensure the given data can represent the population to be analyzed. Finally, *Mutual Data Quality Assurance (MDQA)* attempts to confirm the identified CDQA by comparison with another yet similar data.

A. Process

We propose a process which aims to achieve a high degree of data quality (also known as data fitness) for the data mining goal. As shown in Figure 1, the process consists of numerous recursive steps. The essential steps are the three assurance steps: IDQA, CDQA and MDQA and the corresponding case study is described in Section III. The four of symbol # in the *Decision* block denoted as a diamond shape need to be defined in advanced. A decision, which of two paths will take, is made based on the defined value. The case study in this paper sets the four thresholds as follows: *Uniqueness* < 0.3, *Completeness* < 0.7, *Recency* $<$ 10-years and *Kappa score* > 0.61.

B. Intrinsic Data Quality Assurance

IDQA is aligned to the popular concept in the computer science called ”Garbage In, Garbage Out”. The dimensions in this group are independent to the data mining goal, thus native to the data can be measured objectively.

- The dimension of *Completeness* is a degree to the missing data values in a variable or in a table. The metric to measure the completeness is straightforward as follows:

$$Completeness = \frac{\#ofNotNullRecords}{\#ofRecords} * 100$$

It is a ratio between the number of complete items and the total items. For instance, if a variable or a

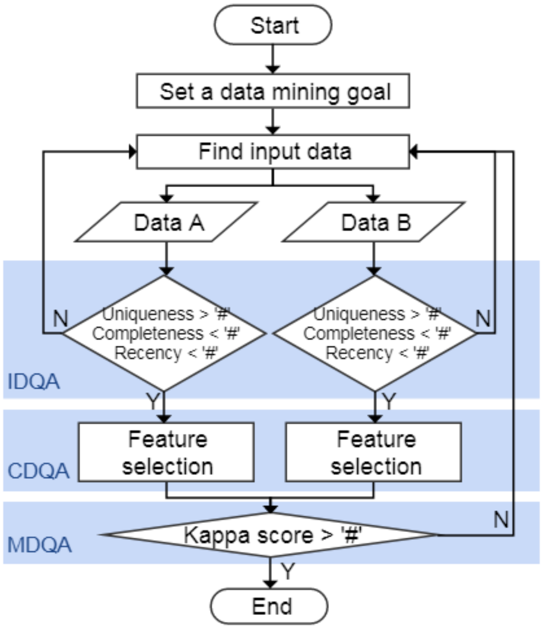

Figure 1. Recursive process of data quality assurance

table is missing 30 percent of values out of 100 of all entries, then the variable or the table achieves 70 percent completeness.

- The dimension of data *Uniqueness* is a column level measurement and it is also known as cardinality. Uniqueness is measured as follows:

$$Uniqueness = \frac{Cardinality}{\#ofRecords} * 100$$

For instance, a variable containing a high cardinality would be a *ID column* which is normally a primary key. In general, high cardinality variables are discarded in predictive modeling if it is not transformed. C Guo et al. [21] and J Moeyersoms et al. [22] introduce an elegant way to transform such variables, but it is out of the scope for this paper.

- The dimension of *Recency* is a degree to how recent data are. It measures how up-to-date the input data are with regard to deliver the data mining goal. According to Bouzeghoub et al. [23], there are two metrics to measure data recency: currency factor [24] and timeliness factor [11]. The currency factor measures the gap between the extraction data and the date of delivery to a data product consumer as follows:

$$Currency = DateofExtraction - DateofDelivery$$

The timeliness factor captures the frequency of the data modification. In this paper, only currency factor is used to measure the *Recency* of data.

C. Contextual Data Quality Assurance

CDQA aims to identify the most relevant variable for the data mining goal. Thus, the result of CDQA varies from the data mining goal unlike IDQA. The identified variable will be later facilitated to conduct a stratified random sampling which is a well-known technique to reduce a bias from data. In order to identify such variable, this study uses *Feature Selection (FS)* methods. FS measures the degree of dependency between every variable and a target variable which is the data mining goal.

FS is a member of dimensionality reduction techniques and it selects a subset of given variables without any reformation. In other words, FS keeps the original values instead of generating a new set of variables using transformation. Thus, it is a suitable methodology in a domain where the data are encouraged to be preserved such as sequence analysis, microarray analysis and single nucleotide polymorphism analysis. In general, there are two taxonomies classifying a subtype of FS techniques.

With an availability of a target variable, FS can be classified either supervised feature selection and unsupervised feature selection [25], which lacks a target variable. Concerning how the model and FS interact, FS can be classified as one of three classes: filter methods, wrapper methods and embedded methods [26]. A number of researchers in the medical domain have applied FS [27], [28], [29] as a preprocessing tool to create a better model, however, there is no one yet who applied FS as a data quality assurance metric to the best of my knowledge.

D. Mutual Data Quality Assurance

In the previous step, the significant features are recognized by FS techniques. However, it is hard to claim it as a ground truth because an observational dataset such as a clinical dataset always has a potential of bias. Perhaps the most significant feature might be valid in the given dataset only. Therefore, the result needs to be cross checked to confirm whether the significant features are still significant in another yet similar dataset. It is carried out in two steps. First, the same procedure of FS is applied to another EHR dataset. Then the level of CDQA inter-agreement between two results is determined based on the result of the Quadratically Weighted Kappa (QWK). QWK is a type of weighted kappa and [30] which is defined as:

$$QWK = \frac{1}{k}\sum_{i=1}^{k}\left(1 - \frac{(a_i - b_i)^2}{(k-1)^2}\right)$$

where k is the total number of elements and a_i and b_i are each element from two comparable lists. It is a tool to estimate inter-agreement between two raters and the output score ranges between zero to one. If the output score approaches to one, it means that the two CDQA results are more identical. QWK is best applied on ratings on an ordinal

Table II. Interpretation of Kappa Score Ranging Between Zero And One. The Closer to One, the More Identical the Two Inputs

Kappa score	Interpretation of the inter-agreement
0	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

scale. Table II[31] and it illustrates the association between kappa score and the corresponding interpretation. It is arbitrary, but it provides a reasonable scale for benchmarks.

E. Tools

Instead of using commercial tools, this paper utilizes only open source libraries. In particular, three Python libraries are facilitated for three steps of data quality assurance. For the IDQA, we use *pandas-profiling* [1]. It generates a comprehensive and user-friendly statistical summary. For the CDQA, we apply *SciKit-Learn Laboratory* [2] and *scikit-learn (sklearn)* [3]. SciKit-Learn Laboratory contains numerous evaluation metrics and QWK metric is used to measure the inter-agreement. *sklearn* contains various machine learning algorithms for FS, stratified sampling, modeling and model evaluation.

In fact, *pandas-profiling* at the step of IDQA can be replaced by any tool. A large number of data cleaning tools are publicly available: *Talend* [4], *TopNotch* [5], *PDQ Tracker* [6] and *DataCleaner* [7]. Additionally, some tools are supporting distributed data systems, but it is out of scope in this paper: *Griffin* [8], *drunken-data-quality* [9] and *Data Quality for BigData* [10].

III. Case Study Part I: Data Quality Assurance

The first form of Electronic Health Record (EHR) data were collected in the 1960s [32] for a billing purpose. EHR data these days need to be kept for a sufficient length of time for compliance with laws and regulations. For instance, the minimum retention periods of EHR data in the United Kingdom is 30 years, while those in Sweden need to be kept permanently [33]. Thanks to the open data initiative, several EHR data sets are publicly accessible. For example, Medical Information Mart for Intensive Care (MIMIC) database [11], the National Comorbidity Survey (NCS-1) [12] and the Patient-Centered Outcomes Research Institute (PCORI) [13] consist of a large amount of data.

A. Datasets

This study demonstrated the data quality assurance using MIMIC-III datasets. It is freely accessible datasets comprising of 40k de-identified patients admitted to intensive critical care units (ICU) in the Beth Israel Deaconess Medical Center in Boston, Massachusetts. MIMIC contains various datasets including medical history, demographics, medication, laboratory results, care notes from care givers, radiology images and more. We prepared an input dataset comprising patients' demographic information and a target variable which consists of two possible states, whether or not a patient died in the hospital. All demographic variables are categorical except *Age* which is a continuous variable. Since MIMIC is de-identified, *Age* variable is not given in the MIMIC datasets. Thus, the age for each patient is derived by difference between their first admission date and the date of their birth.

B. Data Mining Goal

The data mining objective of this case study is a mortality prediction among adult patient populations who are 20-year-old or above at the date of their first admission. Beside of the mortality prediction, there are several data mining applications available in ICU. For instance, *Daily Score* indicates a sequence of patient's statuses over time, *Glasgow Coma Scale* offers a status for the central nervous system, *Length of stay* predicts a duration of inpatient days and *Mortality Prediction* provides a potential mortality.

C. Intrinsic Data Quality Assurance

The IDQA of MIMIC dataset is automatically generated by pandas-profiling library in Python as shown in Table III. The result of *Completeness* shows that there is no missing value in MIMIC data. In the case of a high missing values, a statistical technique called *imputation* can be applied to replace missing values with substituted values. The result of *Uniqueness* indicates that *diagnosis* variable has a high cardinality. It means that there are variety types of diseases in MIMIC data. The temporal coverage of MIMIC is between 2001 and 2012 indicated by the minimum date value and the maximum date value of the admission variable. Thus, the currency factor (*Recency*) is 5 years by subtraction latest year (2012) from the current year (2017).

[1] https://github.com/jospolfliet/pandas-profiling

[2] scikit-learn-laboratory.readthedocs.io

[3] https://github.com/scikit-learn/scikit-learn

[4] https://github.com/Talend/data-quality

[5] https://github.com/blackrock/TopNotch

[6] https://github.com/GridProtectionAlliance/pdqtracker

[7] https://github.com/datacleaner/DataCleaner

[8] https://github.com/apache/incubator-griffin

[9] https://github.com/FRosner/drunken-data-quality

[10] https://github.com/agile-lab-dev/DataQuality

[11] http://mimic.physionet.org

[12] http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/6693

[13] https://www.pcori.org

Table III. An Overview of Intrinsic Data Quality Assurance: Completeness (Missing) and Correctness (Unique)

Variable	**Variable type**	**Missing (%)**	**Uniqueness (%)**
age	Numeric	0.0	0.2
gender	Categorical	0.0	0.0
ethnicity	Categorical	0.0	0.1
admission type	Categorical	0.0	0.0
diagnosis	Categorical	0.0	26.0
discharge location	Categorical	0.0	0.0

D. *Contextual Data Quality Assurance*

In the previous step, MIMIC data comply with three dimensions of IDQA very well. However, it is worth to note that good IDQA does not lead to the better CDQA. To comply with CDQA, data must be relevant to the data mining goal, which is the mortality prediction in use case. The results of four FS methods are illustrated in Table IV. It shows the relevancy level of each feature to the target variable, *mortality*. According to the result, *age* feature is the most significant feature and the second most significant feature turns out *admission type* feature comprising of three categorical values: Emergency, Urgent and Elective.

E. *Mutual Data Quality Assurance*

In the previous step, the *age* feature was identified as the most significant feature, but this may not be the case in another dataset. In this section, MDQA is demonstrated using another yet similar clinical dataset from Geisinger Health System. Geisinger Clinical Data (GCD) contain 171k of de-identified patients who suffered or are still suffering from a mood disorder. GCD contains also various datasets including medical history, demographics, medication, laboratory results and more. Likewise for MIMIC data, GCD are prepared with the same demographic variables and the target variable as MIMIC input data. Table V represent the relevancy level of each feature. To my surprise, the most significant feature remains the same as before. In general, the order of feature relevancy remains similar since the QWK result turns out *0.699*. According to the scale in Table II, the result is in the substantial agreement range, which indicates it is robust and trustful.

Table IV. MIMIC Data Feature Ranks by Four Feature Selection Methods. The Last Column Is The Averaged Ranks

Variable	**Uni-variate**	**Mutual info.**	**Chi square**	**R. forest**	**Avg**
age	**1**	2	**1**	2	**1**
gender	6	6	6	6	6
ethnicity	3	5	5	5	5
admission type	2	**1**	4	**1**	2
diagnosis	5	3	2	3	3
discharge location	4	4	3	4	4

Table V. Geisinger Data Feature Ranks by Four Feature Selection Methods. The Last Column Is The Averaged Ranks

Variable	**Uni-variate**	**Mutual info.**	**Chi square**	**R. forest**	**Avg**
age	**1**	**1**	**1**	**1**	**1**
gender	6	6	4	6	6
ethnicity	2	5	6	5	5
admission type	4	4	5	4	4
diagnosis	3	2	3	2	2
discharge location	5	3	2	3	3

IV. Case Study Part II: Evaluation

This section continues the case study from the previous section. The process of data quality assurance has been completed, but does the better data quality lead a better analysis? How does the better data quality impact on the final decision? This study evaluates these questions in two separated evaluations as shown in Figure 2. Prior to the evaluations, two data inputs need to be prepared: One input is raw MIMIC dataset as a base line, which consists of adult patients older than 19 years of age in ICUs. Another dataset is subset of the raw dataset stratified by *age* which is the most significant feature identified by CDQA and then confirmed by MDQA. Stratified sampling [34] partitions a population (an input dataset) equally into a set of subpopulations called strata. Then, the same number of records are randomly taken from each stratum. Both datasets consist of *mortality status* (the target variable) and eight demographic variables including age, gender, marital status, insurance type, ethnicity, religion, admission type and ICD-9 disease category.

A. *Mortality Rate*

Mortality rate is a descriptive analysis, also known as crude death rate. It is an estimation of the portion of a population that dies during a specified period [35]. Note that the ratio units are of the deaths per 1,000 records, instead of 100 records. Twelve populations of different interest are decomposed into seven strata of the ten-year age stratum (e.g. 20-29 years, 30-39 years and up to 80-89 years). Figure 3 illustrates twelve corresponding distributions of an interest including: all patients, alive patients, deceased patients and nine major International Classification of Diseases (ICD-9) categories as follows:

- (001-139): infectious diseases
- (140-239): neoplasms
- (240-279): endocrine and immunity disorders
- (280-289): blood disease
- (290-319): mental disorders
- (390-459): circulatory diseases
- (460-519): respiratory diseases
- (520-579): digestive diseases
- (800-999): injury

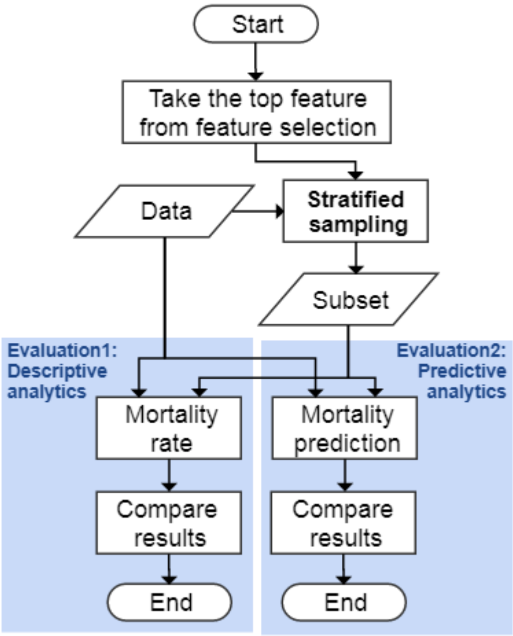
Start
Take the top feature from feature selection
Data
Stratified sampling
Subset
Evaluation1: Descriptive analytics
Evaluation2: Predictive analytics
Mortality rate
Mortality prediction
Compare results
Compare results
End
End

Figure 2. An evaluation process for two analyses: mortality rate and mortality prediction

1,705 records are randomly taken from each stratum because the smallest stratum, the age range 20-29, consists of 1,705 records. The mortality rate comparisons are shown in Figure 4 and it is noteworthy to see the four populations in green color. They may draw two conflicting conclusions before and after applying the sampling method. For instance, 1.9 out of 1,000 patients die if they suffer from the mental disorders (290-319) before applying the sampling method. It is two times lower than the mortality rate of all patients which is conflicting with the result after applying the sampling. The distribution patterns of those four in Figure 3 are also peculiar except the circulatory diseases patients (390-459).

B. Mortality Prediction

Beyond the descriptive data analysis, this section attempts to predict the risk of death for adult patients. Twelve binary classifiers are created based on the combination of two input data and six machine learning algorithms or estimators. For the model evaluation, k-fold cross validation is facilitated. It partitions the given input dataset into k subsets, then it averages of the k times evaluation scores. Two metrics, Area Under the Curve (AUC) and Accuracy, measure how good the classifier is. An AUC score of 1 means a perfect classifier, whereas AUC score of 0.5 indicates a random guess. An Accuracy of 1 indicates a perfect accuracy, whereas an Accuracy of 0 is same as a random guess. The corresponding prediction performances are shown in Table VI.

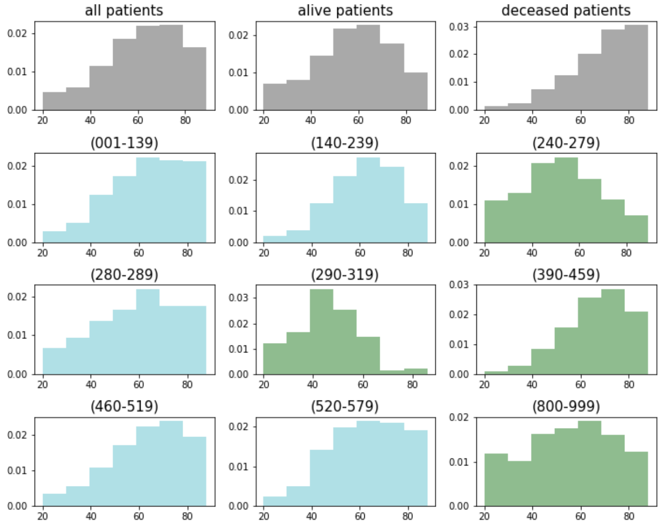

Figure 3. Histograms of twelve populations in MIMIC data. (Y-axis) is mortality rate and (X-axis) is age

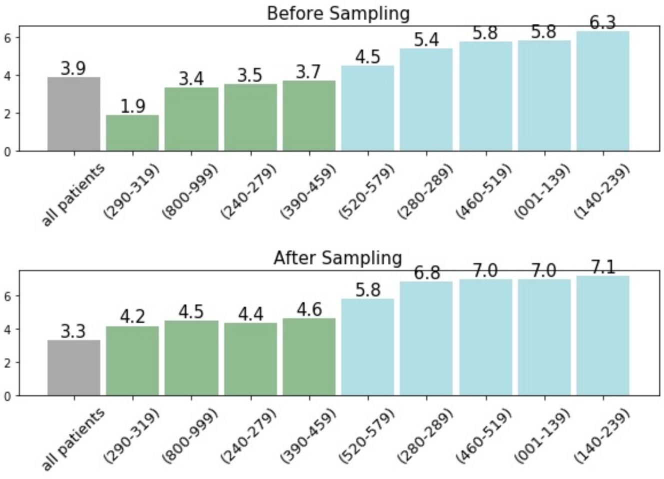

Figure 4. Mortality rate (Y-axis) comparison in MIMIC between raw data and stratified sampled data. (X-axis) is populations of all patients and nine major disease categories

Table VI. Mortality Prediction Comparison between Raw Data and Stratified Sampled Data in MIMIC. Decision Tree (DT), Extra Trees (ET), Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB) and Calibrated Classifier (CC)

Classifier	All patients		Sampled patients	
	AUC	Accuracy	AUC	Accuracy
DT	0.74 ± 0.06	0.72 ± 0.06	0.76 ± 0.06	0.73 ± 0.06
ET	0.77 ± 0.06	0.71 ± 0.06	0.81 ± 0.06	0.76 ± 0.06
RF	0.76 ± 0.06	0.70 ± 0.06	0.80 ± 0.06	0.75 ± 0.06
LR	0.70 ± 0.06	0.65 ± 0.06	0.72 ± 0.06	0.69 ± 0.06
GB	0.77 ± 0.06	0.70 ± 0.06	0.79 ± 0.06	0.73 ± 0.06
CC	0.70 ± 0.06	0.64 ± 0.06	0.71 ± 0.06	0.66 ± 0.06
Average	0.74 ± 0.06	0.69 ± 0.06	0.77 ± 0.05	0.72 ± 0.05

Algorithms with the stratified sampled input are able to improve the prediction. It increases upon the baseline data, all patients, by predicting three percent better for both AUC and Accuracy. Also the standard deviation declines by one percent.

V. Conclusion

Data quality is a corner stone for assuring a high quality of decision making across domains. In particular, it is a critical topic in the context of reusing EHR data, because a false data driven conclusion can harm the life of a human subject. Therefore, it might be dangerous to draw a medical conclusion without conducting data quality assurances because it can minimize the false conclusions. The contributions of this paper are as follows:

- Five conventional data quality dimensions are classified into three categories as: Intrinsic Data Quality Assurance (IDQA), Contextual Data Quality Assurance (CDQA) and Mutual Data Quality Assurance (MDQA). Each category is mutually exclusive, meaning that a high score of IDQA does not imply a high score of MDQA.
- A multi-layered process combining the three categories was proposed and the proposed process has been evaluated compared to the naive approaching. The evaluation result demonstrated that the proposed process is able to draw a more robust conclusion without involving any assumption or hypothesis.

One may say that standardized mortality ratio [36] may carry similar result, but the process proposed in this paper does not need to involve a medical domain expert. This can decrease the complexity and save numerous human resources.

For the future work, one can consider to conceptualize *Velocity* as a new data quality dimension. For instance, Which metrics can be applied to measure the speed of decision and how can we determine the quality or impact of the real-time decision? A real-time clinical decision supportive system could save more lives in certain domains like the emergency department, thus this topic needs to be addressed by more researchers.

Acknowledgment

The author acknowledges Roberto Zicari, Todor Ivanov, Karsten Tolle, Florian Reifschneider, Alex Klein, Jannik Schaaf and Naveed Mushtaq at the Frankfurt Big Data Lab for support. The author also thanks Paul Gortner for his helpful comments and valuable suggestions.

References

[1] G. Morbey, *Data Quality in General*. Wiesbaden: Springer Fachmedien Wiesbaden, 2013, pp. 3–13. [Online]. Available: https://doi.org/10.1007/978-3-658-01823-8_1

[2] J. M. Juran and F. M. Gryna, "Juran s quality control handbook, 4ta," *Edition, pag. AII*, vol. 3, 1988.

[3] R. Zicari, "Q&A with Data Scientists Jeff Saltz," http://www.odbms.org/2017/08/qa-with-data-scientists-jeff-saltz, 2017, [Online; accessed 10-October-2017].

[4] L. A. Celi, M. Csete, and D. Stone, "Optimal data systems: the future of clinical predictions and decision support," *Current opinion in critical care*, vol. 20, no. 5, p. 573, 2014.

[5] J. M. McGinnis, L. Stuckhardt, R. Saunders, M. Smith *et al.*, *Best care at lower cost: the path to continuously learning health care in America*. National Academies Press, 2013.

[6] J. van der Lei *et al.*, "Use and abuse of computer-stored medical records." *Methods Archive*, vol. 30, pp. 79–80, 1991.

[7] J. Green, G. Czanner, G. Reeves, J. Watson, L. Wise, and V. Beral, "Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a uk primary care cohort," *Bmj*, vol. 341, p. c4444, 2010.

[8] C. R. Cardwell, C. C. Abnet, M. M. Cantwell, and L. J. Murray, "Exposure to oral bisphosphonates and risk of esophageal cancer," *Jama*, vol. 304, no. 6, pp. 657–663, 2010.

[9] M. J. Schuemie, P. B. Ryan, W. DuMouchel, M. A. Suchard, and D. Madigan, "Interpreting observational studies: why empirical calibration is needed to correct p-values," *Statistics in medicine*, vol. 33, no. 2, pp. 209–218, 2014.

[10] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.

[11] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, vol. 12, no. 4, pp. 5–33, 1996.

[12] B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications," *International Journal of Production Economics*, vol. 154, pp. 72–80, 2014.

[13] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "Aimq: a methodology for information quality assessment," *Information & management*, vol. 40, no. 2, pp. 133–146, 2002.

[14] D. P. Ballou and H. L. Pazer, "Modeling data and process quality in multi-input, multi-output information systems," *Management science*, vol. 31, no. 2, pp. 150–162, 1985.

[15] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 16, 2009.

[16] R. Blake and P. Mangiameli, "The effects and interactions of data quality and problem complexity on classification," *Journal of Data and Information Quality (JDIQ)*, vol. 2, no. 2, p. 8, 2011.

[17] A. Haug and J. Stentoft Arlbjørn, "Barriers to master data quality," *Journal of Enterprise Information Management*, vol. 24, no. 3, pp. 288–303, 2011.

[18] J. L. Malin, K. L. Kahn, J. Adams, L. Kwan, M. Laouri, and P. A. Ganz, "Validity of cancer registry data for measuring the quality of breast cancer care," *Journal of the National Cancer Institute*, vol. 94, no. 11, pp. 835–844, 2002.

[19] A. Parssian, "Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions," *Decision Support Systems*, vol. 42, no. 3, pp. 1494–1502, 2006.

[20] M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, and C. Batini, "Managing data quality in cooperative information systems," *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pp. 486–502, 2002.

[21] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," *arXiv preprint arXiv:1604.06737*, 2016.

[22] J. Moeyersoms and D. Martens, "Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector," *Decision Support Systems*, vol. 72, pp. 72–81, 2015.

[23] M. Bouzeghoub, "A framework for analysis of data freshness," in *Proceedings of the 2004 international workshop on Information quality in information systems*. ACM, 2004, pp. 59–67.

[24] A. Segev and W. Fang, "Currency-based updates to distributed materialized views," in *Data Engineering, 1990. Proceedings. Sixth International Conference on*. IEEE, 1990, pp. 512–520.

[25] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature Selection: A Data Perspective," *arXiv:1601.07996 [cs]*, Jan. 2016, arXiv: 1601.07996. [Online]. Available: http://arxiv.org/abs/1601.07996

[26] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[27] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[28] H. Liu and H. Motoda, *Feature extraction, construction and selection: A data mining perspective*. Springer Science & Business Media, 1998, vol. 453.

[29] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 2, pp. 44–49, 1998.

[30] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and psychological measurement*, vol. 33, no. 3, pp. 613–619, 1973.

[31] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[32] M. C. Data, *Secondary Analysis of Electronic Health Records*. Springer, 2016.

[33] A. Stanković and H. Stančić, "Development of health care e-services in the european union," in *INFuture2015: e-Institutions-Openness, Accessibility, and Preservation*, 2015.

[34] R. M. Groves, "Research on survey data quality," *The Public Opinion Quarterly*, vol. 51, pp. S156–S172, 1987.

[35] M. Porta, *A dictionary of epidemiology*. Oxford University Press, 2014.

[36] A. J. McMichael, "Standardized mortality ratios and the'healthy worker effect': Scratching beneath the surface." *Journal of Occupational and Environmental Medicine*, vol. 18, no. 3, pp. 165–168, 1976.