University of Nairobi

School of Mathematics

ISSN: 2410-1397

Master Project in Biometry

A Comparative Analysis of Unsupervised Outlier Detection Methods for Data Quality Assurance

Research Report in Biometry, Number 49, 2020

Mercy Chepkirui Terer, 156/12362/2018

October 2020

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Biometry

Master Project in Biometry **University of Nairobi**

October 2020

A Comparative Analysis of Unsupervised Outlier Detection Methods for Data Quality Assurance

Research Report in Biometry, Number 49, 2020

Mercy Chepkirui Terer, 156/12362/2018

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Biometry

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

Data quality assurance is a key component in research. It is almost impossible to routinely check for errors in large datasets if automated smart mechanisms are not put in place. The quality of results from data analysis heavily relies on the underlying state of data. Quality data leads to effective and unbiased reporting. Errors introduced into the data are inevitable hence the need to have error-checking mechanisms.

Error checking mechanisms such as the use of range checks, quantile ranges and z-scores are limited to continuous data types and effective for small feature space data. Errors in dichotomous and character data types are easily omitted hence the need to use methods that scan anomalies for all data types and for extremely large datasets. Two pass verification on the other hand is a gold standard method for checking the quality state of data. It involves random sampling of observations to be re-entered from similar source documents to measure the level of accuracy and consistency of data. It is an accurate process; however, it is a tedious and manual process that relies on random sampling for larger datasets.

We propose possible alternative methods for error checking by applying machine learning outlier detection algorithms. The observations that are outlying are subjected to cross-referencing for possible errors instead of randomly selecting a set of observations.

We evaluated k-means clustering and isolation forest unsupervised machine learning algorithms to detect outliers. The outliers form the sample of observations to be validated and verified. We then compared two pass verification anomaly scores, k-means anomaly scores and isolation forest anomaly scores. Normalized mutual information score and the coefficient of determination metrics were used to determine the strength of the correlation. The results indicated that unsupervised machine learning methods can be possible alternatives for data quality assurance with a flexibility for future considerations and improvements. Isolation forest performed better than k-means clustering.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature Date

Mercy Chepkirui Terer
Reg No. I56/12362/2018

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Signature Date

Dr. Timothy Kamanu
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: tkamanu@uonbi.ac.ke

Signature Date

Mr Paul Mwaniki
KEMRI - Wellcome Trust
P.O Box 43640 – 00100, Nairobi, Kenya.
E-mail: PMwaniki@kemri-wellcome.org

Dedication

I dedicate this project to God for a good health , strength, knowledge and inspiration. He has been the source of my strength throughout my studies and on His wings only have I soared. I also dedicate this work to my dad; Joseph Kipngetich Terer who encouraged me all the way and whose encouragement has made sure that I give it all it takes to finish that which I started. To my husband, Sylvester Mwambeke and child, Danson Mrhongo Mwambeke, who have been greatly affected in every way possible by this quest,thank you. My love for you all can never be quantified. God bless you.

List of Tables

List of Figures

Acknowledgments

I am thankful to God for giving me good health and guidance to do this project. I would like to appreciate Dr. Timothy Kamanu and Mr. Paul Mwaniki for the invaluable guidance, professional suggestions and support throughout the process of my research project writing and analysis. To all the lecturers who taught me during the course work period, their guidance and ideas provided the framework and foundation upon which this research project is structured. To my classmates too, big thank you for your endless support.

Mercy Chepkirui Terer

Nairobi, 2020.

Contents

1 Introduction

1.1 Background

Data quality is a critical element in research (Cai & Zhu, 2015). Data management utilizes data quality assurance as a pre-condition to achieve effective data-driven decision making.

Limited trust in the data is explained by data quality issues such as missing values for critical variables in the datasets and incorrect data formats. Decisions and interventions made out of poor quality data are less effective and biased(Redman, 1998). The major implications include users abandoning the data and a significant waste of resources invested in obtaining the data (Haug et al., 2011). Anomalies in the data (Foorthuis, 2018) take the form of invalid data values, missing data, values in corrupted format, duplicate instances, inconsistent unit measures and incomplete cases. Data anomalies vary from one domain to the other (Azeroual et al., 2018)) hence the different modes of processes to control quality. For instance, an intrusion attack on a computer network, a suspicious money transaction on a credit card and unexpected geographical event.

Errors in the data are introduced in a routine data collection setting by poor implementation of electronic data collection systems (Bowman, 2013). Software upgrading (Rodríguez-Pérez et al., 2020) may introduce bugs that modifies the original expected format of the data. Manual data entry process, extensively used across most disciplines, introduces typographical errors (Ley et al., 2019). Using incorrect source document (Bargaje, 2011) and mixing data from different sources introduce errors into the data. Unintended data manipulation while data mining and natural novelties in the data are potential sources of anomalies.

Two pass verification (Paulsen et al., 2012) is used for determining the quality of data in domains where manual data entry is done. Other disciplines implement automated (Sodemann et al., 2012) systems that scan the data to flag potential anomalies. Automated systems, however, use predefined known patterns and ranges to determine anomalous values. Statistical methods (Seo, 2006) such as the use of quantile ranges and z-score values are often used during data pre-processing to detect outliers. Training users and policy implementation creates awareness of the importance of data quality and minimizes the chances of committing errors during data entry but does not eliminate errors.

This work focuses on data quality for routine data collected for research. Two pass verification (Büchele et al., 2005) is a gold standard for measuring data quality. Multiple users key in similar forms of data into the electronic system at different times. Data forms are randomly sampled for large datasets and entirely repeated for smaller datasets based on the investigator's preference. The aim is to perform a pairwise comparison for each observation then determine the level of agreement for the two datasets. Discordant observations are eventually reviewed and resolved. Two pass verification is tedious, expensive and there is no way of correcting errors for records not sampled.

Machine Learning(ML) as defined by (Panch et al., 2018) is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. ML methods for outlier detection have been widely used in various disciplines such as;(1)The financial sector for streaming transactional data to detect credit card fraud (Dhankhad et al., 2018) (Dornadula & Geetha, 2019) (Randhawa et al., 2018). (2) Manufacturing industries utilized ML methods as a quality control measure to achieve defect-free products (Escobar & Morales-Menendez, 2018). (3) Wireless Sensor Networks (Di & Joo, 2007; Kumar et al., 2019) dynamic nature require automated methods for detecting faults and (4) in military surveillance for enemy related attacks. Outlier detection may be useful in ensuring good quality data in medical research. Unsupervised outlier detection methods could be used to detect problems in the data such that values found to be outliers are further examined for error-checking purposes and resolution. Outliers may not necessarily imply an error, but outliers could be carrying underlying useful detail that could aid the data quality process.

Machine learning methods are cheaper to apply in large datasets compared to traditional methods such as two pass verification and univariate methods like z-scores and quantile ranges. Continuous and categorical data types can be applied on the ML models without any limitation. Unsupervised ML methods does not require prior knowledge and distribution of the dataset hence no prior training of the data is required. Unsupervised learning is a critical feature suitable for outlier detection.

This work will seek to explore unsupervised ML outlier detection techniques in medical research data as a component of data quality assurance. The output from ML methods will be compared with the scores from two pass verification process. Isolation Forest(iForest) and k-means clustering models will be evaluated in this analysis.

1.2 Statement of problem

Over the last decade, there has been a huge embrace of technology in the health sector (Galetsi et al., 2020) and in medical research. The use of Electronic Health Records (EHR) systems for data collection increases the need for effective data quality methods. Existence of outliers in the data can indicate observations that have a unique behavior from the rest of the observations. Outlying observations could have a critical element that requires immediate focus especially when it comes to routine patient management in routine care settings.

Outliers are often eliminated to improve the accuracy of the estimators. Examining outliers and exceptions in data mining has not received as much attention as other topics like classification and clustering.

Error checking methods such as range checks, quantile ranges and z-scores are limited to small feature space datasets. Range checks are suitable for continuous variables hence not suitable for dichotomous and unstructured data types. Two pass verification employs an effective method of re-entering sampled data then performing a level of agreement for each pair of observations. However, there is no way of checking errors in data not sampled.

Unsupervised machine learning anomaly detection algorithms can be possible alternatives for effective error checking. Unsupervised learning involves detecting patterns in the data without prior knowledge which makes it a critical feature for outlier detection. Unsupervised learning are cheap for large datasets and they can be applied on both continuous and categorical data.

1.3 Objectives

1.3.1 Overall objective

The overall objective is to determine the correlation between unsupervised machine learning outlier detection methods and two pass verification outlier scores.

1.3.2 Specific objectives

1. Derive outlier scores using k-means, isolation forest and two pass verification methods.
2. Calculate the correlation between machine learning outlier scores and two pass verification scores.

1.4 Justification of the study

Error investigation in large datasets seems an impossible task especially for extremely large datasets growing with time. There is no perfect data source(Brown et al., 2018), and mistakes are inevitably made in one way or another. The inherent nature of Big Data calls for urgent need for best measures of ensuring good quality data for analysis and ultimately accurate reports needed for decision making.

Automated unsupervised (Ghahramani, 2004) measures are needed for real-time screening of data at the point of data capture. Error flagging at the point of entry is easier to manage and handled than at the point of analysis. Real time error rate reporting can be used to recommend and implement techniques of improving data quality overtime.

The outcome of this analysis ,if recommended, will inform the routine care settings case management protocol of routine data quality.

Anomaly detection algorithms have been widely applied in various fields such as in military surveillance, cyber security, fraud detection and faulty detection in critical health care systems (Ding & Fei, 2013; S. Hawkins et al., 2002; Sodemann et al., 2012). Unsupervised anomaly detection methods could be used as an alternative method of data quality assurance in routine clinical data such as inpatient routine data collection in health care and in medical research.

2 Literature review

This section describes approaches of outlier detection and their applications in literature. The review demonstrates different outlier detection algorithms. We highlight their differences, strengths, weaknesses and their application domains.

Datasets with plenty of variables recorded involve sampling during analysis. One of the first steps is to check outlaying observations. Outliers are eliminated (Williams et al., 2002) even though they could potentially carry critical information. Outlier contain data points deviating from the norm that could lead to biased estimates and misspecification of the model (Ben-Gal, 2005) and inappropriate results.

Researchers have defined outliers dynamically based on hidden assumptions in the datasets and the methods used. (D. M. Hawkins, 1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. (Pincus, 1995) and (Johnson et al., 2002) defined outliers as a data point that appears to deviate markedly from other members of the same sample and as an observation in the dataset which appears to be inconsistent with the rest of the set respectively. (Liu et al., 2008) describes it as data points that are few and different. Outlier detection methods have been classified based on the number of features in a dataset or based on the underlying distribution of the data.

Statistical outlier detection methods (Ben-Gal, 2005) assume an underlying distribution of observations. Outliers then becomes those values that deviates from the model assumptions. Parametric methods are not suitable for high-dimensional datasets if prior information about the underlying distribution is not known (Papadimitriou et al., n.d.). Non-parametric methods are distribution-free hence they can be applied on large datasets; no prior assumptions about the dataset are made. Distance-based based methods (Ester et al., 1996; Knorr & Ng, 1997) capable of handling large databases falls in this category together with clustering techniques (Acuna & Rodriguez, 2004; Barbará & Chen, 2000; Ramaswamy et al., 2000) where clusters with less dense patterns than the rest of the clusters are labelled then further partitioned into non-overlapping groups of outliers and inliers.

Univariate methods studies one feature at a time while multivariate methods scans more than one feature at a time. Methods such as histograms and boxplots are mostly used for their simplicity even though most of the surveys are multivariate. Non-robust univariate methods were not considered by (Templ et al., 2020) since they cannot adequately detect outliers; quantiles specific methods falsely classified a number observations. Further, Box-cox transformation was used to account for skewness in the dataset then outlier detection methods were applied. Pareto tail modelling (Dupuis & Victoria-Feser, 2006; Ziegel, 2004) which copes with rightly skewed data by using a cut-off point from Van Kerm's rule of the thumb (Alfons et al., 2010; Vanpaemel et al., 2008) and adjusted boxplot was used to better accommodate skewed data (Hubert & Vandervieren, 2008). Methods which did not account for skewed data detected lower outliers and outliers detected by these methods did not account for skewness. Adjusted box plot did not perform well compared to pareto tail modelling. Univariate methods must be adapted for skewness; they do not perform well without transformation. Precaution needs to be taken if used in practice especially if data is skewed. Multivariate methods showed better results compared to univariate methods which could be improved by choosing better tuning constants. They detected true/positive outliers and flagged only few false/positive outliers. Further study is needed for outlier detection with complex survey designs since multivariate methods do not consider sampling weights.

Mahalanobis distance (Filzmoser, 2004) is a critical element in multivariate methods. Parameters estimated are compared with a critical chi-square value such that values high than the critical values are assumed outliers. These values may not necessarily be outliers but data points forming part of distribution. To solve this gap (Garrett, 1989) came up with a chi-square plot which used empirical distribution Mahalanobis distance against the chi-square value such that a break in the tail indicated an outlier point. This method however needs continuous interaction which implied a tedious impossibility for large-dimensional datasets. The use of robust distances (RD) by (Rosseeuw & Van Zomeren, 1990) such that squared RD for an observation is higher than the critical value of a record is considered an anomaly as an outlier detection method did not account for underlying data structure hence some outliers could turn out as false/positives. Automated multivariate method (Filzmoser, 2004) which accounted for different data dimensions and sample sizes was the best alternative. However, it did demonstrate the performance of the method on real data and data with more than 2-dimensions.

Majority of model-based anomaly detection methods construct a profile of normal instances then identify data points that do not conform with the profile as anomalies. Such methods include Replicator Neural Network(RNN) (S. Hawkins et al., 2002; Williams et al., 2002), classification-based methods (Abe et al., 2006), one-class SVM (Shahid et al., 2015) and clustering methods (Loureiro et al., 2004). These approaches are not optimized for outlier detection hence their output has many false/positives anomalies. Additionally, they are constrained to small-dimensional datasets since they were not originally designed for anomaly detection but for other purposes (classification and clustering).

(Ester et al., 1996) developed a distance-based clustering approach called DBSCAN, an outlier detection algorithm, which can detect anomalies even for values that are less extreme and even for highly extreme values as demonstrated by a study done by (Çelik et al., 2011). However, finding the epsilon and minpts for each cluster for a dataset can be very difficult. (Thang & Kim, 2011) introduced DBSCAN-MP algorithm with a way of finding DBSCAN parameters for multiple clusters while utilizing DBSCAN approach and that could be applied on dynamic data updated overtime. This method however had high false positive rate when data environment changed overtime. A brilliant idea to introduce automated process (Akbari & Unland, 2016) to determine the input parameters (Eps and Minpts) works perfectly for datasets with known distributions.

K-means (Wu, 2012; Zhong, 2005) like DBSCAN is a clustering approach method applicable as an outlier detection method. It is a popular method for clustering huge datasets. K-means out-perform k-means++ and the mini-batch K-means both at quality approximation and relationship between number of distance computations (Capó et al., 2017). (Wishart, 2003) addressed practical issues in k-means cluster analysis on segmentation with mixed types of variables and missing values. K-means can effectively perform clustering and outlier detection concurrently(Chawla & Gionis, 2013).

Isolation Forest (Liu et al., 2008) is a model based approach that performed favorably in terms of AUC and processing time compared to one-class SVM, ORCA, Local Outlier Factor(LOF) and random forests for large datasets. This method isolates instances instead of profiling them irrespective of distance and density. It exploits sub-sampling not utilized in any of the existing methods (depth-based, distance-based, density-based, model-based and link-based) which handles the masking and swamping problems (Chiang et al., 2007). It does not use distance or density measures to detect anomalies and hence eliminates the computation cost of distance and density computation. (Ding & Fei, 2013) describes it as the best performing method to achieve linear and space complexities. It can detect anomalies surrounded by normal points. No further adjustments are done on the basic measure to detect scattered or clustered anomalies unlike distance and density-based method.

One of the strengths of unsupervised learning(Ghahramani, 2004) is the ability to flag patterns without prior information. (Yamanishi et al., 2004) used Smart Sifter engine utilizing the unsupervised techniques on on-line data and some of the advantages included adaptation to non-stationary data, low computational costs and the ability to handle both categorical and continuous data. Such methods are most appropriate for changing data environment. A good example is identifying network intrusion attacks, (Zhang & Zulkernine, 2006) used unsupervised random forest algorithm to overcome the drawbacks of supervised learning – using anomaly free training data not applicable in real-data and the need to have a predefined pattern which is depended on network vendor testing protocol (Jyothsna et al., 2011). Changing patterns and network environment dynamics favors this approach.

Outlier detection methods have been suggested for numerous applications, such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, severe weather prediction, geographic information systems, athlete performance analysis, and other data-mining tasks (D. M. Hawkins, 1980; Barnett & Lewis, 1984; Acuna & Rodriguez, 2004).

3 Methods

3.1 Data description

The data used for analysis comprised of two sets obtained from a Clinical Information Network (CIN) database. Clinical Information Network is a collaborative project between KEMRI – Wellcome Trust Research Programme and the Ministry of Health. CIN comprises of 20 county referral hospitals in Kenya. The data is routinely collected in each hospital from the paediatric inpatient unit for children under the age of 5 years. Data collection began in September 2013 to present.

The attributes in the data include bio data, history of illness, admission diagnosis, discharge diagnosis and treatment indicators.

The target period for the analysis is September 2013 to December 2019. The period is based on the initial data collection period and the latest double data entry period performed.

The data sets are

1. **Original dataset**: Original dataset contains the data keyed into electronic system by the data clerks in each hospital. The source document of the original dataset is the Paediatric Admission Record(PAR).
2. **Double data entry dataset:** Double data entry dataset contains randomly sampled data from the original dataset that were re-entered by an auditing clerk. The data was periodically collected after every quarter in a year. Double data entry set represents 0.7% of the original dataset. The small sample size is explained by occurrence strikes in hospitals that interrupted periodical audit periods. Additionally, some hospitals were introduced into the Clinical Information Network as late as 2019.

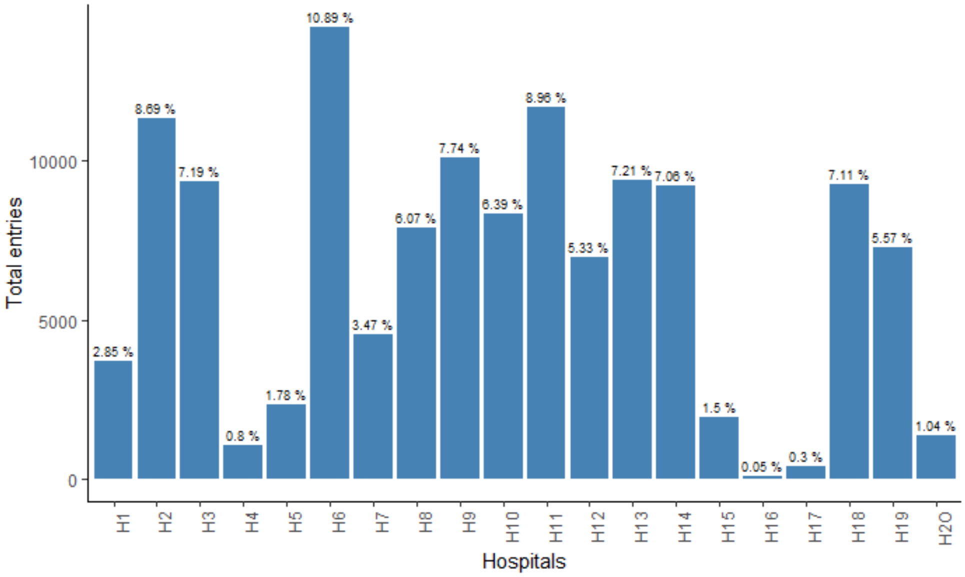
The table 1 describes the size of the datasets, number of observations and type of variables.

Table 1. Data structure

	Original dataset	Double data entry dataset
Observations	130426	907
Variables	288	288
Variables data types'		
Categorical	213	213
Continuous	53	53
Dates	22	22

Figure 1 shows the number of records per hospitals. The x axis displays the hospitals' identifiers and the y axis represents the total number of records. Each bar is labelled with the percentage of the hospital records when compared to the cumulative records.

Figure 1. Total records per hospital from September 2013 to December 2019



3.2 Data preprocessing

Data preprocessing is a process of transforming data into a state that can be interpreted by an algorithm. The analysis data consists of categorical, continuous and dates data types.

3.2.1 Categorical data

Each categorical variable was transformed into dummy variables. The dummy variables expanded the feature-space to **936** variables(McNamara & Horton, 2018; McKinney, 2012). For each categorical variable with k levels, k-1 dummy variables were defined.

3.2.2 Date types

All dates were converted into numeric values by computing the number of days difference from 2013-09-01. Each of the values was normalized to have a range from 0 and 1. We converted dates into numeric since our K-means and isolation forests required standardized values.

The normalization equation is:

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where;

- x is the data point value
- x_{scaled} is the normalized value ranging from 0 and 1
- $max(x)$ is the maximum value of the variable
- $min(x)$ is the minimum value of the variable

3.2.3 Continuous variables

Continuous data points were normalized using equation 1. Unique identifiers of each record were removed from the dataset prior to analysis then later merged with the scores of each observation. The identifiers removed were Inpatient ID, data clerk ID and record ID.

3.2.4 Handling missing values

Missing numerical values were replaced with -1 while missing categorical values were replaced with a category called missing. The value -1 was based on the data entry protocol document where each variable missing documentation was keyed in as -1. Variables with more than 10% missingness were excluded from the analysis dataset. 38.9% of all the variables had more that 10% missing values.

The table 5 shows the percentage of variables with missing values.

Table 2. Percentage of variables with missing values

p_na > 90	p_na < 90
252(38.9%)	648(61.1%)

Multiple imputation (Sterne et al., 2009) for missing values was not done on both datasets since we expected some variables to be missing. These variables are dependent on other variables' branching logic at the point of entry. For instance, admission diagnosis is captured captured if a patient had severe key symptoms.

3.2.5 Feature scaling

Feature scaling is performing transformations on the data such that it can be easily accepted as input for machine learning algorithms while still retaining its original meaning.

Machine Learning(ML) algorithms consider all features on an even range of values hence the need to transform all data points to the same scale. This process is called feature scaling. This is a significant step for unsupervised machine learning models because uneven data values have a significant impact on performance of the algorithm. Equation 1 elaborates mathematical logic of this process where values are normalized to have a range of 0 and 1.

3.2.6 Dimensionality reduction

The curse of dimensionality (Shultz et al., 2011; Johnstone & Lu, 2009) refers to the phenomena that data analysis tasks become significantly harder as the dimensionality of the data increases. As the dimensionality increases, the number planes occupied by the data increases thus adding more and more sparsity to the data which is difficult to model and visualize(Marimont & Shapiro, 1979).

Dimensionality reduction maps the dataset to a lower-dimensional space. The objective is to reduce the dimensions of a dataset by creating new features which are a combination of old features.

We used Principal Component Analysis (PCA) to reduce the feature space. We find the optimal number of components which capture the greatest amount of variance in the data.

Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction technique used to emphasize variation of principal components. The steps for computing PCA is outlined in the appendix B section.

3.3 Outlier detection

3.3.1 K – Means implementation

$k-means$ clustering is an unsupervised machine-learning technique used to identify clusters in a dataset. Clustering is the process of partitioning data into groups while clusters are groups of data objects that have homogeneous properties in their group. (Li & Wu, 2012; Pamula et al., 2011; Wu, 2012). A variable k is defined as the number of clusters.

K-means clustering was implemented using Python 3.7 programming language using scikit-learn 0.23.2 package.

The figure 2 and 3 (Scikits-learn) shows a representation of data before clustering and clustered data.

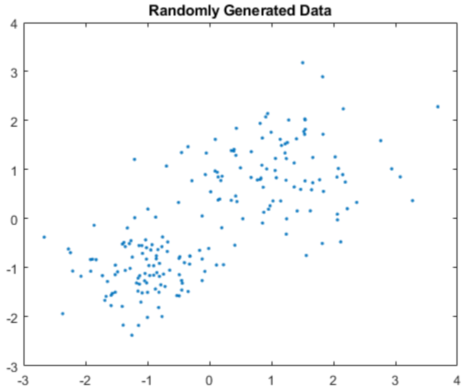


Figure 2. Unclustered data.

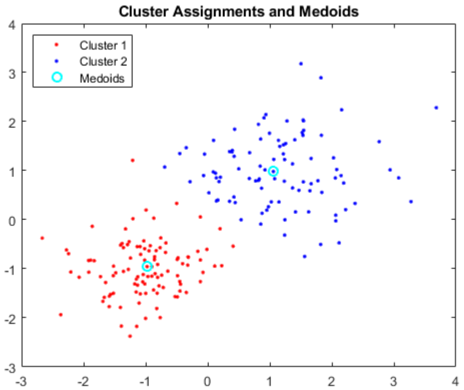


Figure 3. Clustered data.

The aim of aim of $k-means$ is to minimize the squared error function given in the equation 2.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i - \mu_j||^2 \tag{2}$$

where;

- n is the number of data points in the j^{th} cluster
- k is the number of cluster centers
- $||x_i - \mu_j||$ is the euclidean distance between x_i and the centroid μ_j

Steps for $k - means$ clustering

1. Randomly select k centroids, where k is equal to the number of clusters. Centroids represent the mean of each cluster (data point representing the center of a cluster).

$$\mu_1, \mu_2, \ldots, \mu_k$$

2. Calculate the euclidean distance between each data point and cluster centers

$$||x_i - \mu_j||$$

3. Assign each data point to the cluster whose distance from the cluster center is minimum of all clusters centers.

4. Recalculate the new cluster center using the equation 3

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

where;

- x_i is a data point in cluster j
- n represents the number of data points in the j^{th} cluster.

5. Recalculate the distance between each data point and the obtained clusters in step 4.

6. Repeat the steps 3,4 and 5 until no data point is reassigned and no change of the centroid.

Choosing the appropriate number of clusters

A combination of elbow, silhouette coefficient and grid search hyper parameter tuning were used to determine the optimal value of k.

Grid search hyper parameter tuning

Hyperparameter refers to a model configuration argument that guides the learning process for a dataset. The hyperparameters are manually set.

Grid search defines a search space as a grid of hyperparameter values and evaluate every position in the grid. Grid-search is used to find the optimal hyperparameters of a model which results in the most 'accurate' predictions.

In order to achieve an optimal model architecture, we iterated a range of possible estimators.

We varied different number of principal components against different clusters to study the pattern of the r-squared scores. We then picked the pair of inputs that gave the highest r-squared value.

3.3.2 Isolation Forest and its implementation

Isolation forest is an unsupervised ML method for anomaly detection that isolates anomalies instead of profiling normal points. Isolation forest was initially proposed by (Liu et al., 2008). According to the author, anomalous data points are few and different from other data points. Anomalies are easier to isolate compared to normal points.

Figure 4 indicates that x_0 is easier to isolate compared to x_i. Thus, x_0 is an anomaly while x_i is a normal data point.

Isolation forest builds an ensemble of isolates Trees(iTrees) from the dataset. In every tree, anomalies are points that have shorter average path lengths.

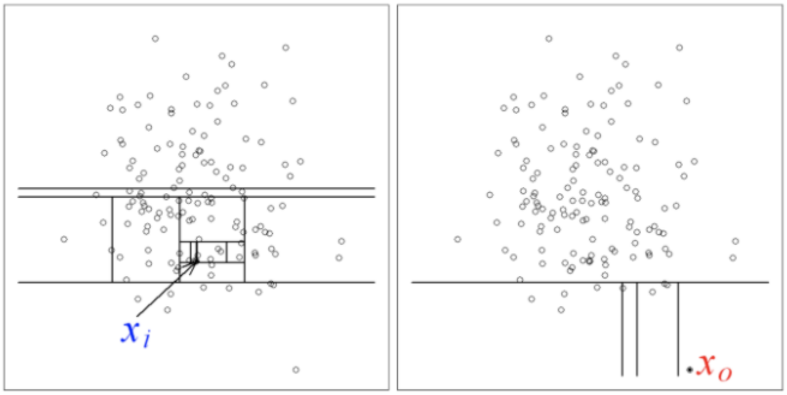


Figure 4. Isolation forest partitioning

Data points in a sample are partitioned repeatedly in a recursive manner by selecting an attribute randomly then randomly selecting a split value for the attribute. Values between minimum and maximum values are allowed for that attribute. This random partitioning results in shorter paths for anomalies such that

- The fewer instances of anomalies result in a smaller number of partitions – shorter paths in a tree structure
- The instances with distinguishable attribute-values are more likely to be separated in early partitioning.

Random trees that collectively produce shorter path lengths for data points in a forest are possible anomalies.

Isolation forest steps

iTree is a structure representing the recursive partitioning. ***Path length*** represents the number of partitions required to isolate a point within a tree. It is the length to reach a terminating node from the root of the iTree.

We use the path length to measure the degree of susceptibility to isolation (Liu et al., 2012), such that short path length means high susceptibility to isolation and long path length means low susceptibility to isolation.

Let $X = x_1, x_2, \ldots, x_n$ of n observations with a set of $d - dimensional$ points and $X^{'} \subset X$. T as a node in the iTree such that T is an external node with no child or an internal with one test and exactly 2 daughters(T_l, T_r). A test has attributes q and p such that a data point is divided into T_l or T_r depending on the test $q < p$.

X is recursively divided by randomly selecting q and a split value p until either the iTree reaches a height limit or $|X| = 1$ or all data points in X have the same values.

The path point $h(x)$ of a point x is measured by the number of edges x traversed an iTree from the root until it is terminated at an external node.

Average path length equation,

$$c(n) = 2(H(x)) - 2\frac{(n-1)}{n}$$

$H(x)$ is a harmonic number that is estimated by Euler's constant and n as the size of a sample set:

$$H(x) = \ln(i) + 0.5772$$

The anomaly score s of a point x is defined as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{4}$$

Where $E(H(x))$ is the average of $h(x)$ from a collection of iTrees.

The anomaly score for each tree and average them out across different trees and get the final anomaly score for an entire forest for a given data point

The evaluation stages are explained in the appendix C section.

3.3.3 Two pass verification

Two pass verification is a data quality assurance method that uses two passes of data entry. The first pass, records are entered to an electric system. The second pass involves keying the same set records by a verifier. The outcome is two comparable datasets.

Let $X_1 = x_1, x_2, \ldots, x_n$ and $X_2 = x_1, x_2, \ldots, x_n$ represent two datasets where X_1 is the first pass dataset has n observations and X_2 is the second pass dataset with n observations and d dimensions.

A dimension score of s_d for each pair of observation is obtained using equation 5.

$$s_d = \begin{cases} 1 \; if \; X_{1i} = \; X_{2i} \\ 0 \quad otherwise \end{cases} \tag{5}$$

The anomaly score, S, for an observation is given by equation 6.

$$S = \frac{1}{d} \sum_{d=1}^{d} s_d \tag{6}$$

3.4 Evaluation metrics

Evaluation metrics are used to measure the quality of a statistical model. The coefficient of determination and normalized mutual information metrics were used in our analysis. The objective was to determine the strength of correlation between two pass verification scores, k-means scores and isolation forest scores.

3.4.1 The coefficient of determination

The coefficient of determination is a statistical measurement that assesses how strong the linear relationship is between two variables.

The coefficient of determination values that tends towards 1 indicates a strong correlation between two variables.

An observation had three scores,

- $K-means$ score, K_S
- Two pass verification score, D_s
- Isolation forest score, F_s

A linear regression model was used to determine the correlation coefficient of the anomaly scores. Equation 7 shows a linear regression model for two pass verification anomaly scores and k-means anomaly scores.

$$D_s = \beta_0 + \beta_1 K_s + \varepsilon \tag{7}$$

Equation 8 shows a linear regression model for two pass verification anomaly scores and isolation forest anomaly scores.

$$D_s = \beta_0 + \beta_1 F_s + \varepsilon \tag{8}$$

The coefficient of determination equations were computed for equations using 9 and 10.

$$r_k = \frac{n\left(\sum K_s D_s\right) - \left(\sum K_s\right)\left(\sum D_s\right)}{\sqrt{\left[n\sum {K_s}^2 - \left(\sum K_s\right)^2\right]\left[n\sum {D_s}^2 - \left(\sum D_s\right)^2\right]}} \tag{9}$$

$$r_f = \frac{n\left(\sum F_s D_s\right) - \left(\sum D_s\right)\left(\sum F_s\right)}{\sqrt{\left[n\sum {D_s}^2 - \left(\sum D_s\right)^2\right]\left[n\sum {F_s}^2 - \left(\sum F_s\right)^2\right]}} \tag{10}$$

where:

- r_k is the coefficient of determination value of $k-means$ anomaly scores and two pass verification anomaly scores.
- r_f is the coefficient of determination value of isolation forest scores and two pass verification anomaly scores.

3.4.2 Normalized mutual information

Mutual information (Corso et al., 2020) is a measure of similarity between two labels of the same data. Mutual information is symmetric and independent of the absolute values of the two sets in comparison. Mutual information is a good measure of the level of agreement of two independent labels assignments on the same dataset when the ground truth is not known (Amelio & Pizzuti, 2015).

$Normalized\ Mutual\ Information\ (NMI)$ is a normalization to scale the results between 0 and 1. 0 for no mutual information and 1 for perfect correlation.

Perfect labels are both homogeneous and complete if the NMI scores are 1.0. NMI scores that tends towards zero indicate incomplete and lack of homogeneity while NMI scores that tends closer to 1 indicates a strong correlation.

Equation 11 was used to compute the NMI scores.

$$NMI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|} \tag{11}$$

Where:

- U represents two pass verification anomaly scores
- V represents $K-means$ anomaly scores or isolation forest anomaly scores
- $|U_i|$ is the number of samples in set U_i
- $\left|V_j\right|$ is the number of samples in set V_j

4 Data analysis and results

4.1 Exploratory Data Analysis(EDA)

This section indicates general attributes of the data. The structure of the two datasets used are as outlined below.

Data types of variables in the dataset before and after data pre-processing are shown in table 3 and table 4 respectively.

Table 3. Raw data types

character	factor	integer	logical	numeric
136(21%)	425(66%)	28(4%)	3(0.46%)	56(8.6%)

Table 4. Processed data types

Categorical	Continuous
213(73.9%)	75(26.04%)

Table 3 indicates the data types in the dataset before data cleaning and feature scaling and Table 4 indicates the final data types used for analysis. 213 categorical variables were expanded as dummy variables to form additional 648 variables. 136 character variables were excluded from the analysis.

4.1.1 Data distribution over time

The original dataset has data entry rate over time per hospital as shown in the figure 5.

Figure 5 indicates the distribution of data per over the selection period (Sept 2013 – Dec 2019).

The colored trends on the chart shows the number of entries captured over time per hospital. The x-axis indicates the time of data entry and the y-axis shows the total entries per day.

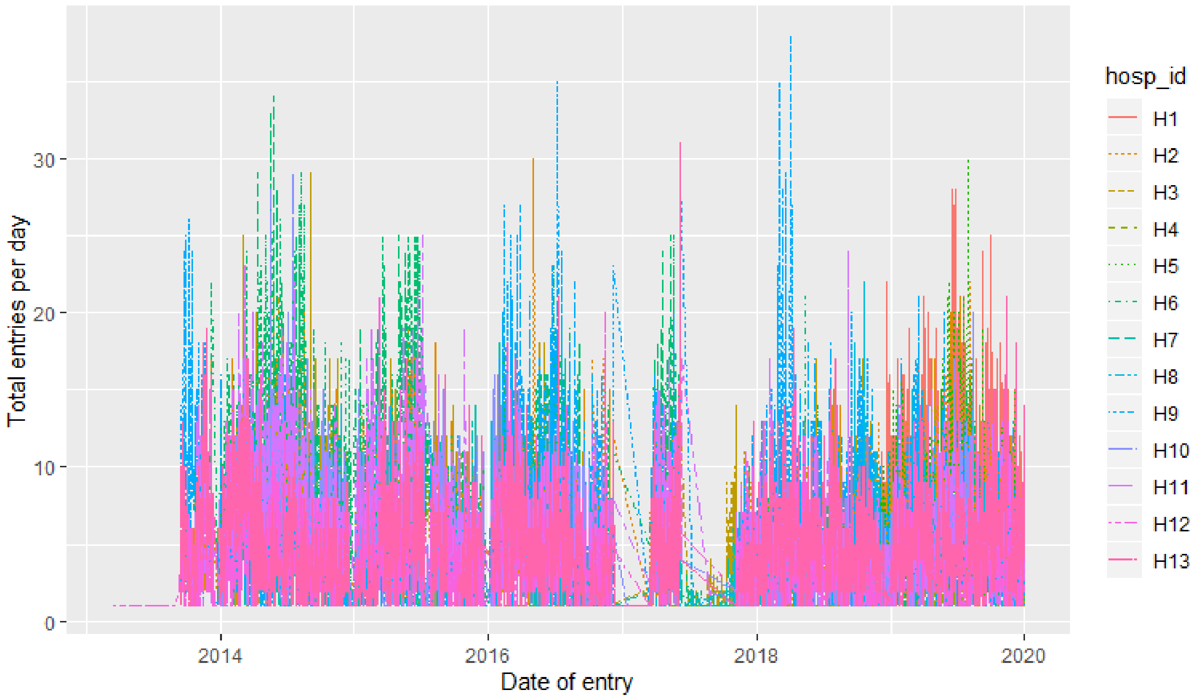


Figure 5. Data collection rate since september 2013

We explored the distribution of observations for each hospital per season. See the facet figure 6 .

Hospital H6 is the first hospital to be introduced into the Clinical Information Network(CIN) while H5 is a hospital introduced later in the year 2019 hence the fewer observations.

There is a gap for the year 2017 since this is the year that Kenyan hospitals experienced strikes mostly throughout the year.

Figure 6, indicates the period some hospitals were introduced into the Clinical Information Network.

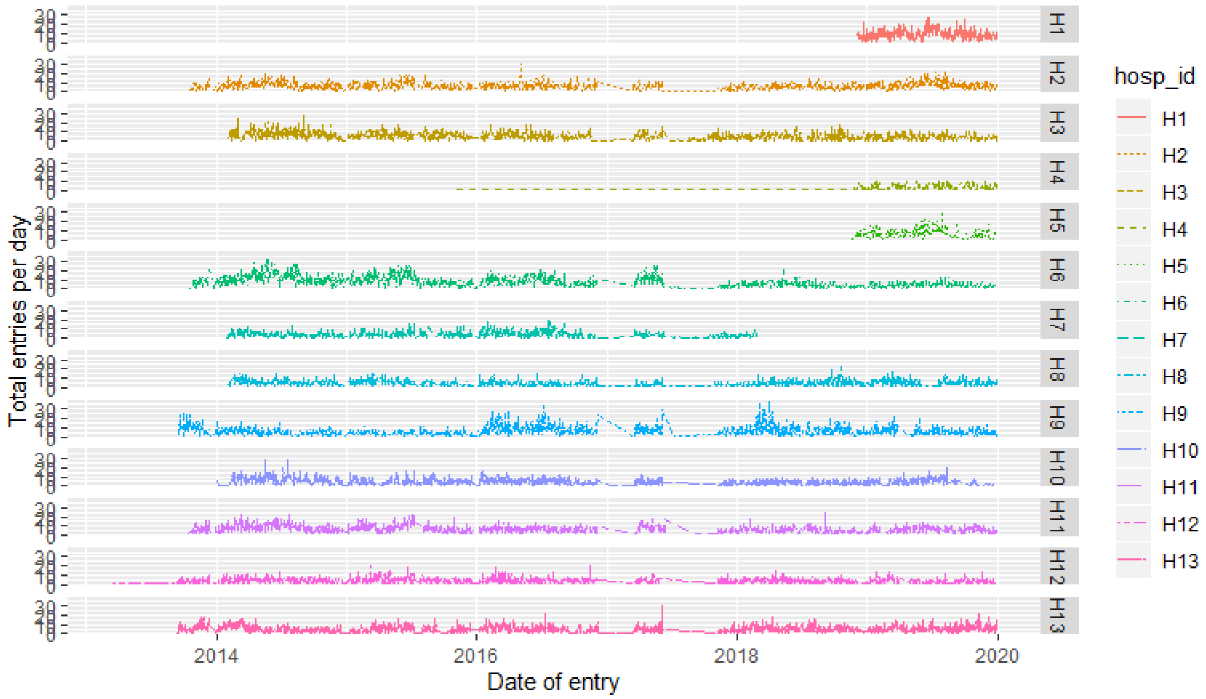


Figure 6. Seasonal data entry per hospital

4.1.2 Missingness

Variables with missing values $> 90\%$ represented 38.8% of all the variables.

Table 5. Missing values percentage per variable

p_na > 90	p_na < 90
252(38.9%)	648(61.1%)

From table 6,the column of interest is *p_na* which represents the percentage of missing data points from all observations for each specific variable.

Table 6. Percentage of missingness per variable

	variable <chr>	q_zeros <int>	p_zeros <dbl>	q_na <int>	p_na <dbl>	type <fctr>	unique <int>
1	id	0	0	0	0.00	integer	137243
2	doc_source	0	0	3154	2.30	factor	3
3	surgical_burns	0	0	3154	2.30	factor	2
4	date_adm	0	0	0	0.00	character	2266
5	date_discharge	0	0	0	0.00	character	2231
6	hosp_id	0	0	9	0.01	factor	20
643	dsc_rx5	0	0.00	28331	20.64	character	249
644	dsc_rx_other1	0	0.00	40523	29.53	character	1082
645	dsc_rx_other2	0	0.00	77700	56.61	character	849
646	rx_nt_listed	0	0.00	91803	66.89	factor	2
47	rx_free_text	0	0.00	77910	56.77	character	675
648	discharge_information_complete	87	0.06	0	0.00	integer	3

4.2 Two pass verification scores

The level of agreement scores for the double data entry dataset and the original dataset were visualized using a histogram and box plot.

The overall mean value for two pass verification scores is 91.4% indicating that 91.4% of all data points matched and only 9% of all the values were discordant and were subjected to cross-validation.

The pair-values with a score of less than 0.7 as seen in figure 7 would be considered outlying observations. Variable specific score for each observation is compared across all study sites. Poor performing variables are used to determine the measures taken after audit period.

Figure 7 shows anomaly scores in the x-axis and hospitals in the y-axis. Outlying observations are colored red while inliers are colored green.

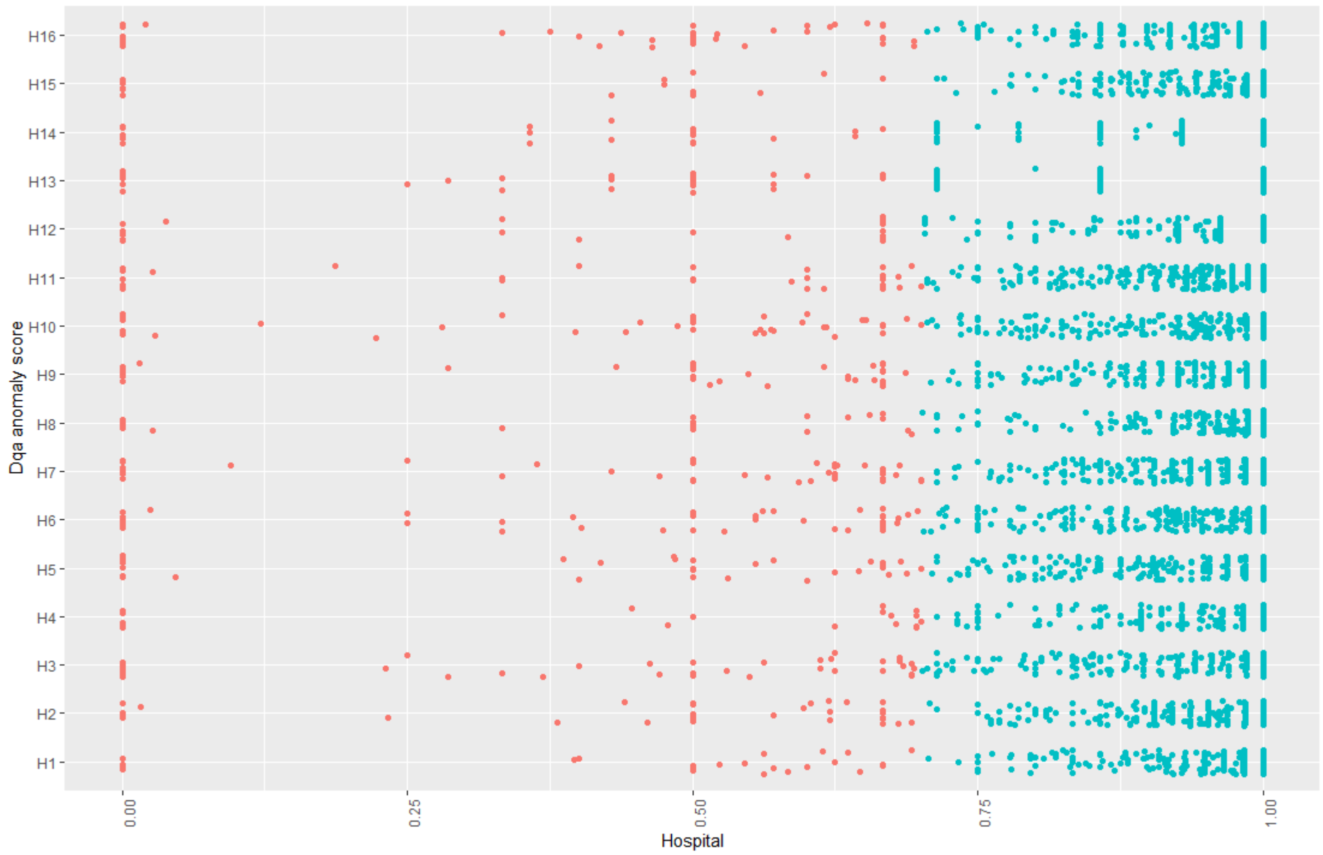


Figure 7. Outlying observations

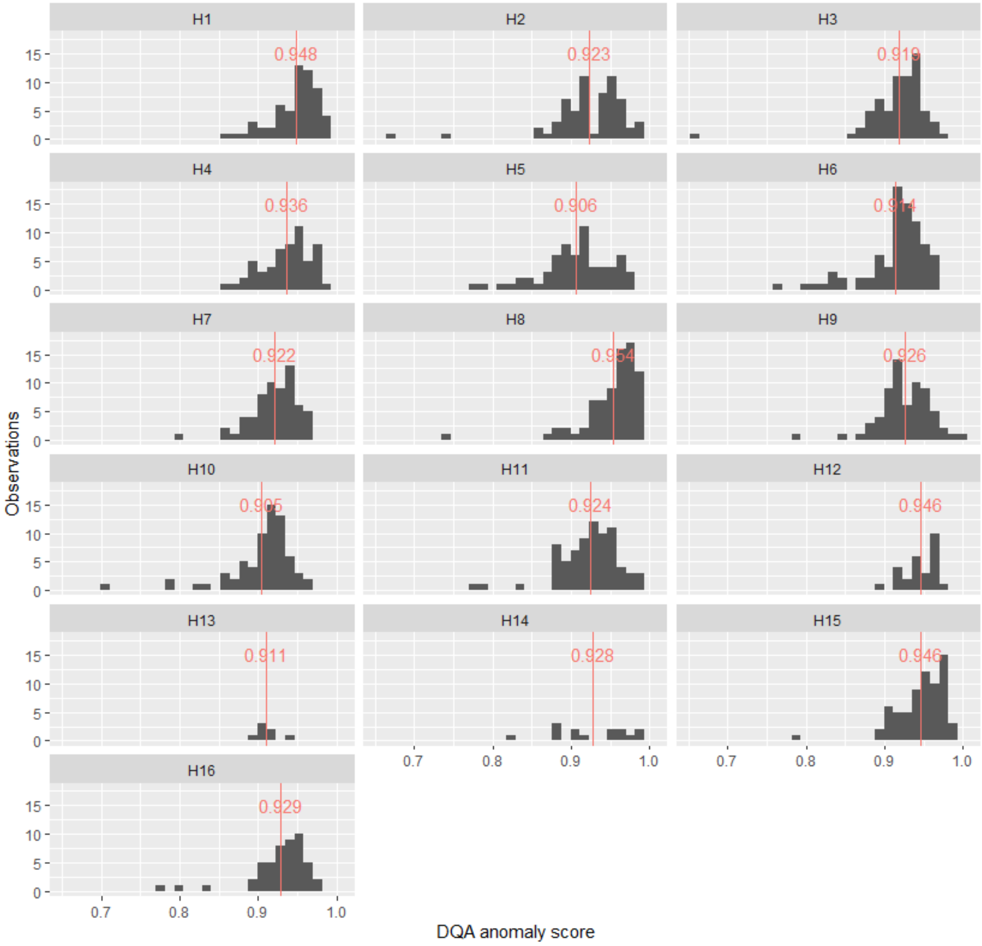


Figure 8. Distribution of two pass verification anomaly scores per hospital

We compared the anomaly score for each hospital as shown in figure 8. The x-axis on the facet describes the anomaly score with the red line showing the mean anomaly score per hospital. The y-axis shows the total observations.

If the observations scores were plotted for each hospital using boxplots, each hospital has an average score of more than 87%. Approximately 13% of all observations in each hospital are anomalous. Outlying observations are shown dotted in figure 9.

The x-axis in figure 9 shows anomaly scores per year. The overall anomaly score is shown by the red line. The y-axis shows hospitals represented in each year facet. There is an improvement of the average data quality scores from 2013 to 2020 as shown in figure 9.

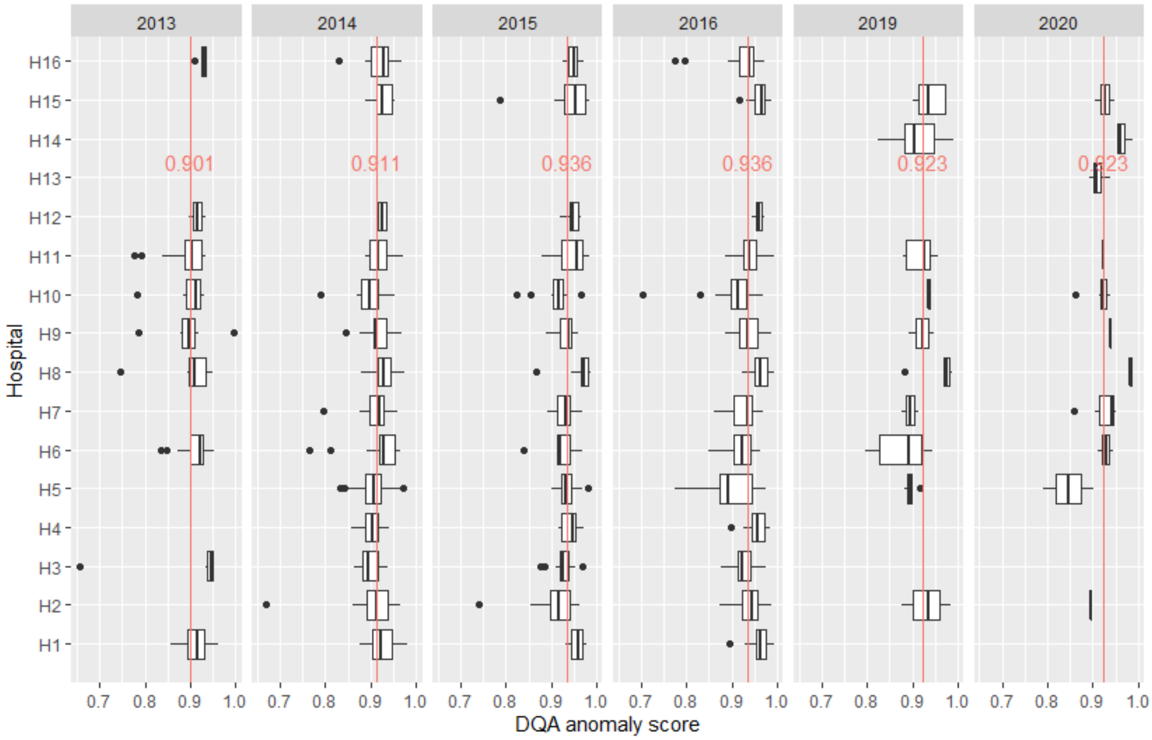


Figure 9. Box plot representation of two pass verification scores per year across all hospitals

Appendix A shows variable specific scores per hospital. Values missing in both entries are indicated as NA. Hospital specific score of each variable are indicate in each column.

4.3 K-means clustering

In this section, we present the correlation scores for k-means algorithm and two pass verification scores.

Hyper-parameter tuning results and how the estimators were chosen are shown in table 7.

Table 7 shows the coefficient of determination scores for each set of parameters. Principal components were varied from 50 to 300 while incrementing clusters from 50 to 500.

Table 7. The coefficient of determination values for principal components and their corresponding cluster size

clusters	pca_50	pca_100	pca_150	pca_200	pca_250	pca_300
50	0.533686	0.497862	0.429951	0.411575	0.349443	0.360413
100	0.576042	0.554289	0.518916	0.398479	0.42901	0.376849
150	0.639485	0.596997	0.555006	0.539908	0.4117	0.539908
200	0.616659	0.624934	0.578606	0.509724	0.429164	0.426788
250	0.641621	0.639828	0.584587	0.506819	0.461169	0.368224
300	0.631581	0.634313	0.642524	0.524931	0.500223	0.462206
350	0.694614	0.636302	0.581976	0.548	0.503347	0.40689
400	0.66686	0.621541	0.599777	0.533459	0.554847	0.398209
450	0.689356	0.634779	0.594055	0.553857	0.479269	0.533014
500	0.662366	0.630557	0.605851	0.588783	0.48916	0.457992

We picked 50 principal components in combination with 350 clusters since it had the highest score of 0.694614.

Figure 10 to figure 15 were used to visualize the behavior of the coefficient of determination scores. The x-axis in each plot represent the number of clusters while the y-axis represent the number of principal components. The plotted line shows the coefficient of determination scores for each pair of principal component and cluster size.

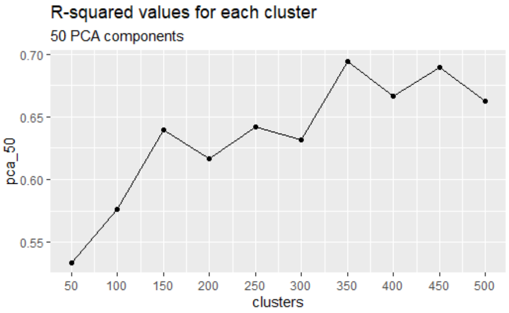


Figure 10. R-squared scores for 50 Principal components

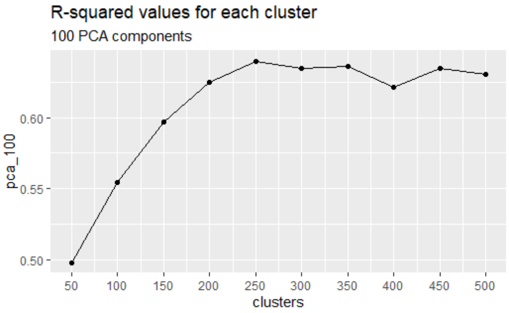


Figure 11. R-squared scores for 100 Principal components

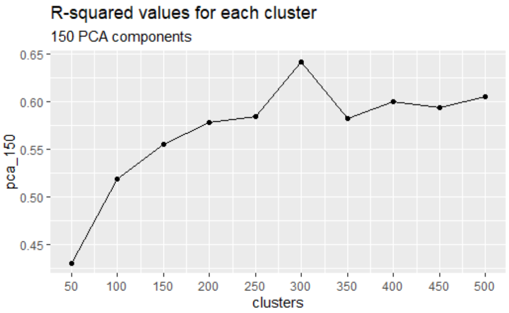


Figure 12. R-squared scores for 150 Principal components

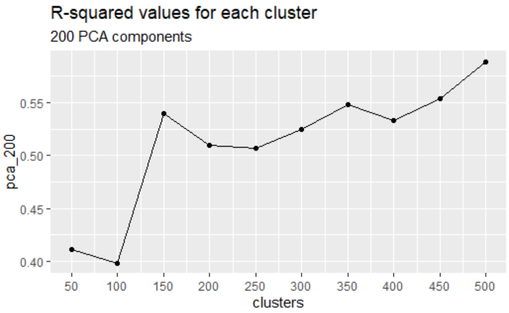


Figure 13. R-squared scores for 200 Principal components

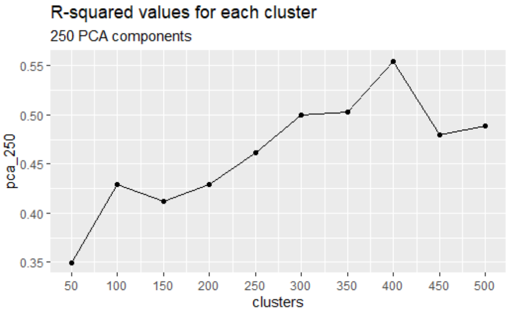


Figure 14. R-squared scores for 250 Principal components

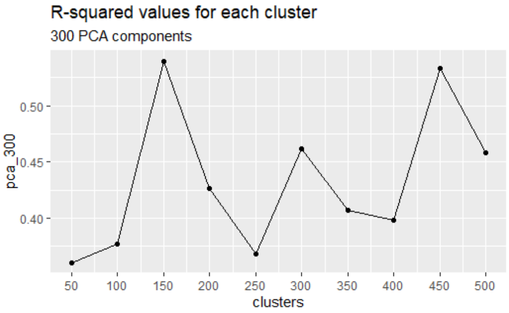


Figure 15. R-squared scores for 300 Principal components

k-means and two pass verification anomaly scores correlation

Figure 16 shows scatter plot of k-means anomaly scores and two pass verification anomaly scores. The x-axis shows the k-means scores while the y-axis represent the two pass verification scores. The linear regression line is plotted through the plot with an r-squared value of 0.694641.

There is a relationship as demonstrated in the figure 16. We expected to have a strong linear relationship to indicate a stronger correlation, r-squared score of *0.694614* indicates a good relationship but not perfect.

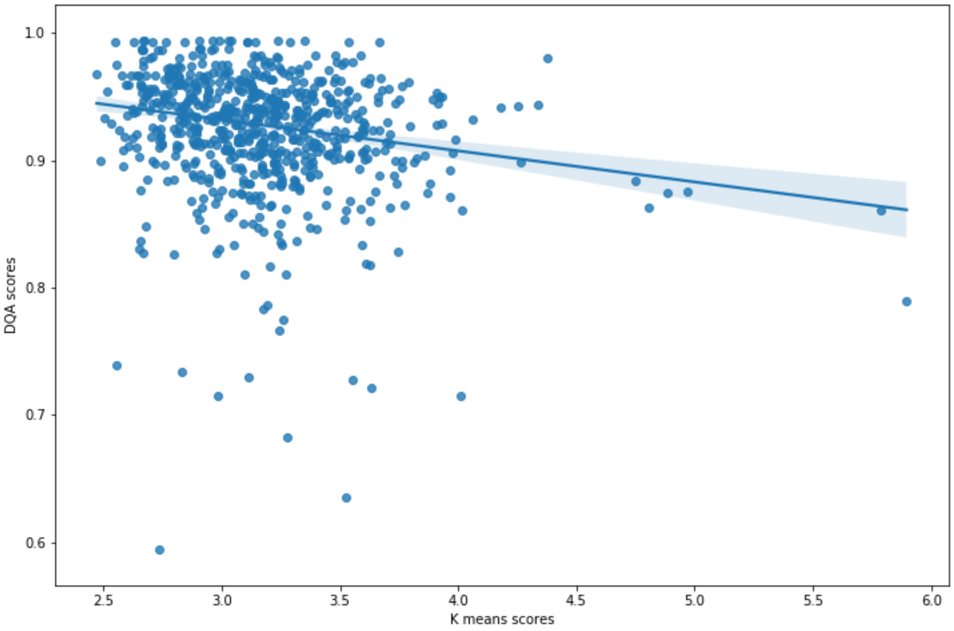


Figure 16. K-means vs two pass verification scores

We re-evaluated the scores using an alternative machine learning metric, normalize mutual information score.

We used Normalized Mutual Information (NMI) score to evaluate the level of agreement between the two pair of scores.

We obtained an NMI score of 0.937.

$$NMI\ score = \ 0.937$$

This indicates that there is a strong correlation between $k - means$ anomaly scores and two pass verification anomaly scores.

Table 8 shows the tabular comparison of the correlation strength.

Table 8. Evaluation metrics for k-means anomaly scores vs two-pass verification

R-squared score	NMI score
0.694641	0.937

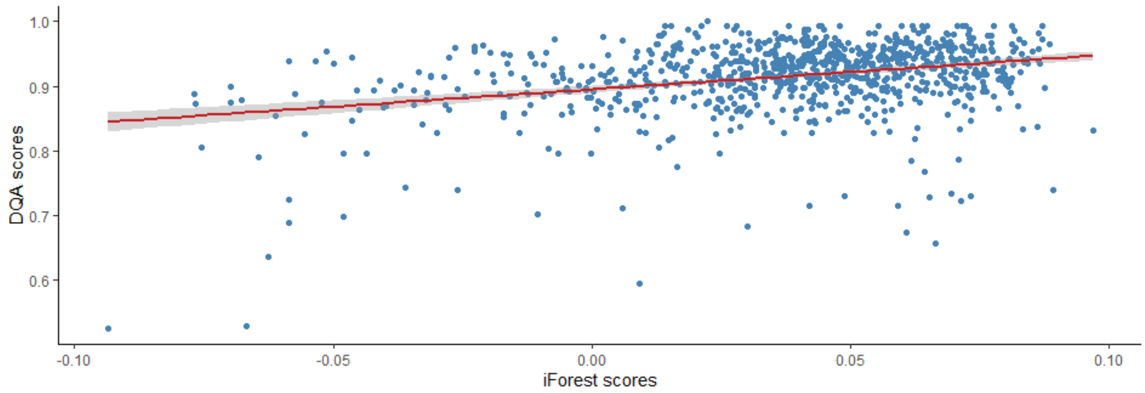


Figure 17. Isolation forest vs two pass verification scores correlation plot

4.4 Isolation Forest

Isolation forest anomaly scores are plotted against two pass verification anomaly scores. The coefficient of determination was computed to determine the strength of the relationship.

The figure 17 shows a plot for *iForest scores* and two pass verification scores with an r-squared value of *0.7189*. The x-axis shows isolation forest anomaly scores while the y-axis shows the two pass verification anomaly scores.

Table 9. Evaluation metrics for isolation forest anomaly scores vs two-pass verification

R-squared score	NMI score
0.7189	0.9843

The coefficient of determination score of 0.7189 indicate a strong relationship between isolation forest scores and two pass verification scores.

Normalized mutual information score metric gave a score of 0.9843 when two pass verification scores were compared with isolation anomaly scores.

$$NMI\ score = 0.9843$$

NMI score indicates that isolation forest anomaly scores have a stronger level of agreement with two pass verification scores.

Table 10 shows a comparison of the normalized mutual information metric and the coefficient of determination. DDE scores represent the two pass verification anomaly scores.

Isolation forest had a higher correlation score for both metrics compared to k-means clustering.

Table 10. K-means clustering vs Isolation forest

Relationship	R^2 *scores*	*NMI scores*
k-means vs DDE scores	0.694614	0.9370
iForest vs DDE scores	0.7189	0.9843

5 Conclusions and recommendations

Two pass verification is a gold standard method that is used to determine the quality of a dataset. Two pass verification demonstrated data quality improvement over time from the year 2013 to 2020. Data quality assurance leads to good quality data.

From the results, unsupervised machine learning outlier detection methods can be alternative methods for ensuring good quality data. The outlying observations obtained from k-means clustering or isolation forest can be subjected to further verification process. The verification process will aim to find the reasons for outlying observations. Checking outlying observations narrows down the number of observations to be cross-validated against the source document. This will minimize time taken to check for errors in a dataset.

The use of k-means and isolation forest methods are less tedious and can be applied to large datasets with continuous, categorical and date-time data types.

We compared k-means clustering with isolation forest performance and found that isolation forest gave a higher correlation scores. K-means clustering performance relies on the choice of the optimal number of clusters for each dataset and how accurate hyper-parameter searching is done.

Normalized mutual information metric proved to be the best metric to determine the level of agreement between two groups of datasets with different labels.

K-means and isolation forest for outlier detection for data quality assurance can still be improved. This study recommends the use of unsupervised machine learning algorithms for data quality assurance in future.

Further work

Further work would be to examine the outlying observations obtained from k-means clustering and isolation forest. It will be crucial to know if the observations have any similarity.

Additionally, isolation forest can be tested on streaming large datasets to detect anomalies.

Bibliography

Abe, N., Zadrozny, B., & Langford, J. (2006). Outlier detection by active learning. In *Proceedings of the 12th acm sigkdd international conference on knowledge discovery and data mining* (pp. 504–509).

Acuna, E., & Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez*, 1–25.

Akbari, Z., & Unland, R. (2016). Automated determination of the input parameter of db-scan based on outlier detection. In *Ifip international conference on artificial intelligence applications and innovations* (pp. 280–291).

Alfons, A., Templ, M., Filzmoser, P., & Holzer, J. (2010). A comparison of robust methods for pareto tail modeling in the case of laeken indicators. In *Combining soft computing and statistical methods in data analysis* (pp. 17–24). Springer.

Amelio, A., & Pizzuti, C. (2015). Is normalized mutual information a fair measure for comparing community detection methods? In *Proceedings of the 2015 ieee/acm international conference on advances in social networks analysis and mining 2015* (pp. 1584–1585).

Azeroual, O., Saake, G., & Schallehn, E. (2018). Analyzing data quality issues in research information systems via data profiling. *International Journal of Information Management*, *41*, 50–56.

Barbará, D., & Chen, P. (2000). Using the fractal dimension to cluster datasets. In *Proceedings of the sixth acm sigkdd international conference on knowledge discovery and data mining* (pp. 260–264).

Bargaje, C. (2011). Good documentation practice in clinical research. *Perspectives in clinical research*, *2*(2), 59.

Barnett, V., & Lewis, T. (1984). Outliers in statistical data. *osd*.

Ben-Gal, I. (2005). Outlier detection. In *Data mining and knowledge discovery handbook* (pp. 131–146). Springer.

Bowman, S. (2013). Impact of electronic health record systems on information integrity: quality and safety implications. *Perspectives in health information management*, *10*(Fall).

Brown, A. W., Kaiser, K. A., & Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences*, *115*(11), 2563–2570.

Büchele, G., Och, B., Bolte, G., & Weiland, S. K. (2005). Single vs. double data entry. *Epidemiology*, *16*(1), 130–131.

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, *14*.

Capó, M., Pérez, A., & Lozano, J. A. (2017). An efficient approximation to the k-means clustering for massive data. *Knowledge-Based Systems*, *117*, 56–69.

Çelik, M., Dadaşer-Çelik, F., & Dokuz, A. Ş. (2011). Anomaly detection in temperature data using dbscan algorithm. In *2011 international symposium on innovations in intelligent systems and applications* (pp. 91–95).

Chawla, S., & Gionis, A. (2013). k-means–: A unified approach to clustering and outlier detection. In *Proceedings of the 2013 siam international conference on data mining* (pp. 189–197).

Chiang, J.-T., et al. (2007). The masking and swamping effects using the planted mean-shift outliers models. *Int. J. Contemp. Math. Sciences*, *2*(7), 297–307.

Corso, G., Ferreira, G. M., & Lewinsohn, T. M. (2020). Mutual information as a general measure of structure in interaction networks. *Entropy*, *22*(5), 528.

Dhankhad, S., Mohammed, E., & Far, B. (2018). Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In *2018 ieee international conference on information reuse and integration (iri)* (pp. 122–125).

Di, M., & Joo, E. M. (2007). A survey of machine learning in wireless sensor netoworks from networking and application perspectives. In *2007 6th international conference on information, communications & signal processing* (pp. 1–5).

Ding, Z., & Fei, M. (2013). An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings Volumes*, *46*(20), 12–17.

Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. *Procedia Computer Science*, *165*, 631–641.

Dupuis, D. J., & Victoria-Feser, M.-P. (2006). A robust prediction error criterion for pareto modelling of upper tails. *Canadian Journal of Statistics*, *34*(4), 639–658.

Escobar, C. A., & Morales-Menendez, R. (2018). Machine learning techniques for quality control in high conformance manufacturing environment. *Advances in Mechanical Engineering*, *10*(2), 1687814018755519.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, pp. 226–231).

Filzmoser, P. (2004). *A multivariate outlier detection method.* na.

Foorthuis, R. (2018). A typology of data anomalies. In *International conference on information processing and management of uncertainty in knowledge-based systems* (pp. 26–38).

Galetsi, P., Katsaliaki, K., & Kumar, S. (2020). Big data analytics in health sector: Theoretical framework, techniques and prospects. *International Journal of Information Management*, *50*, 206–216.

Garrett, R. G. (1989). The chi-square plot: a tool for multivariate outlier recognition. *Journal of Geochemical Exploration*, *32*(1-3), 319–341.

Ghahramani, Z. (2004). Unsupervised learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced lectures on machine learning: Ml summer schools 2003, canberra, australia, february 2 - 14, 2003, tübingen, germany, august 4 - 16, 2003, revised lectures* (pp. 72–112). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from `https://doi.org/10.1007/978-3-540-28650-9_5` doi: 10.1007/978-3-540-28650-9_5

Haug, A., Zachariassen, F., & Van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management (JIEM)*, *4*(2), 168–193.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). Springer.

Hawkins, S., He, H., Williams, G., & Baxter, R. (2002). Outlier detection using replicator neural networks. In *International conference on data warehousing and knowledge discovery* (pp. 170–180).

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, *52*(12), 5186–5201.

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis* (Vol. 5) (No. 8). Prentice hall Upper Saddle River, NJ.

Johnstone, I. M., & Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, *104*(486), 682–693.

Jyothsna, V., Prasad, V. R., & Prasad, K. M. (2011). A review of anomaly based intrusion detection systems. *International Journal of Computer Applications*, *28*(7), 26–35.

Knorr, E. M., & Ng, R. T. (1997). A unified approach for mining outliers. In *Proceedings of the 1997 conference of the centre for advanced studies on collaborative research* (p. 11).

Kumar, D. P., Amgoth, T., & Annavarapu, C. S. R. (2019). Machine learning algorithms for wireless sensor networks: A survey. *Information Fusion*, *49*, 1–25.

Ley, B., Rijal, K. R., Marfurt, J., Adhikari, N. R., Banjara, M. R., Shrestha, U. T., … Ghimire, P. (2019). Analysis of erroneous data entries in paper based and electronic data collection. *BMC research notes*, *12*(1), 537.

Li, Y., & Wu, H. (2012). A clustering method based on k-means algorithm. *Physics Procedia*, *25*, 1104–1109.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413–422).

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *6*(1), 1–39.

Loureiro, A., Torgo, L., & Soares, C. (2004). Outlier detection using clustering methods: a data cleaning application. In *Proceedings of kdnet symposium on knowledge-based systems for the public sector.*

Marimont, R., & Shapiro, M. (1979). Nearest neighbour searches and the curse of dimensionality. *IMA Journal of Applied Mathematics*, *24*(1), 59–70.

McKinney, W. (2012). *Python for data analysis: Data wrangling with pandas, numpy, and ipython.* " O'Reilly Media, Inc.".

McNamara, A., & Horton, N. J. (2018). Wrangling categorical data in r. *The American Statistician*, *72*(1), 97–104.

Pamula, R., Deka, J. K., & Nandi, S. (2011). An outlier detection method based on clustering. In *2011 second international conference on emerging applications of information technology* (pp. 253–256).

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of global health*, *8*(2).

Papadimitriou, S., Kitagawa, H., Gibbons, P., & Faloutsos, C. (n.d.). Loci: Fast outlier detection using the local correlation. In *19th international conference on data engineering* (pp. 315–326).

Paulsen, A., Overgaard, S., & Lauritsen, J. M. (2012). Quality of data entry using single entry, double entry and automated forms processing–an example based on a study of patient-reported outcomes. *PloS one*, *7*(4), e35087.

Pincus, R. (1995). Barnett, v., and lewis t.: Outliers in statistical data. j. wiley & sons 1994, xvii. 582 pp.,£ 49.95. *Biometrical Journal*, *37*(2), 256–256.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 acm sigmod international conference on management of data* (pp. 427–438).

Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using adaboost and majority voting. *IEEE access*, *6*, 14277–14284.

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, *41*(2), 79–82.

Rodríguez-Pérez, G., Robles, G., Serebrenik, A., Zaidman, A., Germán, D. M., & Gonzalez-Barahona, J. M. (2020). How bugs are born: a model to identify how bugs are introduced in software components. *Empirical Software Engineering*, 1–47.

Rosseeuw, P., & Van Zomeren, B. (1990). Unmasking multivariate outliers and leverate points. *Journal of the American Statistical Association*, *85*, 633–639.

Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets* (Unpublished doctoral dissertation). University of Pittsburgh.

Shahid, N., Naqvi, I. H., & Qaisar, S. B. (2015). One-class support vector machines: analysis of outlier detection for wireless sensor networks in harsh environments. *Artificial Intelligence Review*, *43*(4), 515–563.

Shultz, T. R., Fahlman, S., Craw, S., Andritsos, P., Tsaparas, P., Silva, R., & Mueen, A. (2011). Curse of dimensionality. *en. In: Encyclopedia of Machine Learning. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US*, 257–258.

Sodemann, A. A., Ross, M. P., & Borghetti, B. J. (2012). A review of anomaly detection in automated surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 1257–1272.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., … Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, *338*.

Templ, M., Gussenbauer, J., & Filzmoser, P. (2020). Evaluation of robust outlier detection methods for zero-inflated complex data. *Journal of Applied Statistics*, *47*(7), 1144–1167.

Thang, T. M., & Kim, J. (2011). The anomaly detection by using dbscan clustering with multiple parameters. In *2011 international conference on information science and applications* (pp. 1–5).

Vanpaemel, D., Hubert, M., & Dierckx, G. (2008). A robust estimator for the extreme value index of pareto-type distributions. *Computational Statistics & Data Analysis*, *51*(12), 6252–6268.

Williams, G., Baxter, R., He, H., Hawkins, S., & Gu, L. (2002). A comparative study of rnn for outlier detection in data mining. In *2002 ieee international conference on data mining, 2002. proceedings.* (pp. 709–712).

Wishart, D. (2003). K-means clustering with outlier detection, mixed variables and missing values. In *Exploratory data analysis in empirical research* (pp. 216–226). Springer.

Wu, J. (2012). *Advances in k-means clustering: a data mining thinking*. Springer Science & Business Media.

Yamanishi, K., Takeuchi, J.-I., Williams, G., & Milne, P. (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, *8*(3), 275–300.

Zhang, J., & Zulkernine, M. (2006). Anomaly based network intrusion detection with unsupervised outlier detection. In *2006 ieee international conference on communications* (Vol. 5, pp. 2388–2393).

Zhong, S. (2005). Efficient online spherical k-means clustering. In *Proceedings. 2005 ieee international joint conference on neural networks, 2005.* (Vol. 5, pp. 3180–3185).

Ziegel, E. R. (2004). Statistical size distributions in economics and actuarial sciences. *Technometrics*, *46*(4), 499.

Appendices

A Variable specific anomaly scores per hospital

Table 11. Variable specific anomaly score

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
doc_source	1	0.952	1	1	0.906	0.957	0.968	1	1	0.983	1	0.958	1	1	1	1
surgical_burns	1	1	1	1	1	1	1	1	1	0.983	1	1	1	1	1	1
date_adm	0.983	0.873	0.955	0.911	0.875	0.817	0.968	0.909	0.896	0.824	0.92	0.889	1	1	0.912	0.918
date_discharge	0.983	0.937	0.896	0.875	0.812	0.927	0.937	0.896	0.836	0.75	0.84	0.704	1	0.857	0.853	0.837
leave_period	1	1	1	1	1	1	1	1	1	1	1	1	NA	1	1	1
random	NA	NA	NA	NA	NA	0.962	NA	NA	NA	0.8	NA	NA	NA	NA	NA	NA
depid	1	NA	NA	NA	NA	0.824	NA	NA	1	0.75	1	1	NA	NA	NA	0.75
is_minimum	1	1	1	1	0.98	1	1	1	1	0.98	1	1	1	1	1	1
date_today	0	0.016	0	0	0.047	0.024	0	0	0.015	0.029	0.027	0	0	0	0	0
timestamp	1	0.857	1	1	0.875	0.878	0.841	0.831	0.881	0.882	0.88	1	0	0	0.838	1
ipno	0.847	0.921	0.955	0.982	1	0.841	0.921	0.688	0.97	0.397	0.773	0.741	0.286	1	0.971	0.959
child_sex	0.983	0.968	0.925	0.929	0.953	0.963	0.984	0.961	0.985	0.838	0.905	0.923	1	0.929	0.941	0.875
age_recorded	1	1	1	1	1	1	1	1	1	0.926	1	1	1	1	0.974	1
age_less1mnth	1	1	0.985	0.982	0.984	1	1	1	0.985	1	1	1	1	1	1	1
age_days	NA	1	NA	NA	1	1	NA	NA	1	0.5	NA	NA	NA	NA	NA	NA
age_years	0.915	0.758	0.97	0.964	0.857	0.89	0.762	0.961	0.985	0.879	0.947	1	0.857	1	0.91	0.918
age_mths	0.915	0.887	0.925	0.911	0.825	0.72	0.889	0.883	0.94	0.652	0.813	0.926	1	1	0.955	0.837
res_loc	0.797	0.8	0.866	0.768	0.891	0.5	0.841	0.896	0.896	0.649	0.8	0.852	1	1	0.882	0.898
res_dst	0.797	0.8	0.896	0.839	0.938	0.474	0.935	0.974	0.896	0.919	0.88	1	1	1	0.731	1
ref_hosp	0.949	0.81	0.701	0.714	0.625	0.561	0.889	0.922	0.761	0.559	0.933	0.889	NA	0.929	0.868	0.898
ref_hosp_spec	0.966	0.96	0.985	1	1	1	1	0.987	0.94	0.946	0.867	1	1	1	0.971	0.98
readmin_hosp_dcs	1	0.667	NA	1	0	1	0.25	NA	0.5	1	1	NA	1	NA	1	1
readm_hosp	0.831	0.794	0.758	0.732	0.75	0.679	0.825	0.935	0.836	0.706	0.878	0.889	NA	1	0.897	0.735
weight	0.932	0.952	0.896	0.911	0.734	0.805	0.857	0.948	0.821	0.882	0.84	0.926	0.857	0.929	0.941	0.918
height	0.983	0.905	0.97	0.982	0.903	0.959	0.937	0.987	0.896	0.833	0.933	1	1	1	0.985	0.958
whz	0.397	0.381	0	0.446	0.484	0	0.095	0.027	0.433	0.121	0.187	0.038	0	0.571	0.559	0.021
muac	0.966	0.921	0.985	0.964	0.984	0.951	0.952	0.922	0.955	0.779	0.96	0.852	1	0.929	0.956	0.939
vacc_source	0.78	0.857	0.621	0.732	0.891	0.395	0.714	0.883	0.687	0.824	0.8	0.778	NA	0.929	0.882	0.5
vacc_status_text	1	1	0.941	0.963	1	1	0.917	1	1	1	0.875	0.75	NA	1	1	0.9
par	1	1	1	1	0.786	0.972	0.98	1	1	0.967	1	0.958	NA	1	0.983	1
biodata_complete	1	1	1	1	1	1	1	1	1	1	0.987	1	1	1	1	1
lo_illness_	0.914	0.873	0.851	0.982	0.742	0.88	0.825	0.974	0.881	0.864	0.84	0.963	1	1	0.971	0.898
fever	0.948	0.905	0.881	0.946	0.871	0.88	0.905	0.961	0.94	0.955	0.947	0.815	1	0.929	0.985	0.959
fever_dur	0.957	0.786	0.843	0.977	0.6	0.923	0.867	0.929	0.843	0.905	0.902	0.944	0.857	1	0.942	0.969
cough	0.966	0.889	0.925	0.982	0.903	0.907	0.921	0.921	0.955	0.894	0.88	1	0.714	1	0.941	0.959
cough_dur	0.929	0.842	0.833	0.917	0.778	0.95	0.833	0.944	0.795	0.927	0.864	0.917	0.25	0.857	0.895	0.786
cough_2wks	0.897	0.947	1	0.861	0.917	1	0.931	0.972	0.909	0.95	0.814	1	1	1	0.919	0.857
tb_contact	0.828	0.707	0.889	0.9	0.818	0.857	0.73	0.886	0.659	0.85	0.735	0.5	1	1	0.911	0.52

Table 11 continued from previous page

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
diff_breath	0.983	0.921	0.851	0.911	0.823	0.827	0.937	0.934	0.896	0.848	0.905	0.926	0.857	1	0.941	0.939
diarrhoea	0.948	0.905	0.94	0.911	0.839	0.88	0.952	0.961	0.925	0.924	0.932	0.963	0.714	1	0.985	0.959
diarrhoea_dur	0.92	0.958	0.815	0.7	0.789	0.92	0.667	0.941	0.815	0.923	0.9	1	1	1	0.96	0.905
diarrhoea_14d	0.96	0.917	1	1	0.947	1	0.682	1	0.889	1	0.931	1	1	1	0.957	1
diarrhoea_bloody	1	0.917	1	0.95	0.947	1	0.864	0.971	0.926	0.962	0.933	1	1	1	0.958	1
vomits	0.931	0.921	0.896	0.911	0.855	0.893	0.952	0.974	0.91	0.939	0.878	0.778	0.857	1	0.971	0.918
vomit_everything	1	0.914	0.763	0.963	0.833	0.958	0.824	0.935	1	0.969	0.943	0.882	0.8	1	0.882	1
vomit_freq	1	0.667	0.667	0.946	0.889	0.857	0.964	0.5	0.895	0.946	0.939	0.4	NA	NA	0.885	0
diff_feed	0.948	0.905	0.761	0.893	0.726	0.867	0.921	0.947	0.896	0.894	0.892	0.889	0.571	1	0.897	0.918
convulsions	0.966	0.905	0.896	0.893	0.839	0.867	0.937	0.947	0.925	0.894	0.905	0.926	0.714	1	0.926	0.939
convulsions_no	0.941	0.769	0.833	0.667	0.667	0.571	0.7	0.867	0.636	0.8	0.8	0.875	0	1	0.833	0.667
fits	1	0.846	0.722	0.917	0.778	0.857	0.9	0.867	1	1	1	0.875	0.5	0	1	0.833
history_complete	1	1	1	1	1	1	1	1	1	1	0.987	1	1	1	1	1
vacc_opv_penta	1	0.944	1	0.857	0.778	0.722	0.364	0.955	1	0.828	0.882	1	NA	NA	0.947	1
vacc_opv	0.909	1	1	0.857	0.9	NA	0.778	1	0.957	0.562	1	0.909	NA	NA	1	1
vacc_penta	0.909	1	1	0.857	0.9	NA	0.778	1	0.955	0.625	1	0.909	NA	NA	1	1
rotavirus	0.92	1	0.875	1	0.8	0.667	0.471	0.655	0.935	0.71	0.939	0.933	NA	NA	0.826	1
pcv10	0.946	0.962	1	0.909	0.947	0.765	0.565	0.949	0.771	0.771	0.945	0.941	NA	NA	0.982	1
bcg	0.974	0.963	0.926	0.955	0.895	0.944	0.913	1	0.979	0.979	0.945	0.882	1	NA	0.982	0.941
vacc_ipv	0.636	0.667	0.857	0.857	0.7	NA	0.857	0.857	0.81	0.938	1	0.818	NA	NA	0.794	1
measles	0.838	0.769	0.741	0.864	1	0.944	0.609	0.923	0.812	0.792	0.873	0.824	0	NA	0.911	0.706
measles_dose	0.857	1	1	1	1	NA	0.667	1	1	1	1	1	NA	NA	0.955	1
temp	0.897	0.921	0.881	0.929	0.758	0.92	0.937	0.908	0.955	0.939	0.946	0.926	1	0.857	0.941	0.857
resp_rate	0.983	0.952	0.985	0.929	0.71	0.867	0.952	0.961	0.97	0.909	0.946	0.963	0.857	1	0.985	0.959
pulse_rate	0.983	0.921	0.985	0.911	0.806	0.88	0.937	0.921	0.94	0.879	0.919	1	1	1	0.971	0.939
oxygen_sat_done	0.983	0.968	0.985	0.964	0.903	0.933	0.984	0.921	0.985	0.97	0.946	1	1	1	1	0.98
oxygen_sat	1	0.917	0.897	1	0.895	0.786	1	0.976	1	0.836	0.892	NA	1	1	0.962	1
bp_done	0.979	1	0.966	1	1	1	0.913	0.981	1	0.98	1	1	NA	NA	0.957	1
bp_syst	NA	NA	NA	NA	NA	1	NA	1	NA	1	NA	NA	NA	NA	NA	NA
bp_diast	NA	NA	NA	NA	NA	1	NA	0	NA	1	NA	NA	NA	NA	NA	NA
thrush	1	0.937	0.896	0.893	0.79	0.787	0.952	0.934	0.866	0.955	0.986	0.963	1	0.714	0.985	0.939
lymph_nd	0.983	0.952	0.925	0.804	0.742	0.88	0.921	0.947	0.896	0.97	1	0.963	1	0.786	0.985	0.857
wrist_sign	0.948	0.873	0.746	0.696	0.806	0.8	0.841	0.961	0.866	0.788	0.865	1	0.714	0.857	0.926	0.98
jaundice	1	0.937	0.955	0.929	0.887	0.933	0.952	0.961	0.925	0.97	0.973	0.926	0.857	0.857	0.971	0.98
sev_wasting	0.979	0.864	0.845	0.812	0.756	0.926	0.935	0.981	0.94	0.94	0.912	1	NA	NA	0.809	0.953
oedema	0.983	0.952	0.94	0.893	0.855	0.933	0.937	0.921	0.955	1	0.973	0.963	0.714	0.857	0.985	0.939
umbil	NA	1	NA	NA	NA	0.5	NA	NA	1	1	NA	NA	NA	NA	NA	NA
stridor	1	0.968	0.97	0.946	0.839	0.947	0.921	0.974	0.97	1	0.973	1	0.857	1	1	0.959
c_cyanosis	1	0.968	0.955	0.929	0.887	0.973	0.952	0.974	0.97	1	1	0.963	0.857	1	1	0.959
indrawing	0.983	0.921	0.881	0.893	0.823	0.933	0.905	0.961	0.955	0.909	0.946	0.963	0.857	1	0.956	0.959
grunting	1	0.968	0.896	0.911	0.903	0.973	0.937	0.974	0.97	0.955	0.946	0.963	1	1	1	0.939

Table 11 continued from previous page

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
acidotic__breathing	0.966	0.944	0.955	0.911	0.909	0.97	0.981	0.984	0.949	0.931	0.986	0.963	NA	NA	1	0.918
wheeze	1	0.921	0.94	0.893	0.903	0.933	0.937	1	0.97	0.939	0.973	0.926	1	1	0.985	0.959
crackles	0.931	0.873	0.925	0.821	0.855	0.853	0.921	0.974	0.925	0.955	0.905	0.889	0.857	1	0.971	0.878
pulse	1	0.984	0.985	0.929	0.839	0.947	0.984	0.961	0.94	0.955	0.973	1	1	0.857	1	0.959
cap_refill_cat	0.944	0.857	0.788	0.893	0.742	0.689	0.825	0.893	0.75	0.569	0.918	0.958	1	0.929	0.838	0.911
cap_refill	0.778	1	0.4	NA	NA	0.868	1	0.714	1	0.784	0.955	0.963	NA	NA	NA	0.959
skin_temp	1	0.889	0.925	0.804	0.79	0.827	0.968	0.934	0.91	0.894	0.919	1	1	0.714	0.985	0.939
pallor	1	0.952	0.91	0.875	0.823	0.827	0.921	0.934	0.94	0.924	0.932	0.926	0.714	0.929	0.941	0.939
sunk_eyes	0.983	0.937	0.866	0.821	0.645	0.787	0.984	0.921	0.91	0.773	0.946	0.963	0.857	0.786	0.985	0.918
skin_pinch	0.879	0.952	0.836	0.857	0.855	0.827	0.968	0.921	0.925	0.758	0.932	0.963	0.714	0.643	0.985	0.898
avpu	1	0.968	0.925	0.929	0.952	0.973	0.921	0.974	0.97	0.97	1	1	0.857	0.857	0.985	0.959
can_drink	0.966	0.873	0.851	0.929	0.774	0.88	0.921	0.947	0.94	0.97	0.973	0.926	1	0.929	1	0.939
stiff_neck	1	0.952	0.97	0.929	0.903	0.947	0.905	0.987	0.955	0.955	0.932	1	1	0.857	0.985	0.939
bulging_font	0.966	0.937	0.94	0.929	0.903	0.907	0.921	0.947	0.955	0.955	0.946	0.926	1	0.929	0.985	0.959
irrit	NA	1	1	NA	0	1	1	NA	0.75	0	1	NA	NA	NA	NA	NA
red_mov	NA	1	1	NA	0	1	1	NA	0.75	1	1	NA	NA	NA	NA	NA
examination_complete	1	1	1	1	1	1	1	1	1	1	0.987	1	1	1	1	1
mal1_order	1	0.921	0.866	0.964	0.812	0.939	0.937	0.961	0.91	0.809	0.933	0.963	0.857	0.929	0.926	1
mal1_result_avail	0.815	0.903	0.792	1	1	0.923	0.945	0.985	0.892	0.806	0.881	0.846	1	1	0.933	0.762
mal1_result	0.833	0.862	0.755	1	1	0.923	0.927	0.909	0.865	0.871	0.905	0.846	1	0.75	1	0.786
other_mal_test1	1	0.952	0.955	1	0.968	1	0.905	0.974	0.954	0.955	0.851	1	1	1	0.94	0.918
other_mal_result1	NA	NA	NA	NA	NA	1	NA	1	1	1	1	NA	NA	NA	1	1
other_mal_date1	1	0.952	0.955	1	0.969	1	0.905	0.948	0.955	0.971	0.84	1	1	1	0.941	0.918
other_mal_test2	1	1	1	1	1	0.986	1	1	1	1	1	1	1	1	1	1
other_mal_result2	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	1	NA	NA	NA	NA	NA
other_mal_date2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
hb1_order	0.914	0.905	0.821	0.929	0.839	0.813	0.73	0.947	0.879	0.939	0.947	1	0.857	0.929	0.765	0.878
hb1_test	0.75	1	NA	0.75	0.667	0.75	1	0.8	0.286	1	1	1	NA	NA	1	1
hb1_result_avail	0.87	0.783	0.8	1	0.952	0.923	1	1	0.838	0.949	0.846	1	1	1	0.778	0.778
hb1_result	0.818	1	0.923	1	1	1	0.923	0.889	0.85	0.795	0.75	1	1	0.9	0.8	1
hb_units	1	NA	NA	NA	1	1	1	1	1	1	1	NA	NA	NA	NA	1
gluc1_order	0.983	0.889	0.925	0.911	0.902	0.932	0.968	0.947	0.833	0.769	0.824	1	0.857	1	0.97	0.98
gluc1_test	NA	0	NA	0.5	1	0.75	NA	0.5	0.667	0.615	0.667	NA	NA	NA	1	1
gluc1_results	0.833	0.786	0.5	0.625	0.714	0.333	0.857	0.5	0.725	0.486	0.682	NA	0	1	0.93	0.75
gluc_test_units	0.8	0.857	1	0.857	1	0.667	1	0.75	0.889	0.95	0.706	NA	NA	1	0.907	0.5
chemistry	0.983	0.887	1	0.929	0.917	0.918	0.968	1	0.879	0.742	0.784	1	1	1	0.926	0.959
chem_test___1	1	1	1	0.964	0.953	1	1	1	0.836	0.75	0.827	1	1	1	0.897	0.98
chem_test___2	0.983	1	0.97	0.982	0.953	1	1	1	0.866	0.824	0.893	1	1	1	0.838	0.98
chem_test___3	0.983	0.952	0.985	0.964	0.922	0.976	1	1	0.896	0.824	0.92	1	1	1	0.926	0.98
chem_test___4	0.983	0.984	1	1	1	1	1	1	1	0.971	0.973	1	1	1	1	0.98
chem_test___5	0.966	1	1	1	0.984	0.988	1	1	0.985	0.971	0.947	1	1	1	0.985	0.959

Table 11 continued from previous page

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
chem_test___6	0.983	0.968	1	0.964	0.953	0.939	0.968	1	0.881	0.853	0.787	1	1	1	0.779	1
chem_test____1	1	0.968	1	0.982	1	1	1	1	1	1	1	1	1	1	1	0.98
hiv1_order	0.881	0.905	0.612	0.696	0.75	0.72	0.81	0.922	0.836	0.779	0.787	0.852	0.857	1	0.824	0.694
hiv1_test	0.875	0.667	0.368	0.933	0.8	0.667	0.625	0.977	0.75	0.562	0.949	0.778	NA	NA	0.878	0.522
hiv1_result	0.9	0.972	0.789	1	0.76	0.909	0.878	1	1	0.844	0.821	1	NA	NA	0.918	1
hiv_inpt_order	1	1	1	0.852	0.913	0.917	0.7	1	0.821	1	0.826	1	1	1	0.857	0.917
hiv_inpt_test	NA	NA	1	1	NA	1	NA	NA	1	NA	1	NA	1	NA	NA	NA
hiv_inpt_result	NA	NA	1	1	NA	1	NA	NA	1	NA	1	NA	1	NA	NA	NA
micro_order	0.982	0.952	0.924	0.893	0.951	0.905	0.887	1	0.848	0.952	0.958	1	1	1	0.956	0.936
micro_tests___1	1	0.952	0.925	0.893	0.938	0.902	0.889	1	0.851	0.926	0.947	1	1	1	0.956	0.939
micro_tests___2	0.966	1	1	1	1	0.988	1	1	0.955	1	0.987	1	1	1	0.985	1
micro_tests_date	0.949	0.88	0.91	0.893	0.938	0.895	0.968	1	0.94	0.838	0.933	1	NA	NA	0.912	0.98
lp1_bedside___1	1	1	1	1	1	1	1	1	1	0.985	0.987	1	1	1	1	1
lp1_bedside___2	1	0.984	1	1	1	0.988	0.968	1	0.985	1	0.973	1	1	1	0.985	1
lp1_bedside___3	1	1	1	1	1	1	1	0.987	1	0.985	0.987	1	1	1	1	1
lp1_bedside___4	0.983	1	0.985	0.982	1	1	0.984	1	0.97	1	0.987	1	1	1	1	1
lp1_bedside___5	0.983	0.984	0.97	0.946	1	0.976	0.952	0.974	0.836	0.971	0.96	1	1	1	0.971	0.898
lp1_bedside___6	0.966	0.937	0.836	0.929	0.938	0.915	0.952	0.987	0.955	0.926	0.933	1	1	1	0.956	0.898
lp1_result	0.6	1	0.714	0.929	0.667	0.75	1	0.667	0.636	0.6	0.833	NA	NA	NA	0.615	0.667
csf_other	0.949	1	0.896	0.982	0.984	0.974	1	0.987	0.97	1	0.96	1	NA	NA	1	0.98
xray___1	0.983	0.937	0.985	1	0.984	0.927	0.937	1	0.925	0.985	0.907	1	1	0.929	0.912	0.918
xray___2	1	0.952	1	1	0.984	0.939	1	1	1	0.985	1	1	1	0.929	0.971	1
xray___3	1	1	1	0.982	1	0.988	0.984	1	0.985	0.956	0.973	1	1	1	0.971	1
xray___4	0.898	0.889	0.836	0.893	0.922	0.866	0.841	0.896	0.925	0.824	0.92	0.963	1	0.857	0.912	0.959
urine	0.96	1	0.982	0.976	0.891	0.986	0.868	0.985	0.95	0.949	0.864	1	0.857	1	0.932	0.957
urine_test___1	0.932	1	0.985	1	0.906	0.988	0.921	0.987	0.985	0.956	0.88	1	0.857	1	0.926	0.959
urine_test___2	0.966	1	1	0.982	1	1	0.984	1	0.97	1	0.96	1	1	1	0.985	1
urine_test_date	0.983	NA	1	NA	NA	1	1	1	0.985	1	0.939	NA	NA	NA	NA	1
tb_test	1	1	1	1	1	0.968	1	1	0.935	1	0.839	1	1	1	1	1
date_tb_ordered	1	1	1	1	1	0.988	1	1	0.97	1	0.933	1	1	1	0.985	1
tb_test_type___1	1	1	1	1	1	1	1	1	0.985	1	0.987	1	1	1	1	1
tb_test_type___2	1	1	1	1	1	0.988	1	1	0.97	1	0.947	1	1	1	0.985	1
tb_test_type___3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
tb_test_type___4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
tb_test_type____1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
xpert_date_done	1	1	1	1	1	1	1	1	1	1	0.947	1	1	1	0.985	1
mantoux_date_done	1	1	NA	NA	1	1	1	NA	1	1	1	NA	NA	NA	0.952	NA
date_tb_done	1	1	1	1	1	0.981	1	1	0.97	1	0.973	NA	NA	NA	1	1
tb_specimen___1	1	1	1	1	0.984	1	1	1	1	1	1	1	1	1	1	1
tb_specimen___2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
tb_specimen___3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 11 continued from previous page

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
tb_specimen___4	1	1	1	1	1	1	1	1	1	1	0.947	1	1	1	1	1
tb_specimen___5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
tb_specimen___6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
tb_specimen____1	1	1	1	1	0.984	1	1	1	1	1	0.987	1	1	1	1	1
tb_result_xpert	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA
investigations_complete	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dx1_primary	0.932	0.81	0.806	0.75	0.781	0.716	0.825	0.948	0.727	0.734	0.904	0.852	0.429	1	0.941	0.837
dx1_malaria	0.947	0.667	0.878	NA	1	1	0.885	0.952	0.5	0.5	0.857	0.947	1	1	NA	0.947
dx1_malaria_sev	0.647	NA	0.471	NA	NA	0.5	0.545	1	NA	NA	0.667	0.727	1	1	NA	0.6
dx1_malaria_non_sev	0.4	NA	0.857	NA	NA	NA	0.75	0.75	NA	1	NA	1	NA	NA	NA	0.75
dx1_malaria_no_class	0.667	0	0	NA	NA	NA	0.667	0.333	NA	NA	NA	NA	NA	NA	NA	NA
dx1_pneum	1	0.853	0.929	0.957	0.906	0.947	1	0.947	1	0.833	0.955	1	1	1	1	0.917
dx1_diarrhoea	0.929	0.812	0.857	0.952	0.714	0.789	0.917	0.96	1	0.952	0.947	0.667	NA	NA	0.944	1
dx1_dehydrat	1	1	0.933	0.889	0.833	0.857	0.923	1	0.833	0.833	0.952	0.75	1	NA	1	1
dx1_hiv	1	NA	NA	NA	NA	NA	NA	NA	NA	1	1	NA	NA	NA	NA	NA
dx1_malnutr	1	1	1	1	1	0.75	1	1	1	1	1	NA	0.5	NA	1	1
dx1_tb	NA	NA	NA	NA	NA	1	NA	NA	1	NA	1	1	NA	NA	1	1
dx1_tb_status	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
dx1_anaemia	1	1	1	NA	NA	NA	0.857	1	NA	1	1	NA	1	NA	1	NA
dx1_meningitis	1	1	1	1	1	1	1	1	1	0.9	1	1	NA	1	1	1
dx1_asthma	1	1	1	1	NA	0.75	1	NA	NA	1	0.667	NA	NA	NA	1	1
dx1_rickets	NA	NA	NA	1	NA	1	NA	NA	0.667	NA	1	1	NA	1	NA	1
dx1_sepsis	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA
dx1_pre_lbw	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA
dx1_sickle_cell	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	0.923	NA	NA
dx1_other_1	0	0	0	0	0	0	0	NA	0	0	0	NA	NA	NA	0	0
dx1_other_3	0.983	0.81	0.94	0.75	0.672	0.866	0.857	0.935	0.896	0.794	0.973	0.926	0.857	0.786	0.824	0.878
dx1_other_4	0.983	0.968	1	1	0.969	0.976	0.968	0.987	0.97	0.971	1	1	1	1	0.941	1
dx1_other_3_text	0.983	0.984	1	0.964	0.938	0.927	0.937	0.987	1	0.926	0.973	1	1	1	0.941	0.959
sec_dx	0.971	0.906	0.85	0.929	0.893	0.824	0.882	0.958	0.8	0.815	0.923	0.947	NA	1	0.891	0.844
dx2_malaria	1	NA	1	NA	NA	1	1	1	NA	NA	1	1	NA	NA	1	1
dx2_malaria_sev	0	NA	0.75	NA	NA	NA	0.5	NA	NA	NA	NA	NA	NA	NA	NA	NA
dx2_malaria_non_sev	1	NA	1	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	0.5
dx2_malaria_non_class	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	1
dx2_pneum	1	1	0.8	1	NA	NA	1	1	1	1	1	0	NA	1	1	1
dx2_diarrhoea	1	1	0.8	0	NA	NA	1	1	1	NA	1	1	NA	0	1	1
dx2_dehydrat	0.75	1	1	1	1	1	1	1	0.667	NA	1	NA	NA	1	0.857	1
dx2_hiv	NA	NA	1	NA	NA	NA	1	1	NA	NA	NA	NA	NA	NA	1	NA
dx2_malnutr	NA	NA	1	1	1	NA	NA	NA	1	NA	1	1	NA	1	1	1
dx2_anaemia	1	NA	1	NA	NA	NA	1	1	NA	NA	1	1	NA	1	NA	1
dx2_meningitis	NA	NA	1	NA	1	1	NA	1	1	1	1	NA	NA	1	NA	NA

Table 11 continued from previous page

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
dx2_asthma	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1
dx2_rickets	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.75	NA	NA	1	NA	NA
dx2_tb	NA	NA	1	NA	NA	NA	NA	NA	1	NA	1	NA	NA	NA	NA	NA
dx2_sepsis	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA
dx2_pre_lbw	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA
dx2_sickle_cell	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	1	NA	NA
dx2_other_1	0	NA	NA	NA	NA	NA	NA	0	NA	0	NA	NA	NA	NA	0	0
dx2_other_3	0.932	0.921	0.985	0.982	0.906	0.976	0.905	0.987	0.955	0.912	0.987	0.963	1	0.714	0.868	0.939
dx2_other_4	0.915	0.937	0.985	0.982	0.969	0.988	0.984	1	0.925	0.971	1	0.926	1	1	0.897	1
dx2_other_3_text	0.949	1	1	1	0.984	0.988	0.984	1	0.985	1	0.973	1	1	1	0.985	0.959
admission_diag	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
tsheet_prsnt	1	0.965	1	0.977	0.927	1	0.981	1	1	0.96	1	0.952	1	1	0.967	1
pen_pres	1	1	0.97	0.926	0.911	0.947	0.934	1	0.97	0.952	1	1	1	1	0.967	0.98
pen1_route	1	0.897	1	0.778	0.912	1	0.962	1	1	0.941	0.866	1	1	1	0.857	1
pen1_dose	0.952	0.897	0.952	0.857	0.853	0.782	0.846	0.968	0.868	0.885	0.896	0.333	0.667	0.857	0.867	0.931
pen1_unit	0.857	0.931	1	0.852	0.882	0.926	0.923	0.935	0.947	0.941	0.925	1	1	1	0.867	0.966
pen1_freq	0.952	0.931	0.857	0.929	0.912	0.927	0.923	0.968	0.947	1	0.94	0.667	1	0.714	1	1
pen1_days	0.524	0.862	0.667	0.893	0.818	0.527	0.769	0.871	0.658	0.596	0.821	0.333	0.667	0.714	0.867	0.621
pen1_date	0.915	0.952	0.925	0.875	0.844	0.854	0.873	0.948	0.91	0.75	0.907	0.963	1	1	0.971	0.857
pen1_date_started	1	1	1	0.982	0.953	0.976	1	1	0.985	0.941	0.973	1	1	1	0.985	0.98
gent1_pres	1	0.967	0.985	0.981	0.964	0.907	0.951	0.986	0.955	0.921	0.973	1	1	1	0.967	0.98
gent1_route	0.833	0.812	1	0.75	1	0.939	0.81	0.733	1	1	0.772	NA	1	1	1	1
gent1_dose	1	0.938	1	1	0.923	0.909	0.905	1	0.862	0.971	0.93	NA	1	1	1	0.952
gent1_unit	1	1	1	1	1	0.952	1	1	0.952	1	0.976	NA	NA	NA	1	1
gent1_freq	1	1	1	1	1	0.879	0.905	1	0.966	1	0.965	NA	1	1	1	0.905
gent1_days	0.583	0.875	0.769	1	0.846	0.697	0.81	0.867	0.862	0.441	0.825	NA	0.5	0.714	0.889	0.81
genta1_date	0.949	0.937	0.955	0.982	0.953	0.866	0.889	0.948	0.925	0.735	0.92	1	1	1	0.956	0.898
genta1_date_started	1	0.984	1	0.982	0.969	0.963	1	1	0.97	0.882	0.96	1	1	1	0.985	0.98
amox1_pres	0.983	0.967	1	0.981	0.964	0.96	0.967	0.986	1	1	0.973	1	0.857	1	0.934	0.98
amox1_dose	1	1	0.818	1	0.8	0.889	0.875	0.786	1	1	1	0.8	NA	NA	1	0.5
amox1_unit	1	1	0.909	1	0.8	0.889	0.875	1	1	1	1	0.8	NA	NA	0.875	0.5
amox1_formulation	0.5	1	0.833	0.889	0.556	0.625	0.625	0.692	0.667	0.5	0.333	0.75	NA	NA	1	1
amox1_freq	1	1	0.727	1	0.9	0.778	1	0.929	1	0.5	1	0.8	NA	NA	0.917	1
amox1_days	1	0.917	0.727	1	0.8	0.778	0.625	0.714	1	1	0.75	0.8	NA	NA	0.913	0
amox1_date	0.949	0.937	0.955	0.964	0.875	0.951	0.937	0.974	1	1	0.973	0.889	0.857	1	0.838	0.959
amox1_date_started	0.983	0.984	0.985	0.964	0.969	0.939	0.984	0.987	1	0.985	0.987	0.963	0.857	1	0.956	0.98
ceftri1_pres	0.983	0.983	0.985	0.981	0.982	0.933	0.984	0.986	1	0.984	0.92	1	1	1	0.967	0.98
ceftri1_route	0.857	0.9	0.955	0.8	1	1	1	1	0.957	1	0.714	1	1	1	1	1
ceftri1_dose	1	1	0.909	1	1	1	1	1	1	0.864	0.952	1	1	1	1	0.8
ceftri1_freq	1	1	0.864	1	1	0.667	1	0.9	1	0.955	0.952	1	1	1	0.857	0.8
ceftri1_days	0.857	0.818	0.682	1	0.5	0.833	0.733	1	0.565	0.455	0.85	1	0.6	1	1	0.4

Table 11 continued from previous page

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
ceftri1_date	0.983	0.968	0.91	0.982	0.984	0.939	0.873	0.987	0.955	0.882	0.96	1	0.714	1	0.956	0.939
ceftri1_date_started	1	1	1	1	1	0.988	0.984	0.987	0.985	0.971	1	1	0.714	1	0.985	1
caf1_pres	1	0.983	1	0.944	1	0.947	1	1	0.97	0.937	0.987	0.962	1	1	0.984	1
caf1_route	NA	1	1	0.667	1	1	NA	1	1	0.857	0.692	1	NA	NA	1	1
caf1_dose	NA	0.5	1	0.9	0.857	1	NA	1	1	1	0.923	1	NA	NA	1	0.75
caf1_units	NA	1	NA	1	1	1	NA	1	1	1	1	NA	NA	NA	NA	1
caf1_freq	NA	1	1	1	0.857	0.833	NA	0.8	1	0.714	0.923	1	NA	NA	1	1
caf1_days	NA	0.5	1	0.8	0.571	0.833	NA	0.6	0	0.571	0.615	0	NA	NA	1	0.75
caf1_date	1	0.968	1	0.946	0.969	0.927	1	1	0.955	0.941	0.933	0.963	1	1	1	0.959
metr1_pres	0.983	0.933	1	1	1	0.987	1	0.986	0.97	1	0.96	1	1	0.929	1	1
metr1_route	1	NA	NA	NA	NA	1	1	NA	1	NA	0.333	NA	NA	NA	NA	NA
metr1_dose	1	NA	NA	NA	NA	1	1	NA	1	NA	0.667	NA	NA	NA	NA	NA
metr1_unit	1	NA	NA	NA	NA	1	1	NA	1	NA	0.667	NA	NA	NA	NA	NA
metr1_freq	1	NA	NA	NA	NA	1	1	NA	1	NA	1	NA	NA	NA	NA	NA
metr1_days	0.8	1	0.5	0	NA	1	0.5	0.5	0.5	0	0.6	NA	NA	NA	NA	0
metr1_date	0.983	0.937	0.97	0.982	1	1	1	0.974	0.97	1	0.947	1	1	0.929	1	1
metr1_date_started	1	1	1	1	1	1	1	1	0.97	1	1	1	1	0.929	1	1
cotrimox1_pres	0.983	1	0.985	0.963	1	0.987	1	1	1	1	0.973	1	1	1	1	1
cotrimox1_route	1	NA	NA	1	NA	1	NA	NA	1	NA	1	1	NA	NA	1	0.333
cotrimox1_dose	1	NA	NA	1	NA	1	NA	NA	1	NA	1	1	NA	NA	1	1
cotrimox1_unit	1	NA	NA	1	NA	0.667	NA	NA	1	NA	1	1	NA	NA	1	1
cotrimox1_freq	1	NA	NA	1	NA	0.667	NA	NA	1	NA	1	1	NA	NA	1	1
cotrimox1_days	1	NA	NA	0	NA	0.333	NA	NA	1	NA	1	1	NA	NA	1	0.667
cotrimox1_date	0.983	1	0.985	0.964	1	0.976	1	1	0.985	1	0.973	0.963	1	1	1	1
cotrimox1_date_started	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
anti_tb1_pres	1	1	1	1	1	1	1	1	0.985	1	0.973	1	1	1	1	1
anti_tb_presc	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
ant_tb_date	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
anti_malarials	0.983	1	0.97	1	1	1	0.951	0.986	0.985	0.984	0.987	1	1	1	1	0.959
quinl1_pres	0.976	1	1	NA	NA	0.75	0.964	1	1	1	1	1	1	1	NA	1
quinl1_route	0.818	NA	0.857	NA	NA	1	1	NA	NA	NA	0.833	1	NA	NA	NA	0.857
quinl1_dose	1	NA	0.929	NA	NA	1	1	NA	NA	NA	1	0.857	NA	NA	NA	1
quinl1_date	0.966	1	0.94	1	1	0.988	0.968	1	1	0.985	0.947	0.889	1	1	1	0.959
quinm1_pres	0.976	1	0.982	NA	NA	0.75	0.964	1	1	1	0.944	1	1	1	NA	1
quinm1_route	1	NA	0.846	NA	NA	1	0.625	NA	NA	NA	0.6	0.667	NA	NA	NA	0.857
quinm1_dose	0.909	NA	1	NA	NA	1	1	NA	NA	NA	1	1	NA	NA	NA	0.857
quinm1_freq	0.909	NA	0.333	NA	NA	1	0.625	NA	NA	NA	0.4	0.667	NA	NA	NA	0.571
quinm1_days	0.545	NA	0.231	NA	NA	1	0.625	NA	NA	NA	0.6	0	NA	NA	NA	0.714
quinm1_date	0.949	1	0.896	1	1	0.988	0.937	1	1	0.973	0.96	0.889	1	1	1	0.98
arte_pres	0.976	1	0.964	NA	NA	1	0.964	1	1	1	1	1	1	1	NA	0.957
arte_route	1	0.5	1	NA	NA	1	0.895	1	1	1	0.909	1	1	1	NA	0.938

Table 11 continued from previous page

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
arte_dose	1	1	0.973	NA	NA	1	0.818	0.966	1	1	0.7	0.917	NA	NA	NA	1
arte_dose1	NA	NA	NA	NA	NA	NA	1	1	NA	NA	1	NA	1	1	NA	NA
arte_dose2	NA	NA	NA	NA	NA	1	0	1	NA	NA	NA	NA	1	NA	NA	NA
arte_dose3	NA	NA	NA	NA	NA	NA	1	1	1	NA	NA	NA	NA	NA	NA	NA
arte_dose4	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA
arte_dose6	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
arte_dose8	NA	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	1	NA	NA	NA
arte_freq	0.808	0.5	0.919	NA	NA	0.667	0.842	0.944	0.5	0	0.889	0.583	1	1	NA	0.438
arte_days	0.692	0.5	0.694	NA	NA	0.667	0.842	0.889	1	0.5	0.5	0.833	0.5	1	NA	0.375
arte_date	0.949	1	0.866	1	1	1	0.937	0.896	0.985	1	0.933	0.815	0.714	1	1	0.857
arte_date_started	0.983	1	0.985	1	1	1	1	0.987	1	1	0.987	0.963	0.714	1	1	0.959
artemether	1	1	0.981	NA	NA	0.75	1	1	0.667	1	0.944	1	1	1	NA	1
coart1_pres	0.902	1	0.945	NA	NA	0.75	0.893	0.978	1	1	1	0.95	1	1	NA	0.957
coart1_dose	1	NA	1	NA	NA	1	0.923	0.941	1	1	1	1	1	0	NA	1
coart1_units	1	NA	1	NA	NA	1	0.923	1	1	1	1	1	1	0	NA	1
coart1_freq	1	NA	0.882	NA	NA	1	0.846	1	0.667	1	0.667	1	1	1	NA	0.95
coart1_days	0.889	NA	0.706	NA	NA	0	0.769	0.882	1	0	0.778	1	1	1	NA	0.75
coart1_date	0.932	1	0.821	1	1	0.976	0.873	0.974	1	0.985	1	0.963	0.857	1	1	0.878
coart1_date_started	0.983	1	0.985	1	1	0.988	0.937	1	0.985	0.985	1	1	0.857	1	1	1
ceta1_pres	0.879	0.883	0.97	0.833	0.893	0.92	0.852	0.973	0.94	0.794	0.84	0.923	1	1	1	0.837
salb_pres	1	0.983	0.985	0.963	0.982	0.987	0.984	1	0.97	0.952	0.947	1	1	0.929	0.967	0.98
salb1_route	0	1	0.8	1	1	1	1	1	0.667	0.889	0.857	NA	NA	0.667	1	1
pred1_pres	1	0.983	0.985	1	1	0.987	1	1	1	0.984	0.947	0.962	1	0.929	0.984	1
vita	0.966	0.967	0.894	0.963	0.982	0.973	0.984	0.986	0.955	0.937	0.96	1	1	1	0.918	1
zinc1_pres	0.948	0.883	0.939	0.944	0.875	0.933	0.902	0.959	0.881	0.905	0.893	0.923	1	0.929	0.984	0.959
dextrose_10	0.931	1	0.894	1	1	0.973	0.951	0.986	0.985	1	1	0.923	0.857	1	1	0.878
dextrose_vol	0.75	NA	0.5	NA	NA	1	NA	1	NA	NA	NA	NA	1	NA	NA	1
adm_rx	0.86	0.833	0.848	0.741	0.786	0.867	0.902	0.986	0.879	0.857	0.787	0.885	1	1	0.852	0.918
adm_rx1	0.797	0.603	0.716	0.679	0.531	0.756	0.603	0.844	0.806	0.735	0.587	0.704	0.714	0.714	0.721	0.837
adm_rx1_date_presc	0.966	0.921	0.955	1	0.891	0.927	0.873	1	0.94	0.971	0.933	1	1	0.714	0.956	0.959
adm_rx1_date_given	0.966	0.937	0.955	1	0.875	0.939	0.905	1	0.955	0.971	0.933	1	0.714	0.714	0.897	0.959
adm_rx2	0.881	0.635	0.836	0.964	0.781	0.878	0.714	0.922	0.761	0.897	0.68	0.889	0.429	0.5	0.868	0.878
adm_rx2_date_presc	1	0.905	0.94	1	0.969	1	0.937	1	0.881	0.985	0.973	1	0.571	0.714	0.956	0.98
adm_rx2_date_given	1	0.921	0.94	1	0.969	1	0.921	1	0.866	0.985	0.987	1	0.571	0.714	0.956	0.98
adm_rx3	0.983	0.714	0.955	0.982	0.891	0.939	0.857	0.961	0.925	0.941	0.88	0.963	0.857	0.643	0.985	0.98
adm_rx3_date_presc	1	0.952	0.985	1	0.969	1	0.952	1	0.985	1	0.987	1	0.857	0.5	1	1
adm_rx3_date_given	1	0.952	0.985	1	0.969	1	0.968	1	0.985	1	1	1	0.857	0.5	0.985	1
adm_rx4	1	0.81	0.985	1	0.953	0.988	0.905	0.987	0.97	0.971	0.92	1	0.857	0.714	0.985	1
adm_rx4_date_presc	1	1	1	1	0.984	1	0.968	1	1	1	1	1	0.857	0.429	0.985	1
adm_rx4_date_given	1	1	1	1	0.984	1	0.968	1	1	1	1	1	0.857	0.429	0.985	1
adm_rx5	1	1	1	1	0.984	1	1	1	1	1	1	1	1	0.786	1	1

Table 11 continued from previous page

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
adm_rx5_date_presc	1	1	1	1	0.984	1	1	1	1	1	1	1	1	0.357	1	1
adm_rx5_date_given	1	1	1	1	0.984	1	1	1	1	1	1	1	1	0.357	1	1
adm_rx6	1	1	1	1	0.984	1	1	1	1	1	1	1	1	0.786	1	1
adm_rx6_date_presc	1	1	1	1	0.984	1	1	1	1	1	1	1	1	0.357	1	1
adm_rx6_date_given	1	1	1	1	0.984	1	1	1	1	1	1	1	1	0.357	1	1
adm_rx_other1	0.966	1	0.94	0.929	0.984	0.947	0.984	0.987	0.925	0.946	0.893	0.926	NA	NA	0.956	0.959
adm_rx_nt_listed	NA	0.8	NA	NA	1	1	1	1	1	NA	0	NA	1	0.889	1	NA
adm_rx_free_text	1	0.984	1	1	1	1	1	1	1	1	0.96	1	1	0.929	1	1
treatment_complete	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
oxy_order	1	0.937	0.97	1	0.984	0.96	0.984	0.974	0.985	0.894	0.893	1	1	1	0.985	0.959
oxy_rate	1	1	0.857	1	1	0.909	1	1	0.933	0.5	0.952	NA	NA	NA	1	0.5
oxy_route	1	1	0.714	1	0.5	0.909	1	0.75	0.733	0.7	0.95	NA	NA	NA	1	1
oxy_date	1	0.937	0.955	1	0.984	0.963	0.984	1	0.94	0.824	0.88	1	1	1	0.985	0.959
transf_order	0.948	1	0.955	1	1	1	0.984	1	1	1	0.987	1	1	1	0.985	0.98
blood_comp	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	1	NA	NA	NA
transf_vol	0.75	1	0.727	NA	1	1	0.667	1	0	1	0	NA	1	NA	NA	1
transf_hrs	1	1	1	NA	0	1	0.667	1	0	1	1	NA	0.5	NA	NA	1
transf_date_pres	0.949	1	0.94	1	0.984	1	0.952	0.987	1	1	0.987	1	1	1	0.985	0.98
transf_date_gvn	0.932	1	0.881	1	1	0.988	0.968	0.974	0.985	1	0.987	1	0.857	1	0.985	0.959
photo_therap_presc	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
fluid_bolus	0.913	1	1	1	0.98	1	0.964	0.985	0.929	0.962	0.966	1	0.714	1	1	0.902
fluid_bolus_type	1	1	1	NA	NA	1	0.5	1	1	1	1	NA	NA	0.5	1	1
number_boluses	NA	NA	NA	NA	NA	NA	1	1	NA	1	NA	NA	NA	1	1	1
bolus_volume	NA	NA	NA	NA	NA	NA	0	1	NA	1	NA	NA	NA	1	1	1
fluid_bolus_dura	1	1	1	NA	NA	1	0	0.6	1	1	0.75	NA	NA	0.5	1	1
dehyd_fluid	0.845	0.905	0.806	0.982	0.919	0.907	0.825	0.987	0.821	0.848	0.84	0.926	0.571	1	0.926	0.857
iv_fluid	0.933	0.957	0.682	0.958	0.909	0.926	0.867	0.972	0.828	0.8	0.92	1	1	1	1	0.737
fluid_pres1	0.667	0.833	1	1	1	0.9	1	0.889	0.75	0.933	0.909	1	1	NA	0.714	0.875
other_fluid_presc	1	1	0.985	1	1	1	1	1	0.985	1	1	1	1	1	0.985	1
total_vol1	0.5	0.667	0.714	1	0.4	0.636	0.714	0.667	0.5	0.667	0.818	0	0	NA	1	0.625
fluid_time1	0.667	0.833	0.286	1	0.8	0.818	0.429	0.889	0.75	0.8	0.727	1	0	NA	0.429	0.75
fluid_step1_2	0.957	1	0.964	1	0.92	0.862	0.889	1	0.893	0.962	0.966	0.95	0.857	0.929	0.967	0.854
oral_fluid	0.867	1	0.955	0.96	1	0.964	0.867	0.972	0.793	0.923	0.88	1	1	1	1	0.947
fluid_pres2	1	1	1	1	1	0.96	0.778	1	1	0.938	1	1	NA	1	1	1
total_vol2	0.833	0.905	0.692	1	0.889	0.708	0.889	0.97	0.708	0.688	0.824	0.667	NA	0.5	0.875	1
fluid_time2	0.917	0.952	0.923	1	1	0.833	1	0.97	0.875	0.875	0.824	0.667	NA	0.5	0.833	1
vol_stool	0.917	0.905	0.692	0.958	0.947	0.88	1	0.879	0.875	0.875	0.706	0.667	NA	1	0.917	0.6
fluid_maint	0.877	0.952	0.746	0.929	0.967	0.92	0.873	0.973	0.91	0.788	0.863	0.963	0.714	1	0.912	0.816
fluid_maint_vol	NA	NA	0.25	NA	NA	NA	NA	NA	1	1	1	1	NA	NA	0.5	1
malnourished	0.983	1	0.955	1	0.984	0.92	0.968	1	0.91	0.97	0.88	1	0.857	1	0.985	0.98
feeds_after_adm	1	0.979	1	1	1	1	1	1	0.967	1	1	1	1	1	1	1

Table 11 continued from previous page

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
feed_pres	1	0.75	1	NA	1	0.75	1	0.5	0.929	1	1	NA	NA	1	1	1
other_feed_pres	1	1	1	1	1	1	1	1	0.985	1	1	1	1	1	1	1
feed_vol	0.5	0.75	0	NA	0.5	0.25	0	1	0.643	1	0.875	NA	NA	1	1	1
feed_frequency	0	0	1	NA	0.5	0.333	0	1	0.667	1	0	NA	NA	1	1	1
freq_24hrs	0	NA	NA	NA	NA	0	NA	0	0.5	NA	0.5	NA	NA	NA	NA	NA
date_feeds_start	1	1	1	1	1	0.963	1	0.987	0.896	0.946	0.893	1	NA	NA	1	0.98
date_post_adm_feeds	0.983	0.984	1	1	1	0.988	1	1	0.985	1	0.973	1	1	1	1	1
fluid_feed_mon	0.707	0.46	0.881	0.982	0.419	0.68	0.905	0.776	0.791	0.894	0.88	1	0.857	0.929	0.896	0.816
fluid_feed_monpres	0.614	0.233	0.917	1	0.387	0.556	0.905	0.783	0.548	0.556	0.747	1	NA	0	0.853	0.667
supportive_care_complete	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
vitals_chart	0.879	0.905	0.97	0.964	0.952	0.933	0.984	0.987	0.985	0.894	0.787	0.926	1	1	0.897	0.918
vital_monit_48hrs	0.922	0.962	0.846	0.944	0.949	0.971	1	0.959	1	0.847	0.965	NA	0.429	1	1	0.956
temp_chart	0.562	0.44	0.625	0.478	0.732	0.403	0.678	0.792	0.615	0.28	0.711	NA	0.333	1	0.5	0.465
resp_chart	0.625	0.62	0.562	0.696	0.857	0.762	0.593	0.736	0.523	0.222	0.921	NA	1	1	0.475	0.465
pulse_chart	0.562	0.62	0.625	0.674	0.839	0.721	0.627	0.736	0.516	0.333	0.947	NA	1	1	0.475	0.419
bp_moni	1	1	NA	NA	NA	NA	1	1	1	NA	NA	1	NA	NA	1	NA
oxy_sat_moni	1	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
monitoring_complete	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
disch_death_summ	0.966	0.984	0.881	0.964	0.984	0.915	0.952	0.987	0.955	0.926	0.946	0.963	1	1	1	0.98
outcome	1	0.984	1	0.981	1	0.975	0.984	1	1	1	1	1	1	1	0.985	1
outcome_res	1	0.964	0.982	1	1	0.982	1	0.984	1	1	0.981	1	1	1	0.983	1
dsc_condition	0.895	0.597	0.692	0.75	0.746	0.703	0.77	0.944	0.821	0.868	0.781	0.667	1	1	0.91	0.755
follow_up	0.825	0.887	0.53	0.696	0.683	0.597	0.787	0.778	0.791	0.75	0.877	0.704	0.714	1	0.882	0.837
follow_up_days	1	0.68	0	1	0.889	0.556	0.833	1	0.909	0.667	0.812	1	0.5	1	0.857	1
follow_up_days_lw	NA	0	NA	NA	NA	NA	NA	1	1	1	1	1	NA	NA	NA	NA
dsc_dx1_primary	0.881	0.857	0.833	0.75	0.812	0.79	0.762	0.855	0.806	0.735	0.77	0.926	1	1	0.882	0.959
dsc_dx1_malaria	0.909	NA	0.867	NA	NA	1	0.76	0.825	1	0.667	0.857	0.941	1	1	0	0.95
dsc_dx1_malaria_sev	0.733	NA	0.55	NA	NA	0.5	0.333	1	1	1	0.5	0.9	1	1	NA	0.545
dsc_dx1_malaria_non_sev	0.571	NA	0.75	NA	NA	NA	0.5	0.778	NA	NA	NA	NA	NA	NA	NA	0.5
dsc_dx1_malaria_no_class	1	NA	1	NA	NA	NA	1	NA	NA	NA	1	0	1	NA	NA	NA
dsc_dx1_pneum	1	0.867	0.684	0.947	0.929	0.806	1	1	0.967	0.848	0.8	1	1	1	0.962	1
dsc_dx1_diarrhoea	0.833	0.692	1	1	0.947	0.846	0.8	0.96	1	0.923	0.938	1	NA	NA	0.923	0.857
dsc_dx1_dehydrat	1	0.769	1	1	0.923	0.667	0.875	0.889	1	0.8	1	NA	NA	NA	0.9	0.833
dsc_dx1_hiv	1	NA	NA	NA	1	NA	1	NA	1	NA	NA	NA	NA	NA	1	NA
dsc_dx1_malnutr	1	1	NA	1	1	0.833	1	1	1	0.75	0.818	NA	NA	NA	1	1
dsc_dx1_anaemia	1	1	0.714	NA	NA	1	0.667	1	NA	NA	0.857	NA	0.333	NA	1	NA
dsc_dx1_meningitis	NA	1	1	1	1	1	1	1	1	1	1	1	NA	1	1	1
dsc_dx1_rickets	NA	NA	NA	1	1	1	NA	NA	NA	NA	1	NA	NA	1	NA	NA
dsc_dx1_asthma	1	1	NA	0	1	0.25	NA	NA	NA	1	1	NA	NA	NA	0.667	0.5
dsc_dx1_tb	NA	NA	NA	NA	NA	NA	NA	1	1	NA	NA	NA	NA	NA	1	NA
dsc_dx1_sepsis	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA

Table 11 continued from previous page

variable	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
dsc_dx1_pre_lbw	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA
dsc_dx1_sickle_cell	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	0.889	NA	NA
dsc_dx1_other_1	NA	0	0	0	0	0	0	NA	0	0	0	NA	NA	NA	0	0
dsc_dx1_other_2	NA	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
dsc_dx1_other_3	0.974	0.96	1	0.982	0.969	0.988	1	1	1	1	1	1	NA	NA	1	0.98
dsc_dx1_other_4	0.983	0.81	0.91	0.821	0.656	0.866	0.857	0.961	0.896	0.794	0.947	1	1	0.786	0.779	0.837
dsc_dx1_other_5	1	0.984	1	0.964	0.906	0.976	0.968	1	0.94	0.971	0.987	1	1	1	0.941	0.98
dsc_dx2	0.904	0.912	0.962	0.884	0.786	0.812	0.815	0.879	0.768	0.909	0.796	0.9	1	1	0.896	0.889
dsc_dx2_malaria	0	NA	NA	NA	NA	NA	NA	1	NA	NA	1	NA	NA	NA	NA	0.5
dsc_dx2_malaria_non_sev	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	1
dsc_dx2_malaria_non_class	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	NA
dsc_dx2_pneum	1	1	NA	NA	NA	1	1	1	1	1	1	NA	NA	1	1	0.833
dsc_dx2_diarrhoea	1	1	NA	1	NA	NA	1	NA	0.75	NA	1	1	NA	0	0.5	0.5
dsc_dx2_dehydrat	1	1	NA	NA	NA	NA	NA	NA	1	NA	1	1	NA	1	1	1
dsc_dx2_hiv	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1
dsc_dx2_malnutr	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	1	1
dsc_dx2_anaemia	1	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	1	NA	0.667
dsc_dx2_meningitis	1	NA	NA	NA	NA	NA	NA	1	1	NA	NA	NA	NA	1	NA	NA
dsc_dx2_rickets	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA
dsc_dx2_tb	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA
dsc_dx2_sepsis	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA
dsc_dx2_pre_lbw	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA
dsc_dx2_sickle_cell	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA
dsc_dx2_other_1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0	NA
dsc_dx2_other_4	0.949	0.952	1	0.946	0.875	0.988	0.921	0.987	0.94	0.971	0.933	0.926	1	0.786	0.838	0.98
dsc_dx2_other_5	1	1	1	1	0.969	1	1	1	1	0.985	0.987	1	1	1	0.985	1
dsc_rx	1	0.984	0.866	0.836	0.952	0.917	0.836	0.986	0.969	0.939	0.958	1	0.857	0.857	0.941	0.959
dsc_rx1	0.763	0.571	0.463	0.732	0.484	0.646	0.619	0.701	0.761	0.618	0.733	0.667	0.857	0.786	0.838	0.694
dsc_rx2	0.78	0.619	0.612	0.821	0.688	0.732	0.651	0.636	0.761	0.75	0.893	0.704	0.857	0.857	0.882	0.653
dsc_rx3	0.78	0.762	0.791	0.982	0.828	0.915	0.794	0.714	0.896	0.912	0.933	0.667	1	0.857	0.956	0.857
dsc_rx4	0.898	0.762	0.925	0.982	0.953	0.976	0.921	0.857	0.97	0.985	0.96	0.889	0.857	0.857	1	0.98
dsc_rx5	0.949	0.778	0.955	1	0.984	1	0.984	0.922	0.97	1	0.987	1	1	0.929	1	0.98
dsc_rx_other1	0.932	0.778	0.896	0.857	0.953	0.915	0.937	0.935	0.896	0.811	0.827	0.926	1	1	0.985	0.959
dsc_rx_other2	0.983	0.794	0.985	0.982	0.984	0.921	0.984	0.948	0.985	0.973	0.907	0.963	1	1	0.985	0.98
rx_nt_listed	NA	1	NA	NA	0.857	1	1	1	1	1	0	NA	0.8	1	1	NA
rx_free_text	1	1	1	1	0.984	1	1	1	1	1	0.933	1	0.857	1	1	1
discharge_info	1	1	1	1	1	1	1	0.987	1	1	0.987	1	1	1	1	1

B Principal Component Analysis steps

Step 1

Mean of all dimensions in the dataset are calculated, then the data is scaled so that each variable contributes equally to the analysis. The equation below explains the scaling step.

$$z = \frac{x - \mu}{\sigma}$$

z is the scaled value, x is the original value, σ is the standard deviation while μ is the mean.

Step 2

Compute the covariance of the two variables X and Y using the formula below.

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Xi - \overline{x})(Yi - \overline{y})$$

Step 3

Compute Eigenvectors and their corresponding Eigenvalues. An eigenvector of a matrix B is a vector such that:

$$B\overrightarrow{\vartheta} = \lambda\overrightarrow{\vartheta}$$

Where λ is a scalar value called the eigenvalue. If we transform, equation (??) as defined by λ it becomes:

$$B\overrightarrow{\vartheta} - \lambda\overrightarrow{\vartheta} = 0$$
$$\Rightarrow \overrightarrow{\vartheta}(B - \lambda I) = 0$$

Where I is the identity matrix. The eigenvectors provide the patterns in the data for us to extract the most useful ones.

Step 4

Choose the k eigenvectors with the largest eigenvalues. These values are sorted with respect with decreasing order of eigenvalues and k is choosen where k is the number of dimensions you wish to have in the new dataset. K

Principal components are the new variables constructed for the initial features such that the new variables are uncorrelated.

We rank the principal components in order of their eigenvalues.

C Isolation forest evaluation stages

A 2-stage process is employed when detecting anomalies using iForest

1. The **training stage** builds iTrees using sub-samples of the training set
2. The **evaluation stage** passes the test instances through the iTrees to generate anomaly score for each instance.

The training stage

Each iTree is constructed using a sample $X^{'}$ randomly selected without replacement from $X, X' \subset X$.

Algorithm Steps:

Let $iForest(X,t,\psi)$ be a function that takes X as the input data, t as the numer of trees and ψ as the sub-sampling size.

1. Initialize an empty *Forest*
2. Iterate through X to create random samples as shown below:

for $i\ to\ t$ do

$$X' \leftarrow sample(X,\psi)$$

$$Forest \ \leftarrow Forest\ \cup iTree(X^{'})$$

end for. The output is a Forest.

Let $iTree(X^{'})$ be a function that takes subsample $X^{'}$ as an input parameter. To get an iTree, the following steps are taken;

1. $if\ X^{'}$ cannot be split then

$$return\ exNode\ Size\ \leftarrow \left|X^{'}\right|$$

2. *else*

(a) $let\ Q\ be\ a\ list\ of\ atrributes\ in\ X^{'}$

(b) $randomly\ select\ an\ attribute\ q\ \in Q$

(c) randomly select a split value point p between the maximum and minimum values of the attribute q in X'.

(d) $T_l \leftarrow filter(X',\ q < p)$

(e) $T_r \leftarrow filter(X',\ q \geq p)$
return $inNode\{Left \leftarrow iTree(T_l), Right \leftarrow iTree(T_r), SplitAtt \leftarrow q, SplitValue \leftarrow p\}$

(f) $end\ if.$

The evaluation stage

This is the stage for computing anomaly score for each observation.

Let $Pathlength(x,T,hlim,e)$ be a function of x instances, T iTrees, $hlim$ height limit and e as the current path length.

All these parameters are initialized to zero at first.

To achieve an output of x as the score, we follow the steps outlined below;

$if\ T$ is an external node or $e\ \geq hlim$ then

Return $e + c(T.size)$ as defined in equation 4

$$end\ if.$$

$$a \leftarrow T.splitAtt$$

$$if\ x_a < T.splitValue\ then$$

Return $PathLength(x, T.left, hlim, e+1)$

$$else\ \{x_a\ \geq\ T.splitValue\}$$

Return $PathLength(x, T.right, hlim, e+1)$

$$end\ if.$$

D Data dictionary

This section describes the variables analyzed in the data.

Variable / Field Name	Form Name	Field Type	Field Label
id	biodata	text	Unique ID
doc_source	biodata	radio	Document Source?
surgical_burns	biodata	yesno	Surgical/Burns Patient?
date_adm	biodata	text	Admission Date
date_discharge	biodata	text	Date of discharge/death
hosp_id	biodata	dropdown	Hospital ID
random	biodata	radio	Randomized?
depid	biodata	text	Data Entry Person ID
date_today	biodata	text	Today's Date
ipno	biodata	text	Patients IPNO
age_years	biodata	text	Age (years)
age_mths	biodata	text	Age (months)
age_less1mnth	biodata	yesno	Age less than 1 month
age_days	biodata	text	Age (Days)
res_loc	biodata	text	Residence - Location/Sub-Location
res_dst	biodata	text	Residence - District
ref_hosp	biodata	radio	Referred to hospital?
ref_hosp_spec	biodata	text	Referred from which facility?
readm_hosp	biodata	radio	Re-admission to this hospital?
weight	biodata	text	Weight (kgs)
height	biodata	text	Height / Length (cm)
whz	biodata	dropdown	Weight-Height Z score
muac	biodata	text	MUAC (Mid-upper arm circumference) in cm
child_sex	biodata	radio	Gender
vacc_source	biodata	radio	Vaccination data source
vacc_status_text	biodata	dropdown	Vaccination status from text
vacc_opv_penta	biodata	radio	Number of doses of OPV/Penta
pcv10	biodata	radio	Number of doses of PCV 10 (Pneumococcal vaccine)
rotavirus	biodata	radio	Number of doses of Rota virus given
bcg	biodata	radio	BCG given
measles	biodata	radio	Measles given
par	biodata	yesno	PAR used
lo_illness_	history	text	Length of illness (days)
fever	history	radio	Fever

fever_dur	history	text	Fever duration
cough	history	radio	Cough
cough_dur	history	text	Cough duration
cough_2wks	history	radio	Cough >2 weeks
diff_breath	history	radio	Difficulty breathing
diarrhoea	history	radio	Diarrhoea
diarrhoea_dur	history	text	Diarrhoea duration
diarrhoea_14d	history	radio	Diarrhoea > 14d
diarrhoea_bloody	history	radio	Diarrhoea bloody
convulsions	history	radio	Convulsions
convulsions_no	history	text	Number of fits
fits	history	radio	Partial / focal fits
vomits	history	radio	Vomiting
vomit_freq	history	text	Vomiting frequency
vomit_everything	history	radio	Vomiting everything
diff_feed	history	radio	Difficulty feeding
tb_contact	history	radio	History of TB contact
temp	examination	text	Temperature (degrees celsius)
resp_rate	examination	text	Respiratory rate- RR (per minute)
pulse_rate	examination	text	Enter pulse value/Heart rate(HR)
oxygen_sat_done	examination	yesno	Oxygen saturation measured
oxygen_sat	examination	text	Oxygen saturation
bp_done	examination	yesno	Bp measured
bp_syst	examination	text	Systolic blood pressure (mmHg)
bp_diast	examination	text	Diastolic blood pressure (mmHg)
thrush	examination	radio	Thrush
lymph_nd	examination	radio	Lymph nodes > 1cm
wrist_sign	examination	radio	Wrist / rib signs for rickets
jaundice	examination	dropdown	Jaundice

sev_wasting	examination	radio	Visible severe wasting
oedema	examination	dropdown	Oedema of Kwashiorkor
umbil	examination	dropdown	Umbilicus
stridor	examination	radio	Stridor
c_cyanosis	examination	radio	Central cyanosis
indrawing	examination	radio	Indrawing
grunting	examination	radio	Grunting
acidotic__breathing	examination	radio	Acidotic breathing
wheeze	examination	radio	Wheeze
crackles	examination	radio	Crackles / crepitations
pulse	examination	dropdown	Pulse strength
cap_refill_cat	examination	dropdown	Cap refill
cap_refill	examination	text	CAP Refill
skin_temp	examination	dropdown	Extremities warm up to
pallor	examination	dropdown	Pallor / Anaemia
sunk_eyes	examination	radio	Sunken eyes
skin_pinch	examination	dropdown	Skin pinch (sec)
avpu	examination	dropdown	Disability (AVPU)
can_drink	examination	radio	Can drink / breastfeed?
stiff_neck	examination	radio	Stiff neck
bulging_font	examination	radio	Bulging fontanelle
irrit	examination	radio	Irritable
red_mov	examination	radio	Reduced movement / tone
mal1_order	investigations	yesno	Malaria test ordered at admission

mal1_result_avail	investigations	yesno	Malaria admission test results documented in the clinicians notes
mal1_result	investigations	radio	Malaria admission test results (from file or lab)
other_mal_test1	investigations	yesno	Other post-admission malaria test 1
other_mal_result1	investigations	radio	Other post-admission malaria test results1
other_mal_date1	investigations	text	Date other post-admission malaria test 1 done
other_mal_test2	investigations	yesno	Other post-admission malaria test 2
other_mal_result2	investigations	radio	Other post-admission malaria test results2
other_mal_date2	investigations	text	Date post-admission other malaria test 2 done
hb1_order	investigations	yesno	Hb test ordered at admission
hb1_test	investigations	radio	Test used to request Hb
hb1_result_avail	investigations	yesno	Hb results available
hb1_result	investigations	text	Hb results
hb_units	investigations	radio	Units for Hb results
gluc1_order	investigations	yesno	Glucose (RBS) ordered at admission
gluc1_test	investigations	radio	Type of glucose test requested
gluc1_results	investigations	text	Results
gluc_test_units	investigations	radio	Glucose test results units
chemistry	investigations	yesno	Chemistry investigations
chem_test	investigations	checkbox	Chemistry test requested
hiv1_order	investigations	yesno	HIV test ordered at admission
hiv1_test	investigations	radio	HIV test type
hiv1_result	investigations	dropdown	Results
hiv_inpt_order	investigations	yesno	Other HIV test orderd during inpatient stay
hiv_inpt_test	investigations	radio	HIV test type

hiv_inpt_result	investigations	dropdown	Results
micro_order	investigations	yesno	Microbiology test order
micro_tests	investigations	checkbox	Type of microbiology test ordered
micro_tests_date	investigations	text	Date microbiology test done
lp1_bedside	investigations	checkbox	Bed side exam of CSF
lp1_result	investigations	dropdown	Results (microscopy/culture)
csf_other	investigations	text	Other results of microscopy/culture
xray	investigations	checkbox	Any X-Ray done
urine	investigations	yesno	Investigations for urine ordered
urine_test	investigations	checkbox	Type of urine test ordered
urine_test_date	investigations	text	Date urine test was done
desctx2	admission_diagnosis	descriptive	<h1><font color="green">ADMISSION DIAGNOSIS</font></h1>
dx1_primary	admission_diagnosis	yesno	Clear primary admission diagnosis
dxg_pri_pres	admission_diagnosis	descriptive	<h1><font color=blue>Primary diagnosis (Enter ONLY the first diagnosis i.e ticked 1)</font></h1>
dx1_malaria	admission_diagnosis	dropdown	Malaria
dx1_pneum	admission_diagnosis	dropdown	Pneumonia
dx1_diarrhoea	admission_diagnosis	dropdown	Diarrhoea/ Acute GE (Gastro-Enteritis)
dx1_dehydrat	admission_diagnosis	dropdown	Dehydration
dx1_hiv	admission_diagnosis	dropdown	HIV / AIDS
dx1_malnutr	admission_diagnosis	dropdown	Malnutrition
dx1_anaemia	admission_diagnosis	dropdown	Anaemia
dx1_meningitis	admission_diagnosis	yesno	Meningitis
dx1_asthma	admission_diagnosis	dropdown	Asthma
dx1_rickets	admission_diagnosis	yesno	Rickets

dx1_tb	admission_diagnosis	yesno	Suspected TB
dx1_other_1	admission_diagnosis	dropdown	Other Diagnoses 1
dx1_other_3_text	admission_diagnosis	text	Other Diagnoses 3 in text
sec_dx	admission_diagnosis	yesno	Is there a secondary diagnosis?
dx2_malaria	admission_diagnosis	dropdown	Malaria
dx2_pneum	admission_diagnosis	dropdown	Pneumonia
dx2_diarrhoea	admission_diagnosis	dropdown	Diarrhoea/ Acute GE (Gastro-Enteritis)
dx2_dehydrat	admission_diagnosis	dropdown	Dehydration
dx2_hiv	admission_diagnosis	dropdown	HIV / AIDS
dx2_malnutr	admission_diagnosis	dropdown	Malnutrition
dx2_anaemia	admission_diagnosis	dropdown	Anaemia
dx2_meningitis	admission_diagnosis	yesno	Meningitis
dx2_asthma	admission_diagnosis	dropdown	Asthma
dx2_rickets	admission_diagnosis	yesno	Rickets
dx2_tb	admission_diagnosis	yesno	Suspected TB
dx2_other_1	admission_diagnosis	dropdown	Other primary Diagnoses 1
dx2_other_2	admission_diagnosis	dropdown	Other primary Diagnoses 2
dx2_other_3_text	admission_diagnosis	text	Other Diagnoses 3 not listed above
desctx3	treatment	descriptive	<h1><font color="green">TREATMENT - get information for this section from the treatment sheet</font></h1>
pen_pres	treatment	yesno	Xpen(Benzyl/Crystalline Penicillin) prescribed
pen1_route	treatment	radio	<i>route<i>
pen1_dose	treatment	text	<i>dose<i>
pen1_unit	treatment	radio	<i>units<i>
pen1_freq	treatment	dropdown	<i>frequency<i>
pen1_days	treatment	text	<i>duration (days)<i>
pen1_date	treatment	text	<i>Date Xpen was prescribed<i>
gent1_pres	treatment	yesno	Gentamicin prescribed

gent1_route	treatment	radio	<i>route<i>
gent1_dose	treatment	text	<i>dose<i>
gent1_unit	treatment	radio	<i>units<i>
gent1_freq	treatment	dropdown	<i>frequency<i>
gent1_days	treatment	text	<i>duration (days)<i>
genta1_date	treatment	text	<i>Date gentamicin was prescribed <i>
amox1_pres	treatment	yesno	Amoxicillin (Amoxyl) prescribed
amox1_dose	treatment	text	<i>dose<i>
amox1_unit	treatment	radio	<i>units<i>
amox1_freq	treatment	dropdown	<i>frequency<i>
amox1_days	treatment	text	<i>duration (days)<i>
amox1_date	treatment	text	<i>Date amoxicillin was prescribed<i>
ceftri1_pres	treatment	yesno	Ceftriaxone prescribed
ceftri1_route	treatment	radio	<i>route<i>
ceftri1_dose	treatment	text	<i>dose<i>
ceftri1_freq	treatment	dropdown	<i>frequency<i>
ceftri1_days	treatment	text	<i>duration (days)<i>
ceftri1_date	treatment	text	<i>Date ceftreaxone was prescribed<i>
caf1_pres	treatment	yesno	chloramphenical(CAF) prescribed
caf1_route	treatment	radio	<i>route<i>
caf1_dose	treatment	text	<i>dose<i>
caf1_freq	treatment	dropdown	<i>frequency<i>
caf1_days	treatment	text	<i>duration (days)<i>
caf1_date	treatment	text	<i>Date chloramphenical was prescribed<i>
metr1_pres	treatment	yesno	Metronidazole(flagyl) prescribed
metr1_route	treatment	radio	<i>route<i>
metr1_dose	treatment	text	<i>dose<i>
metr1_unit	treatment	radio	<i>units<i>
metr1_freq	treatment	dropdown	<i>frequency<i>
metr1_days	treatment	text	<i>duration (days)<i>
metr1_date	treatment	text	<i>Date metronidazole was prescribed<i>
cotrimox1_pres	treatment	yesno	Cotrimoxazole (Septrin) prescribed

cotrimox1_route	treatment	radio	<i>route<i>
cotrimox1_dose	treatment	text	<i>dose<i>
cotrimox1_days	treatment	text	<i>duration (days)<i>
cotrimox1_date	treatment	text	<i>Date cotrimoxazole was prescribed<i>
anti_tb1_pres	treatment	yesno	Anti-TBs prescribed
anti_malarials	treatment	yesno	Anti-Malarials prescribed (Qinine, Artesunate, Artemether, Coartem/AL)
quinl1_pres	treatment	yesno	Quinine loading dose prescribed
quinl1_route	treatment	radio	<i>route<i>
quinl1_dose	treatment	text	<i>dose<i>
quinl1_date	treatment	text	<i>date<i>
quinm1_pres	treatment	yesno	Quinine Maintenance dose prescribed
quinm1_route	treatment	radio	<i>route<i>
quinm1_dose	treatment	text	<i>dose<i>
quinm1_freq	treatment	dropdown	<i>frequency<i>
quinm1_days	treatment	text	<i>duration (days)<i>
quinm1_date	treatment	text	<i>Date Quinine was prescribed<i>
arte_pres	treatment	yesno	Artesunate prescribed
arte_route	treatment	radio	<i>route<i>
arte_dose	treatment	text	<i>dose<i>
arte_freq	treatment	dropdown	<i>frequency<i>
arte_days	treatment	text	<i>duration (days)<i>
arte_date	treatment	text	<i>Date Artesunate was prescribed<i>
artemether	treatment	yesno	Artemether prescribed
coart1_pres	treatment	yesno	Coartem (AL/Artemether Lumefantrine) prescribed
coart1_dose	treatment	text	<i>dose<i>
coart1_units	treatment	radio	<i>units<i>
coart1_freq	treatment	dropdown	<i>frequency<i>
coart1_days	treatment	text	<i>duration (days)<i>
coart1_date	treatment	text	<i>Date coartem was prescribed<i>

ceta1_pres	treatment	yesno	Paracetamol prescribed
salb_pres	treatment	yesno	Salbutamol / ventolin prescribed
salb1_route	treatment	radio	<i>route<i>
pred1_pres	treatment	yesno	Predinsolone prescribed
vita	treatment	yesno	Vitamin A prescribed
zinc1_pres	treatment	yesno	Zinc prescribed for diarrhoea
dextrose_10	treatment	yesno	10% dextrose bolus prescribed
dextrose_vol	treatment	text	Volume of 10% dextrose prescribed
adm_rx	treatment	yesno	Other admission treatment prescribed
adm_rx1	treatment	dropdown	Admission treatment1
adm_rx2	treatment	dropdown	Admission treatment2
adm_rx3	treatment	dropdown	Admission treatment3
adm_rx4	treatment	dropdown	Admission treatment4
adm_rx_other1	treatment	text	Admission treatment_other1
desctx4	supportive_care	descriptive	<h1><font color="green">SUPPORTIVE CARE</font></h1>
oxy_order	supportive_care	yesno	Oxygen ordered
oxy_rate	supportive_care	text	<i>flow rate<i>
oxy_route	supportive_care	dropdown	<i>route of admin<i>
oxy_date	supportive_care	text	<i>Date oxygen prescribed<i>
transf_order	supportive_care	yesno	Blood transfusion given
transf_vol	supportive_care	text	<i>volume of blood<i>
transf_hrs	supportive_care	text	<i>duration of transfusion prescribed<i>
transf_date_pres	supportive_care	text	<i>Date transfusion prescribed<i>
transf_date_gvn	supportive_care	text	<i>Date transfusion given<i>
dehyd_fluid	supportive_care	yesno	Fluids prescribed at admission for dehydration
iv_fluid	supportive_care	yesno	Child given IV fluids for dehydration
fluid_bolus	supportive_care	yesno	Fluid bolus given

fluid_bolus_type	supportive_care	dropdown	Type of fluid given for bolus infusion
fluid_bolus_dura	supportive_care	dropdown	Duration of bolus adminstration
fluid_pres1	supportive_care	dropdown	<i>type of fluid prescribed for dehydration<i>
other_fluid_presc	supportive_care	text	Other fluid prescribed
total_vol1	supportive_care	text	<i>total volume prescribed<i>
fluid_time1	supportive_care	text	<i>total duration prescribed<i>
fluid_step1_2	supportive_care	yesno	<i>Step 1 and 2 used <i>
oral_fluid	supportive_care	yesno	Oral fluids prescribed
fluid_pres2	supportive_care	dropdown	<i>type of fluid prescribed<i>
total_vol2	supportive_care	text	<i>total volume prescribed<i>
fluid_time2	supportive_care	text	<i>duration prescribed<i>
vol_stool	supportive_care	text	<i>volume with each stool<i>
fluid_maint	supportive_care	yesno	Maintenance fluids prescribe
fluid_maint_vol	supportive_care	text	<i>total volume of maintenance fluids<i>
malnourished	supportive_care	yesno	Was the child prescribed feeds at admission
feed_pres	supportive_care	dropdown	<i>type of feeds prescribed<i>
other_feed_pres	supportive_care	text	Other feed prescribed
feed_vol	supportive_care	text	<i>feed volume<i>
feed_vol_packets	supportive_care	text	<i>Number of packets in 24 hours<i>
freq_24hrs	supportive_care	text	<i>frequecy in 24 hrs<i>
date_feeds_start	supportive_care	text	<i>Date feeds were started<i>
fluid_feed_mon	supportive_care	yesno	fluid/feed monitoring chart availble
fluid_feed_monpres	supportive_care	text	Frequency of fluid/feed monitoring in 24hrs
vitals_chart	monitoring	yesno	vitals signs chart present
vital_monit_48hrs	monitoring	yesno	Vital signs monitored in the first 48 hours

temp_chart	monitoring	dropdown	Number of times temp monitored in 48 hrs
resp_chart	monitoring	dropdown	Number of times respiratory rate monitored in 48 hrs
pulse_chart	monitoring	dropdown	Number of times pulse rate monitored in 48 hrs
bp_moni	monitoring	yesno	Bp monitored
bp_charting	monitoring	dropdown	Number of times BP monitored in 48hrs
oxy_sat_moni	monitoring	yesno	Oxygen saturation monitored
oxy_sat_chart	monitoring	dropdown	Number of times oxygen saturation monitored in 48 hrs
disch_death_summ	discharge_information	yesno	Death / discharge summary present
outcome	discharge_information	dropdown	Outcome at discharge
dsc_condition	discharge_information	dropdown	Condition on discharge
follow_up	discharge_information	dropdown	Follow up care
dsc_dx1_primary	discharge_information	yesno	Clear primary discharge diagnosis
dsc_dxc_dig_sec	discharge_information	descriptive	Primary diagnosis (Enter ONLY the first diagnosis/ticked 1)
dsc_dx1_malaria	discharge_information	dropdown	Malaria
dsc_dx1_pneum	discharge_information	dropdown	Pneumonia
dsc_dx1_diarrhoea	discharge_information	dropdown	Diarrhoea / Acute GE (Gastro-Enteritis)
dsc_dx1_dehydrat	discharge_information	dropdown	Dehydration
dsc_dx1_hiv	discharge_information	dropdown	HIV / AIDS
dsc_dx1_malnutr	discharge_information	dropdown	Malnutrition
dsc_dx1_anaemia	discharge_information	dropdown	Anaemia
dsc_dx1_meningitis	discharge_information	yesno	Meningitis
dsc_dx1_asthma	discharge_information	dropdown	Asthma
dsc_dx1_tb	discharge_information	dropdown	TB
dsc_dx1_other_1	discharge_information	dropdown	Other Primary discharge diagnoses 1
dsc_dx1_other_2	discharge_information	text	Other Primary discharge diagnoses 2

dsc_dx2	discharge_information	yesno	Is there a secondary diagnosis?
dsc_dx2_malaria	discharge_information	dropdown	Malaria
dsc_dx2_pneum	discharge_information	dropdown	Pneumonia
dsc_dx2_diarrhoea	discharge_information	dropdown	Diarrhoea / Acute GE (Gastro-Enteritis)
dsc_dx2_dehydrat	discharge_information	dropdown	Dehydration
dsc_dx2_hiv	discharge_information	dropdown	HIV / AIDS
dsc_dx2_malnutr	discharge_information	dropdown	Malnutrition
dsc_dx2_anaemia	discharge_information	dropdown	Anaemia
dsc_dx2_meningitis	discharge_information	yesno	Meningitis
dsc_dx2_asthma	discharge_information	dropdown	Asthma
dsc_dx2_tb	discharge_information	dropdown	TB
dsc_dx2_other_1	discharge_information	dropdown	Other secondary discharge diagnoses 1
dsc_dx2_other_2	discharge_information	dropdown	Other secondary discharge diagnoses 2
dsc_dx2_other_3	discharge_information	text	Other secondary discharge diagnoses 3
dsc_dx2_other_4	discharge_information	text	Other secondary discharge diagnoses 4
dsc_rx	discharge_information	yesno	Discharge treatment prescribed
dsc_rx1	discharge_information	dropdown	Discharge treatment1
dsc_rx2	discharge_information	dropdown	Discharge treatment2
dsc_rx3	discharge_information	dropdown	Discharge treatment3
dsc_rx4	discharge_information	dropdown	Discharge treatment4
dsc_rx5	discharge_information	dropdown	Discharge treatment5
dsc_rx_other1	discharge_information	text	Discharge treatment_other1
dsc_rx_other2	discharge_information	text	Discharge treatment_other2

E Affiliations

- The University of Nairobi
- KEMRI - Wellcome Trust

F Supervisors

- Dr. Timothy Kamanu
- Mr. Paul Mwaniki