Leveraging Machine Learning and Process Mining to Treat Anaemia with the Help of Prescription Records

ARDA SATICI, University of Twente, The Netherlands

ABSTRACT

According to WHO statistics, the global anaemia prevalence is around 30%, making anaemia one of the most encountered diseases [2]. Hence, I wanted to focus on the effect of machine learning and process mining on the treatment studies of this common disease in my research paper. This research paper has the aim of working on the extension of a project that Mike Pingel has done before and to compare my findings with the findings of his project and share the results with the reader. The limitation of this research paper is that the methodology of this research paper should be kept the same as the project that I accepted as the foundation. That is, I can use the techniques used in that project in exactly the same way and should not make additional variants myself. In a branch such as machine learning where new developments are experienced every day, it would be a future improvement for my research if this project is repeated in the future with new machine learning algorithms and more recent MIMIC datasets and my deficiencies in this research paper are determined accordingly.

Additional Key Words and Phrases:
Anaemia, Process Mining, Machine Learning, Regression, Random Forest, Prescription, Treatment

1 INTRODUCTION

According to the definition of the World Health Organization, anaemia is a disease in which the number of red blood cells or the concentration of hemoglobin in red blood cells is lower than it should be [1]. And as a result of anaemia, some symptoms such as fatigue, weakness, dizziness and shortness of breath appear in the patient. Anaemia is one of the most common diseases of today and this disease affects young children, menstruating adolescent girls and women, and pregnant and postpartum women, and the incidence of this disease is 40% in children 6-59 months of age [1].

According to Interactive Process Mining in Healthcare[12] and IBM website [7], Process Mining is the analysis of event-logs in such a way that events occurring in a particular procedure in real life can be understood by humans. It is the integral application of data science and it helps humans to discover, validate and improve workflows by analysing operational processes. As a result, the processes, bottlenecks, and areas which are open to improvements can be discovered by the organizations.

While improving system performance, help can be obtained from computational methods. In this way, experience is learned and the system decides according to these experiences. This is called machine learning. In other words, machine learning develops new models from existing data with learning algorithms, and the purpose of this is to improve performance. [21]

Machine learning and process mining can be used during the treatment studies of anaemia. Finding the best drugs for anaemia treatment, and classification according to the different anaemia types are important for the health of patients, and details about this will be shared in the following sections.

TScIT 39, July 7, 2023, Enschede, The Netherlands

2 PROBLEM STATEMENT

A lot of research has been done so far on the subject of 'anaemia treatment', which is the subject this research is focused on. The research that is considered fundamental in this research is Mike Sven Pingel's research. However, Machine Learning is a constantly evolving branch and it is very possible to get different results with different datasets. Thus, the main purpose in this research paper is to extend Mike's research a little further and see how valid his research is. Hence, a different dataset will be used in this analysis. In this way, the best treatment methods for the determined anaemia species will be determined. For this research paper, *CITI's Data or Specimens Only Research* training has been completed and MIMIC-III v1.4 Clinical Database has been used.

2.1 Research Question

For the solution of the above-mentioned problem statement, the research question was determined as follows, and this research question is also the main title of this research:
'To what extent it is effective to use machine learning and process mining to predict the best anaemia treatment with the help of prescription records?'

2.2 Sub-Research Questions

This research question is divided into 3 sub-questions for more accurate and detailed research and those questions are as follows:

(1) To what extent can treatment predictions be made for different types of anaemia with machine learning algorithms?
(2) With the help of process mining logs, how much information about the treatment of anaemia can be obtained?
(3) How much do the results from the machine learning and process mining logs match each other? If it doesn't match, what could be the reason?

3 RELATED WORK

3.1 Main Sources Used

Before I started doing this research, my knowledge of anaemia and the anaemia studies of machine learning was very limited. Therefore, in this section, research on anaemia, process mining, and the impact of machine learning on anaemia will be shared with the reader.

The main source used in this research is the research of Mike Pingel [16], from which this research was inspired. In particular, in the introduction, related works and methodology sections, Mike's thesis helped a lot in writing this research and the methods and algorithms he used were also used in this research.

Besides Mike's research, some online databases were made use while researching. Scopus, the largest abstract and citation database of peer-reviewed literature – scientific journals, books and conference proceedings [5], and FindUT (University of Twente's online database) [6] were used to gather information for research.

Besides Mike's research, some help was received from Carlos Fernandez-Llatas' publication [12] to learn about how process mining is applied in the healthcare industry. Thanks to this publication, more information was obtained about process mining and inspiration was obtained for the methodology of this research.

Apart from that, for a better understanding of the process mining graphs, help is taken from van der Aalst's article [19]. In this way, a better analysis will be possible after the process mining operations are completed. The publication [11] named 'A Classification of Process Mining Bottleneck Analysis Techniques for Operational Support' by Rob Bemthuis, Niels van Slooten, Jeewanie Jayasinghe Arachchige, Jean Paul Sebastian Piest and Faiza Allah Bukhsh was inspirational when writing about process iteration, getting assistance in the research methodology and literature review procedure.

A lot of resources have also been used about the techniques used during the machine learning trainings (LASSO, Random Forest, k-Fold) and these resources are given in later sections.

3.2 Sources Available Online

As mentioned before, Mike Pingel's research [16], which this research is fundamentally based on, also benefited from Scopus and UTFind databases and determined the keywords as **"Process Mining" AND "Machine Learning"** during the research, and a total of 253 papers were found.

The number of papers with the same keywords ("Process Mining" AND "Machine Learning") that were shared in these 2 databases since August 2021 was also checked. For the findings, only peer-reviewed papers were taken into account and other papers were neglected, as the research that is considered fundamental did. Also, only papers in English were considered.

In Scopus, initially there were 305 results before applying any filtering. Then, from the Document Type section, Conference Paper, and Article are chosen. This filtering reduced the total count to 249. Then, all the papers which are not in English are removed. But this filtering contains papers from 2003 to 2023. Since the focus is on the papers published after the fundamental research has been done (August 2021), all the papers between 2003 and 2020 are removed, and only the papers which were published in 2021, 2022 and 2023 are kept. After removing peer-reviewed papers (journals) and removing duplicates, the final remaining paper count is 53 papers.

In FindUT, initially there were 195 papers found in University of Twente. Then, since the focus is only on articles and papers, and not books, the books are removed and the count becomes 141 papers. Then after removing non-English publications, it reduces to 120. Then, the time filtering is set as year>2020 and year<2024, and the count decreases to 64. Then, the feature of 'remove duplicates' and the filter called 'limit to peer-reviewed' are applied, and at the end, the final publication count becomes 7 papers.

	Scopus	UTFind	Total
Initially	305	195	500
After removing books	249	141	431
After removing non-English ones	243	120	404
After applying the year-filter	118	64	197
After removing duplicates	53	54	107
After getting peer-reviewed papers	53	7	60

Table 1. Overview of task division

For the analysis of the articles found in FindUT, the reader can check the **respective section** in Appendix. Since analyzing 53 articles (articles from Scopus) for this research is out of the scope, instead of choosing most cited articles which are not related to medicine, only the articles which are related to medicine branch are analyzed and the reader can find the analysis of Scopus articles in **Appendix**. Hence, some information about how machine learning and process mining are applied to medicine related research can be observed, which is the main focus of this research.

After finding the articles in FindUT and Scopus, first, the subject of these articles in general was searched. Therefore, as can be found below, the keywords in these articles were determined and the distribution of these keywords in the pie chart was observed. The pie chart below shows the distribution of 13 different keywords in the filtered articles in the 2 base research databases.

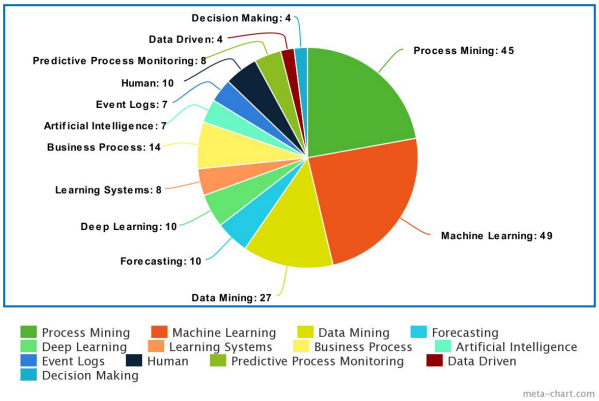


Fig. 1. Keywords on a pie chart

Since the dataset in Scopus is larger and the Scopus website itself offers analysis, the analytical figures in Scopus ('document by subject area' and 'document by year') have been added. Only filtering for these figures is setting the year-filter as 2021-2023, and 150 documents are returned. According to these figures, Process Mining and Machine Learning are mostly discussed in the computer science branch and only 3 percent of the papers focus on the medicine branch (11 documents). The topics covered by these 11 publications are COVID-19, heart failure, pediatric oncology and diabetes. This situation shows the number of sources related to anaemia is relatively low compared to other diseases.

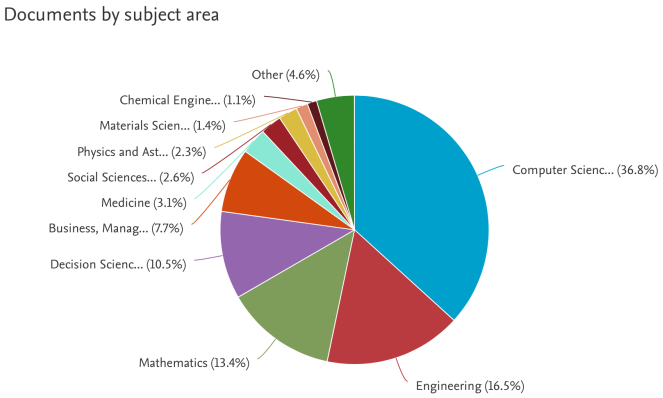


Fig. 2. Documents by Subject Area - Scopus

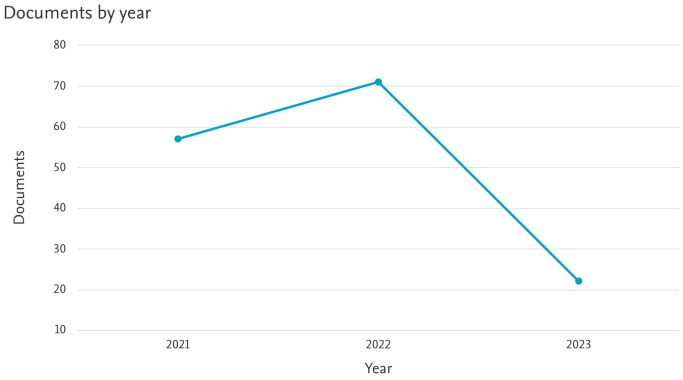


Fig. 3. Documents by Year - Scopus

Apart from the research papers in Scopus and FindUT, a training has been completed to work on this project more effectively. I took the 'Introduction to Process Mining with ProM' course from the University of Eindhoven and used it to search the MIMIC-III v1.4 Clinical Database, which can also be found on the PhysioNet website [14].

4 METHODS OF RESEARCH

This section explains how to answer the 3 research questions that are mentioned in **Section 2 - Problem Statement**. The first part of this section (Section 4.1) focuses on the Cross-industry standard process for data mining (CRISP-DM) process. CRISP-DM is an approach to process mining projects, and the research question (and sub-research questions) is answered with this approach. Then, the next subsection (Section 4.2) of this section focuses on the MIMIC-III v1.4 Clinical Database which is obtained from PhysioNet. Additionally, in this section, how DBeaver is used and how SQL Query is applied on the dataset are discussed. At the end, in Section 4.3, the machine learning training process on the dataset for the anaemia treatment prediction, which serves as the main purpose of this research paper, is explained to the reader.

4.1 CRISP-DM Process Model

CRISP-DM, the abbreviation of Cross-industry standard process for data mining, is considered as a proven guide for the data mining projects [9]. The CRISP-DM life cycle consists of 6 different phases:

(1) Business Understanding
(2) Data Understanding
(3) Data Preparation
(4) Modelling
(5) Evaluation
(6) Deployment

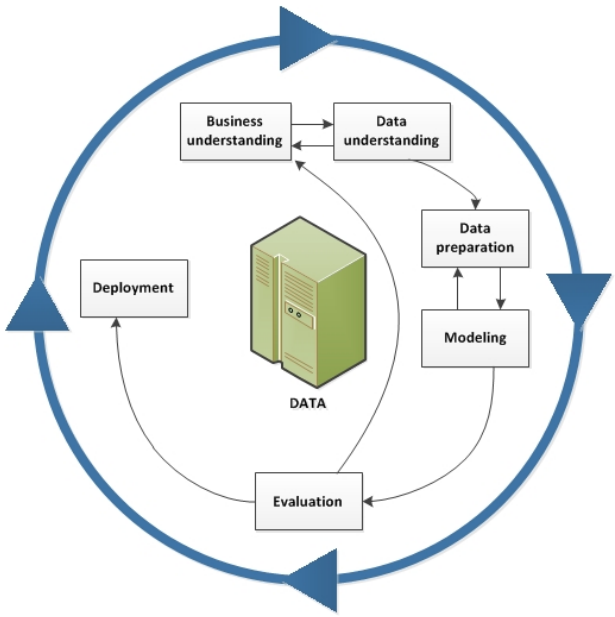


Fig. 4. CRISP-DM Lifecycle [9]

Cycle (RQ)	Cycle 1	Cycle 2	Cycle 3
Business Understanding	**Section 3**	**Section 3**	**Section 3**
Data Understanding	**Section 5**	**Section 5**	**Section 6**
Data Preparation	**Section 5**	**Section 5**	**Section 6**
Modelling	**Section 6**	**Section 6**	**Section 6**
Evaluation	**Section 6**	**Section 6**	**Section 8**
Deployment	**NA**	**NA**	**NA**

Table 2. CRISP-DM Lifecycle with Research Questions

Each cycle seen in Figure 3 was applied one by one for the 3 research questions mentioned in Section 2.2. Since this research does not include deployment, the focus is on the first 5 steps. That means, the 5-step process that compose the CRISP-DM partly was applied 3 times.

Business understanding of all sub questions was completed in Section 3, during related works was being researched. While using

Machine Learning and Process Mining, data understanding and preparation were completed with the modifications made in MIMIC-III Dataset, and these are explained in Section 5. For the third sub question, which aims to compare these two branches, the table in the results section in Section 6 helped for data understanding and preparation. Modeling was done in Section 6 for each subquestion and evaluation was made in the 6th and 8th sections according to these modelings.

4.2 MIMIC-III Clinical Database

For this research paper, the clinical database used for this research paper is MIMIC-III v1.4 Clinical Database [14] .The generators of this database are defining it as following: "MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012." [14]. This database consists of 26 tables (all of them are csv-files), and all of those tables are connected to each other with FOREIGN KEYS (more specifically, IDs of the tables). Each table has an ID and these IDs are unique. Hence, they are PRIMARY KEYS. During this project, some help was obtained from SQL while searching and filtering on this dataset. However, since the 'compressed' size of the MIMIC-III dataset is 16 GB, this dataset is added to a database management tool for more practical processing. DBeaver is chosen because it is easy to learn, and ran SQL queries in that application. Attributes of dataset tables used during this project can be seen from the photo attached to **Appendix B**.

When the query that returns the number of all patients with the disease description containing 'anaemia' is executed, it is seen that there are 13922 registered patients and 6713 of them are female (48.2%) and 7209 are male (51.7%).

	Count	%
Female	6713	48.2 %
Male	7209	51.7 %
Total	13922	100 %

Table 3. Gender Analysis

4.3 Machine Learning Training for the Drug Prediction

During machine learning training, instead of applying a single selection technique, many different techniques were used to detect the most suitable drugs for anaemia, and all of these techniques were looked at as a whole. I will use the techniques that are used Mike's research, and those techniques are Random Forest, LASSO regression, and Pearson's correlation coefficient. On top of those, Linear Regression will be also considered during machine learning training.

Each technique has some advantages over the others. For example, Random Forest generates a more accurate, more understandable result than the other algorithms. [10] Lasso-regression has the advantage of having large number of covariates in the model, and it can regulate unnecessary and uninfluential covariates in the model by setting their coefficients to 0. [15] Pearson's correlation coefficient is a widely used method of measuring the relationship between two variables, and it generates a result between -1 and 1. The farther the value is from 0, the stronger the correlation exists between these two variables. [18]

Since the variance of the machine learning model is high, in order to deal with this, k-fold cross-validation is used during the machine learning training. K-fold cross-validation evaluates the inputted model by dividing the model into k pieces, and treating each piece as validation data set [20].

5 STEPS

After gaining access to the MIMIC-III dataset, some filtering had to be done in the dataset to obtain the data that should be used for machine learning. For this, help was taken from SQL. Although I have access to 26 tables, only 5 of these tables have been used.

- The prediction has been started by finding the different anaemia species in the dataset and their ICD9 codes. The result can be found **here**. After that, the drugs prescribed for each type of anaemia are returned, and the SQL query is added in the **Appendix**. But the result did not produce useful output because it was varying a lot and complex.
- So, the opposite of the previous step is done and an SQL code that showed how many different types of anaemia each drug type was used to treat is written, and as an example this code shows that Potassium Chloride was used in the treatment of 32 different types of anaemia.
- After that, the PRESCRIPTIONS.csv file is imported into the jupyter notebook, this file contains the time when a drug was started to be delivered and the delivery was stopped. So, thanks to this csv file, how long it takes for each drug to heal the patient can be seen. A new column is added to this csv file, the difference between enddate and startdate is calculated, and this difference is inserted into the new CSV file.
- As a result, for each prescription, it was calculated that which drug was prescribed and how long it was used. This process has a limitation, the available records only store the day, month, year and the time is always 00:00:00. This situation can sometimes reduce the accuracy of the analysis. The step I did with this file was to calculate the average duration of the 30 drugs used in the treatment of anaemia in the previous steps, and the output is added in the **Appendix**.
- And at the end, the Python code is modified in a way that it can predict the predicted treatment time for each drug-anaemia type groups. The final result prints the best drugs and their treatment times for each anaemia type. This result can be found **in the next section**.
- After the machine learning prediction was finished, the modified dataset was imported into Disco, a process mining application, and ran the application by selecting **Case ID** as ROW_ID, SUBJECT_ID, HADM_ID, **Activity** as DRUG and LONG_TITLE (anaemia type) and **Timestamp** as dates.

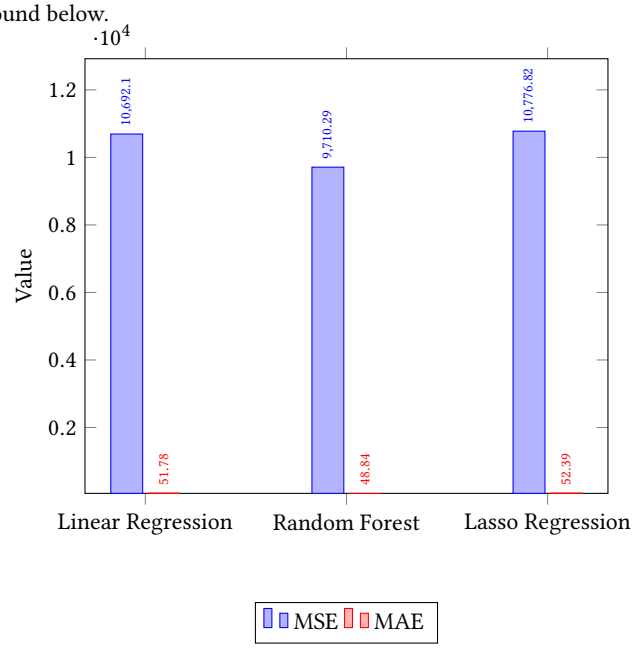
6 RESULTS

6.1 Machine Learning Results

The aim of using machine learning in this project is to find the best way to treat anaemia with the help of machine learning models according to the 12150 patients (filtered) in the MIMIC-III Dataset and the treatment records of these patients and the drugs they use.

The size of the dataset which is used is (11805336 rows × 22 columns) and this dataset is filtered with 29 drugs. In other words, the focus for the prediction is for the most used drugs.

3 different training algorithms were used to find the best machine learning regression training round, they are Linear Regression, Random Forest and LASSO Regression. It is concluded that the pearson correlation coefficient is not the best method for the purpose because it only measures the linear relationship between pairs of variables (X,Y) [13] and this is not exactly suitable. Also, to increase the accuracy of the training, 5-fold cross validation was applied before starting the regression trainings. The results of the trainings can be found below.



Having high values of Mean Squared Error (MSE) and Mean Absolute Error (MAE) shows that the model's predictions deviate significantly from the real values. According to the bar graph given above, it is observable that Lasso Regression has a high MSE value compared to Random Forest Regression and Linear Regression, which suggests that Lasso Regression not a very good fit for the dataset and the research. Also, the graph suggests that Random Forest Regression is a better choice compared to Linear Regression because it has low MSE and MAE values. Therefore, Random Forest was used when calculating the best drug for each anaemia type and the predicted time of this drug in the treatment process.

After following the steps described in **Section 5** one by one with Random Forest, for each type of anaemia, the best drug and the predicted time for the treatment of this anaemia with this drug were calculated. If the results of Random Forest are combined in a figure, it can be concluded that 5 different drugs are included in the results.

Anemia Type	Drug Name	Treatment Time
Acquired hemolytic anemia, unspecified	Lorazepam	28.19 hours
Acute posthemorrhagic anemia	Morphine Sulfate	3.25 hours
Anemia associated with other specified nutritional deficiency	SW	57.47 hours
Anemia in chronic kidney disease	Morphine Sulfate	11.39 hours
Anemia in neoplastic disease	Lorazepam	3.25 hours
Anemia of mother, antepartum condition or complication	Morphine Sulfate	24.00 hours
Anemia of mother, delivered, with mention of postpartum...	Morphine Sulfate	31.11 hours
Anemia of mother, delivered, with or without mention of...	Lorazepam	36.24 hours
Anemia of mother, postpartum condition or complication	Fentanyl Citrate	35.63 hours
Anemia of other chronic disease	Morphine Sulfate	5.08 hours
Anemia, unspecified	Morphine Sulfate	4.43 hours
Anemias due to disorders of glutathione metabolism	Morphine Sulfate	29.99 hours
Aplastic anemia, unspecified	Fentanyl Citrate	24.04 hours
Autoimmune hemolytic anemias	Morphine Sulfate	21.51 hours
Family history of anemia	Lorazepam	32.24 hours
Hereditary hemolytic anemia, unspecified	Lorazepam	46.26 hours
Iron deficiency anemia secondary to blood loss (chronic)	Morphine Sulfate	5.08 hours
Iron deficiency anemia, unspecified	Fentanyl Citrate	9.02 hours
Other non-autoimmune hemolytic anemias	Morphine Sulfate	24.67 hours
Other specified anemias	Lorazepam	24.00 hours
Other specified aplastic anemias	Morphine Sulfate	37.64 hours
Other specified iron deficiency anemias	Lorazepam	34.31 hours
Other vitamin B12 deficiency anemia	Magnesium Sulfate	29.69 hours
Pernicious anemia	Morphine Sulfate	8.97 hours
Unspecified deficiency anemia	Lorazepam	5.08 hours

Fig. 5. For each anaemia, best drugs and predicted treatment times

6.2 Process Mining Results

Two different types of graphs were obtained in process mining graphs. These are the lasagna process and the spaghetti process. The difference between these two processes is that lasagna processes have a more specific structure than spaghetti processes, and the flow of work is much more understandable in lasagna processes [19].

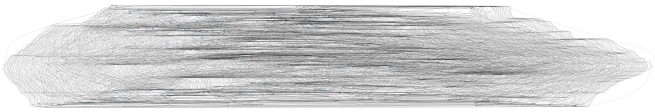


Fig. 6. Process Graph - Lasagna

The lasagna process graph given above consists of 3 main different layers. But when we look at this graph in detail, it doesn't give much information about drugs and treatment. Therefore, the case_ids have been changed and the following new process graph has been created.

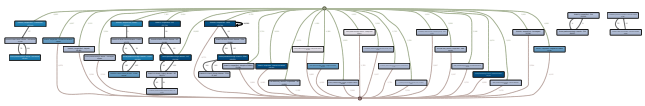


Fig. 7. Process Graph - Spaghetti

For the Spaghetti Graph, these are the chosen IDs:
- 'LONG_TITLE' → Activity
- 'ROW_ID' → Case ID
- 'SUBJECT_ID' → Case ID
- 'HADM_ID' → Case ID
- 'STARTDATE' → Timestamp (Pattern: 'yyyy-MM-dd')
- 'ENDDATE' → Timestamp (Pattern: 'yyyy-MM-dd')
- 'DRUG' → Activity

Anaemia type and drug are chosen as 'Activities' because these are attributes that are focused on. Row_ID, Subject_ID and HADM_-ID are unique and they are marked as 'Timestamps'. This spaghetti graph tells more information compared to the lasagna graph. From top to bottom, the graph is readable and the darkness of the activity's color shows how frequent that activity is. For example, (Anemia, unspecified - D5W) is shown in a very dark blue, indicating that its frequency weight is too high. Also, some activities seem to have loops. These loops show that that activity is repeated. Due to the large number of activities, it is more practical to analyze in the 'Activity' tab in Disco than to analyze in the process graph. Analysis can be found in the table below.

Anemia Type	Drug Name	Frequency	Treatment Time (mean)
Acquired hemolytic anemia, unspecified	D5W	3,114	1 day, 19 hours
Acute posthemorrhagic anemia	NS	184,014	1 day, 22 hours
Anemia associated with other specified nutritional deficiency	NS	540	17 hours, 36 minutes
Anemia in chronic kidney disease	Insulin	86,220	2 days, 8 hours
Anemia in neoplastic disease	NS	17,676	3 days, 10 hours
Anemia of mother, antepartum condition or complication	NS	396	1 day, 7 hours
Anemia of mother, delivered, with mention of postpartum...	D5W	558	19 hours, 21 minutes
Anemia of mother, delivered, with or without mention of...	NS	1,260	1 day, 17 hours
Anemia of mother, postpartum condition or complication	Potassium Chloride	468	17 hours, 32 minutes
Anemia of other chronic disease	NS	64,368	1 day, 14 hours
Anemia, unspecified	D5W	176,796	1 day, 19 hours
Anemias due to disorders of glutathione metabolism	NS	342	1 day, 1 hour
Aplastic anemia, unspecified	NS	414	1 day, 1 hour
Autoimmune hemolytic anemias	NS	4,176	2 day, 12 hours
Family history of anemia	NS	144	9 hours
Hereditary hemolytic anemia, unspecified	Insulin	108	2 day, 12 hours
Iron deficiency anemia secondary to blood loss (chronic)	NS	61,560	1 day, 23 hours
Iron deficiency anemia, unspecified	NS	43,020	1 day, 10 hours
Other non-autoimmune hemolytic anemias	D5W	3,582	2 day, 6 hours
Other specified anemias	Insulin	3,204	1 day, 21 hours
Other specified aplastic anemias	0.9% Sodium Chloride	990	1 day, 17 hours
Other specified iron deficiency anemias	NS	684	20 hours, 50 mins
Other vitamin B12 deficiency anemia	Potassium Chloride	2,412	23 hours, 17 minutes
Pernicious anemia	Potassium Chloride	1,692	1 day, 8 hours
Unspecified deficiency anemia	Potassium Chloride	6,318	1 day, 1 hour

Fig. 8. Process Graph - Statistics with a Table

In the activity statistics section, the Disco application offers attributes such as 'Activity', 'Frequency', 'Relative Frequency', 'Median Duration', 'Mean Duration' and 'Duration range'. However, I chose 'Activity', 'Frequency' and 'Mean Duration' in this dataset because I focused on which drug is used most often for the treatment of each type of anaemia and how long this drug treats the patient on average. As can be seen in the paragraph above, our pair selected as activity is anaemia type.

As can be seen from Figure 8, for every anaemia species in the dataset, the drugs that are most frequently used with their average treatment time are determined by process mining. As an example, in real life, for the treatment of 'anemia of mother, postpartum condition or complication', the most used drug is potassium chloride, and according to the records in the dataset, the average time for this drug to treat this type of anaemia is 17 hours and 32 minutes.

6.3 Comparison of Machine Learning and Process Mining Results

Anemia Type	Drug Name (PM)	Treatment Time (mean)	Drug Name (ML)	Treatment Time
Acquired hemolytic anemia, unspecified	D5W	43 hours	Lorazepam	28.19 hours
Acute posthemorrhagic anemia	NS	46 hours	Morphine Sulfate	3.25 hours
Anemia associated with other specified nutritional deficiency	NS	18 hours	SW	57.47 hours
Anemia in chronic kidney disease	Insulin	56 hours	Morphine Sulfate	11.39 hours
Anemia in neoplastic disease	NS	82 hours	Lorazepam	3.25 hours
Anemia of mother, antepartum condition or complication	NS	31 hours	Morphine Sulfate	24.00 hours
Anemia of mother, delivered, with mention of postpartum...	D5W	19 hours	Morphine Sulfate	31.11 hours
Anemia of mother, delivered, with or without mention of...	NS	41 hours	Lorazepam	36.24 hours
Anemia of mother, postpartum condition or complication	Potassium Chloride	17 hours	Fentanyl Citrate	35.63 hours
Anemia of other chronic disease	NS	36 hours	Morphine Sulfate	5.08 hours
Anemia, unspecified	D5W	43 hours	Morphine Sulfate	4.43 hours
Anemias due to disorders of glutathione metabolism	NS	25 hours	Morphine Sulfate	29.99 hours
Aplastic anemia, unspecified	NS	25 hours	Fentanyl Citrate	24.04 hours
Autoimmune hemolytic anemias	NS	60 hours	Morphine Sulfate	21.51 hours
Family history of anemia	NS	9 hours	Lorazepam	32.24 hours
Hereditary hemolytic anemia, unspecified	Insulin	60 hours	Lorazepam	46.26 hours
Iron deficiency anemia secondary to blood loss (chronic)	NS	47 hours	Morphine Sulfate	5.08 hours
Iron deficiency anemia, unspecified	NS	34 hours	Fentanyl Citrate	9.02 hours
Other non-autoimmune hemolytic anemias	D5W	54 hours	Morphine Sulfate	24.67 hours
Other specified anemias	Insulin	45 hours	Lorazepam	24.00 hours
Other specified aplastic anemias	0.9% Sodium Chloride	41 hours	Morphine Sulfate	37.64 hours
Other specified iron deficiency anemias	NS	21 hours	Lorazepam	34.31 hours
Other vitamin B12 deficiency anemia	Potassium Chloride	23 hours	Magnesium Sulfate	29.69 hours
Pernicious anemia	Potassium Chloride	32 hours	Morphine Sulfate	8.97 hours
Unspecified deficiency anemia	Potassium Chloride	25 hours	Lorazepam	5.08 hours

Fig. 9. Results in one table

7 DISCUSSION

This research has investigated the subject of making predictions for the anaemia treatment with machine learning and process mining with the help of prescription records. This research was developed around the research question 'To what extent it is effective to use machine learning and process mining to predict the best anaemia treatment with the help of prescription records?' and in order to follow an effective way when answering this question, this question is divided into 3 sub-research questions. These questions and their answers can be found below.

7.1 Answering the first sub-research question

With the help of Random Forest Regression and information about the drug prescription in the dataset, predictions could be made for the anaemia types, and the best drug for each type was determined. The results of this can be found **here**.

7.2 Answering the second sub-research question

With Disco, a process mining application, the application of drugs included in the MIMIC-III dataset on patients was detected and Lasagna and Spaghetti graphs were created with the output. In addition, thanks to graphs, it was seen how frequent the activities (drugs applied to anaemia types) were and how much each drug type was applied against each type of anaemia in real life. The results of this can be found **here**.

7.3 Answering the third sub-research question

When the machine learning and process mining results are compared, it is concluded that these two methods actually generate different outputs. However, the reason for these differences has been investigated and explained with reasons and examples. These results and explanations can be found in **Section 6.3** and **Section 8**.

8 CONCLUSION

The table in **Section 6.3** shows that the drugs used in real life do not correspond with the drugs which are predicted as 'best' by the Random Forest Algorithm. Additionally, the table contains only 5 different drugs for process mining (D5W, NS, Insulin, Potassium Chloride, 0.9% Sodium Chloride) and machine learning prediction results (Lorazepam, Morphine Sulfate, SW, Fentanyl Citrate, Magnesium Sulfate). These results may be explained by the following reasons why both treatments are treated with only 5 drugs out of 30 drugs (that is, these 5 drugs are more advantageous than the other 25 drugs).

- Cost: The drug recommended for anaemia treatment by machine learning is perhaps much more expensive than the drug used in real life generated by process mining logs. Therefore, due to the price difference, a more affordable drug may be used instead of the most efficient drug in the treatment of that particular anaemia type.
- Availability of Drugs: Finding some drugs can be very difficult. Their production may be limited or there may be problems while passing through customs. Therefore, the drug revealed by the process mining logs may be used instead of the drugs that are the best in the treatment predicted by the algorithm.
- Side-effects of Drugs: Although the drugs recommended by the algorithm are the best drugs in terms of treatment duration, the Random Forest Algorithm does not consider the side-effects of the drug when making predictions. Maybe these drugs have some negative effects on patients that these drugs are not used that much in real life.
- Legal Reasons: Some drugs may be prohibited or restricted for use in Israel. Therefore, drugs that are allowed by the government are used, not the fastest drugs in treatment.

To support the points above, some comparisons are given below.

The reason why NS is used more than Fentanyl Citrate is the price difference. I was able to find a 100 ml bottle of NS for 10 Rupees, that is, 0.11 Euros, **in a store that sells NS on the internet**. On the other hand, the price of **100 mcg Fentanyl Citrate with the brand Johnson & Johnson** can go up to 3595 Rupee, that is, 40 Euros.

Lorazepam's side effects may be the reason why D5W is preferred over Lorazepam, although Lorazepam is predicted to be a better drug. Side effects of D5W (dextrose) are ordinary side-effects such as headache, changes in skin color, swelling that can occur after taking any medication [3]. However, Lorazepam has side effects such as life-threating breathing problems, or coma (if taken together with some drugs) and other than that, it can addict the patient [4].

The MIMIC-III dataset I'm using contains data up to 2012. A study that started in 2013 investigates drug shortages in Israel [17] and according to the results of that study, there is a Morphine Sulfate deficiency in Israel. On the other hand, insulin is a very easily accessible drug that can be bought without a prescription, even from supermarkets. Therefore, due to the shortage of morphine sulfate in Israel, morphine sulfate is a better drug according to the machine learning prediction, process mining logs have determined that insulin is used more in the treatment of anaemia.

9 LIMITATIONS

9.1 PC Memory Storage

I got a lot of 'Memory Space' warnings, especially while training Random Forest algorithm with k-fold and calculating predictions. Training was wasting so much memory on the computer that it was impossible for me to do other activities during that time.

9.2 Size of the dataset

The size of the dataset and the large number of tables required me to double-check every query I executed. Since the wrong use of many tables and foreign keys will cause wrong results, I have tried to make sure that the queries I have are correct. However, due to the size of the dataset, this situation caused me a great loss of effort.

9.3 Physionet problems

For this research, I needed to solve the Mimic-III dataset trainings and tests required by Physionet. After spending some time on these, I waited for a while for my application to be approved. The importance of the dataset for this research is indisputable, and I couldn't do anything on the project when I didn't have access to the dataset. I waited for this permission for about 1 week and this time was a loss for me. After the permission was granted, I wanted to add this dataset to Google BigQuery. Despite following the instructions they shared with me (along with my supervisor and other knowledgeable people), I was unable to add the dataset to GoogleBigQuery in any way. Despite sending an e-mail to Physionet, I did not receive any response. Due to the size of the Dataset, not every database management system was suitable for this. Because of this, I had a loss of 4-5 days.

9.4 Algorithm training times

The dataset I am using has millions of rows of data and it takes time to train. K-fold is used to increase the accuracy and to randomize the data, and the random forest's computation for each decision tree slows down the whole process even more [8]. Even though I optimized the code and dataset, this situation still hasn't changed much. As a result, I had to wait 3-4 hours for each training.

10 FUTURE IMPROVEMENTS

10.1 Same research with different datasets

In order to generalize my conclusions I have made as a result of my research, the same research should be conducted with the same attributes and algorithms, perhaps with data from another country. If, for example, data from France also match the results I have now found, then this means I have successfully researched about anaemia treatment.

10.2 Slightly expanding research's scope

The research I'm doing right now is all about the treatment of anaemia. Another dataset can be obtained and a research can be done on anaemia diagnosis.

10.3 Updating machine learning algorithms and repeating research

As I said in the abstract, since machine learning is a branch that constantly evolves and an algorithm with higher accuracy emerges, repeating this research will perhaps lead to a more accurate prediction.

10.4 Repeating research with more recent datasets

Keeping everything constant, just repeating the same research with Physionet's more recent MIMIC-III dataset and the new results matching the current results would help generalize the results I've found now.

ACKNOWLEDGMENTS

Although I was doing research on a health-themed subject for the first time in my life, I had less difficulty than I expected. In this section, I would like to express my gratitude to my supervisor dr. Faiza Allah Bukhsh, who supported me in every problem I faced and supported me with the extra research materials they provided. Certainly, such detailed and successful research at the end of these 11 weeks had not been became real without the periodic meetings with Faiza.

REFERENCES

[1] [n. d.]. Anaemia. https://www.who.int/health-topics/anaemia#tab=tab_1
[2] [n. d.]. Anaemia in woman and children. https://www.who.int/data/gho/data/themes/topics/anaemia_in_women_and_children
[3] [n. d.]. Dextrose (Intravenous Route). https://www.mayoclinic.org/drugs-supplements/dextrose-intravenous-route/side-effects/drg-20073387
[4] [n. d.]. Lorazepam. https://medlineplus.gov/druginfo/meds/a682053.html
[5] [n. d.]. Scopus. https://www.scopus.com/search/form.uri?display=basic#basic
[6] [n. d.]. University Library. https://www.utwente.nl/en/service-portal/university-library
[7] [n. d.]. What is process mining? https://www.ibm.com/topics/process-mining#:~:text=the%20next%20step-,What%20is%20process%20mining%3F,and%20other%20areas%20of%20improvement.
[8] [n. d.]. What is random forest? https://www.ibm.com/topics/random-forest
[9] 2021-08-17. CRISP-DM Help Overview. https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview
[10] 2022-10-25. Random Forest Algorithm. https://www.mygreatlearning.com/blog/random-forest-algorithm/
[11] Rob Bemthuis, Niels van Slooten, Jeewanie Jayasinghe Arachchige, Jean Paul Sebastian Piest, and Faiza Allah Bukhsh. 2021-01. A Classification of Process Mining Bottleneck Analysis Techniques for Operational Support. (2021-01). https://doi.org/10.5220/0010578601270135
[12] Carlos Fernandez-Llatas. 2021. Interactive Process Mining in Healthcare. In *Interactive Process Mining in Healthcare*. https://link.springer.com/book/10.1007/978-3-030-53993-1
[13] Debbie L. Hahs-Vaughn. 2023. Pearson Correlation Coefficient. (2023). https://www.sciencedirect.com/topics/social-sciences/pearson-correlation-coefficient
[14] A. Johnson, T. Pollard, and R. Mark. [n. d.]. MIMIC-III Clinical Database (version 1.4). https://doi.org/10.13026/C2XW26
[15] Archana J. McEligot, Valerie Poynor, Rishabh Sharma, and Anand Panangadan. 2020-08-31. Logistic LASSO Regression for Dietary Intakes and Breast Cancer. In *Logistic LASSO Regression for Dietary Intakes and Breast Cancer*. https://doi.org/10.3390/nu12092652
[16] Mike Pingel. 2021-08. Leveraging Machine Learning and Process Mining to Predict Anaemia with the Help of Biomarker Data. (2021-08).
[17] Eyal Schwartzberg, Denize Ainbinder, Alla Vishkauzan, and Ronni Gamzu. [n. d.]. Drug shortages in Israel: regulatory perspectives, challenges and solutions. 6 ([n. d.]). https://doi.org/10.1186/s13584-017-0140-9
[18] Shaun Turney. 2022-12-05. Pearson Correlation Coefficient (r) | Guide & Examples. https://www.scribbr.com/statistics/pearson-correlation-coefficient/
[19] Wil van der Aalst. 2021. Process mining: discovering and improving Spaghetti and Lasagna processes. (2021). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6129461
[20] Xinyu Zhang and Chu-An Liu. 2023-04-26. Model averaging prediction by K-fold cross-validation. In *Model averaging prediction by K-fold cross-validation*. https://doi.org/10.1016/j.jeconom.2022.04.007
[21] Zhi-Hua Zhou. 2021-08-21. Machine Learning. In *Machine Learning*. Springer, Singapore, 1–24. https://doi.org/10.1007/978-981-15-1967-3_1

A APPENDIX A: ARTICLE ANALYSIS

A.1 Article Analysis - FindUT

	Name	Keywords	Branch	Health Related?	Citations
1	Data-driven dynamic causality analysis of industrial systems using interpretable machine learning and process mining	Causality analysis · Interpretable machine learning · Process mining · Petri nets · Discrete event systems · Supervisory control	No	Black Box Prediction	?
2	LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models	Interpretable machine learning, Post-hoc interpretation, Probabilistic inference, Bayesian network, Predictive analytics, Explainable AI	No	Business Management	24
3	Business process variant analysis: Survey and classification	Process mining, Machine learning, Business process management	No	Event Prediction	28
4	Process data properties matter: Introducing gated convolutional neural networks (GCNN) and key-value-predict attention networks (KVP) for next event prediction with deep learning	Process mining, Predictive process monitoring, Machine learning, Deep learning, Gated convolutional neural network, Key-value-predict attention network	No	Business Process Management	18
5	Detecting anomalies in business process event logs using statistical leverage	Business process event log, Anomaly score, Case anomaly detection, Statistical leverage, Information-theoretic measure	No	Business Process Management	7
6	Leveraging Small Sample Learning for Business Process Management	Small Sample Learning BPM, Process Prediction, Machine Learning, Process Mining	No	Business Process Management	2
7	Enhancing change mining from a collection of event logs: Merging and Filtering approaches	Business process, variability, configurable process, collection of event logs, Filtering, Merging, variable fragments.	No	Business Process Management	5

Fig. 10. FindUT - Articles

A.2 Article Analysis - Scopus

	Name	Keywords	Branch	Health Related?	Citations
1	Improving the In-Hospital Mortality Prediction of Diabetes ICU Patients Using a Process Mining/Deep Learning Architecture	deep learning; diabetes; in-hospital mortality; intensive care; Process mining; risk assessment	Yes	Medicine	21
2	Mining tasks and task characteristics from electronic health record audit logs with unsupervised machine learning	audit logs; clinician activities; electronic health records; human-computer interaction; metrics; tasks; Unsupervised learning	Yes	Medicine	10
3	Prediction of unplanned 30-day readmission for ICU patients with heart failure	Deep learning; Heart failure; Hospital readmission; Process mining	Yes	Medicine	4
4	A process mining-deep learning approach to predict survival in a cohort of hospitalized COVID-19 patients	COVID-19 prediction; Deep learning; Machine learning; Mortality prediction; Process mining; SARS-CoV-2	Yes	Medicine	3
5	Composition of a Service-Oriented Clinical Decision Support System using Machine Learning	Artificial Intelligence; Clinical Guideline; Computer Interpretable Guideline; Path Diagnostic Therapeutic Care; Pathways	Yes	Medicine	0

Fig. 11. Scopus - Articles

B APPENDIX B: THE MIMIC-III DATASET

Fig. 12. Tables and Attributes of the MIMIC-III Dataset

C APPENDIX C: SQL QUERIES AND OUTPUTS

C.1 Getting ICD9 CODES AND EXPLANATIONS FOR ANAEMIA

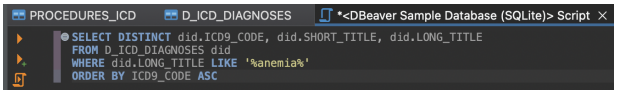


Fig. 13. SQL Query for returning the ICD9 Codes

ICD9_CODE	SHORT_TITLE	LONG_TITLE
2800	Chr blood loss anemia	Iron deficiency anemia secondary to blood loss (chronic)
2801	Iron def anemia dietary	Iron deficiency anemia secondary to inadequate dietary iron intake
2808	Iron defic anemia NEC	Other specified iron deficiency anemias
2809	Iron defic anemia NOS	Iron deficiency anemia, unspecified
2810	Pernicious anemia	Pernicious anemia
2811	B12 defic anemia NEC	Other vitamin B12 deficiency anemia
2812	Folate-deficiency anemia	Folate-deficiency anemia
2813	Megaloblastic anemia NEC	Other specified megaloblastic anemias not elsewhere classified
2814	Protein defic anemia	Protein-deficiency anemia
2818	Nutritional anemia NEC	Anemia associated with other specified nutritional deficiency
2819	Deficiency anemia NOS	Unspecified deficiency anemia
2822	Glutathione dis anemia	Anemias due to disorders of glutathione metabolism
2823	Enzyme defic anemia NEC	Other hemolytic anemias due to enzyme deficiency
2828	Hered hemolytic anem NEC	Other specified hereditary hemolytic anemias
2829	Hered hemolytic anem NOS	Hereditary hemolytic anemia, unspecified
2830	Autoimmun hemolytic anem	Autoimmune hemolytic anemias
2839	Acq hemolytic anemia NOS	Acquired hemolytic anemia, unspecified
2849	Aplastic anemia NOS	Aplastic anemia, unspecified
2850	Sideroblastic anemia	Sideroblastic anemia
2851	Ac posthemorrhag anemia	Acute posthemorrhagic anemia
2853	Anemia d/t antineo chemo	Antineoplastic chemotherapy induced anemia
2858	Anemia NEC	Other specified anemias
2859	Anemia NOS	Anemia, unspecified
7735	NB late anemia:isoimmun	Late anemia of fetus or newborn due to isoimmunization
7740	Perinat jaund-hered anem	Perinatal jaundice from hereditary hemolytic anemias
7765	Congenital anemia	Congenital anemia
7766	Anemia of prematurity	Anemia of prematurity
28310	Nonauto hem anemia NOS	Non-autoimmune hemolytic anemia, unspecified
28319	Oth nonauto hem anemia	Other non-autoimmune hemolytic anemias
28409	Const aplastc anemia NEC	Other constitutional aplastic anemia
28489	Aplastic anemias NEC	Other specified aplastic anemias
28521	Anemia in chr kidney dis	Anemia in chronic kidney disease
28522	Anemia in neoplastic dis	Anemia in neoplastic disease
28529	Anemia-other chronic dis	Anemia of other chronic disease
64820	Anemia in preg-unspec	Anemia of mother, unspecified as to episode of care or not applicable
64821	Anemia-delivered	Anemia of mother, delivered, with or without mention of antepartum condition
64822	Anemia-delivered w p/p	Anemia of mother, delivered, with mention of postpartum complication
64823	Anemia-antepartum	Anemia of mother, antepartum condition or complication
64824	Anemia-postpartum	Anemia of mother, postpartum condition or complication
V182	Family hx-anemia	Family history of anemia
V780	Screen-iron defic anemia	Screening for iron deficiency anemia
V781	Screen-defic anemia NEC	Screening for other and unspecified deficiency anemia

Fig. 14. ICD9 Codes and Explanations for Anaemia

C.2 Getting the procedure data with ICD9 codes

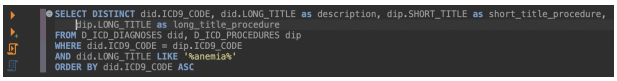


Fig. 15. SQL Query for returning the procedure per ICD9 Codes

ICD9_CODE	description	short_title_procedure	long_title_procedure
2811	Other vitamin B12 deficiency anemia	Tonsil&adenoid biopsy	Biopsy of tonsils and adenoids
2819	Unspecified deficiency anemia	Tonsil&adenoid dx op NEC	Other diagnostic procedures on tonsils and adenoids
7735	Late anemia of fetus or newborn due to isoimmunization	Femoral division NEC	Other division of bone, femur
7740	Perinatal jaundice from hereditary hemolytic anemias	Bone biopsy NOS	Biopsy of bone, unspecified site
7765	Congenital anemia	Loc exc bone les femur	Local excision of lesion or tissue of bone, femur
7766	Anemia of prematurity	Loc exc bone les patella	Local excision of lesion or tissue of bone, patella

Fig. 16. Procedure for each anaemia type

C.3 Getting the drugs prescribed

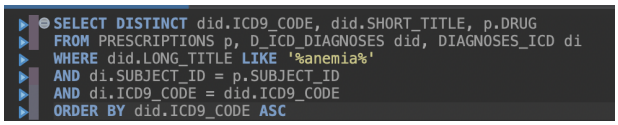


Fig. 17. Anaemia Type and Prescribed Drug

C.4 Drugs with number of times they used

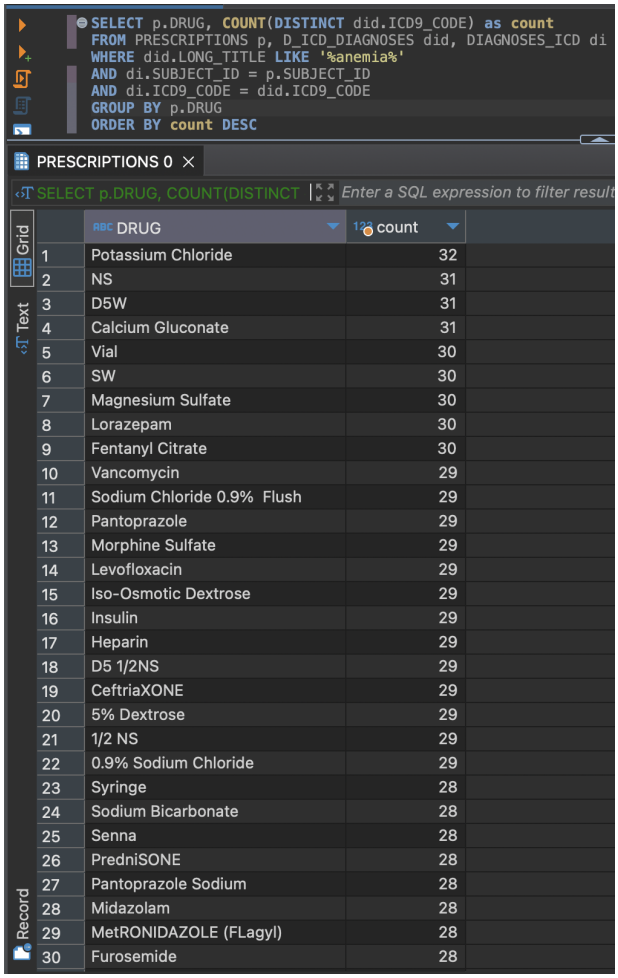


Fig. 18. Drug vs Anaemia Type

C.5 Most popular drugs with their average prescription duration

```
SELECT dft.DRUG, AVG(dft.treatment_time_in_hours) as average
FROM datasetForTime dft
WHERE dft.DRUG = 'Potassium Chloride' OR dft.DRUG = 'NS' OR
dft.DRUG = 'D5W' OR dft.DRUG = 'Calcium Gluconate' OR dft.DRUG = 'Vial' OR dft.DRUG = 'SW' OR
dft.DRUG = 'Magnesium Sulfate' OR dft.DRUG = 'Lorazepam' OR dft.DRUG = 'Fentanyl Citrate' OR dft.DRUG = 'Vancomycin' OR
dft.DRUG = 'Sodium Chloride 0.9%  Flush' OR dft.DRUG = 'Pantoprazole' OR dft.DRUG = 'Morphine Sulfate' OR
dft.DRUG = 'Levofloxacin' OR dft.DRUG = 'Iso-Osmotic Dextrose' OR dft.DRUG = 'Insulin' OR dft.DRUG = 'Heparin'
OR dft.DRUG = 'D5 1/2NS' OR dft.DRUG = 'CeftriaXONE' OR dft.DRUG = '5% Dextrose' OR dft.DRUG = '1/2 NS' OR
dft.DRUG = '0.9% Sodium Chloride' OR dft.DRUG = 'Syringe' OR dft.DRUG = 'Sodium Bicarbonate' OR dft.DRUG = 'Senna' OR
dft.DRUG = 'PrednISONE' OR dft.DRUG = 'Pantoprazole Sodium' OR dft.DRUG = 'Midazolam'
OR dft.DRUG = 'MetRONIDAZOLE (FLagyl)' OR dft.DRUG = 'Furosemide'
GROUP BY dft.DRUG
HAVING average > 0
ORDER BY average DESC
LIMIT 30
```

Fig. 19. SQL query for: Drugs and How many minutes they are used in descending order

	DRUG	average
1	Senna	135,208287596
2	Sodium Chloride 0.9% Flush	134,3328856485
3	Heparin	103,5103420461
4	Pantoprazole Sodium	96,1281285878
5	Pantoprazole	91,2367284961
6	Vial	91,0074279651
7	Calcium Gluconate	80,8253310305
8	SW	70,760321006
9	Insulin	70,2934931865
10	MetRONIDAZOLE (FLagyl)	67,0298417785
11	Magnesium Sulfate	64,7464142347
12	Levofloxacin	64,6352005367
13	Lorazepam	63,3043792456
14	Fentanyl Citrate	54,1437999305
15	Iso-Osmotic Dextrose	50,7159818401
16	Morphine Sulfate	50,5825473976
17	CeftriaXONE	49,9937117326
18	NS	49,432440974
19	Potassium Chloride	44,4569492158
20	Vancomycin	43,8050382324
21	Midazolam	41,6765309112
22	0.9% Sodium Chloride	41,1627467402
23	D5W	40,6115817521
24	Syringe	37,9484420534
25	5% Dextrose	36,809925639
26	PrednISONE	33,1783893986
27	Furosemide	33,1585162483
28	Sodium Bicarbonate	28,8755844575
29	D5 1/2NS	23,7390022118
30	1/2 NS	19,0363318438

Fig. 20. Results

D PROCESS MINING GRAPHS

D.1 Drug Frequency from Activities of Spaghetti Graph

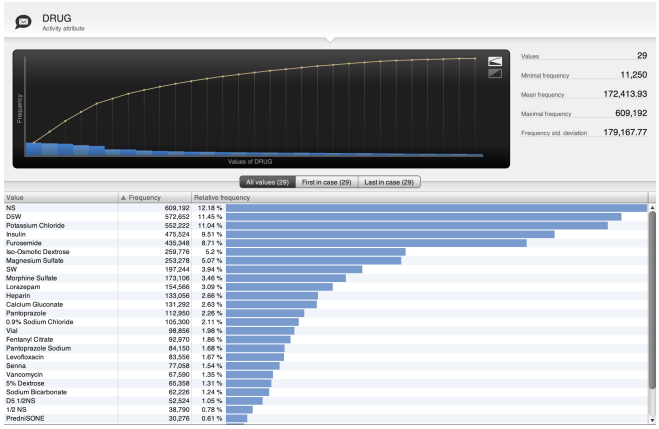


Fig. 21. Drug Frequencies