EISSN: 2707-0425

*East African Journal of Science, Technology and Innovation, Vol. 2 (2): March 2021.*



# Suitability of Electronic Health Record Data for Computational Phenotyping of Diabetes Mellitus at Nairobi Hospital, Nairobi City County, Kenya

[1*]OLWENDO A., [2]OTIENO G., [2]RUCHA K.

[1]Department of Health Information Management, Jomo Kenyatta University of Agriculture & Technology P.O. BOX 62000-00200, Nairobi, Kenya.
[2]Department of Health Management and Informatics, Kenyatta University P.O. Box 43844 00100 Nairobi
***Corresponding author**: aolwend@gmail.com

**Abstract**

This research aims to determine the applicability of routine healthcare in clinical informatics research. One of the key areas of research in precision medicine is computational phenotyping from longitudinal Electronic Health Record (EHR) data. The objective of this research was to determine how the interplay of EHR software design, the use of a data dictionary, the process of data collection, and the training and motivation of the human resource involved in the collection and entry of data into the EHR affect the quality of EHR data thus the suitability of such data for utility in computational phenotyping of diabetes mellitus. This research employed a prospective/retrospective study design at the diabetes clinic in Nairobi Hospital. The first source of data was from interviews with 32 staffs; nurses, doctors, and health record officers using a referenced peer-reviewed usability questionnaire. Thereafter, a sample of EHR data collected during routine care between January 2012 and December 2016 was also analyzed by looking into the quality of clusters identified in the data using a density-based clustering algorithm and Statistical Package for Social Sciences (SPSS) version 21. Regression analysis shows that software design and the utility of a data dictionary explained 50.7% and 32.3% respectively in the improvement of the suitability of EHR data for computational phenotyping of diabetes mellitus. Also, EHR software was rated useful (82%) in accomplishing users' daily tasks. However, EHR data were found to be unsuitable for utility in computational phenotyping of diabetes. Despite the fact that 88% of EHR data were clustered as noise, the clustering algorithm identified a total of 23 clusters from the diabetes dataset. However, with improved quality of EHR data, sub-phenotyping tasks would be achievable. This research concludes that the poor quality of EHR data are as a result of employees' unmet intrinsic factors of motivation.

**Keywords**: *EHR; Computational Phenotyping; Data Quality; Density-based Clustering; Usability*

**Cite as:** Olwendo *et al., (2021).* Suitability of Electronic Health Record Data for Computational Phenotyping of Diabetes Mellitus at Nairobi Hospital, Nairobi City County, Kenya. *East African Journal of Science, Technology and Innovation 2(2).*

Received: 21/09/20
Accepted: 03/12/20
Published: 25/03/21

**Introduction**

Precision Medicine (PM) initiative aims at improving healthcare through the utilization of personal healthcare data. One of the main goals of PM is the discovery of disease sub-phenotypes defined by their distinct molecular. Moreover, critical disease subtypes distinctions may be impacted by nonmolecular factors such as socioeconomic status. One of the key areas of research in PM is computational phenotyping which is the identification and extraction of useful clinical representations from longitudinal EHR data (Tenenbaum & Avillach, 2016; Tutty et al., 2019; Yadav et al., 2017).

The use of technology in healthcare has undoubtedly grown over the years in various departments of the hospital setting. Furthermore, the collection and storage of medical records in their electronic versions has also grown from the use of various non-standardized Electronic Medical Record (EMR) to standardized Electronic Health Records (EHR) over time. (Adler-Milstein et al., 2017; van der Bij et al., 2017). The use of EHR in managing healthcare data has led to an explosion of electronic patient records collected during routine care and stored over time. Such data could be put to secondary use such as in clinical informatics research. Secondary use of EHR data has a great potential for solving a number of problems experienced in medicine. Also, the use of EHR data to conduct retrospective study designs would undoubtedly reduce research costs and promote patient-centered research (Yadav et al., 2017).

However, EHR data are characterized by; inconsistencies, missing or incomplete observations, presence of noise and outliers Consequently, without appropriate interventions, EHR data may not be "fit for use" in clinical informatics research (Kahn et al., 2016; Reimer et al., 2016; Yadav et al., 2017). Therefore, data-driven research using EHR data can only be beneficial in research if quality measures are put in place to ensure the validity and reliability of such data (Fox et al., 2018b; Verheij, Curcin, Delaney, & McGilchrist, 2018). A number of healthcare institutions have come up with data quality assurance teams that work with healthcare workers that generated such data to ensure that outliers, noise, inconsistencies and incomplete observations in data are minimized. However, the presence of noise and outliers in data is usually less addressed. Also, handling incompleteness and inconsistencies in data also results into the generation of at least some noise whose effect may range from insignificant to very serious. Noise is a random error or variance in a measured variable. On the other hand, an outlier in an object that significantly deviates from the normal objects as if it were generated by a different mechanism.

One of the determinants of the quality of EHR data is the design of the EHR software itself. Software design may make the software 'easy to use' or complicated hence affect data quality. Software that is made 'easy to learn' lessens the user's cognitive requirements hence reduced mental load. Furthermore, ease of use also adds onto making software suitable for its intended users. Moreover, well developed user interfaces need to include opportunities for error detection and/or correction. Functionalities within a given EHR system would affect the quality of its data. These include but not limited to; successful task completion, difficulty in completion of tasks in a proper sequence, number of errors identified/detected and/or corrected, and risks according to the users that arise due to confusion or misunderstanding when carrying out directed scenarios of use. In countries like the Netherlands, all EHR software are certified hence standardized. However, this is not the case in developing countries such as Kenya where there are varieties of EHR software used in hospitals across the country without any quality control mechanisms in place (Farrell et al., 2017; Keny et al., 2015; van Engen-Verheul et al., 2016; Verheij, Curcin, Delaney, & McGilchrist, 2018).

The second determinant of the quality of EHRR data collected during routine care is utility of a data dictionary during data entry. A data dictionary provides a comprehensive definition of all data elements in an EHR hence increased data quality. As a result, it is recommended that all EHR should come with an inbuilt data dictionary which is an outline of data design and management and semantic interoperability and collaboration (Keny et al., 2015; Verheij, Curcin, Delaney, & McGilchrist, 2018). However, a

number of health care institutions in Kenya have acquired varieties of EHR systems. Therefore, in the effort to create a pool of common terminology, the Ministry of Health constituted a Concept Dictionary Working Group in 2014 tasked to come up with a solution for Kenya to enhance semantic interoperability (Keny et al., 2015).

The human resource responsible for collection and entry of data from patients into the EHR should be equipped with the right knowledge and skills regarding the use of the EHR. Users of EHR should be provided with adequate opportunities for training to be able to comfortably use the EHR software to accomplish their tasks. Trainings should be subject-oriented and designed to meet the needs of both expert and novices in the use of computers. Moreover, EHR users need to be motivated to perform their duties and not simply work to keep their jobs. Employee motivation is key to performance. Also, it's good to note that employee workload definitely affects the quality of their work output. (Longhurst et al., 2019; Lopez et al., 2018; Verheij, Curcin, Delaney, & McGilchrist, 2018).

Finally, the data collection process encompasses all the activities and tools used for data collection and entry into the EHR software. For example, in healthcare, patients may have to fill some forms and the forms further forwarded to the nurses/doctors as part of evaluation. Thereafter, the content of this form later becomes a component of the patient file or a source of the specific patient health records. Given the heavy workload as experienced in several healthcare settings, inadequate time dedicated to an activity may result into poor data quality. Furthermore, healthcare workers are usually collaborative in their work activities. As a result, an EHR should not only provide opportunities for data entry but also teamwork and collaboration amongst various users (Farrell et al., 2017; Verheij, Curcin, Delaney, & McGilchrist, 2018).

In unsupervised learning, the clustering function determines important properties about the distribution of the data input x. After the construction of the model, the target function is supplied with the input x and it's upon it to determine the output y. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is the density-based clustering algorithm and it works best for the case of data with much noise (Ester et al., 1996; Shickel et al., 2017). DBSCAN has much applicability in computational phenotyping since it learns latent relationships of arbitrary shapes from raw data without human intervention. In computational phenotyping, this would involve identifying disease phenotypes and sub-phenotypes based on healthcare data for adequate management of patient cases. Computational phenotyping tasks include; discovering and stratifying new disease sub-phenotypes and; discovering specific phenotypes for improving classification under existing disease boundaries and definitions. This research was limited to the development of an unsupervised learning model from the sampled data to determine the model's ability to discover and stratify diabetes mellitus cases to help in improving classifications under existing boundaries (Che & Liu, 2017; Denaxas et al., 2017; Ghosh et al., 2016; Richesson et al., 2014; Tenenbaum & Avillach, 2016; Yadav et al., 2017).

**Materials and methods**

This research employed a prospective /retrospective study design conducted at the diabetes clinic in Nairobi Hospital located in Nairobi City County, Kenya. The first source of data was from interviews with the healthcare staff involved in the collection and entry of data into the EHR. The study conducted a census and interviewed all the staff at the clinic that totalled to 32 and comprised of nurses, doctors, and health record officers working at the clinic that handle clients with cases of diabetes mellitus. Interviews were conducted using a referenced peer-reviewed usability questionnaire with each interview lasting 15 minutes.

On the other hand, the second dataset was obtained from confirmed cases of diabetes mellitus collected during routine clinical care between January 2012 and December 2016. A sample of 652 records was obtained through stratified sampling considering both gender and age. The EHR dataset was subjected to pre-processing which included data cleaning, resolving cases of incompleteness in records,

discovery and handling of outliers that may be in the dataset. Finally, data were subjected to transformation and smoothing for the removal of noise. Descriptive statistics and regression analysis were conducted using SPSS version 21. Linkert scaled results were summarized using the formular

% = 100 * ∑ (respondent scores * weights) / (Total number of respondents * Highest possible score on the scale * Number of questions).

Alternatively, the EHR dataset was subjected to DBSCAN for cluster analysis.

## Results

### *Characteristics of Participants*

The participants in this study comprised of 81% (26/32) females of whom 69% (22/32) were nurses and 62% (20/32) of the participants had worked at the same clinic for at least 5 years. The other of the details are as summarized in Table 1.

*Table 1: Characteristics of Participants.*

| Characteristic | Category | Frequency (n = 32) | Percentage (%) |
|---|---|---|---|
| Gender | Male | 6 | 19% |
| | Female | 26 | 81% |
| Specialization | Nurse | 22 | 69% |
| | Physician | 4 | 13% |
| | HRIO | 4 | 13% |
| | Key Informant | 2 | 5% |
| Number of Years working at the same clinic | Less than 5 Years | 12 | 38% |
| | 5 – 10 Years | 12 | 38% |
| | 11 – 15 Years | 7 | 22% |
| | More than 15 Years | 1 | 2% |

### *Effect of the EHR software design on the suitability of EHR data for computational phenotyping of diabetes mellitus*

The regression analysis showed that the adjusted $r^2$ was 0.507 thus, the independent variable software design explained 50.7% of the improvement in the suitability of EHR data for computational phenotyping of diabetes. The sections below summarize results from the evaluation of the usability of the EHR software design based on the measure of its user's perceived usefulness, ease of use, and ease of learning.

### *Participants' perceived usefulness of the EHR software*

The participants in this study 82% (26/32) agreed that the EHR software is useful to them in accomplishing their daily tasks. Moreover, participants 100% (32/32) reported that the EHR software makes them more productive. On the other hand, 56% (18/32) of participants denied that the EHR provides them with all the features that they need to perform their duties. Also, 75% (24/32) of participants denied that the EHR software provides them with good error messages whenever they make errors. The other of the details are as summarized in Table 2.

### *Participants' perceived ease of use of the EHR software*

The participants in this study 94% reported that he EHR software was easy to use. Also, participants 100% (32/32) reported that they can use the EHR without instructions and that both regular and occasional users of the EHR would like to use the software. Other details are as summarized in Table 3.

***Participants' perceived ease of learning to use the EHR software***

The participants 99% perceived the EHR software as easy to learn to use. Participants 91% (29/32) quickly learnt to use the EHR. Also, 100% (32/32) of participants reported that they easily remembered how to use the EHR and quickly became skilful at using the EHR as summarized in Table 4.

***Effect of the use of a Data Dictionary on the suitability of EHR data for computational phenotyping of diabetes mellitus***

The regression analysis showed that the adjusted $r^2$ was 0.323 meaning that the independent variable, use of a data dictionary explained 32.3% of the improvement in the suitability of EHR data for computational phenotyping of diabetes. Also, the participants 100% (32/32) reported that the EHR had an inbuilt data dictionary in addition to other software tools that enabled them to accomplish their tasks easily.

.

*Table 2: Participants' evaluation of the usefulness of the HER*

| Index | Subject | Agree | Disagree |
|---|---|---|---|
| 1 | It helps me to be more effective | 100% | 0% |
| 2 | It helps me to be more productive | 100% | 0% |
| 3 | It is useful | 100% | 0% |
| 4 | It makes accomplishing tasks easier | 100% | 0% |
| 5 | It saves me time when I use it | 100% | 0% |
| 6 | It meets my needs | 100% | 0% |
| 7 | It provides all features I need to perform my duties | 44% | 56% |
| 8 | It minimizes user memory load | 100% | 0% |
| 9 | It provides properly marked shortcuts | 100% | 0% |
| 10 | It provides feedback | 75% | 25% |
| 11 | It provides good error messages in case I make an error | 25% | 75% |
| 12 | It prevents errors from occurring | 19% | 81% |
| 13 | It provides help and documentation | 100% | 0% |

*Table 3: Participants' evaluation of the Ease of Using the EHR.*

| Index | Subject | Agree | Disagree |
|---|---|---|---|
| 1 | It is easy to use | 91% | 9% |
| 2 | It is simple to use | 91% | 9% |
| 3 | I require fewer steps possible to accomplish tasks | 92% | 8% |
| 4 | Using it is effortless | 81% | 19% |
| 5 | I can use it without written instructions | 100% | 0% |
| 6 | I don't notice any inconsistencies as I use it | 81% | 19% |
| 7 | Both occasional and regular users would like it | 100% | 0% |
| 8 | I can recover my mistakes quickly and easily | 81% | 19% |
| 9 | I can use it successfully every time | 81% | 19% |

*Table 4: Participants' evaluation of Ease of learning to use the EHR.*

| Index | Subject | Agree | Disagree |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 1 | I learnt to use it quickly | 91% | 9% |
| 2 | I easily remember how to use it | 100% | 0% |
| 3 | It is easy to learn to use | 100% | 0% |
| 4 | I quickly become skillful with it | 100% | 0% |

***Effect of the Human Resource on the suitability of EHR data for computational phenotyping of diabetes mellitus***

The regression analysis showed that the adjusted $r^2$ was 0.166 indicating that the independent variable, human resource, explained 16.6% of the improvement in the suitability of EHR data for the computational phenotyping of diabetes. Moreover, results from participant evaluation of the provisions for training and motivation for the continued utility of the EHR are in the sections below.

***Participants' perceived provision of adequate opportunities for training to use the EHR software***

The participants 93% (29/32) reported that they had been provided with adequate opportunities for training to be able to effectively use the EHR. The participants also reported that they had other support tools necessary for them to accomplish their tasks effectively.

***Participants' motivation for continued utility of the EHR software***

The participants 90% (28/32) reported that they were motivated to continue utilizing the EHR software since they believed it made their work much easier. Table 3 shows a summary of participants' motivation with the use of the EHR.

***Effect of the process of data collection on the suitability of EHR data for computational phenotyping of diabetes mellitus***

The regression analysis results showed that the adjusted $r^2$ was 0.226 indicating that the intermediate variable process of data collection explained 22.6% of the improvement in the suitability of EHR data for computational phenotyping of diabetes. Also, the participants 98% (31/32) believed that their data collection process, contributed to the collection of good quality.

***Description of the EHR Dataset***

A total of 653 records with confirmed cases of diabetes mellitus were extracted. The attributes of the data comprised of; - Age, Gender, Weight, Height, BMI (Body Mass Index), BSA, Pulse, Systolic, Diastolic, RBS, $SPO_2$ (Oxygen saturation), Temperature, and Respiration. The diagnoses of diabetes mellitus (DM) present in the data included gestational, Prediabetes, Type 1 DM, and Type 2 DM. Co-morbidities identified in the data were variations of Hypertension. The other details are summarized in Table 5 and Table 6.

*Table 5: Descriptive statistics of the attributes of the diabetes dataset from the EHR*

| Attribute | Min | Max | Mean | Std. deviation | Variance | Skewness |
|---|---|---|---|---|---|---|
| Age | 0 | 1 | 0.531 | 0.182 | 0.033 | -0.306 |
| Gender | 0 | 1 | 0.451 | 0.498 | 0.248 | 0.198 |
| BMI | 0 | 1 | 0.046 | 0.059 | 0.003 | 12.493 |
| BSA | 0 | 1 | 0.249 | 0.048 | 0.002 | 5.706 |
| Pulse | 0 | 1 | 0.440 | 0.178 | 0.032 | 0.138 |
| Systolic | 0 | 1 | 0.536 | 0.121 | 0.015 | 0.187 |
| Diastolic | 0 | 1 | 0.544 | 0.162 | 0.026 | -0.192 |
| Random BS | 0 | 1 | 0.304 | 0.189 | 0.036 | 1.228 |
| SPO2(%) | 0 | 1 | 0.970 | 0.043 | 0.002 | -17.778 |
| Temperature | 0 | 1 | 0.924 | 0.052 | 0.003 | -11.539 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Respiration | 0 | 1 | 0.038 | 0.039 | 0.002 | 23.065 |
| HTN | 0 | 1 | 0.667 | 0.472 | 0.222 | -0.711 |

***Evaluation for Quality Based on Statistical Methods***

The data attribute Temperature was leading with the number of outliers present at 26.84% (175/652) followed by oxygen saturation $SPO_2$ 20.55% (134/652). The data attribute Random BS had 5.52% (36/652), and BMI 3.07% (20/652) outliers present in the records respectively. On the other hand, respiration 0.15% (1/652) and pulse 0.61% (4/652) had the lowest number of outliers present in the diabetes dataset. The other details are as summarized in Figure 5.

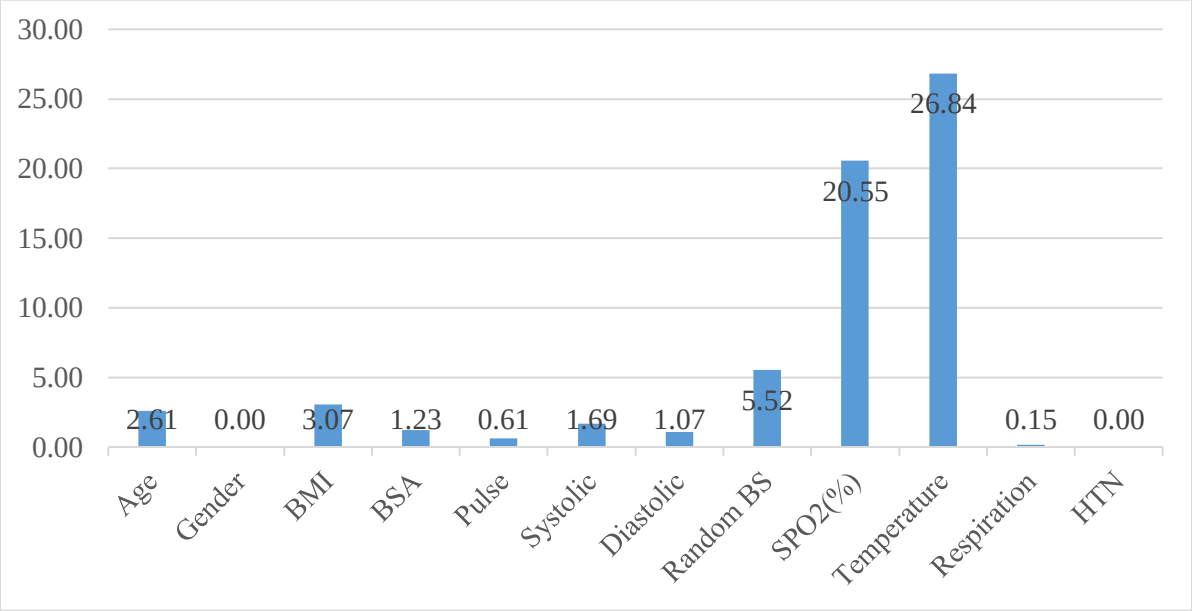


*Figure 1: Number of outliers present in the dataset per given attribute.*

***Evaluation for Quality Based on Dimensions of EHR Data Quality***

The quality of the EHR data was most affected by incompleteness at 25%. On the other hand, the consistency of the data values was rated at 99% thus both value and computational conformance of data values were also rated at 99% respectively. Others details are as summarized in Table 7.

*Table 6: Distribution of cases of diabetes mellitus in the EHR data set*

| Category | Number of Cases | Percentage (%) |
|---|---|---|
| Prediabetes | 3 | 0.4% |
| Gestational | 7 | 1% |
| Hypertension only | 25 | 4% |
| T1DM only | 19 | 3% |
| T1DM with Hypertension | 0 | 0% |
| T2DM only | 171 | 26% |

| | | |
|---|---|---|
| T2DM with Hypertension | 255 | 39% |
| T1DM and T2DM without Hypertension | 18 | 3% |
| T1DM and T2DM with Hypertension | 155 | 24% |

*Table 7: Summary of the evaluation of the quality of EHR data.*

| Dimension | Parameter | Frequency N= 652 | Percentage | Remarks |
|---|---|---|---|---|
| Conformance | Value conformance | 646 | 99% | Values were within range except for BMI and temperature |
| | Relational conformance | 652 | 100% | Values conformed to relational constraints |
| | Computational conformance | 646 | 99% | Fewer conformances were realized in the determination of BMI which is as a result of arithmetic error. |
| Consistency | Procedure for measurement | 652 | 100% | Procedure for measurement was evaluated by interviewees to be consistent for all data elements. |
| | Data measure | 646 | 99% | Data measures were consistent for all attributes except temperature and BMI. |
| | Data Granularity | 642 | 98% | Granularity was not consistent for attributes such as Height, Weight, Temperature and BMI. |
| Completeness | Completeness | 489 | 75% | Incompleteness was mainly due to data entry errors where an attribute had an unexpected value or an attribute with no value at all. |

***Evaluation for Quality through Density-Based Clustering***

The DBSCAN algorithm categorized 88% (573/652) records from the EHR dataset as noise. In addition, the remaining 12% of the EHR dataset were grouped into 23 clusters. The descriptions of the 23 clusters were verified by the ICD 10 codes that identified each diagnosis and summarized as shown in Table 8.

*Table 8: Summary of clusters identified from the diabetes dataset using the DBSCAN algorithm*

| Cluster | ICD 10 Codes | Complication |
|---|---|---|
| 1 | E11 – T2DM<br>E10 – T1DM<br>I10 – Essential Hypertension<br>I75 - Atheroembolism<br>H52.4 - Presbyopia-long-sightedness | Eye damage |
| 2 | E11 – T2DM | None |
| 3 | E11 – T2DM<br>K30 – Functional dyspepsia<br>H40 – Glaucoma – eye condition | Eye damage |
| 4 | E11 – T2DM<br>M75.4 – impingement syndrome of shoulder<br>M21.6 – deformities of foot | Neuropathy<br>Foot damage |
| 5 | E11 – T2DM<br>E10 – T1DM<br>I10 – Essential hypertension<br>I79 - Disorders of arteries, arterioles and capillaries… | Cardiovascular disease |
| 6 | E11 – T2DM<br>E10 – T1DM<br>I10 – Essential hypertension<br>I86 - Varicose veins of other sites<br>E78 - Disorders of lipoprotein metabolism and other lipidaemias | Cardiovascular disease<br>Lipids |
| 7 | E11 – T2DM<br>E10 – T1DM<br>I10 – Essential hypertension | None |
| 8 | E10 – T1DM | None |
| 9 | E11 – T2DM<br>E10 – T1DM<br>I10 – Essential hypertension<br>I74 –<br>I78 - Diseases of capillaries | Cardiovascular disease |
| 10 | E11 – T2DM<br>M54.5 - Low back pain | Neuropathy |
| 11 | E11 – T2MD<br>F31 - Bipolar affective disorder<br>B35.3 - Athlete's foot, also known as tinea pedis | Depression<br>Foot damage |

| | | |
|---|---|---|
| 12 | E11 – T2DM<br>I10 – Essential hypertension<br>E78 - Disorders of lipoprotein metabolism and other lipidaemias | Lipids |
| 13 | E11 – T2DM<br>I10 – Essential hypertension<br>E78 - Disorders of lipoprotein metabolism and other lipidaemias<br>E66 – Obesity | Obesity<br>Lipids |
| 14 | E11 – T2DM<br>E10 – T1DM<br>I10 – Essential hypertension<br>F31- Bipolar affective disorder<br>M13.9 – arthritis – unspecified | Depression<br>Neuropathy |
| 15 | E11 – T2DM<br>I10 – Essential hypertension<br>M47 – Spondylosis- Spondylosis-arthritis to the spine | Neuropathy |
| 16 | E11 – T2DM | None |
| 17 | E11 – T2DM | None |
| 18 | E11- T2DM<br>E10 – T1DM<br>H52.4 – Presbyopia | Eye damage |
| 19 | E11 – T2MD<br>E10 – T1DM<br>I10 – Essential hypertension | None |
| 20 | E11 – T2DM | None |
| 21 | E11 – T2DM | None |
| 22 | E11 – T2DM<br>E10 – T1DM<br>I10 – Essential hypertension | None |
| 23 | E11 – T2DM<br>E10 – T1DM<br>I10 – Essential hypertension | None |

**Discussion**

Secondary utility of EHR data in clinical informatics research will undoubtedly reduce research costs and also stimulate innovation in healthcare thus solve a number of challenges in medicine. However, EHR data are presently associated with poor quality as a result of the

presence of noise, outliers, incompleteness, and inconsistencies in records. Therefore, there is need for improving the quality of EHR data to promote their uptake in research. The hypothesis of this research upheld is that EHR software design, the utility of a data dictionary, training and motivation of the human resource responsible for the collection and entry of patient data into the EHR, and the process of data collection collectively affect the quality of EHR data hence the suitability of such data in computational phenotyping of diabetes mellitus.

The EHR data were found unsuitable for utility in a computational phenotyping task given the presence of much noise that affected the quality of clusters determined by the DBSCAN algorithm. The clusters identified from the dataset had a number of similarities and differences. The variations could simply mean that with improved quality of EHR data, clusters would be distinct enough.

Based on Table 8, Clusters 7, 19, 22, and 23 show similarities in the disease and associated complications based on the ICD 10 codes. However, the algorithm determined that these clusters were significantly distinct hence form distinct sub-groups. Similarly, clusters 2, 16, 17, and 20, based on the characteristics of the dataset look similar. However, the algorithm identified each of them as distinct. This could mean there is an underlying difference in the data that is not obvious thus only known to the learning algorithm. Moreover, clusters; 1, 3, and 18 are distinct sub-groups of retinopathy. On the hand, clusters; 5, 6, and 9 are distinct sub-groups of cardiovascular complications. Moreover, clusters; 12 and 13 are distinct sub-groups of lipids. This shows that the task of computational phenotyping from routine healthcare data is achievable since as the algorithm has not only identified the phenotypes but also the sub-types of the given phenotypes. As a result, clinical decision support systems could easily be developed from such backgrounds to assist physicians in the delivery of healthcare services. Moreover, with minimal efforts to improve the quality of EHR data, computation phenotyping and sub-phenotyping of diseases would easily be realizable.

The appropriateness of the EHR software design was evaluated based on whether the software is comprehensive and covers the complete picture of the required data attributes for the diagnosis and management of diabetes mellitus. Furthermore, this research evaluated the usability of the software and whether the EHR software design addresses security needs of the users whenever errors occur. Participants reported (91%) that the EHR software was comprehensive and adequately covers all the parameters required for the diagnosis and management of diabetes mellitus. However, participants mentioned that the EHR does not provide a field for the measurement of the circumference of the waists for patients. While going through the records of diabetes patients, this study observed that the circumference of the waists of the patients were not present in the records. However, waist size is an essential attribute in the diagnosis and management of diabetes mellitus. Thus the EHR should provide an opportunity for recording the waist measurements for patients (Conti et al., 2017; Daga et al., 2015; Mwangi et al., 2017).

Participants also reported that the EHR does not provide opportunity for nurses to record patient reported drug allergies. Instead, the EHR has pre-recorded drug allergies for the nurses to choose from. Nonetheless, EHR design should cater for opportunities for recording patient specific allergies. Lastly, participants also reported the need for the EHR software to raise alerts when triage parameters such as blood pressure (BP) and pulse rate are high beyond expectation. Conversely, EHR software with the capability to raise alerts for measures that borderline to the worst cases is a great plus. But in actual sense, alerts are usually functions of the clinical decision support systems (Conti et al., 2017; Daga et al., 2015; Garets & Davis, 2006; Weir et al., 2015).

The usability of the EHR software was evaluated by looking into participants' perceived usefulness, ease of use, and ease of learning to use the EHR software. Participants (94%) agreed that the software is very easy to use by both regular and intermittent users and they would easily and quickly recover from their mistakes using the EHR. Participants rated the EHR software to be very useful to them in performing

their tasks. This means that the software had all the key data elements and functions that are required by users to complete their various tasks. Additionally, participants appraised the EHR to be very easy to use. As a matter of fact, one of the nurses asserted that the EHR is "a mimic of our workflow and I do not have to be trained to use it." Also, two recently hired nurses that have never had the chance to be trained to use the EHR reported that they have been using the software comfortably and hardly come across a scenario where they need assistance. Finally, users of any software product in an organization are usually categorized as super-users and the rest as ordinary users and novices. However, in the case of this study site, 99% of participants believed the software is easy to learn to use hence all users believed they were in the category of super-users. As a matter of fact, all users believed they are comfortable using the EHR (Alqahtani, Crowder & Wills 2017; van Engen-Verheul et al. 2016; Lund 2001; Verheij et al, 2018.).

A well-designed software system provides users with adequate mechanisms for error detection, avoidance, and/or correction at various levels of complexity. Such mechanisms help boost user confidence and outcomes when completing tasks. One of the reasons why users make errors is when the software has no clearly marked shortcuts. As a result, users are usually forced to recall (high memory load that results into stress) the steps towards completing a given task rather than recognition that reduces the memory load hence less stress. For the given EHR, 100% of participants believed that the memory load is greatly reduced, and the software had clearly marked shortcuts. Participants' position relatively supported their earlier position that the EHR is easy to learn and easy to use respectively (Associates et al., 2009; Lund, 2001; Mullin et al., 2017; Vehmas & Kaipio, 2018).

On the other hand, 27% of the participants believed the EHR does not provide feedback yet feedback is the method utilized to boost user confidence and participation through the workflow to accomplish tasks. Also, 73% of participants believed that the EHR does not provide good error messages yet provision of good error messages is a positive step towards guiding users off the path of making errors. These results are also supported 82% of users who believed that the EHR hardly prevents errors from occurring. Despite the fact that participants believed that they find the EHR software useful in performing their daily tasks, participants also believed their EHR software design poorly addressed the needs for error detection, avoidance, and/or correction (Fox et al., 2018a; Weir et al., 2015).

A data dictionary is a help tool for EHR users to understand their working environment and the parameters they are working with. Availability of such tools not only makes the users informed but also creates a standard in recording of data (Bruland et al., 2017; Feder, 2017; Hicken et al., 2004; Keny et al., 2015). At this study site, physicians had long queues hence did not get to interact much with the EHR software. Therefore, physicians depended on support from nurses and health record officers. However, results show that 100% of participants approved that the EHR software had a number of useful diagnostic and documentation tools in addition to the data dictionary. Some of the tools that come in addition to a data dictionary included ICD 10, Risk of fall, and Pain score to support their work and data entry. As a matter of fact, this research utilized the same data dictionary to retrieve the dataset from the EHR database.

The data collection process at the clinic involves taking measurements and readings and the data are directly entered into the EHR by the staff that took such measurements without prior recording before entry into the EHR. When asked whether participants believed that their data collection process adequately resulted in the recording of data of good quality, 86% of the participants responded "yes". Furthermore, participants also cited that their working stations were equipped with diagnostic and documentation tools to support the care process. However, 25% of the data in the EHR had incomplete recordings of data elements for patient records. The attributes commonly affected by incompleteness were measures for Temperature, BMI, Pulse Rate, BSA, and Oxygen saturation in the blood ($SPO_2$). The effect of the incompleteness of records is also reflected by the presence of outliers as attributes most affected by outliers had also experienced cases of incompleteness and inconsistencies, As a

matter of fact, 9% of records had cases of inaccurate recording of data (Farrell et al., 2017; Verheij, Curcin, Delaney, & McGilchrist, 2018; Zozus et al., 2014).

Finally, the participants in this study reported (93%) that they had been provided with adequate trainings to be able to effectively use the EHR and tools essential for successful task completion. Also, users reported (90%) that they are motivated to continue with the use of the EHR to complete their various tasks. According to social cognitive theory, learning occurs in a social context with a dynamic reciprocal interaction of the person, environment and behaviour. Also, in the Capability Opportunity model (COM-B), Capability is individual's psychological and physical capacity to engage in the activity concerned which includes having necessary knowledge and skills. Opportunity refers to all factors that lie outside the individual that make the behaviour possible. The Opportunity factors include availability of work space and tools for the staff to successfully perform the expected tasks. Finally, motivation are the brain processes that energize and direct behaviour while motivation relates to attitudes and aspirations (Mayne, 2016; Ory et al., 2010; Wood & Bandura, 1989).

According to Fredrick Herzberg's two factor theory, there are factors in the workplace that cause job satisfaction and other factors which only decrease job dissatisfaction. A working condition refers to a work environment that is meant to promote the efficient performance of tasks by employees. The belief that there is a link between working condition and employee performance at work implies that the key to motivation is within an employee's job itself. "Hygiene factors" only reduce employee dissatisfaction but cannot create satisfaction.

On the other hand, "motivation factors" stimulate satisfaction within the employee provided that the minimum levels of hygiene factors are reached. Herzberg disagreed with the view that money and compensation are the most effective ways of motivating employees. According to Herzberg's motivation hygiene model, employee motivation is achieved when employees are faced with challenging but enjoyable work where one can achieve, grow, and demonstrate responsibility and advance in the organization. Moreover, intrinsic factors for motivation are very effective in creating and maintaining more durable positive effects on employee's performance towards their jobs as these factors are basic human needs for psychological growth. Thus, intrinsic factors increases the employee's quality of work unlike the extrinsic factors that simply permit employee's willingness to work (Bigirimana et al., 2018; Dartey-Baah & Amoako, 2011; Fauziah et al., 2013; Gawel, 1997). Therefore, behaviour, which is the quality of EHR data (outcome), is determined by the combined force of extrinsic and intrinsic factors. Given that the external factors are well addressed by the employer yet the quality of the EHR data is discovered to be poor, the outcome of this research is that the behaviour (poor quality of EHR data) is associated by the unmet needs of the intrinsic factors of motivation.

**Conclusion**

The main goal of this research was to determine the applicability of routine healthcare data in clinical informatics research. The objectives of this research was to determine how the interplay of EHR software design, the use of a data dictionary, the process of data collection, and the training and motivation of the human resource involved in the collection and entry of data into the EHR affect the quality of EHR data thus the suitability of such data for utility in computational phenotyping of diabetes mellitus.

The results from regression analysis show that the EHR software design explained 50.7% of the improvement on the suitability of EHR data for computational phenotyping of diabetes mellitus. On the other hand, use of data dictionary explained 32.3%, the process of data collection explained 22.6%, and staff development explained 16.6% of the suitability of EHRR data for computational phenotyping of diabetes mellitus. Furthermore, the EHRR software was reported to be comprehensive and the usability of the EHR software was acceptable to the participants. Moreover, the EHR software has an embedded data dictionary that enabled participants to accomplish tasks satisfactorily.

Moreover, participants confirmed that their data collection process and the utilization of a data dictionary during data collection and entry into the EHR contributed to the collection of data of good quality. Finally, participants confirmed that their employer had provided them with adequate opportunities for training to use the EHR thus participants were motivated to continue using the EHR.

The second dataset comprised of historical cases of diabetes mellitus collected during routine clinical care at the study site and stored in the EHR software. Results show that the quality of dataset was affected by the presence of outliers and cases of incompleteness and inconsistencies in the data. Furthermore, a density-based clustering algorithm; DBSCAN identified 23 sub-types of diabetes mellitus from 12% of the dataset. However, DBSCAN categorized 88% of the dataset as noise. As a result, with minimal efforts to improve the quality of the EHR data, the task of computational phenotyping from EHR data will be easily realizable.

Therefore, the EHR data were found to be unsuitable for utility in computational phenotyping of diabetes mellitus. However, with proper measures to minimize cases of inconsistencies, incompleteness, and the presence of outliers in the EHR data, computational phenotyping would be realizable from routine healthcare data. This research asserts that the main reason for the poor quality of EHR data is the unsatisfied intrinsic motivational factors among the human resource who are directly involved in the collection and entry of data into the EHR. The Nairobi Hospital is known to be a good employer that has sufficiently met the "hygiene-factors" which are known to only reduce employee dissatisfaction.

The proposal was submitted for ethical approval from Kenyatta University Ethical and Research Committee. Thereafter, the study sought a permit to conduct the research from National Commission for Science, Technology and Innovation (NACOSTI). Moreover, the study sought relevant permissions and approvals from The Nairobi Hospital administration. Following of the review of the protocol for conducting research The Nairobi Hospital, the proposal for this research had to be submitted for a second ethical review with The Nairobi Hospital Bioethics & Research Committee and an approval was granted. Finally, consent of participants was sought before their involvement in this research.

**References**


Adler-Milstein, J., Holmgren, A. J., Kralovec, P., Worzala, C., Searcy, T., & Patel, V. (2017). Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. *Journal of the American Medical Informatics Association*, *24*(6), 1142–1148. https://doi.org/10.1093/jamia/ocx080

Alqahtani, A., Crowder, R., & Wills, G. (2017). Journal of health informatics in developing countries. In *Journal of Health Informatics in Developing Countries* (Vol. 11, Issue 2). University of Otago. http://www.jhidc.org/index.php/jhidc/article/view/160

Associates, J. B., Armijo, D., Mcdonnell, C., & Werner, K. (2009). *Electronic Health Record Usability Interface Design Considerations HEALTH IT*. http://www.ahrq.gov

Bigirimana, S., Sibanda, E., & Masengu, R. (2018). The effects of working conditions on academic staff motivation at africa university. Https://www.researchgate.net/publication/305768428

Bruland, P., Doods, J., Storck, M., & Dugas, M. (2017). What Information Does our EHR Contain? Automatic Generation of a Clinical Metadata Warehouse (CMDW) to Support Identification and Data Access ithin Distributed Clinical Research Networks. *International Medical Informatics Association (IMIA) and IOS Press*. https://doi.org/10.3233/978-1-61499-830-3-313

Che, Z., & Liu, Y. (2017). Deep Learning Solutions to Computational Phenotyping in Health Care. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 1100–1109. https://doi.org/10.1109/ICDMW.2017.156

Conti, C., Mennitto, C., Di Francesco, G.,

Fraticelli, F., Vitacolonna, E., & Fulcheri, M. (2017). Clinical Characteristics of Diabetes Mellitus and Suicide Risk. *Article Mini Review*, *8*(40). https://doi.org/10.3389/fpsyt.2017.00040

Daga, R. A., Naik, S. A., Maqbool, M., Laway, B. A., Shakir, M., & Rafiq, W. (2015). Demographic and Clinical Characteristics of Diabetes Mellitus among Youth Kashmir, India. *Int J Pediatr*, *3*(19), 4–1. http://

Dartey-Baah, K., & Amoako, G. K. (2011). Application of Frederick Herzberg's Two-Factor theory in assessing and understanding employee motivation at work: a Ghanaian Perspective. In *European Journal of Business and Management www.iiste.org ISSN* (Vol. 3, Issue 9). Online. www.iiste.org

Denaxas, S., Direk, K., Gonzalez-Izquierdo, A., Pikoula, M., Cakiroglu, A., Moore, J., Hemingway, H., & Smeeth, L. (2017). Methods for enhancing the reproducibility of biomedical research findings using electronic health records. In *BioData Mining* (Vol. 10, Issue 1, pp. 1–19). BioMed Central Ltd. https://doi.org/10.1186/s13040-017-0151-7

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. 226--231. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220

Farrell, L. J., Shimeng, D., Steege, L. M., Cartmill, R. S., Wiegmann, D. A., & Wetterneck, T. B. (2017). Understanding Cognitive Requirements for EHR Design for Primary Care Teams. *Proceedings of the 2017 International Symposium on Human Factors and Ergonomics in Health Care The*. https://doi.org/10.1177/2327857917061005

Fauziah, W., Yusoff, W., Shen Kian, T., & Idris, M. (2013). *herzberg's two factors theory on work motivation: does its work for todays environment*? (vol. 2, issue 5).

Feder, S. L. (2017). Data Quality in Electronic Health Records Research : Quality Domains and Assessment Methods. *Western Journal of Nursing Research*, 1–14. https://doi.org/10.1177/0193945916689084

Fox, F., Aggarwal, V. R., Whelton, H., & Johnson, O. (2018a). *A Data Quality Framework for Process Mining of Electronic Health Record Data*. IEEE. (In Press. http://eprints.whiterose.ac.uk/132110/

Fox, F., Aggarwal, V. R., Whelton, H., & Johnson, O. (2018b). A data quality framework for process mining of electronic health record data. *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, 12–21. https://doi.org/10.1109/ICHI.2018.00009

Garets, D., & Davis, M. (2006). *Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference A HIMSS Analytics TM White Paper Executive Summary*. www.himssanalytics.org

Gawel, J. E. (1997). *Herzberg's theory of motivation and Maslow's hierarchy of needs* (Vol. 5, Issue 11). http://pareonline.net/getvn.asp?v=5&n=11.

Ghosh, S., Cheng, Y., & Sun, Z. (2016). Deep State Space Models for Computational Phenotyping. *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 399–402. https://doi.org/10.1109/ICHI.2016.71

Hicken, V. N., Thornton, S. N., & Rocha, R. A. (2004). *Integration Challenges of Clinical Information Systems Developed Without a Shared Data Dictionary*. https://pdfs.semanticscholar.org/cbe5/d11ace14f287e2b26c717c7b879ce71c7d07.pdf

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S.-T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., & Schilling, L. (2016). A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Washington, DC)*, *4*(1), 1244. https://doi.org/10.13063/2327-9214.1244

Keny, A., Wanyee, S., Kwaro, D., Mulwa, E., & Were, M. C. (2015). *Developing a National-Level Concept Dictionary for EHR Implementations in Kenya*. https://doi.org/10.3233/978-1-61499-564-7-780

Longhurst, C., Davis, T., Maneker, A., Eschenroeder, H., Dunscombe, R., Reynolds, G., Clay, B., Moran, T., Graham, D., Dean, S., & Adler-Milstein, J. (2019). Local Investment in Training Drives Electronic Health Record User Satisfaction. *Applied Clinical Informatics*, *10*(02), 331–335. https://doi.org/10.1055/s-0039-1688753

Lopez, C., Omizo, R., & Whealin, J. (2018). Impact of a tailored training on advanced electronic medical records use for providers in a Veterans Health Administration Medical System. *JAMIA Open*, *1*(2), 142–146. https://academic.oup.com/jamiaopen/article-abstract/1/2/142/5074118

Lund, A. M. (2001). Measuring Usability with the USE Questionnaire 12 General Background. *Researchgate*. https://www.researchgate.net/profile/Arnold_Lund/publication/230786746_Measuring_Usability_with_the_USE_Questionnaire/links/56e5a90e08ae98445c21561c/Measuring-Usability-with-the-USE-Questionnaire.pdf

Mayne, J. (2016). *The Capabilities, Opportunities and Motivation Behaviour-Based Theory of Change Model*.

Mullin, S., Anand, E., Sinha, S., Song, B., Zhao, J., & Elkin, P. L. (2017). Secondary Use of EHR: Interpreting Clinician Inter-Rater Reliability Through Qualitative Assessment. *Studies in Health Technology and Informatics*, *241*, 165–172. http://www.ncbi.nlm.nih.gov/pubmed/28809201

Mwangi, N., Macleod, D., Gichuhi, S., Muthami, L., Moorman, C., Bascaran, C., & Foster, A. (2017). Predictors of uptake of eye examination in people living with diabetes mellitus in three counties of Kenya. *Tropical Medicine and Health*. https://doi.org/10.1186/s41182-017-0080-7

Ory, M. G., Smith, M. L., Mier, N., & Wernicke, M. M. (2010). The science of sustaining health behavior change: The health maintenance consortium. *American Journal of Health Behavior*. https://doi.org/10.5993/AJHB.34.6.2

Reimer, A. P., Milinovich, A., & Madigan, E. A. (2016). Data quality assessment framework to assess electronic medical record data for use in research. *International Journal of Medical Informatics*, *90*, 40–47. https://doi.org/10.1016/j.ijmedinf.2016.03.006

Richesson, R. L., Horvath, M. M., & Rusincovitch, S. A. (2014). Clinical Research Informatics and Electronic Health Record Data. *IMIA and Schattauer GmbH*, 215–223.

Shickel, B., Tighe, P., Bihorac, A., & Rashidi, P. (2017). *Deep EHR: A Survey of Recent Advances on Deep Learning Techniques for Electronic Health Record (EHR) Analysis*. http://arxiv.org/abs/1706.03446

Tenenbaum, J., & Avillach, P. (2016). An informatics research agenda to support precision medicine: seven key areas. *Journal of The*. http://jamia.oxfordjournals.org/content/23/4/791.abstract

Tutty, M., Carlasare, L., Lloyd, S., & Sinsky, C. (2019). The complex case of EHRs: examining the factors impacting the EHR user experience. *Journal of the American Medical Informatics Association*, *26*(7), 673–677. https://academic.oup.com/jamia/article-abstract/26/7/673/5426085

van der Bij, S., Khan, N., ten Veen, P., de Bakker, D. H., & Verheij, R. A. (2017). Improving the quality of EHR recording in primary care: a data quality feedback tool. *Journal of the American Medical Informatics Association*, *24*(1), 81–87. https://doi.org/10.1093/jamia/ocw054

van Engen-Verheul, M. M., Peute, L. W. P., de Keizer, N. F., Peek, N., & Jaspers, M. W. M. (2016). Optimizing the user interface of a data entry module for an electronic patient record for cardiac rehabilitation: A mixed method usability approach. *International Journal of Medical Informatics*, *87*, 15–26. https://doi.org/10.1016/J.IJMEDINF.2015.12.007

Vehmas, N., & Kaipio, J. (2018). Physicians as usability evaluators-first aid for poor EHR usability? *Finnish Journal of EHealth and EWelfare*, *10*(2–3), 297. https://helda.helsinki.fi/bitstream/handle/10138/248732/69162_Article_Text_92362_1_10_20180520.pdf?sequence=1

Verheij, R., Curcin, V., Delaney, B., & McGilchrist, M. (2018). Possible Sources of

Bias in Primary Care Electronic Health Record Data Use and Reuse. *Journal of Medical Internet Research*, *20*(5), e185. https://doi.org/10.2196/jmir.9134

Verheij, R., Curcin, V., Delaney, B., & MM, M. (2018). Possible sources of bias in primary care electronic health record data use and reuse. *Jmir.Org*. https://www.jmir.org/2018/5/e185

Weir, C. R., Staggers, N., Gibson, B., Doing-Harris, K., Barrus, R., & Dunlea, R. (2015). A qualitative evaluation of the crucial attributes of contextual Information necessary in EHR design to support patient-centered medical home care. *BMC Medical Informatics and Decision Making*, *15*(1), 30. https://doi.org/10.1186/s12911-015-0150-x

Wood, R., & Bandura, A. (1989). Social Cognitive Theory of Organizational Management. *Academy of Management Review*. https://doi.org/10.5465/amr.1989.4279067

Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2017). *Mining Electronic Health Records: A Survey*. http://arxiv.org/abs/1702.03222

Zozus, M. N., Hammond, W. E., Green, B. B., Kahn, M. G., Richesson, R. L., Rusincovitch, S. A., Simon, G. E., & Smerek, M. M. (2014). *Assessing Data Quality*. https://www.nihcollaboratory.org/Products/Assessing-data-quality_V1 0.pdf