University of Technology Sydney

IS DATA VISUALISATION EFFECTIVE IN DETECTING ERRORS IN PATIENT-BASED REAL-TIME QUALITY CONTROL?

Name : Thesya Evania Gabrielle

Student Number : 14013371

Supervisor : Xu Wang

15 June 2024

Bachelor of Computing Science

Table of Contents

Abstract 3

Acknowledgement 3

Chapter 1 ..... ***4***

Introduction 4

Background 4

Aims 5

Objectives 5

Argument for your study 5

Future Impact/Significance 6

Chapter 2 ..... ***8***

Moving Average 8

Data Visualisation in Anomaly Detection 8

Chapter 3 ..... ***10***

Methodology 11

Data Acquisition 11

Data Handling 12

Data Pre-processing 12

PBRTQC Configuration 13

Simulated Error 14

Data Visualisation Process 14

Ethical Stance 15

Chapter 4 ..... ***16***

Chapter 5 ..... ***39***

Discussion 39

Future Works 39

Conclusion 40

References ***41***

Appendix ***43***

Ethical Declaration ***46***

Cover Sheet ***48***

Abstract

Clinical laboratories play a significant role in giving accurate results to patients. In ensuring the accuracy and reliability of the patients' test results, clinical laboratories use quality control methods called the Internal Quality Control (IQC). However, this widely used method relies on periodic testing of the control samples, which could lead to potential delays in error detection and increase the chances of giving false results to patients. As the technology emerges, a new method called the Patient-Based Real-Time Quality Control is introduced, which uses previous patient results as the quality control to detect errors.

This dissertation will explore the effectiveness of using data visualisation of PBRTQC to better understand the PBRTQC method and to improve error detection in patient results. The real-time monitoring of PBRTQC methodology enhances the precision and speed of detecting anomalies in the test results. The primary objective of this research is to develop a real-time visualisation and to evaluate the data visualisation framework using Microsoft Power BI to monitor and control PBRTQC with easy to use framework.

This study involves the acquisition of anonymised patient results from the Westmead Hospital Laboratories for a period of three months. The dataset includes details about the patient, the analytes, test results, and the timestamps. These data were also pre-processed to handle missing values and data duplicates. Control limits and moving averages are visualised in this PBRTQC system to monitor the test performance and test anomalies. Simulated errors, with constant, proportional, and random errors, were introduced to evaluate the effectiveness of moving averages in the PBRTQC system.

The results have shown that the data visualisation of PBRTQC helps improve error detection in the patient test result as it gives a better visualisation of the data. This could lead to better enhancement of clinical laboratory practices, reduce cost, and improve the accuracy of patient outcomes. The future work of this dissertation would provide better features such as alert systems for the PBRTQC visualisation. More advanced visualisation techniques and incorporation of machine learning algorithms could also be added to further improve the PBRTQC system in the future.

Acknowledgement

I would like to express my gratitude towards everyone who supported and contributed to completing this dissertation. First and foremost, I would like to thank my supervisor, Dr. Xu Wang, for his guidance, insightful feedback, and encouragement throughout my research journey. I am also grateful to the staff at Westmead Hospital Laboratories, especially Dr. Yusof, for his assistance with the data collection and tour of the lab, who also guided me to understand the analytes better and give valuable feedback on my reflections. On a personal note, I would like to thank my family for their support and encouragement throughout this journey and to my friends who have supported me morally during my challenging times. Finally, I would like to thank everyone who has supported me in various ways throughout my dissertation journey. This dissertation is a testament to your contribution and support.

Thank you all.

Chapter 1

Introduction

Clinical laboratories play a pivotal role in this modern world of healthcare, providing patients with diagnosis, treatment and analysis of their blood samples. These laboratories are entrusted to provide accurate and reliable test results to the patient. Hence, to ensure the accuracy and reliability of the test results, the clinical laboratory used a mechanism called quality control (QC) to validate the test results (Fleming & Katayev, 2015). Traditionally, clinical laboratories use Internal Quality Control (IQC), which involves using known concentration to determine the measurement errors in the machine periodically in fixed intervals, regardless of the machine's performance.

Despite the widespread uses of IQC, this traditional method relies on the periodic testing of the control samples and does not exactly reflect on the patient sample matrix. These limitations lead to delays in error detection and potentially provide patients with unreliable results. As concerns arise among the IQC users, researchers have developed an approach that can mitigate the limitations of IQC called the Patient-Based Real-Time Quality Control. This approach uses a statistical parameter over a defined number of patient results, to detect the anomalies and potential issues regarding assessing the blood samples. As just using the patient samples, PBRTQC have cut down the cost of the IQC and minimizes the delay of delivering crucial patient results.

Despite the power and cost-effectiveness that PBRTQC can give to clinical laboratories, PBRTQC is not commonly implemented as it is complex to understand. This raises a question, “Can there be a way to implement a Patient-Based Real-Time Quality Contol that is easy to understand even by regular people and is not complex to implement?”.

Although recent advances in data visualisation offer a powerful tool for enhancing the understanding and management of quality control. By transforming complex datasets into graphical representations, data visualisation can provide real-time insights into test performance and error patterns. This dissertation explored the application of data visualisation to detect errors in patient testing within a Patient-Based Real-Time Quality Control (PBRTQC) framework, aiming to improve the precision and reliability of patient diagnosis.

Background

Clinical laboratories play pivotal roles in the healthcare industry, performing a wide range of diagnostic tests on patients. There still is a lot of development in the testing process and availability in the laboratories, but in the clinical laboratory, the quality control techniques used remain unchanged. Internal quality control (IQC), such as liquid QC material, is still widely used by the clinical laboratory to assess analytical accuracy periodically (Cervinski, 2021). Although it has been the central QC strategy, the traditional IQC relies on a periodic analysis of QC materials, a known analyte concentration as the QC. However, the QC material does not reflect the actual patient sample matrix. It lacks commutability, leading to a lack of sensitivity and specificity. This could lead to inaccurate patient results (Badrick et al., 2019). Moreover, the use of IQC is costly, and laboratories need to test it at least twice a day,

which also leads to a massive amount of patient results being released before knowing an error in the measurement (Thaler et al., 2015).

The concept of patient-based real-time quality control (PBRTQC) was first introduced in around the 1960s. However, inadequate software or middleware prevents the system from performing complex calculations in real time (Badrick, Bietenbeck, Katayev, van Rossum, Loh, et al., 2020). PBRTQC, suggested by its name, uses patient samples to detect errors. It usually uses the analyte's mean or median to see the sample error.

PBRTQC itself is a form of anomaly detection that identifies a deviation from the expected behaviour of a given set of data (Mulero-Pérez et al., 2023). Furthermore, data visualisation is often associated with anomaly detection. It creates a visual representation of data for analysis and detecting anomalies in the (Vidyapeetham, 2021). The goal of data visualisation is to improve the performance of anomaly detection and avoid the waste of time eyeballing the data (Cui & Wang, 2017).

Aims

The research aim is to develop and evaluate a patient-based real-time quality control (PBRTQC) with its data visualisation to help laboratories understand PBRTQC better and to detect anomalies in the quality control. In assessing the performance of the data visualisation, it aims to generate a data visualisation that enables laboratory workers to better understand the moving distribution of the patient results for every batch.

Objectives

This research are primarily based on “Is using data visualisation of PBRTQC can help to detect errors in patient results?”.

With some objective of this research include:

- To acquire patient results dataset from the Westmead Hospital and understand the dataset.
- To apply data pre-processing on the dataset, including handling missing values, data transformation, feature scaling, and data splitting
- To determine the PBRTQC configuration including the data
- To develop a data visualisation of the implementation of PBRTQC.
- To develop a dashboard that is easily understood by healthcare professionals.
- To evaluate the effectiveness of data visualisation using data validation and performance metrics.

Argument for your study

This research is conducted to improve the current quality control system in the clinical laboratory. Ensure the quality and accuracy of patient test results are being distributed to the patient. As the existing procedures, liquid internal quality control (IQC), are costly and slow to detect errors, patient-based real-time quality controls are claimed to address such issues. The use of PBRTQC as complementary to the existing quality control could improve the confidence of diagnostic and the speed of error detection in laboratory diagnostics. The results given in PBRTQC could affect the decision to use IQC for better confidence in the quality control

process. Hence, a good visualisation of the PBRTQC results is essential to understand the results better and improve the decision-making.

A good statistical process or machine learning embedded in PBRTQC could not be beneficial if no one understands the results or cannot sense the error and alert given. Hence, a good visual dashboard is needed for a laboratory to know that an error has occurred in the machine. Moreover, although implementing PBRTQC is beneficial, many clinical laboratories do not want to change it because it is difficult to understand (Badrick, Bietenbeck, Katayev, van Rossum, Cervinski, et al., 2020). Hence, it is efficient to have a simple and easy-to-understand dashboard for easier interpretation of the data and detection of anomalies. PBRTQC also requires a huge amount of data for it to run effectively. The complexity of huge data can potentially delay the responses to a critical error in the data. Still, through good data visualisation techniques, it aims to equip laboratory workers to navigate through the quality control result seamlessly and enhance the understanding of the data itself.

Furthermore, data visualisation could not only benefit the anomaly detection in the data. It also helps in the data pre-processing. Better data visualisation of the analyte has a major contribution to choosing the statistical process of the PBRTQC. It also opens up an opportunity for improvement in the PBRTQC to make the error detection more accurate. By visualising the distribution, we could also discard analytes that are not suitable for PBRTQC, which analytes could make the error detection inaccurate due to skewed data distribution. Furthermore, through data visualisation, a better test split could be used based on age, sex, or in-patient and out-patient.

Moreover, the use of good data visualisation in PBRTQC could not only benefit the clinical laboratory itself by reducing the time and cost of quality control. An effective data visualisation in PBRTQC could enhance the precision of quality control, which contributes to improved patient outcomes. Data might be useful, but without a good data visualisation, a good insight cannot be drawn that might be crucial for patient outcomes. The research of enhancing quality management through better data visualisation in PBRTQC not only benefits healthcare professionals but also benefits the healthcare outcome as a whole.

Future Impact/Significance

The integration of data visualisation techniques in PBRTQC offers a significant contribution on not only to healthcare quality management, but it could also impact the research industry, ICT knowledge, and business practice. As PBRTQC is implemented in the laboratories, the healthcare system is becoming more data driven. Hence, the ability to extract meaningful insights from such data becomes important. The study of data visualisation in PBRTQC could revolutionise the real-time quality control approach using patient results.

One primary future impact and significance lies in enhancing pattern detection and anomaly detection in the patient dataset. The interactive dashboard and trend analysis will make laboratory workers to identify anomalies or deviations from expected results. A good visual representation can enhance the speed of error detection, which contributes to less false patient results being distributed. Moreover, good data visualisations and storytelling could improve the communication of such complex information to other healthcare workers and improve decision-making.

Moreover, in business practice, implementing PBRTQC could lower the cost of quality control. Implementing good PBRTQC could reduce or eliminate the use of IQC in clinical laboratories. Applying PBRTQC as a complementary quality control has reduced the use and the cost of IQC by a huge percentage. IQC that are run every four hours could be reduced to 1 or 2 hours per day or being tested if error is detected in PBRTQC to enhance the confidence of the error. Hence, PBRTQC has an impact on reducing the quality control cost in clinical laboratories, which could be allocated for other uses. Better enhancement of PBRTQC could also substitute the use of IQC in the clinical laboratory, which further reduces the cost. The reduction of quality control cost in using PBRTQC could benefit not only a big laboratory it could also benefit small laboratories that cannot afford IQC.

Then, in research practice, data visualisation in PBRTQC could address the rarity of literature addressing the use of data visualisation to better understand the PBRTQC. There is not much literature research that addresses the use of data visualisation in machine learning in general or healthcare data. However, through various research, it is known that good data visualisation could improve the understandability of the machine learning results. As the literature research remains rare, this research could contribute more to using data visualisation in enhancing quality control.

Chapter 2

This chapter will give a review of literatures related into the topic to deepening the understanding of this study.

Moving Average

This subsection will discuss one of the most used statistical methods in the PBRTQC field, which is the Moving Average (MA) method. This method will also be implemented in this study.

Moving Average (MA) was first introduced in 1965 (Ng et al., 2016). MA utilises an unweighted average of patient result to assess the performance of a specific test. It works by continuously recalculating its value when a new eligible patient's result is added to the method. When the new result is added, the earliest result in the time block n is removed. The MA rule is typically defined by a block size and control limit (Badrick, Bietenbeck, Katayev, van Rossum, Loh, et al., 2020). The control limit helps in preserving the quality of patient samples used to determine errors in the PBRTQC. Hence, when the patient results are above the control limits, it will not be injected into the MA, and are classified as outlier. This will provide more stable measure of the Moving Average as it smooths out random fluctuations or noise in the patient samples, giving out a more reliable assessment of the overall testing process.

However, this approach also has a drawback, as it can also be slow in detecting error in the quality control. Hence, numerous errors in the patients outcome can be released before the chart is detected. This could be handled by re-analysis and correction following the detection of errors in the process. To minimize the error, (Fleming & Katayev, 2015) has introduced a "release from the back" strategy, where the earliest results of the MA block is released only when the n+1 results or block size + 1 results have been analysed and the block results are found withing the specific measurement of the control limit. These strategies will hold back the patient samples being released before the block size passed the MA QC.

Despite the drawbacks of MA, using MA in real-time monitoring could be highly effective. Due to the usage of previous patient samples, gradual changes in the Moving Average could indicate an error in the testing machine or in the overall health of the population. Since, the error indication plays a crucial part in Moving Average. This study also wants to highlight the effectiveness of using data visualisation to detect the trends on the Moving Average, which can also detect errors in the patient test results.

Data Visualisation in Anomaly Detection

Data visualisation on the other hand, has achieved success in analysing trends and detecting anomalies in different fields. There has been a successful works in anomalies detection on network access by detecting unauthorized IP address using data visualisation. The data visualisation has helps them to identify trends and anomalies and improves their decision making. They created simple scripts and measures to help extract useful features in the dataset. This approach are said to give more valuable insights into the decision making process (Cui & Wang, 2017).

On the other hand, in the agriculture field, data visualisation has also help to get better understand of big time series data in the digital agriculture area. It discusses that big data that are generated through various data sources in this digital area hides a lot of potentially useful information. It shows that data visualisation can help to discover knowledge from big data and is critical to understand complex dataset and communicating insights to stakeholder. Especially in time series data that is collected over time, visualisation is important in detecting pattern, trends, and anomalies that allies in the data (Dhaliwal, et al., 2024).

There are not many literature reviews regarding the effectiveness of data visualisation to detect errors in the systems. Data visualisation usually acts as a complementary to visualise the results of machine learning algorithms. In the healthcare industry itself, there have been no research that are conducted regarding the use of data visualisation that act as a quality control. However, from the literature review it is concluded that data visualisation can be effective to improve decision making a, identifying trends, errors, and anomalies that relies in the data. It also derives potential insights that hides behind the dataset. Hence, this study will access the effectiveness of data visualisation in PBRTQC in the healthcare industry, especially in clinical laboratories.

Chapter 3

This chapter will describe the dataset and the main methodology used in this study.

Population and Sample Description

The sample dataset consists of 440,114 patient results from a Westmead Hospital laboratory. The dataset consists of details on the patient test results being collected. This included the details of the machine of the results being taken including the Facility and Ward. Then, the patient details, which include the MRN, which is the unique identifier and Age. Next, the information on the test being taken including, the Accessin, or the unique identifier of the specimen, the DTA or the type of test results, and the service resources. Then, it also includes the time of test being ordered, the sample being collected, and the test result being recorded by the service resources. The sample dataset features are described as below :

Feature Names	Feature Type	Data Type	Description
Facility	Categorical	String	The facility which the sample was taken from
Ward	Categorical	String	The ward within the facility which sample was taken from
MRN	Categorical	String	The unique identifier for a patient
Age	Numerical	Int	The age of the patient
Service resource	Categorical	String	The source of the measurement
Accessin	Categorical	Int	The unique identifier for the specimen being tested
DTA	Categorical	String	The type of test result
Result	Numerical	Float/Int/String	The result of the test
Ord	Numerical	Datetime	The time the test was ordered
Coll	Numerical	Datetime	The time the sample was collected
Recorded	Numerical	Datetime	The time the sample was recorded by the service resource
Verified	Numerical	Datetime	The time the sample being verified

Table 1: Dataset description

This study's population includes patients being tested in 44 different facilities in 57 different wards from 31 December 2023 at 10.21 p.m. to 31 March 2023 at 11.07 p.m. for 3 months. This includes person with various background, age, and gender. The total population are 21,859.

Random sampling is used in this research. Random sampling itself means the dataset is taken at random, meaning the samples are not taken on a certain group or gender, but it is randomly taken on the facilities available for data collection. The sample itself is diverse in terms of demographic characteristics, with the age range of the population ranging from 1 to 102 years old. The gender is not specified, but it is informed by the hospital staff that collected the samples that the sample is random, meaning it could be a male or female population.

This study adheres to ethical guidelines. The dataset collection has gone through ethical procedures in the Westmead Hospital. Confidentiality of the population is maintained

throughout the research process. Any sensitive data, such as names, is anonymised by using MRN or a unique identifier for the patient to ensure the privacy and well-being of the patient.

Methodology

The methodology first involves the data acquisition from the Westmead Hospital Laboratories. It is then input into Microsoft SQL Server, which handles and pre-processes data. The data will then be transferred using direct query into Microsoft Power BI.
Here is the methodology framework of this study, which extracts the data directly from its sources at the time of the query, which in this case is the Microsoft SQL Server database. In Power BI, the data will be visualised to show the distribution of the dataset. Furthermore, interactive dashboards are also used to visualise the analytes' control limits and moving averages in different service resources.

Simulated errors are injected into the dataset to evaluate the effectiveness of the visualisation and error detection through the moving averages. This involves the use of constant error, proportional error, and random error.

This is the diagram figure of the methodology used in this study.

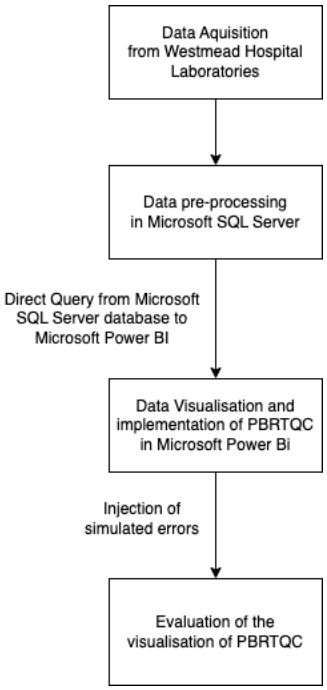

Figure 1: Methodology Diagram

Data Acquisition

The dataset used in this thesis will be based on real-life patient data from several hospitals, which have been anonymised. The dataset contains information about the patient samples taken in different facilities and different wards, containing the details of the machine, analyte and the patients. However, the key information that will be mainly used is the analyte tested, the results and the time it is collected.

In detail, the dataset contains 440 113 patient samples collected over three months from January 2023 to March 2023 from various facilities and wards. It contains a unique identifier for the patients called the Medical Record Number (MRN), the age of the patient, the machine which the sample was taken from, the type of test performed, the result of the test, the time the sample was ordered, obtained, measured, and verified.

Data Handling

The dataset will be handled using a database called the Microsoft SQL Server. The data will be exported into a table. Then, the data will be pre-processed in the Microsoft SQL server using queries. The pre-processed step will be stored in a new table for further pre-processed, which is included in the Data Pre-Processing step in the chapter below. After all of the pre-processing steps, the data will be saved into a view.

The Microsoft SQL Server was chosen because it can handle direct queries in Microsoft Power BI, which is used for the data visualisation of this project. The Power BI provides a direct query that retrieves data directly from its source at the time of the query, allowing real-time connectivity. Since the PBRTQC needs to be as real-time as possible, it is needed for the data handling to handle the data in real time. Hence, additional data added to the SQL Server will also be automatically retrieved by Power BI and updated in the dashboard.

To ensure the security of the data and confidentiality throughout the research process, the MRN of the patient is discarded, and the data remains anonymous.

Data Pre-processing

Data pre-processing is a crucial step in data analysis as it involves preparing and cleaning the data to make it suitable for analysis. The data pre-processing will be done on Microsoft SQL Server. It is first stored in a database, and then queries will be performed to pre-process the data.

The result is the target value, which is the most important value in the data. However, it contains either text data such as cancelled, incorrect or insufficient, decimal data, or an imprecise value such as > 0.1. It is challenging to have different kinds of data types in a column. Hence, we preserve the float data type, which are results that contain number and decimal values and delete the rest that is not a float data type.

Next, it is noticed that in the age field, a huge amount of people have the age of 123, which in this case is identified as missing value, which records are also ignored.

Next, there are two features that are highly correlated with each other, which are the facility and the ward. The facility feature indicates the facility which the sample was taken from, and the ward feature indicates the ward within the facility which the sample was taken from. Hence, only one feature can be kept, and the other can be ignored.

Furthermore, we will focus on the service resources and analytes individually in this data. Due to that factor, we separate the dataset into different service resource, and different DTA. One of the service resources has a lesser amount than others, which is “WMD Centrelink”, which only contributes 0.75%. It will not be used in the visualisation process.

The features that are ignored are not discarded in the dataset as it may be used in the future. It is simply ignored and is not visualised in the dataset; this includes the access.

The queries used in the data pre-processing could be seen in the appendix of this dissertation.

PBRTQC Configuration

In this PBRTQC, control limits and Moving Average visualisation will be used to provide actionable insights. This section will discuss the calculation algorithms behind the control limit and moving averages.

Control limits are statistical boundaries that is set in a control chart to monitor the stability and the performance of the testing machine. Generally, control limits are standard deviation-based, giving support and lower limit for the PBRTQC.

There are three components of the control limits, which include the Central Line (CL), which is the mean value of the test results, Upper Control Limit (UCL) which is set at three standard deviations above the mean, and Lower Control Limit (LCL) that is set at three standard deviations below the mean. In this study, the algorithm of the central line will use the mean of the previous month.

The calculation to calculate the mean, standard deviation, and control limits are as follow :

1. Mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. Standard Deviation

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$$

3. Upper Control Limit (UCL)

$$UCL = \mu + 3\sigma$$

4. Lower Control Limit (LCL)

$$LCL = \mu - 3\sigma$$

Where x_i is the test results and n is the number of test results.

Next, Moving Average (MA) is used as the PBRTQC algorithmn. MA is continuously recalculated when new patient results are introduced. However, when a new patient result is added, the earliest patient result will be dropped (Badrick, Bietenbeck, Katayev, van Rossum, Loh, et al., 2020).

In this study simple moving average is used with the calculation could be seen as below :

$$MA = \frac{1}{n}\sum_{i=0}^{n-1} x_{t-i}$$

Where x is equals the test result, t refers to the batch number, and n refers to the moving average periods. The MA in this study will be calculated in different periods, including 4 hours and 1 day of window size. It will also have a truncation limit to only allow a certain amount to be used in the Moving Average, which is the data that is below the UCL and LCL.

Simulated Error

Simulated errors are introduced to the dataset to evaluate the effectiveness of moving averages in detecting errors or anomalies in the data. There are three types of errors that usually occur in the analytical assay, which are constant error (CE), random error (RE), and proportional error (PE). The simulated errors are introduced to just a certain time frame in the data, which is from 1 February 2023 to 7 February 2023.

These are the equations of the errors :

1. Constant Error (CE)

$$x' = x + CE$$

2. Proportional Error (PE)

$$x' = x\ (1 + PE)$$

3. Random Error (RE)

$$x' = x + \epsilon$$

With x is the original value of the test result and ϵ , which is a random error sampled from a normal distribution $N(0, \sigma)$ with mean 0 and standard deviation σ.

In this study, the CE is 20, the PE is 0.5, and the RE is set to be 0.5.

Data Visualisation Process

The data visualisation process will be conducted in Microsoft Power BI. First, the data will be imported from Microsoft SQL Server using Direct Query. As a direct query retrieves data directly from its source, whenever the data is edited in SQL Server, it will also be updated in Power BI.

Firstly, the analytes are plotted into a bar chart for better understanding. Includes the distribution of DTA in different service resources. The age distribution and the average of analytes in the data source.

To visualise the PBRTQC in Power BI, a few new measures have been added to help create the visualisation: average, standard deviation, UCL, and LCL of every analyte. New measurements are added to the original dataset and the data set with simulated errors for the moving averages.

A new table called the Date table is created to store the Date, Month, and Year data. A relation is then created between the Recorded feature in each dataset and the Date table.

After every measure and relations are created, dynamic and interactive dashboards are created for the control limit and moving averages. A few interactive elements are added to help the visualisation, allowing the user to choose between the service resources, date, and analytes on the moving average dashboard.

Ethical Stance

This research adheres to strict ethical guidelines to ensure the confidentiality and privacy of the patient data. All sensitive information regarding the patient is anonymised, and ethical approval has been obtained from the Westmead Hospital. This project ensures that the well-being of the patients from whom the data is collected is maintained throughout the research process, and the data is handled with precious care to prevent any misuse.

The use of quality control and research usually involves using patients'' data without their consent. However, in return, the patient gets better benefits by enhancement of health outcomes. In this research, normal informed consent will be waived, as the risks of the research are minimal, and there are no rights of the individual will be violated or harm to the patient that data are used for the result.

Moreover, the dataset used in this research remains confidential and adores the patients' privacy. There is no patient information such as name, email, address, or phone number being used in this research. Every patient is anonymised by using a unique identifier for each patient.

Chapter 4

The primary objective of this research is to develop a comprehensive data visualisation dashboard using Power BI that is highly accessible and interpretable to detect anomalies in Patient Based Real-Time Quality Control (PBRTQC). This chapter will represent the visualisation and analysis of the patient dataset from Westmead Hospital Laboratory.

Result

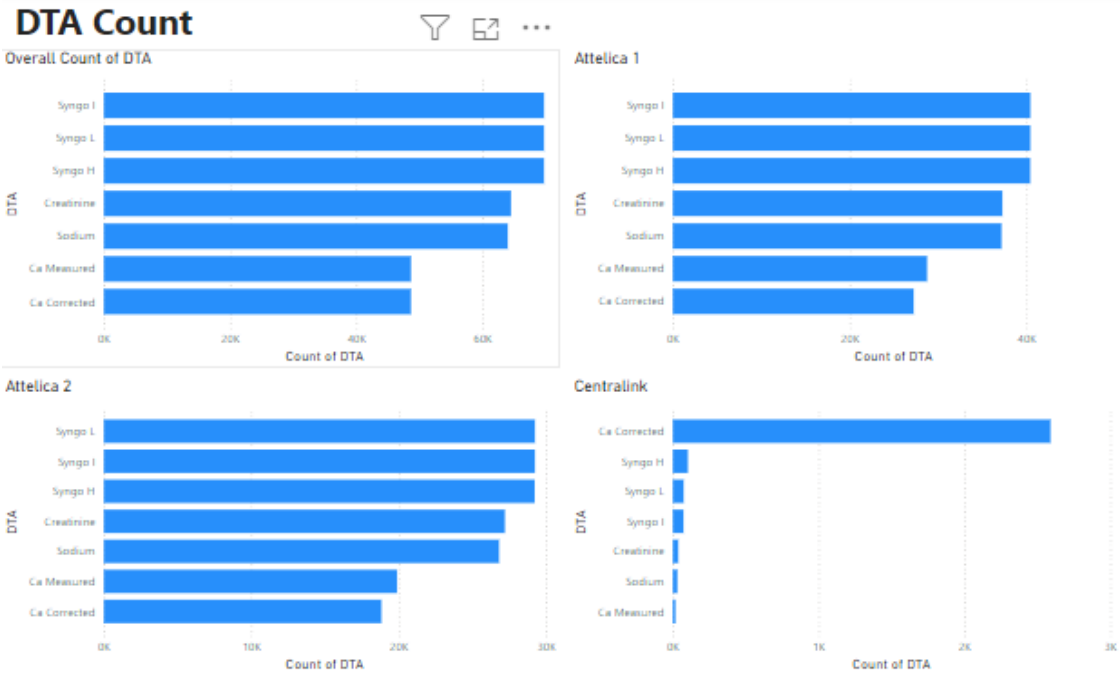

Figure 2 : Distribution of the Analytes in Different Service Resource

Figure 2 describes the distribution of the analytes on different service resources; it is shown that Atellica 1 and Atellica 2 have a similar distribution of the analytes. Hence, just one service resource is used for further visualisation, which is Atellica 1.

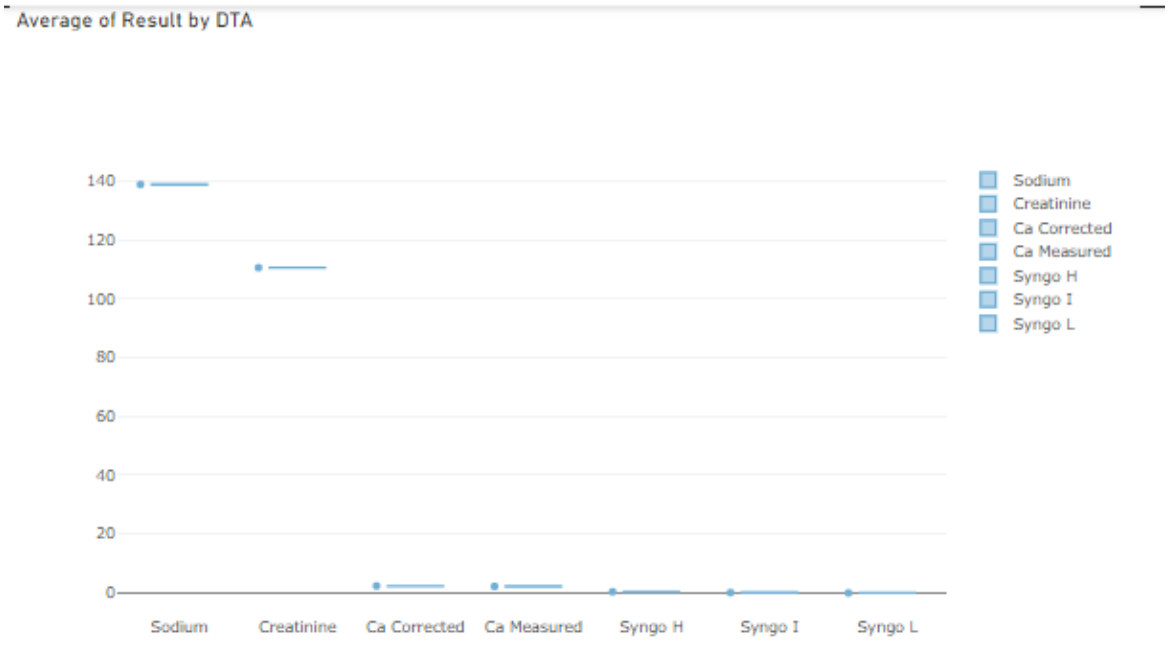

Figure 3 : Average Results of the Analytes

Through the average results, we can see that Syngo H, Syngo I, and Syngo L have an average of nearly 0, which does not give much information in the visualisation. In this study, we will further analyse just 3 of the main analytes: Sodium, Creatinine, and Calcium.

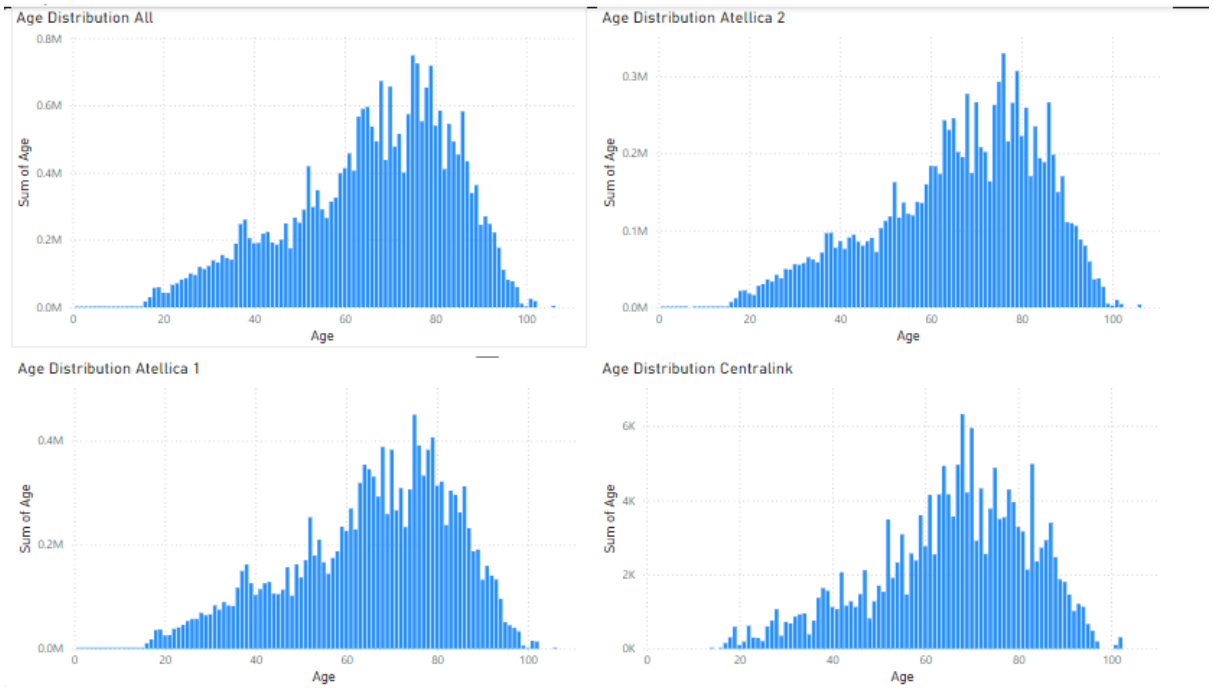

Figure 4 : Age distribution on different service resource

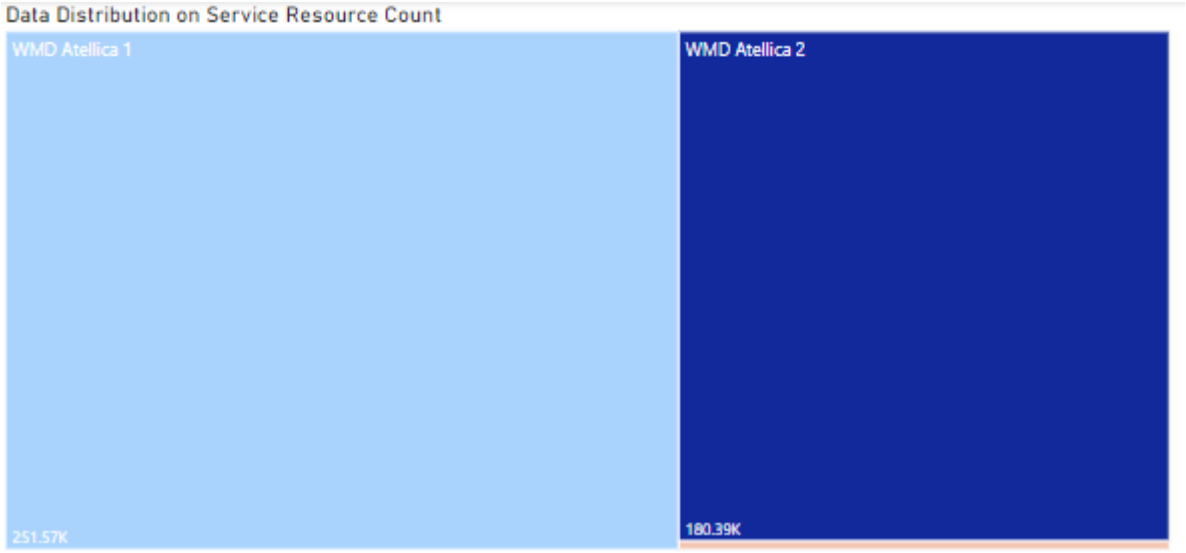

Figure 5 : Data distribution of test results on different service resource

From Figure 5, we can derive that the test results taken on the Centrelink service resources are very little compared to Atellica 1 and Atellica 2. It only contributes below 1% of the total test results. Since Atellica 1 has the most patient results sample, further analysis and visualisation of the control limit and moving averages will focus more on patients' results in the Atellica 1 service resource.

Visual Dashboard of Control Limits Control of Sodium

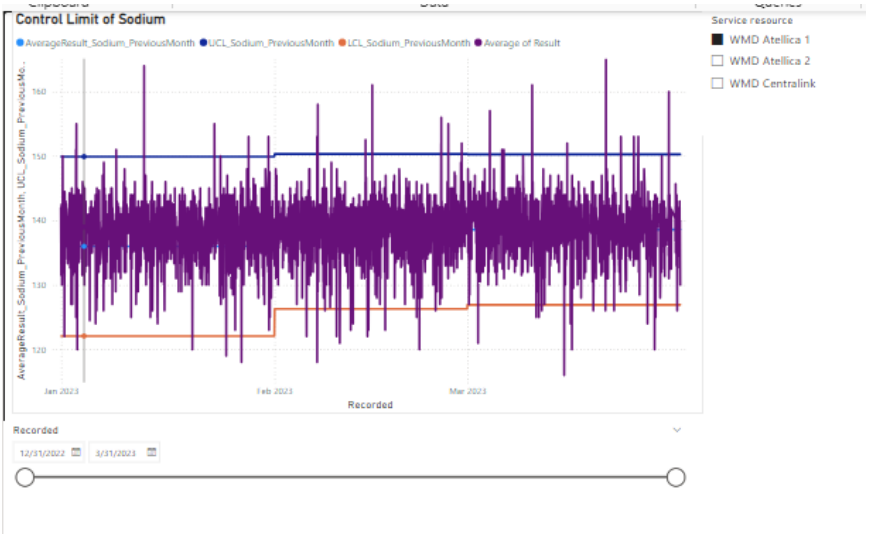

Figure 6 : Control limits of sodium from January 2023 to March 2023

Figure 7 : Control limits of sodium on January 2023

Figure 8 : Control limits of sodium on February 2023

Figure 9 : Control limits of sodium on March 2023

The figures above show the control limit of sodium analytes. The control limit shows where the main distribution lies near the average of the results. It also does not have a lot of outliers. We can see in control limits in January 2023 that the distributions are mainly above the analytes; this is because the averages are derived from the data of the previous months. While the amount of data in December 2022 is really low. This makes the average does not reflect the actual average.

Visual Dashboard of Moving Averages of Sodium

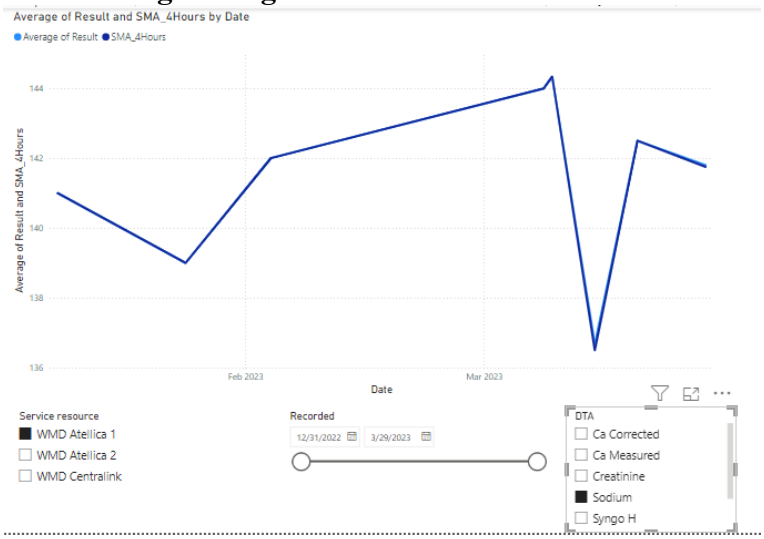

Figure 10 : Moving average of sodium in the time frame of 4 hours

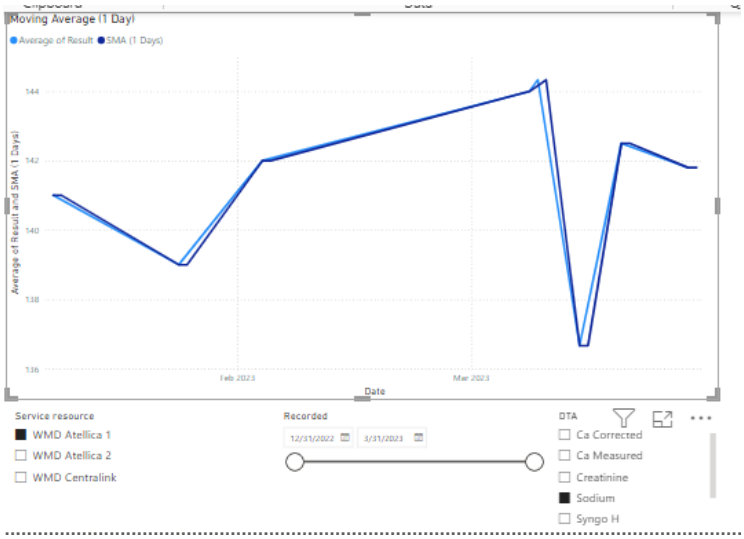

Figure 11 : Moving average of sodium in the time frame of 1 day

From Figures 10 and 11, we can see that there are valleys in the middle of March 2023. This means there are changes in the average of the data in the middle of March. If we verify using the control limit, there are changes in the value in the middle of March. There were a lot of outliers visible during that period. However, the changes are so subtle that we can ignore the error.

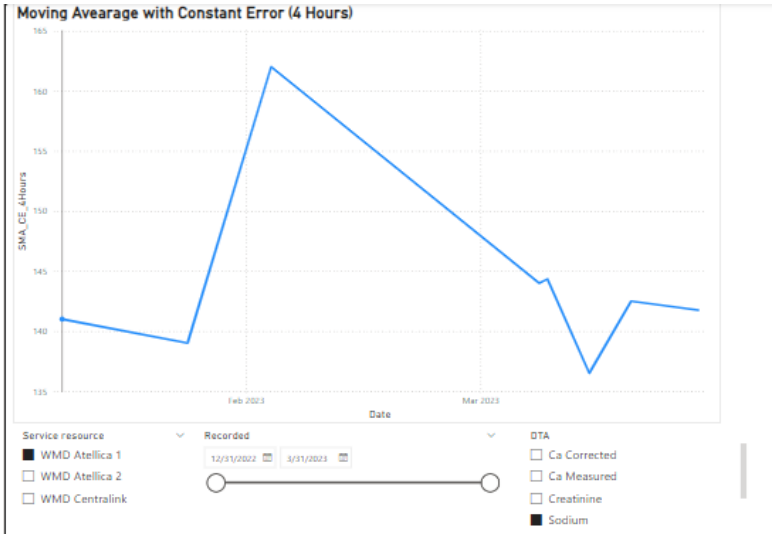

Figure 12 : Moving average of sodium with constant error in the time frame of 4 hours

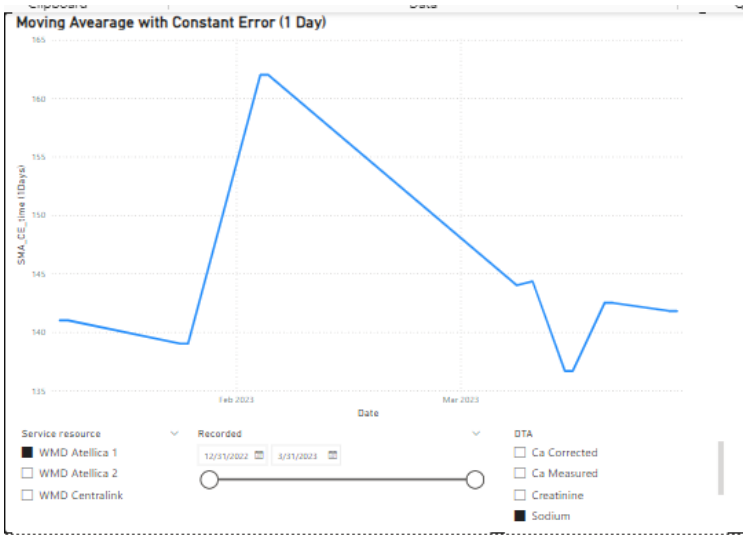

Figure 13 : Moving average of sodium with constant error in the time frame of 1 day

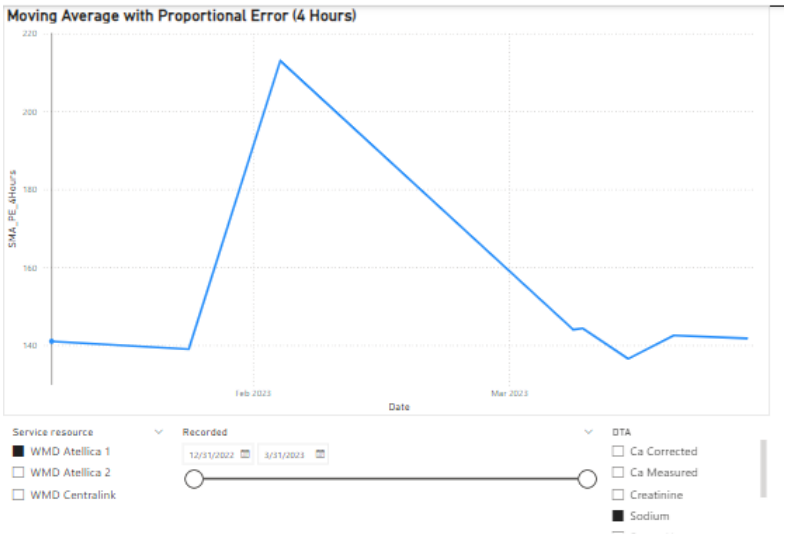

Figure 14 : Moving average of sodium with proportional error in the time frame of 4 hours

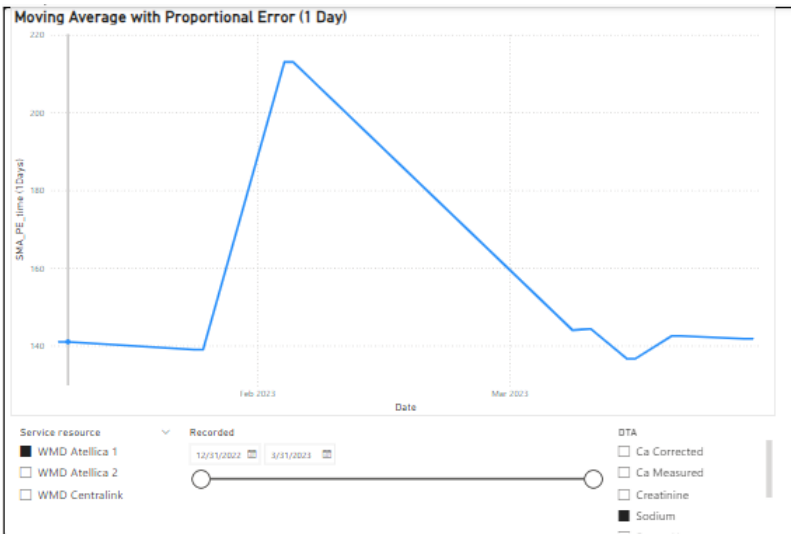

Figure 15 : Moving average of sodium with proportional error in the time frame of 1 day

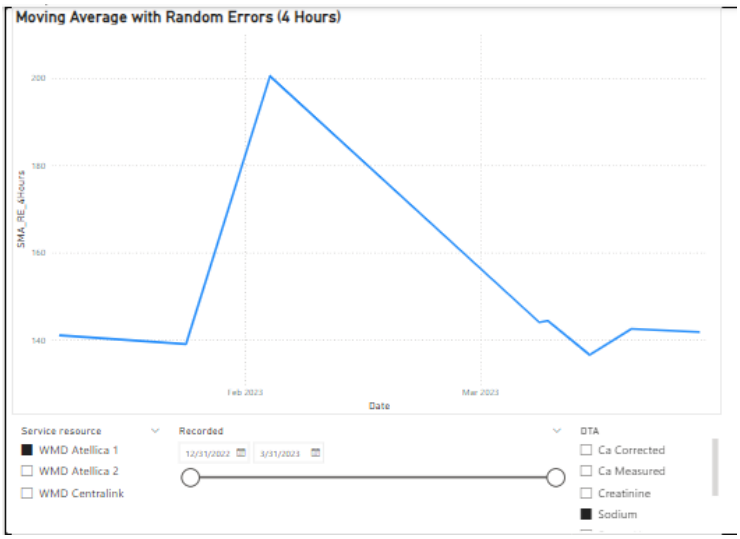

Figure 16 : Moving average of sodium with random error in the time frame of 4 hours

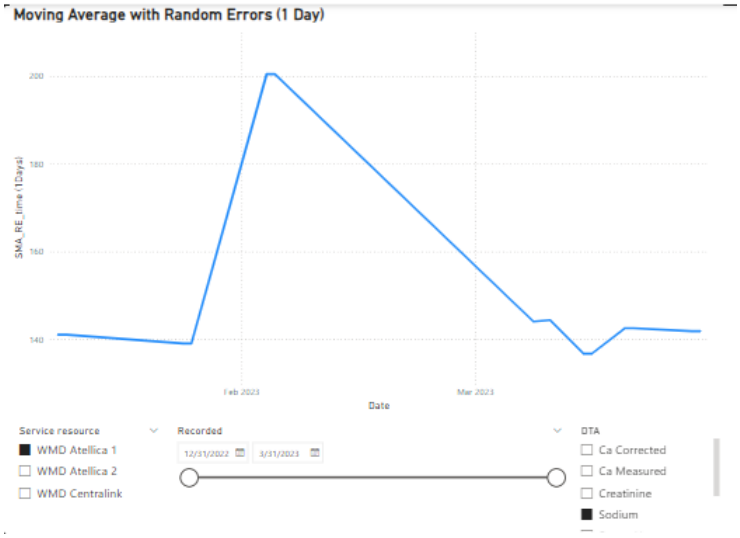

Figure 17 : Moving average of sodium with constant error in the time frame of 1 day

From figures 12 to 17, we can derive that the simulated error has highly affected the moving averages of the sodium analyte. Just a small error added into the dataset has highly changes the average of the distribution. The sudden fluctuations indicate errors in the patient's test results. This is due to a simulated error added from February 1 to February 7. The fluctuation is visible enough that we can derive the error in the data, indicating that it is highly effective to use data visualisation to detect errors.

However, if we look at the moving average on the normal dataset, we can see that there are fluctuations in the moving average. Meanwhile, in the dataset with simulated error, the fluctuations are not as visible, and the most visible fluctuations are the changes that are caused by the errors. Hence, we cannot generalise that fluctuations mean an error in the dataset. We also need to see how high the fluctuation is.

Visual Dashboard on Control Limit of Calcium Measured

Figure 18: Control limits of calcium measured from January 2023 to March 2023

Figure 19 : Control limits of measured corrected January 2023

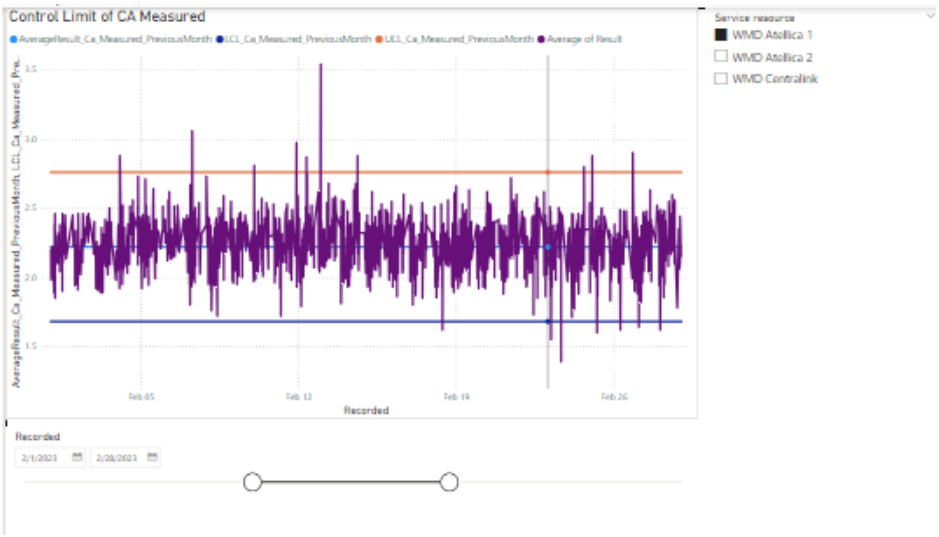

Figure 20 : Control limits of measured corrected February 2023

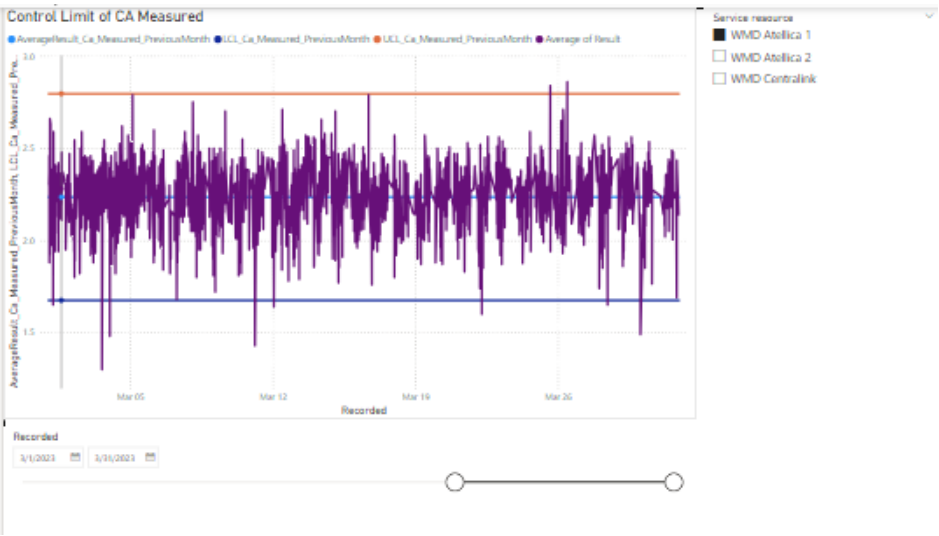

Figure 21 : Control limits of calcium measured January 2023

The figures above show that the calcium analytes distribution has more variance than sodium analytes. We can also see that the averages on the control limits on January 2023 are also off, which is like Sodium, which needs to be adjusted in the future.

Moving Averages of Calcium Measured

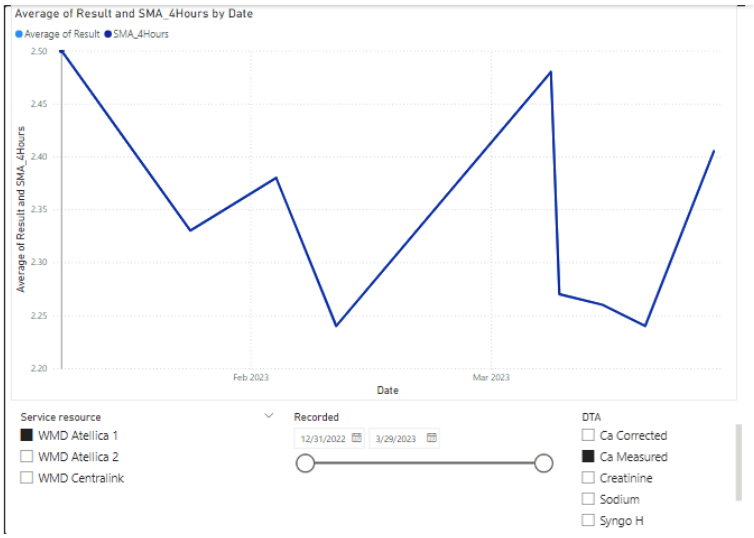

Figure 22 : Moving average of calcium measured within the time frame of 4 hours

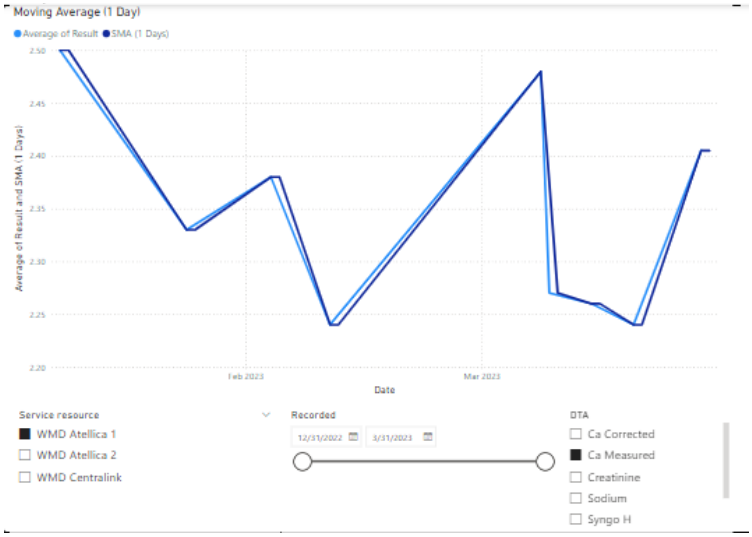

Figure 23 : Moving average of calcium measured within the time frame of 1 day

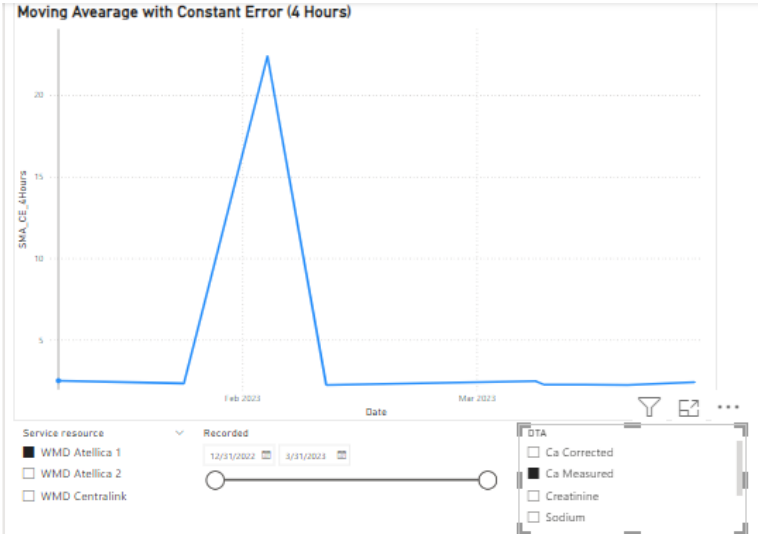

Figure 24 : Moving average of calcium measured with constant error in the time frame of 4 hours

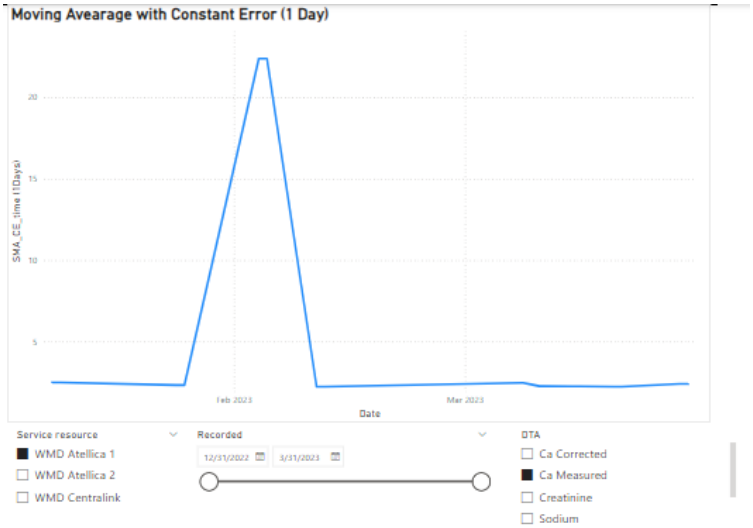

Figure 25 : Moving average of calcium measured with constant error in the time frame of 1 day

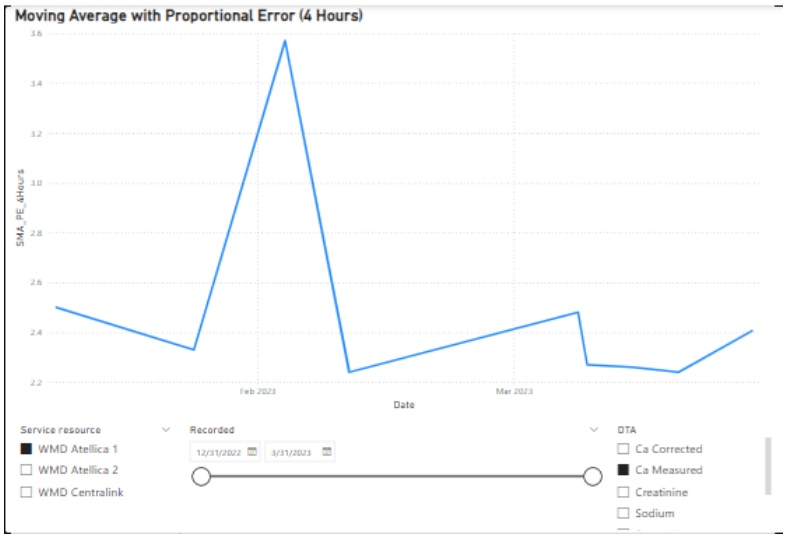

Figure 26 : Moving average of calcium measured with proportional error in the time frame of 4 hours

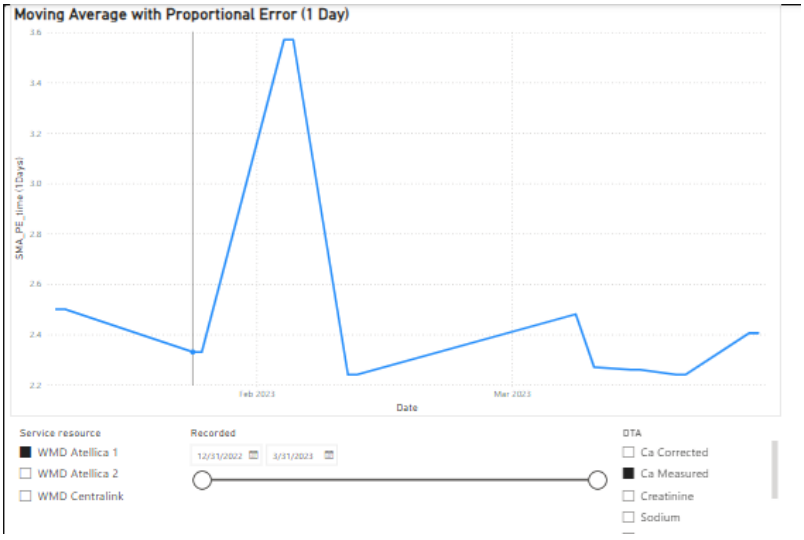

Figure 27 : Moving average of calcium measured with proportional error in the time frame of 1 day

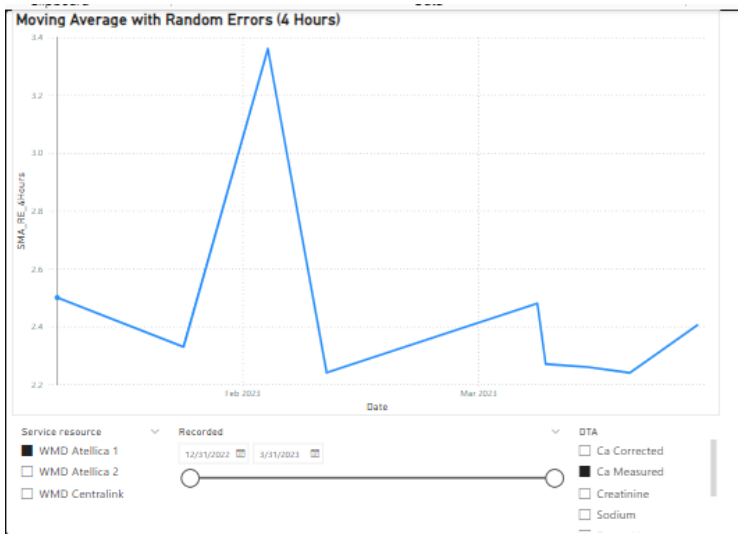

Figure 28 : Moving average of calcium measured with random error in the time frame of 4 hours

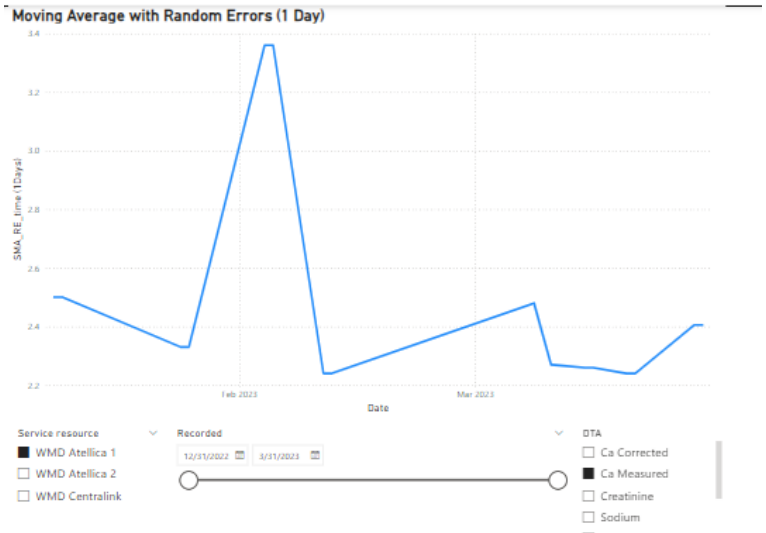

Figure 29 : Moving average of calcium measured with random error in the time frame of 1 day

From the figures above, we can see that there are fluctuations in the moving average, which means there are changes in the control limit. Similar to the moving average of sodium analytes, we can see that errors have highly affected the moving averages in the patient results. From the differences between the moving average of the original dataset and the dataset with simulated error, we can see in data visualization that the fluctuations that are caused by the error are more visible than the fluctuations in the moving average.

Visualisation Dashboard of Ca Corrected Control Limit

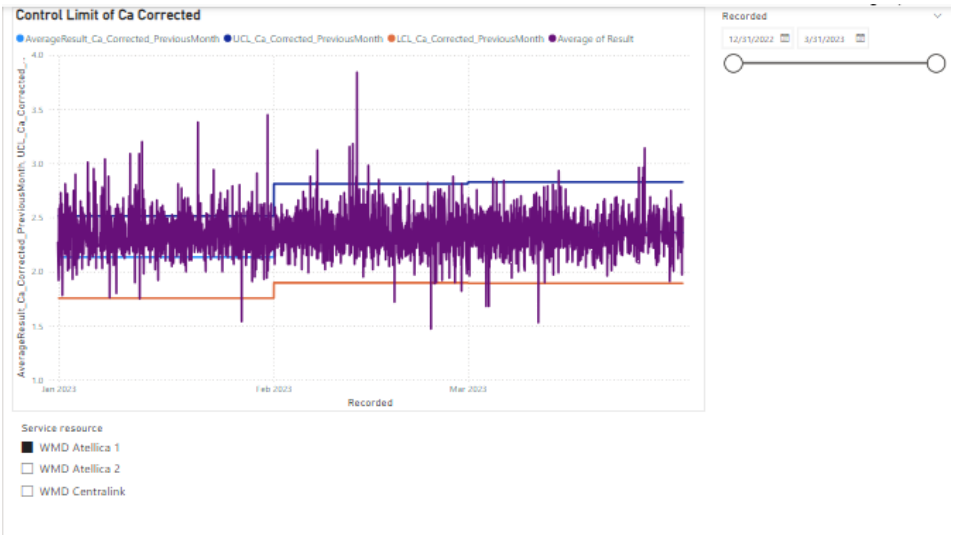

Figure 30 : Control limits of calcium corrected from January 2023 to March 2023

Figure 31 : Control limits of calcium corrected on January 2023

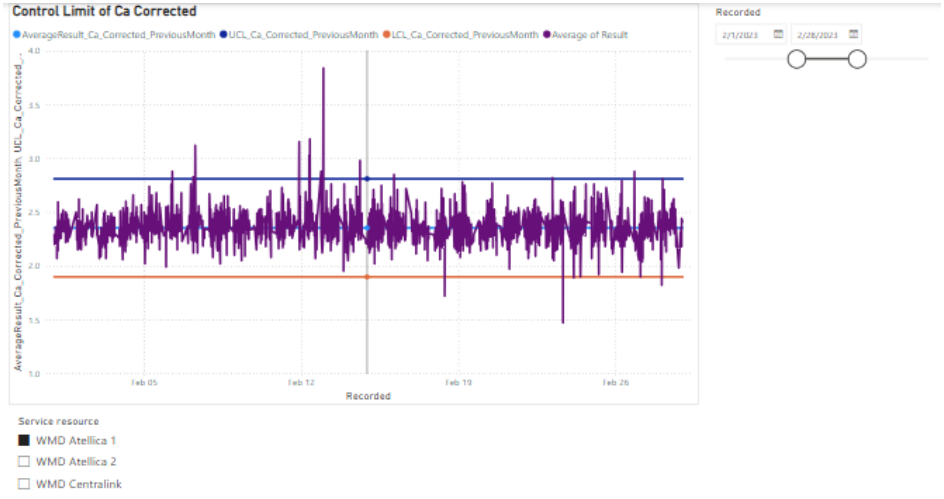

Figure 32 : Control limits of calcium corrected on February 2023

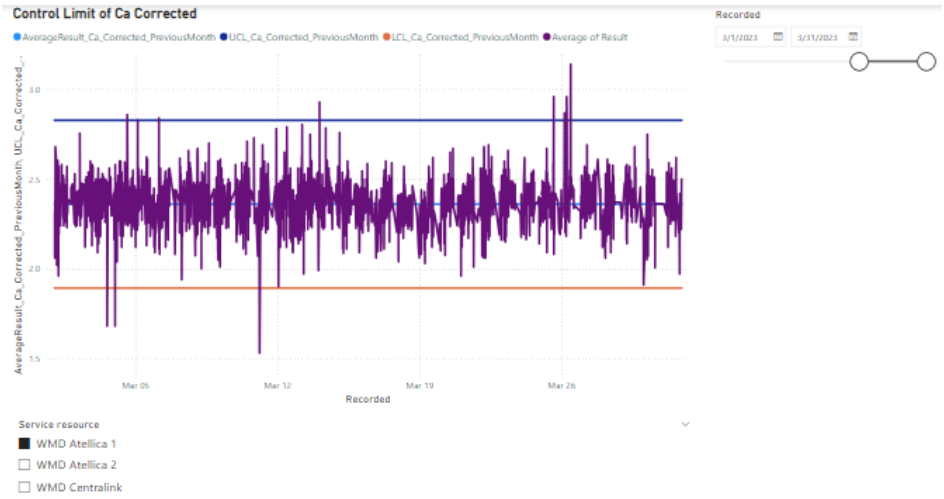

Figure 33 : Control limits of calcium corrected on March 2023

Moving Average of Calcium Corrected

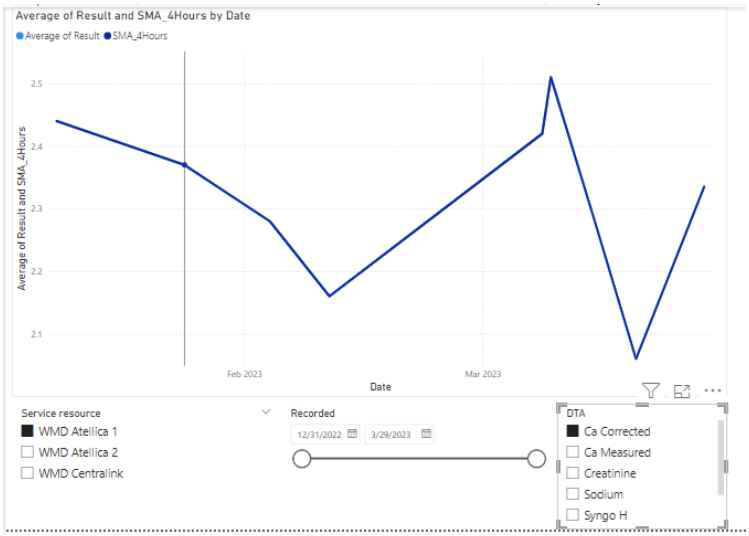

Figure 34 : Moving average of calcium corrected within the time frame of 4 hours

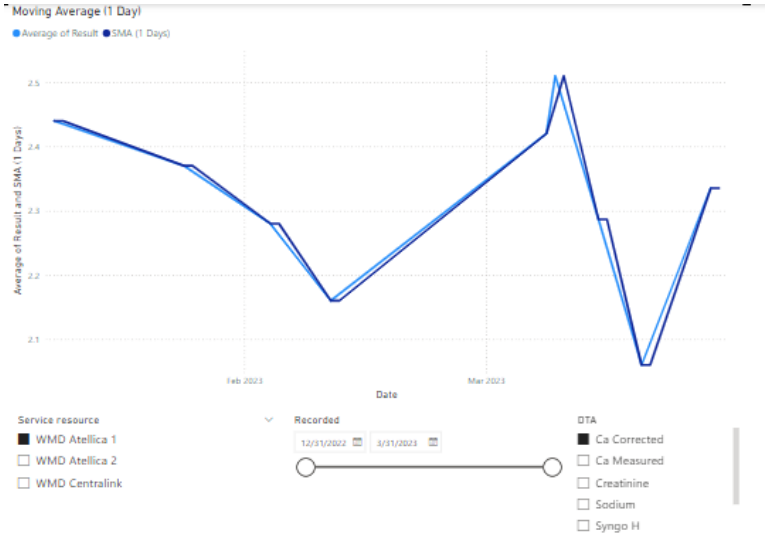

Figure 35 : Moving average of calcium corrected within the time frame of 1 day

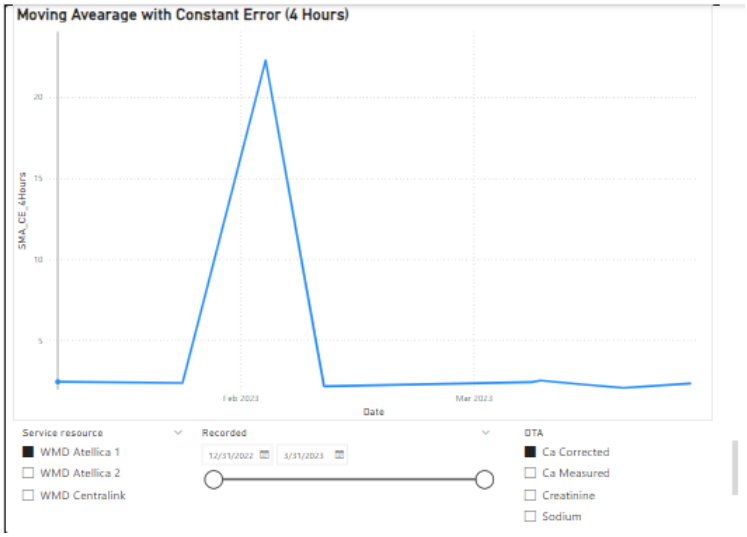

Figure 36 : Moving average of calcium corrected with constant error in the time frame of 4 hours

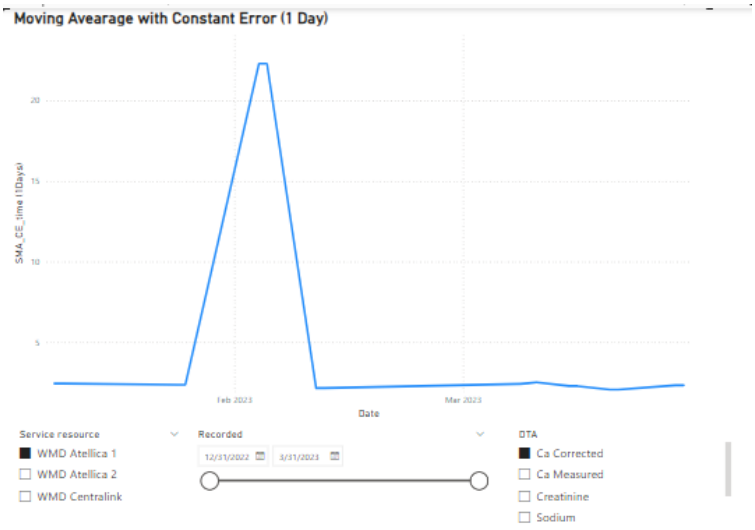

Figure 37 : Moving average of calcium corrected with constant error in the time frame of 1 day

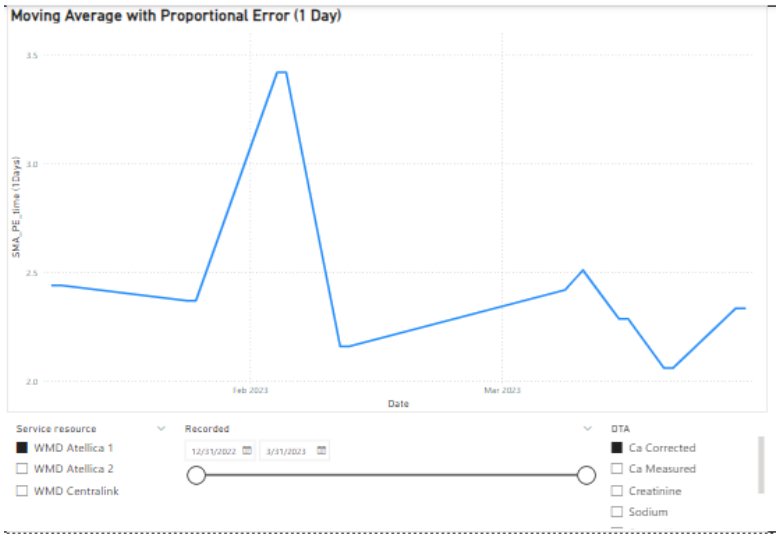

Figure 38 : Moving average of calcium corrected with proportional error in the time frame of 4 hours

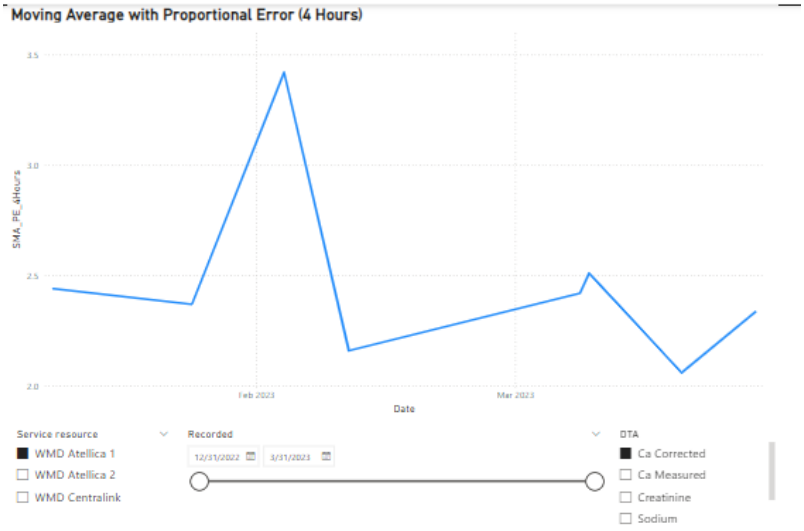

Figure 39 : Moving average of calcium corrected with proportional error in the time frame of 1 day

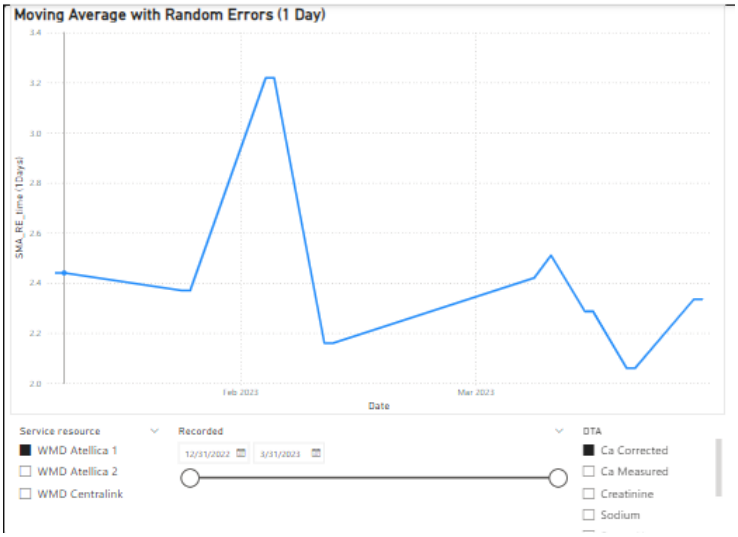

Figure 41 : Moving average of calcium corrected with random error in the time frame of 1

day

The figures of control limits and moving average of the calcium corrected are similar to calcium measured. We can conclude that the insights that we can derive from the Ca Corrected distributions and moving averages are similar to insights that we derived from calcium-measured analytes.

Visual Dashboard of Creatinine Analytes

Figure 42 : Control limits of creatinine on January 2023

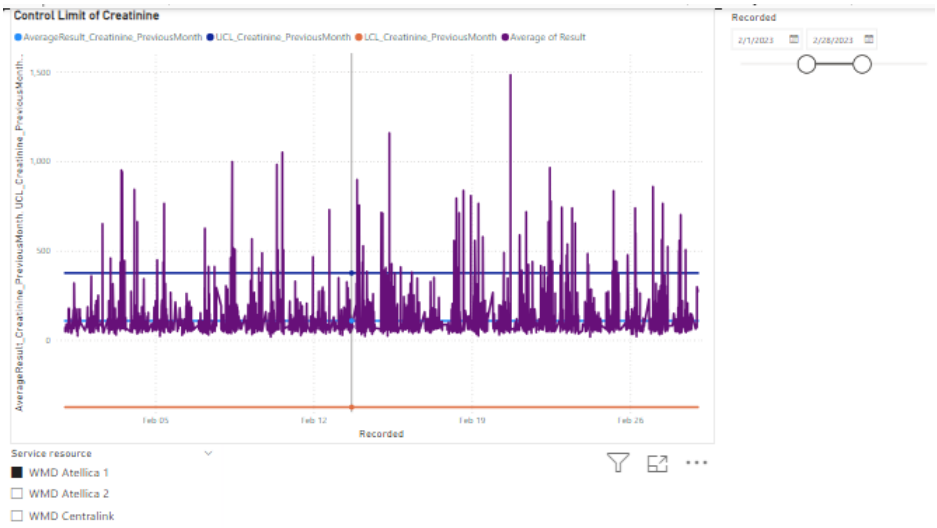

Figure 43 : Control limits of creatinine on February 2023

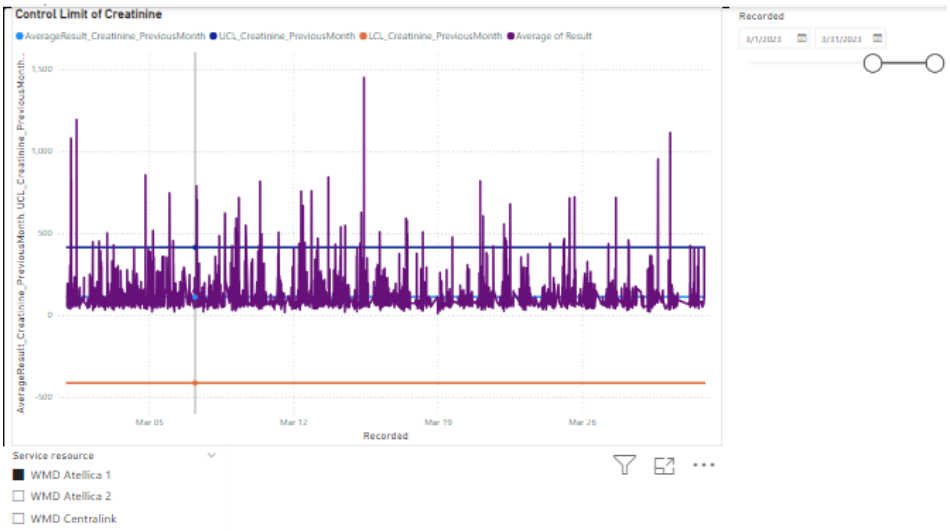

Figure 44 : Control limits of creatinine on March 2023

From the control limits of the creatinine analytes above, we can see that the distribution of the creatinine analytes is skewed. This creates an abnormal distribution in the control limits, which advances the difficulties of deriving insights from the control limits. A skewed distribution of analytes can be hard to use in PBRTQC. Hence, analytes with normal distribution are used for the PBRTQC.

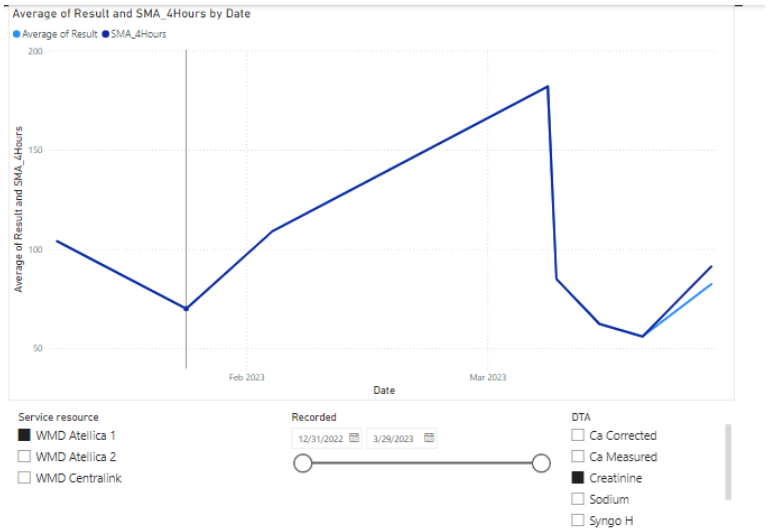

Figure 45 : Moving average of creatinine within the time frame of 4 hours

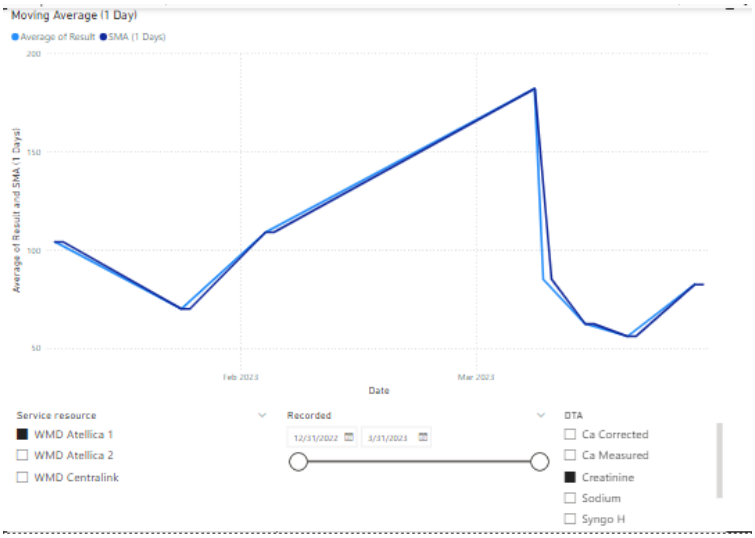

Figure 46 : Moving average of creatinine within the time frame of 1 day

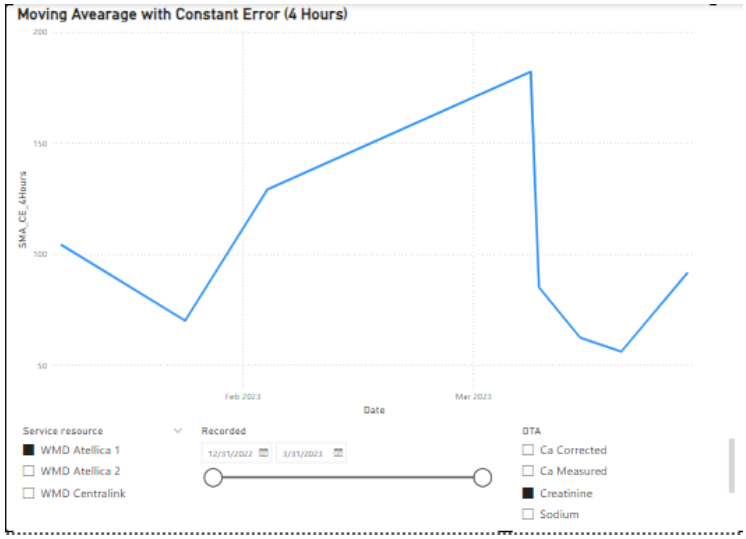

Figure 47 : Moving average of creatinine with constant error in the time frame of 4 hours

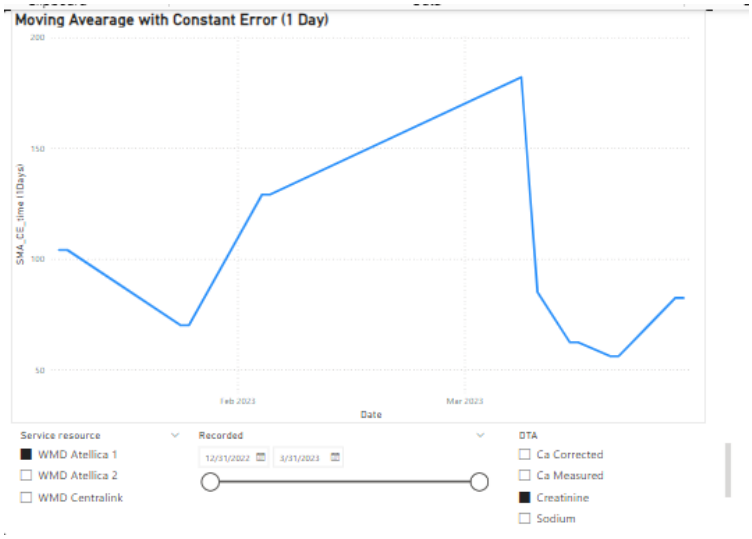

Figure 48 : Moving average of creatinine with constant error in the time frame of 1 day

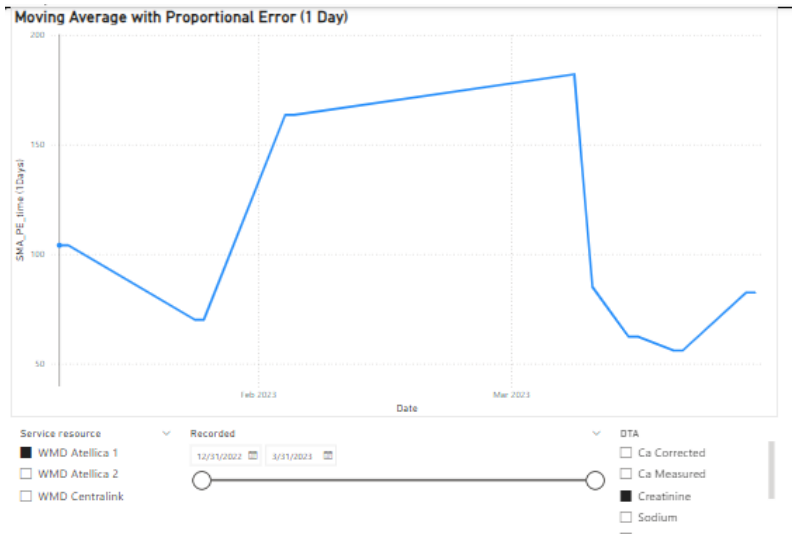

Figure 49 : Moving average of creatinine with proportional error in the time frame of 4 hours

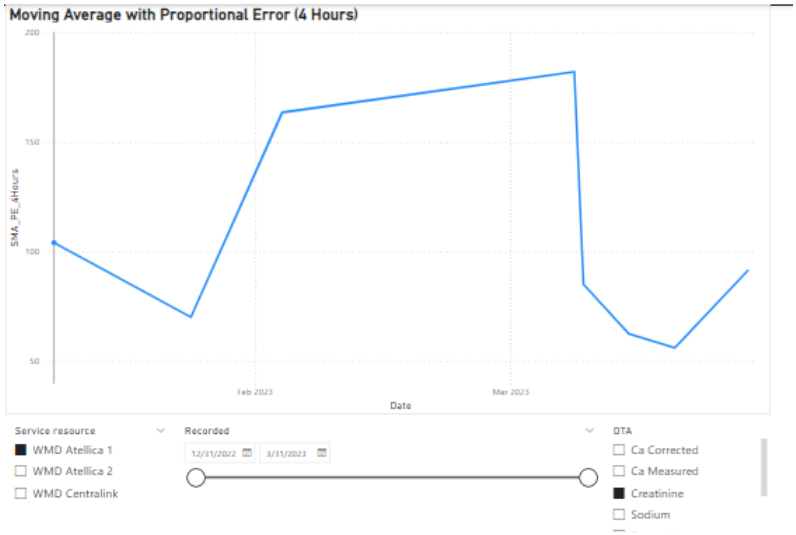

Figure 50 : Moving average of creatinine with proportional error in the time frame of 1 day

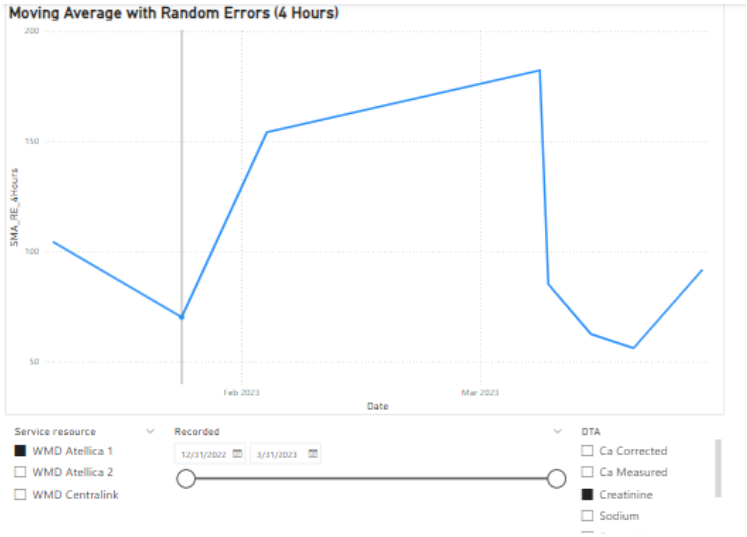

Figure 51 : Moving average of creatinine with random error in the time frame of 4 hours

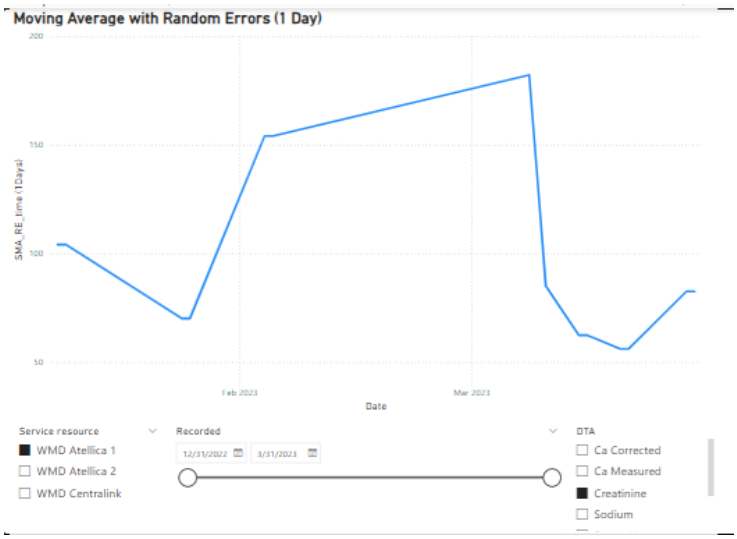

Figure 52 : Moving average of creatinine with random error in the time frame of 1 day

The figures above show the moving average of the creatinine distribution. The moving average shows that there is a downfall in the middle of March. From the control limit, we can see that this is due to the analytes having fewer outliers than before. This should not have affected the moving average if the control limits of the analytes were reduced for the moving averages. From the figures, it seems that the moving average of the normal dataset and the moving average with simulated error do not have a huge difference. However, we can see that the fluctuations on the moving average with simulated errors are more defined than the fluctuations on the moving average with a normal dataset.

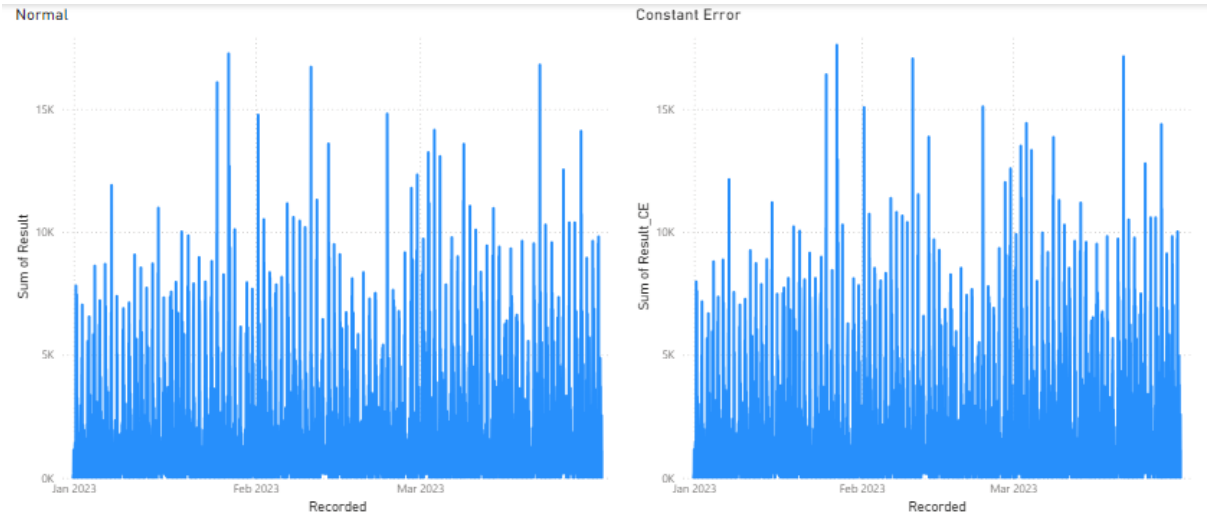

Figure 53: Data distribution of normal dataset and dataset with simulated constant error

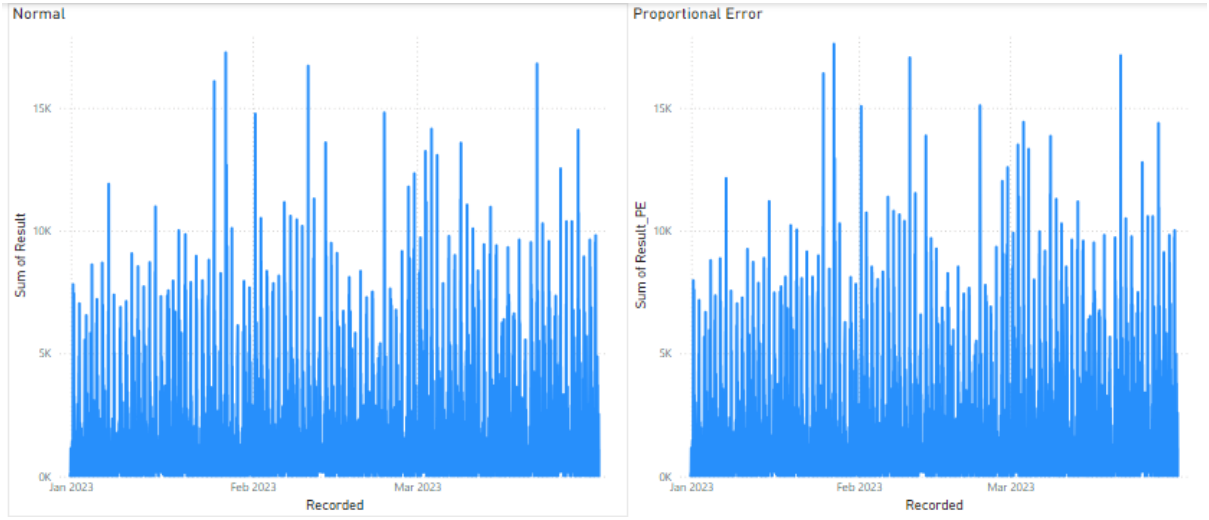

Figure 53: Data distribution of normal dataset and dataset with simulated proportional error

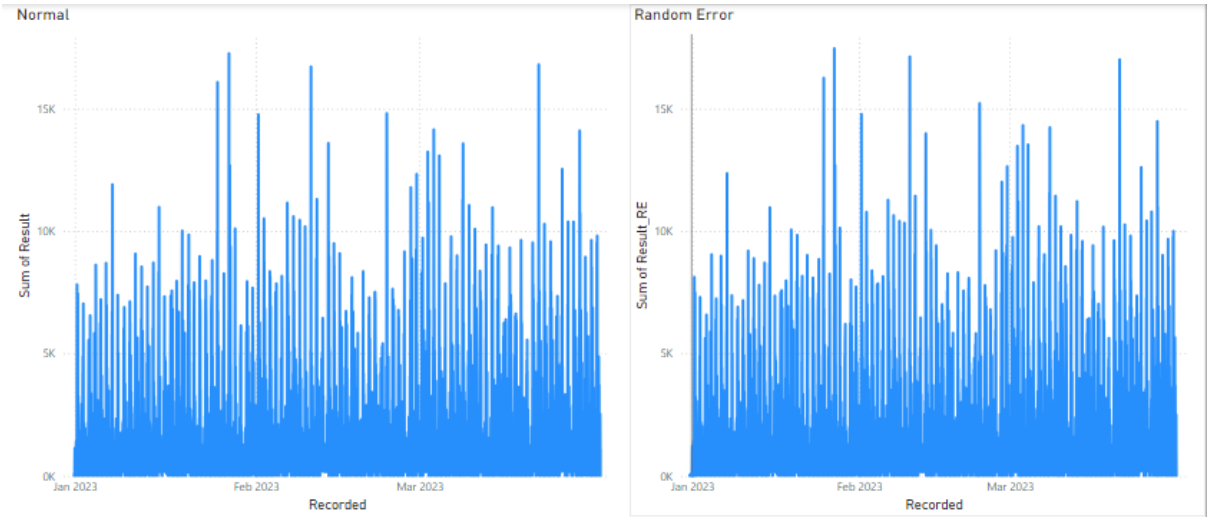

Figure 53: Data distribution of normal dataset and dataset with simulated random error

From the figure above, we can see that the changes of value from the normal dataset with the simulated errors only have the subtle differences, but these subtle differences have shown a huge impact on the moving average of the analytes.

Chapter 5

This chapter will discuss the discoveries that are found throughout the study, as well as outlining limitations and potential for future work.

Discussion

Through the visualisation, we can derive that the data visualisation has provided a better understanding of the data's short-term fluctuations and long-term trends. Although we can see the trends of patient data and outliers in the control limits, it is hard to see the constant changes in the control limit visualisation as there is a huge amount of data. In the real-world scenario, the patient results will be inputted continuously into the as the systems are real-time, making trends harder to see. However, in the Moving Average, we can see trends in the patient test results. The gradual increases in the moving averages can indicate changes in the patient population's health status or an issue with the testing process. Hence, further quality control, such as IQC, can be conducted to confirm the error.

However, in some cases, such as the moving averages of the calcium analytes, there are a lot of fluctuations in the moving averages. After further investigation, we found that the changes in the moving averages are small. Hence, when there are small changes, the moving averages will fluctuate. Compared to the moving averages that are given errors in it, we can see that the errors cause a high fluctuation. The diagram shows fluctuations from 1 February to February 7, where the error is added. Through the visualisation, we can see the differences in fluctuations in normal moving averages and fluctuations caused by machine errors. Therefore, we cannot generalise that fluctuations always mean errors in the machine. There should be a clinical guideline in the future to derive the errors from the moving averages.

It is also derived that PBRTQC is better if it is implemented on analytes with normal distribution than skewed distribution such as creatinine. Hence, better analytical understanding is needed before we implement PBRTQC.

Comparing to moving averages on normal moving averages, we can see that the moving averages on all datasets that are injected by simulated errors on all of the errors give fluctuations in the period of February 1 to February 7, where the error is injected.

Future Works

In the future, more advanced visualisation will be added to understand the data distribution of the analytes better. More statistical methods, such as Moving Median (MM) and Exponential Weighted Moving Average (EWMA), will also be introduced and compared to find a better statistical method. Moreover, as the technology industry grows rapidly, there could also be a chance to integrate Power BI with artificial intelligence (AI) through Python scripts. However, it will be the future work of this data.

On the other hand, alerts can also be given if the moving average or control limits hit a certain value. In this study, the alerts are not configured, as it requires a Power BI

subscription to configure the alerts. However, in the future, alerts can be given as an extensive feature of this visualisation.

Conclusion

This research has demonstrated the potential of using data visualisation to detect errors in patient-based real-time quality control (PBRTQC) and to understand analytes in the dataset better. The developed dashboard of control limits and moving averages has successfully detected errors in the patient test results, providing a valuable tool for clinical laboratories and improving decision-making, giving better outcomes for patients. The data configurations and visualisation that enable real-time monitoring could give real-time insights to laboratory workers, which could detect errors in the dataset faster. Although visualisation is not enough to detect errors in the patient's test results, it can lower internal quality control (IQC) use only when errors are detected in PBRTQC. This could lower the cost of quality control in clinical laboratories.

Future research could focus more on incorporating advanced visualisation techniques and integrating machine learning algorithms to enhance the PBRTQC systems further.

In conclusion, the integration of data visualisation in visualising PBRTQC gives a significant advancement in the clinical laboratories practice, offering an effective solution for real-time quality control and error detection with low cost by using just data visualisation in common data

References

Badrick, T., Bietenbeck, A., Cervinski, M. A., Katayev, A., van Rossum, H. H., & Loh, T. P. (2019). Patient-Based Real-Time Quality Control: Review and Recommendations. *Clinical Chemistry*, *65*(8), 962–971. https://doi.org/10.1373/clinchem.2019.305482

Badrick, T., Bietenbeck, A., Katayev, A., van Rossum, H. H., Cervinski, M. A., & Ping Loh, T. (2020). Patient-Based Real Time QC. *Clinical Chemistry*, *66*(9), 1140–1145. https://doi.org/10.1093/clinchem/hvaa149

Badrick, T., Bietenbeck, A., Katayev, A., van Rossum, H. H., Loh, T. P., Cervinski, M. A., & on behalf of the International Federation of Clinical Chemistry and Laboratory Medicine Committee on Analytical Quality. (2020). Implementation of patient-based real-time quality control. *Critical Reviews in Clinical Laboratory Sciences*, *57*(8), 532–547. https://doi.org/10.1080/10408363.2020.1765731

Cervinski, M. A. (2021). Pushing Patient-Based Quality Control Forward through Regression. *Clinical Chemistry*, *67*(10), 1299–1300. https://doi.org/10.1093/clinchem/hvab155

Cui, W., & Wang, H. (2017). Anomaly detection and visualization of school electricity consumption data. *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*, 606–611. https://doi.org/10.1109/ICBDA.2017.8078707

Fleming, J. K., & Katayev, A. (2015). Changing the paradigm of laboratory quality control through implementation of real-time test results monitoring: For patients by patients. *Clinical Biochemistry*, *48*(7–8), 508–513. https://doi.org/10.1016/j.clinbiochem.2014.12.016

Mulero-Pérez, D., Benavent-Lledó, M., Azorín-López, J., Marcos-Jorquera, D., & García-Rodríguez, J. (2023). Anomaly detection and virtual reality visualisation in

supercomputers. *The International Journal of Advanced Manufacturing Technology*. https://doi.org/10.1007/s00170-023-11255-x

Ng, D., Polito, F. A., & Cervinski, M. A. (2016). Optimization of a Moving Averages Program Using a Simulated Annealing Algorithm: The Goal is to Monitor the Process Not the Patients. *Clinical Chemistry*, *62*(10), 1361–1371. https://doi.org/10.1373/clinchem.2016.257055

Thaler, M. A., Iakoubov, R., Bietenbeck, A., & Luppa, P. B. (2015). Clinically relevant lot-to-lot reagent difference in a commercial immunoturbidimetric assay for glycated hemoglobin A1c. *Clinical Biochemistry*, *48*(16), 1167–1170. https://doi.org/10.1016/j.clinbiochem.2015.07.018

Vidyapeetham, A. V. (2021). *Anomaly Detection using User Entity Behavior Analytics and Data Visualization*.

Appendix

The following attachment is the important souce code snipplets for the experiment documented in this document.

```sql
-- Deduplication - Remove duplicated rows
SELECT DISTINCT *
INTO deduped_lab_results
FROM qcdata;

-- Handling Missing Values - Discard instances where age was recorded as 123
SELECT *
INTO clean_lab_results
FROM deduped_lab_results
WHERE age != 123;

-- Filtering Valid Results - Retain only rows where the result is numeric and convert to FLOAT
SELECT *,
       TRY_CAST(result AS FLOAT) AS Result_FLOAT
INTO valid_lab_results_temp
FROM clean_lab_results
WHERE ISNUMERIC(result) = 1
AND TRY_CAST(result AS FLOAT) IS NOT NULL;
```

Figure 54: SQL query onhandling duplicate data, missing values, and converting data type

```sql
-- Create separate tables for each Service resource
DECLARE @service_resource NVARCHAR(255);
DECLARE service_resource_cursor CURSOR FOR
SELECT DISTINCT [Service resource] FROM valid_lab_results;

OPEN service_resource_cursor;
FETCH NEXT FROM service_resource_cursor INTO @service_resource;

WHILE @@FETCH_STATUS = 0
BEGIN
    DECLARE @query NVARCHAR(MAX);
    SET @query = 'SELECT * INTO service_resource_' + REPLACE(@service_resource, ' ', '_') +
    ' FROM valid_lab_results WHERE [Service resource] = ''' + @service_resource + '''';
    EXEC sp_executesql @query;

    FETCH NEXT FROM service_resource_cursor INTO @service_resource;
END

CLOSE service_resource_cursor;
DEALLOCATE service_resource_cursor;
```

```sql
-- Create separate tables for each DTA
DECLARE @dta NVARCHAR(255);
DECLARE dta_cursor CURSOR FOR
SELECT DISTINCT DTA FROM valid_lab_results;

OPEN dta_cursor;
FETCH NEXT FROM dta_cursor INTO @dta;

WHILE @@FETCH_STATUS = 0
BEGIN
    DECLARE @query_dta NVARCHAR(MAX);
    SET @query_dta = 'SELECT * INTO DTA_' + REPLACE(@dta, ' ', '_') + ' FROM valid_lab_results WHERE DTA = ''' + @dta + '''';
    EXEC sp_executesql @query_dta;

    FETCH NEXT FROM dta_cursor INTO @dta;
END

CLOSE dta_cursor;
DEALLOCATE dta_cursor;
```

Figure 55: SQL query on separating dataset into service resource and analytes

```sql
-- Apply Constant Error (CE) selectively to specific periods
SELECT *,
       CASE
           WHEN Recorded BETWEEN '2023-01-01' AND '2023-01-07' OR
                Recorded BETWEEN '2023-02-01' AND '2023-02-07'
           THEN Result_FLOAT + 5
           ELSE Result_FLOAT
       END AS Result_CE
INTO biased_lab_results_constant
FROM valid_lab_results;
```

Figure 56: SQL query on applying simulated constant error

```sql
-- Apply Proportional Error (PE) selectively to specific periods
SELECT *,
       CASE
           WHEN Recorded BETWEEN '2023-01-01' AND '2023-01-07' OR
                Recorded BETWEEN '2023-02-01' AND '2023-02-07'
           THEN Result_FLOAT * 1.02
           ELSE Result_FLOAT
       END AS Result_PE
INTO biased_lab_results_proportional
FROM valid_lab_results;
```

Figure 57: SQL query on applying simulated proportional error

```sql
-- Apply Random Error (RE) selectively to specific periods
SELECT *,
       CASE
           WHEN Recorded BETWEEN '2023-01-01' AND '2023-01-07' OR
                Recorded BETWEEN '2023-02-01' AND '2023-02-07'
           THEN Result_FLOAT + (0.1 * Result_FLOAT * RAND())
           ELSE Result_FLOAT
       END AS Result_RE
INTO biased_lab_results_random
FROM valid_lab_results;
```

Figure 58: SQL query on applying simulated random error

Ethical Declaration

Date of Decision Notification: **18 Jul 2023**

Dear Dr Yusof Rahman,
Thank you for submitting the following Human Research Ethics Application (HREA) for HREC review;

2023/ETH01071: Patient-Based Real-Time Quality Control (PBRTQC) Using Multiple Simulation Modelling and Machine Learning Algorithms in an Acute Tertiary Laboratory Setting

This Application was reviewed as a **Low or negligible risk review pathway** and was initially considered by the **Western Sydney Local Health District Human Research Ethics Committee**
.
The project was determined to meet the requirements of the National Statement on Ethical Conduct in Human Research (2007) and was **APPROVED**.

This email constitutes ethical and scientific approval only.
This project cannot proceed at any site until separate research governance authorisation has been obtained from the Institution at which the research will take place.

This project has been Approved to be conducted at the following sites:

- **Westmead Hospital**
- **University Technology Sydney**

The following documentation was reviewed and is included in this approval:

- Protocol Version 1.3 dated 23 June 2023
- Data Collection Sheet Version 1.2 Updated 23 June 2023

Application Documents - (link will only be active for 14 days from the decision date. The approved documents are also available to download from forms section of this project in REGIS)

The Human Research Ethics Application reviewed by the HREC was:
Version: 1.03
Date: 12 Jul 2023

The approval is for a period of 5 years from the date of this e-mail **(18 Jul 2023)**

The Committee granted a waiver of the usual requirement of consent for the use of re-identifiable information held by NSW agencies, in line with the State Privacy Commissioner's Guidelines for Research and the Health Records and Information Privacy Act 2002 (NSW) and the Guidelines approved under Section 95/95A of the Privacy Act 1988.

The Coordinating Principal Investigator will:
• provide the HREC with an annual report and the final report when the project is completed at all sites. This will be through the submission of a milestone in REGIS.
• immediately report anything that might warrant review of ethical approval of the project.
• submit proposed amendments to the research protocol, including; the general conduct of the research, changes to CPI or site PI, an extension to HREC approval, or the addition of sites to the HREC before those changes can take effect. This will be through a notification of an amendment in REGIS
• will notify the HREC if the project is discontinued at a participating site before the expected completion date, with reasons provided.

Submission of annual progress/final reports (milestone), amendments and safety reports should be done through the forms provided in REGIS. Guidance on these processes can be found on the REGIS website.

It is noted that the **Western Sydney Local Health District Human Research Ethics Committee** is constituted in accordance with the National Statement on Ethical Conduct in Human Research, 2007 (NHMRC).

The processes used by the HREC to review multi-centre research proposals have been certified by the National Health and Medical Research Council.

Please contact us if you would like to discuss any aspects of this process further, as per the contact details below. We look forward to managing this study with you throughout the project lifecycle.

Yours Sincerely,
Research Office
WSLHD Research and Education Network
Westmead Hospital Cnr Hawkesbury & Darcy Rds Westmead NSW 2145
Tel 02 8890 9007

Cover Sheet

UTS: ENGINEERING & INFORMATION TECHNOLOGY

SUBJECT NUMBER & NAME	NAME OF STUDENT(s) (PRINT CLEARL SURNAME / FIRST NAME	STUDENT ID(s)
31482 Honours Project	Gabrielle Thesya Evania	14013371

STUDENT EMAIL	STUDENT CONTACT NUMBER
Thesya.e.gabrielle@student.uts.edu.au	0434156699

NAME OF TUTOR	TUTORIAL GROUP	DUE DATE

ASSESSMENT ITEM NUMBER & TITLE

☐ I confirm that I have read, understood and followed the guidelines for assignment submission and presentation on page 2 of this cover sheet.
☐ I confirm that I have read, understood and followed the advice in the Subject Outline about assessment requirements.
☐ I understand that if this assignment is submitted after the due date it may incur a penalty for lateness unless I have previously had an extension of time approved and have attached the written confirmation of this extension.

Declaration of originality: The work contained in this assignment, other than that specifically attributed to another source, is that of the author(s) and has not been previously submitted for assessment. I understand that, should this declaration be found to be false, disciplinary action could be taken and penalties imposed in accordance with University policy and rules. In the statement below, I have indicated the extent to which I have collaborated with others, whom I have named.

Statement of collaboration:

Signature of student(s) _____ **Date** _____

✂----------

ASSIGNMENT RECEIPT

To be completed by the student if a receipt is required

SUBJECT NUMBER & NAME	NAME OF TUTOR

SIGNATURE OF TUTOR	RECEIVED DATE