**DANA4830 – Fall 2021**

Submitted by:

Mohamed Ghayaas Anjum

Brief Description of the Dataset:

Acute Pancreatitis (Hereafter abbreviated as AP) is the sudden inflammation of the Pancreas which can be mild or life threatening but usually subsides. In this condition, the inflammation develops quickly and subsides within a few days but can sometimes last for up to a few weeks.

This dataset consists of a set of 165 patients who were diagnosed with hyper triglyceridemic pancreatitis. Out of these, 83 were subjected to Plasma Exchange Therapy (hereafter abbreviated as PEX) and 82 patients were treated according to Vietnam's Ministry of Health's guidelines (2015) for treatment of acute pancreatitis.

All the variable names and the values in the variables which were in Vietnamese language were replaced to English language. A translated dataset was used in the analysis.

The objective of study is to determine effect of various patient characteristics that depend on the two response variables of the study – **Outcome of the patient (Dead or Alive) and Number of days of hospitalization** of the patient.

**Meaning of Variables:**

The meaning of each variable and a brief description of the same is presented in the following table. The following table also contains the reason to exclude those variables from our dataset.

**Table 1: Study of PEX treatment for Acute Pancreatitis – Meaning of Each Variable**

| Var No. | Name | Meaning of the Variable | Remove or Keep | Missing % | Reason to Remove |
|---|---|---|---|---|---|
| 1 | ID | ID variable - a number given to each patient having AP and who were considered in the study | Remove | 0.00 | Id of patients not required for analysis |
| 2 | Age | Patient's age | Keep | 0.00 | |
| 3 | Gender | Patient's gender | Keep | 0.00 | |
| 4 | vv_reason_1 | Major reason of the patient to visit the hospital | Keep | 3.64 | |

| | | | | | |
|---|---|---|---|---|---|
| 5 | vv_reason2 | Additional reasons (if any) for admission into the hospital | Remove | 47.27 | Major reason for hospital admisison is enough. extra variables are redundant |
| 6 | vv_reason3 | Additional reasons (if any) for admission into the hospital | Remove | 96.36 | Major reason for hospital admisison is enough. extra variables are redundant |
| 7 | vv_Others | Additional reasons (if any) for admission into the hospital | Remove | 95.15 | Major reason for hospital admisison is enough. extra variables are redundant |
| 8 | rv_ngaydt | No of days a patient stayed in the hospital. Min =1, Max = 18 in this dataset | Keep | 0.00 | |
| 9 | ts_giadinh | some patients could get AP as hereditary. By this meaning, the variable could mean that is the AP of the patient is hereditary or not. | Keep | 11.52 | |
| 10 | details_ts_giadinh | AP can be caused by rapid lipid metabolism leading to abnormally high levels of triglycerimedia aka hypertriglycerimedia. This rapid metabolism causes swelling in the pancreas and can cause AP. This variable could mention weather this high levels of lipid in the blood is hereditary in the patient or not. | Remove | 71.52 | Extra hereditary information may not have a significant impact on the model |
| 11 | ts_benhmat | Gall stones causes about 40% of the cases of AP. These gallstones pass in the duct connected to pancreas and causes severe pain leading to AP. The variable means that weather the current AP is due to this cause or not | Keep | 0.00 | All values are 'No'. Only 1 value is 'Have'. Keeping this variable, we may not be able to study the gallbladder effect as the values are biased on patients who do not have gallbladder problem. Still lets keep it as there are no missing values |
| 12 | ts_ruou | Alcohol causes about 30% of the cases of AP. The risk of developing AP increases with increasing amounts of alcohol (4 to 7 drinks/day in men and 3 or more drinks/day in women) However less than 10% people who drink alcohol are reported to develop AP. The variable means weather alcohol is the cause for AP in the patient. | Keep | 1.21 | |
| 13 | ts_ruou_nam | The whole number values like 8,10,20 probably indicates how many drinks the patient has consumed in a day | Keep | 53.33 | |
| 14 | ts_ruou_nam_ml | the values in this variable like 500,800,200 probably indicates the amount of alcohol consumed in a day in ml | Remove | 53.33 | Above variable is explainatory of the amount of drinks per day. This variable is redundant |
| 15 | ts_dtd | one meaning is severe AP can cause improper development of insulin in a patient and in severer cases of AP, a patient can develop diabetes due to the existing AP. Second meaning is a patient who admitted in the hospital with AP has a previous diabetes problem or not | Keep | 0.00 | |

| | | | | | |
|---|---|---|---|---|---|
| 16 | ts_vtc | cholesystitis means swelling of gallbladder due to the gallstones blocking the tube, or due to bile juice getting trapped in the gallbladder. The variable means weather the patient has previously reported to cholesystis or not | Keep | 0.61 | |
| 17 | ts_vtc_lancuoi | If the patient has reported for cholesystis, when was the last time it was reported | Remove | 52.12 | Values are not clean and not understandable. And last detection of cholesystis is not important in our study |
| 18 | daubung | Tummy pain is the most common symptom of AP in the patient. The variable means if the patient coming into the hospital is reporting tummy pain or not | Remove | 0.00 | All values are 1. Except 2 values. Cannot study the impact of tummy pain on pex treatment with this distribution of values |
| 19 | non | Vomitting or nausea is another most common symptom of AP. The variable indicated weather the person is reporting to being nauseous or vomitting or not | Keep | 46.67 | |
| 20 | ls_cn_bidaitien | Pale yellow, greasy, foul-smelling stool: malabsorption of fat due to pancreatic insufficiency can be seen in patients with AP. Variable indicates weather these symptoms present in the patient or not | Keep | 72.12 | |
| 21 | ls_cn_ialong | Both acute and chronic pancreatitis can cause the pancreas to produce fewer enzymes that are needed to process nutrients causing diarrhea. Variable indicates weather patient has diarrhea symptom or not | Keep | 75.15 | |
| 22 | ls_tht_bungchuong | Severe abdominal pain is a predominant symptom. A person may also develop some swelling in the upper abdomen. T0 means time at which patient was admitted. | Keep | 10.91 | |
| 23 | ls_tt_lungsuon | The abdominal pain penetrates to the back for almost 50% of the patients. This variable is an indication of penetrating pain in the patients | Remove | 80.61 | Above variable is explainatory of the abdominal distention. This variable is redundant. Almost all values are either 'no' or Nan |
| 24 | ls_tt_alob_t0 | Using the Atlanta criteria, acute pancreatitis is diagnosed when a patient presents with two of three findings, including abdominal pain suggestive of pancreatitis, serum amylase and/or lipase levels at least three times the normal level. Based on the values, this variable should be Lipase level in Blood. normal range <67 U/l | Keep | 41.21 | |
| 25 | ls_tt_bmi_t0 | Should be amylase level aka serum amylase in blood. Normal range 22-80 U/l | Keep | 1.82 | |
| 26 | ls_tn_mach_t0 | Should be systolic blood pressure | Keep | 3.03 | |
| 27 | ls_tn_nhiet_t0 | Body temperature | Keep | 4.24 | |
| 28 | ls_tn_ha_t6 | Blood pressure | Keep | 60.61 | |

| | | | | | |
|---|---|---|---|---|---|
| 29 | ls_tn_spo2_t0 | Oxygen saturation in patient. Denoted by Spo2. lesser oxygen saturation could be a measure of damaged lungs | Keep | 3.64 | |
| 30 | ls_tn_cvp_t0 | Central venous pressure is the blood pressure in the venae cavae, near the right atrium of the heart. CVP reflects the amount of blood returning to the heart and the ability of the heart to pump the blood back into the arterial system. Normal range 8 to 12 mmHg | Remove | 72.73 | 120 values are empty. The center venus pressure was noted on only about 40 patients. This measure is not important in study of AP |
| 31 | ls_diem_apache_t0 | APACHE II is a severity-of-disease scoring systems. Applied within 24 hours of admission of a patient to an ICU. Range = 0 to 71. Higher scores correspond to more severe disease and a higher risk of death | Keep | 15.15 | |
| 32 | ls_diem_ranson_t0 | Ranson criteria isused to predict the severity and mortality of acute pancreatitis. Five parameters are assessed on admission, and the other six are assessed at 48 hours post-admission. One point is given for each positive parameter for a maximum score of 11. The modified criteria have a max score of 10. Five parameters are assessed on admission and the other 5 at the 48-hour mark | Keep | 15.15 | |
| 33 | ls_diem_ct_t0 | CT severity index sums two scores, Balthazar score and grading the extent of pancreatic necrosis. Grading of balthazar - 0 to 4, pancreatic necrosis values are 0,2,4,6. Max ct score can be 10 | Keep | 31.52 | |
| 34 | ls_diem_imrie_t0 | Glasgow imrie score is a modification of Ranson's criteria for AP. Measured at 48 hours after hospital admission. Max score is 8 | Keep | 14.55 | |
| 35 | ls_diem_sofa_t0 | The Sequential Organ Failure Assessment (SOFA) score is a simple and objective score that allows for calculation of both the number and the severity of organ dysfunction in six organ systems (respiratory, coagulatory, liver, cardiovascular, renal, and neurologic). ranges from 0 to 24 | Keep | 15.15 | |
| 36 | cls_sa_tuy_t0 | Patient referred for Ultrasound of pancreas/ type of ultrasound | Remove | 10.91 | The character values in this variable are like 'VTC', does not help us in understanding the patient condition. |
| 37 | CLS_S2 | Has no value | Remove | 100.00 | Has no value |
| 38 | cls_sa_dichob_t0 | Ultrasound finding - Weather patient had abdominal fluid. The pancreas secretes pancreatic fluid into the first part of the small intestine (duodenum). This contains digestive enzymes that help digest food. If a gallstone becomes stuck (where the pancreatic duct empties into the duodenum), the fluid stops flowing. Usually, the blockage is temporary and causes limited damage, which is soon repaired. But if the blockage remains, the enzymes collect in the pancreas and begin to digest the cells of the pancreas, causing severe inflammation | Keep | 11.52 | |

| | | | | | |
|---|---|---|---|---|---|
| 39 | cls_sa_mat_t0 | Weather the patient was referred for bladder ultrasound. The value 'bt' could mean bladder test | Remove | 50.91 | All values are 'bt' which probably means bladder test. Does not inform us anything about patient condition |
| 40 | cls_sa_ketluan_t0 | Conclusion of the ultrasound when the patient was admitted. | Remove | 9.70 | The character values in this variable are like 'VTC', does not help us in understanding the patient condition. |
| 41 | CLS_S1 | Has no value | Remove | 100.00 | Has no value |
| 42 | cls_ct_tuy_lan1 | Weather the patient was referred for computer tomography for diagnosis of AP | Remove | 25.45 | The character values in this variable are like 'VTC', does not help us in understanding the patient condition. |
| 43 | cls_ct_dichob_lan1 | Weather or not abdominal fluid is present, during the computer tomography | Keep | 32.73 | |
| 44 | cls_ct_balthazar_lan1 | Balthazar score calculated from computer tomography. Grading from 0,1,2,3,4 also represented by A,B,C,D,E. | Keep | 26.67 | |
| 45 | cls_ct_ctscore_lan1 | CTSI calculated from computer tomography sums two scores, Balthazar score and grading the extent of pancreatic necrosis. Grading of balthazar - 0 to 4, pancreatic necrosis values are 0,2,4,6. Max ct score can be 10 | Keep | 26.67 | |
| 46 | cls_hh_bc_t0 | WBC level in the blood at the point of admisison to the hospital | Keep | 3.03 | |
| 47 | cls_hh_bc_t6 | WBC level in the blood at 6 hours after hospital admission | Remove | 60.61 | We can see a pattern that WBCs have increased from point of admisison till 72 hours. These 4 variables are just a record of how much they have increased. Imputation will be inappropriate if we keep |
| 48 | cls_hh_bc_t30 | WBC level in the blood at 30 hours after hospital admission | Remove | 41.82 | We can see a pattern that WBCs have increased from point of admisison till 72 hours. These 4 variables are just a record of how much they have increased. Imputation will be inappropriate if we keep |
| 49 | cls_hh_bc_t54 | WBC level in the blood at 54 hours after hospital admission | Remove | 56.97 | We can see a pattern that WBCs have increased from point of admisison till 72 hours. These 4 variables are just a record of how much they have increased. Imputation will be inappropriate if we keep |

| | | | | | |
|---|---|---|---|---|---|
| 50 | cls_hh_bc_t72 | WBC level in the blood at 72 hours after hospital admission | Remove | 54.55 | We can see a pattern that WBCs have increased from point of admisison till 72 hours. These 4 variables are just a record of how much they have increased. Imputation will be inappropriate if we keep |
| 51 | cls_hh_tc_t0 | TC count or total cell count of WBCs. A high white blood cell count may indicate that the immune system is working to destroy an infection | Keep | 2.42 | |
| 52 | cls_hh_tc_t6 | TC count at 6 hours after hospital admission | Remove | 60.00 | Just a record of the total count cells in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 53 | cls_hh_tc_t30 | TC count at 30 hours after hospital admission | Remove | 43.64 | Just a record of the total count cells in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 54 | cls_hh_tc_t54 | TC count at 54 hours after hospital admission | Remove | 55.76 | Just a record of the total count cells in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 55 | cls_hh_tc_t72 | TC count at 72 hours after hospital admission | Remove | 55.15 | Just a record of the total count cells in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 56 | cls_hh_hct_t0 | Hematocrit is the percentage by volume of red blood cells in the blood. Normal levels for men is 41% to 50%. For women is 36% to 48% | Keep | 2.42 | |
| 57 | cls_hh_hct_t6 | Hematocrit at 6 hours after hospital admission | Remove | 58.79 | Just a record of the hematocrit cells in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 58 | cls_hh_hct_t30 | Hematocrit at 30 hours after hospital admission | Remove | 41.82 | Just a record of the hematocrit cells in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |

| | | | | | |
|---|---|---|---|---|---|
| 59 | cls_hh_hct_t72 | Hematocrit at 72 hours after hospital admission | Remove | 55.76 | Just a record of the hematocrit cells in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 60 | cls_hh_hc_t0 | RBC level in the blood. Normal range for men – 4.7 to 6.1 million cells per microlitre (cells/mcL), women – 4.2 to 5.4 million cells/mcL | Keep | 11.52 | |
| 61 | cls_hh_hc_t6 | RBC count at 6 hours after hospital admission | Remove | 63.03 | Just a record of the RBC in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 62 | cls_hh_hc_t30 | RBC count at 30 hours after hospital admission | Remove | 45.45 | Just a record of the RBC in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 63 | cls_hh_hc_t54 | RBC count at 54 hours after hospital admission | Remove | 61.82 | Just a record of the RBC in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 64 | cls_hh_hc_t72 | RBC count at 72 hours after hospital admission | Remove | 57.58 | Just a record of the RBC in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 65 | cls_hh_pt_t0 | Prothrombin is a protein made by the liver. A substance known as clotting (coagulation) factors. | Keep | 6.06 | |
| 66 | cls_hh_pt_t6 | Prothrombin at 6 hours after hospital admission | Remove | 63.03 | Just a record of the Prothrombin in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 67 | cls_hh_pt_t30 | Prothrombin at 30 hours after hospital admission | Remove | 52.73 | Just a record of the Prothrombin in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |

| | | | | | |
|---|---|---|---|---|---|
| 68 | cls_hh_pt_t72 | Prothrombin at 72 hours after hospital admission | Remove | 63.64 | Just a record of the Prothrombin in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 69 | cls_hh_aptt_t0 | aPTT is a screening test that helps evaluate a person's ability to appropriately form blood clots. It measures the number of seconds it takes for a clot to form in a sample of blood after reagents are added | Keep | 6.67 | |
| 70 | cls_hh_aptt_t6 | APTT at 6 hours after hospital admission | Remove | 63.03 | Just a record of the APTT in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 71 | cls_hh_aptt_t30 | APTT at 30 hours after hospital admission | Remove | 52.73 | Just a record of the APTT in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 72 | cls_hh_aptt_t72 | APTT at 72 hours after hospital admission | Remove | 65.45 | Just a record of the APTT in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 73 | cls_hh_fib_t0 | Fibrinogen in Blood | Keep | 6.67 | |
| 74 | cls_hh_fib_t6 | Fibrinogen at 6 hours after hospital admission | Remove | 62.42 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 75 | cls_hh_fib_t30 | Fibrinogen at 30 hours after hospital admission | Remove | 52.12 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 76 | cls_hh_fib_t72 | Fibrinogen at 72 hours after hospital admission | Remove | 64.24 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 77 | cls_sh_ure_t0 | Urea in Blood | Keep | 2.42 | |

| | | | | | |
|---|---|---|---|---|---|
| 78 | cls_sh_ure_t6 | Urea at 6 hours after hospital admission | Remove | 62.42 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 79 | cls_sh_ure_t30 | Urea at 30 hours after hospital admission | Remove | 49.09 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 80 | cls_sh_ure_t72 | Urea at 72 hours after hospital admission | Remove | 54.55 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 81 | cls_sh_cre_t0 | Creatinine in Blood | Keep | 3.64 | |
| 82 | cls_sh_cre_t6 | Creatinine at 6 hours after hospital admission | Remove | 63.64 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 83 | cls_sh_cre_t30 | Creatinine at 30 hours after hospital admission | Remove | 49.09 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 84 | cls_sh_cre_t72 | Creatinine at 72 hours after hospital admission | Remove | 55.15 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 85 | cls_sh_glu_t0 | Glucose in Blood | Keep | 21.21 | |
| 86 | cls_sh_glu_t6 | Glucose at 6 hours after hospital admission | Remove | 79.39 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 87 | cls_sh_glu_t30 | Glucose at 30 hours after hospital admission | Remove | 78.79 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |

| | | | | | |
|---|---|---|---|---|---|
| 88 | cls_sh_glu_t72 | Glucose at 72 hours after hospital admission | Remove | 86.67 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 89 | cls_sh_bil_t0 | Bilirubin level in Blood | Keep | 10.91 | |
| 90 | CLS_S0 | Has no value | Remove | 100.00 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 91 | cls_sh_bil_t6 | Bilirubin at 6 hours after hospital admission | Remove | 70.91 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 92 | cls_sh_bil_t30 | Bilirubin at 30 hours after hospital admission | Remove | 60.61 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 93 | cls_sh_bil_t72 | Bilirubin at 72 hours after hospital admission | Remove | 63.03 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 94 | cls_sh_gan_t0 | AST - aspartate aminotransferase, ALT - alanine transaminase. High AST and ALT levels can indicate heart problems, liver damage or pancreatitis. Normal range of AST - 5 to 40 units per liter of serum, Normal range of ALT - 7 to 56 units per liter of serum | Keep | 15.76 | |
| 95 | cls_sh_gan_t6 | AST/ALT at 6 hours after hospital admission | Remove | 78.79 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 96 | cls_sh_gan_t30 | AST/ALT at 30 hours after hospital admission | Remove | 66.06 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 97 | cls_sh_ck_t0 | Creatinine Kinase in blood. normal range is 22 to 198 U/L (units per liter) | Keep | 39.39 | |

| | | | | | |
|---|---|---|---|---|---|
| 98 | cls_sh_chol_t0 | Cholesterol level in Blood | Keep | 17.58 | |
| 99 | cls_sh_chol_t6 | Cholesterol at 6 hours after hospital admission | Remove | 81.82 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 100 | cls_sh_chol_t30 | Cholesterol at 30 hours after hospital admission | Remove | 79.39 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 101 | cls_sh_chol_t72 | Cholesterol at 72 hours after hospital admission | Remove | 87.27 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 102 | cls_sh_tri_t0 | Triglycerid level in Blood. Normal range-less than 1.7 millimoles per liter (mmol/L), Borderline high - (1.8 to 2.2 mmol/L), High - (2.3 to 5.6 mmol/L) | Keep | 3.03 | |
| 103 | cls_sh_tri_t6 | Triglycerid at 6 hours after hospital admission | Remove | 59.39 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 104 | cls_sh_tri_t30 | Triglycerid at 30 hours after hospital admission | Remove | 56.97 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 105 | cls_sh_tri_t72 | Triglycerid at 72 hours after hospital admission | Remove | 74.55 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 106 | cls_sh_amy_t0 | Amylase level in Blood. Normal range is 40 to 140 units per liter (U/L) or 0.38 to 1.42 microkat/L (µkat/L) | Keep | 37.58 | |
| 107 | cls_sh_amy_t6 | Amylase at 6 hours after hospital admission | Remove | 92.73 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |

| | | | | | |
|---|---|---|---|---|---|
| 108 | cls_sh_amy_t30 | Amylase at 30 hours after hospital admission | Remove | 84.24 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 109 | cls_sh_lip_t0 | Lipase level in Blood. The normal range for adults younger than 60 is 10 to 140 U/L. Normal results for adults ages 60 and older is 24 to 151 U/L. Higher than normal levels of lipase mean problems with pancreas. If the blood has 3 to 10 times the normal level of lipase, then it's likely that you have acute pancreatitis | Keep | 47.88 | |
| 110 | cls_sh_lip_t30 | Lipase at 30 hours after hospital admission | Remove | 84.24 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 111 | cls_sh_pro_t0 | Total protein level in blood. A high total protein level could indicate dehydration or a certain type of cancer, such as multiple myeloma, that causes protein to accumulate abnormally. further tests will be needed to identify which proteins are too high or too low. The normal range is 60 to 83 g/L | Keep | 24.24 | |
| 112 | cls_sh_pro_t6 | Total protein at 6 hours after hospital admission | Remove | 87.27 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 113 | cls_sh_pro_t54 | Total protein at 30 hours after hospital admission | Remove | 81.21 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 114 | cls_sh_alb_t0 | Albumin level in blood. | Keep | 17.58 | |
| 115 | cls_sh_alb_t6 | Albumin at 6 hours after hospital admission | Remove | 83.64 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 116 | cls_sh_alb_t30 | Albumin at 30 hours after hospital admission | Remove | 83.03 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |

| | | | | | |
|---|---|---|---|---|---|
| 117 | cls_sh_na_t0 | Sodium level in blood | Keep | 4.85 | |
| 118 | cls_sh_na_t6 | Sodium at 6 hours after hospital admission | Remove | 62.42 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 119 | cls_sh_na_t30 | Sodium at 30 hours after hospital admission | Remove | 45.45 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 120 | cls_sh_ka_t0 | Potassium level in blood | Keep | 6.67 | |
| 121 | cls_sh_ka_t6 | Potassium at 6 hours after hospital admission | Remove | 61.21 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 122 | cls_sh_ka_t30 | Potassium at 30 hours after hospital admission | Remove | 44.24 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 123 | cls_sh_ka_tn6 | Potassium at 72 hours after hospital admission | Remove | 78.79 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 124 | cls_sh_ca_t0 | Calcium level in blood | Keep | 61.21 | |
| 125 | cls_km_ph_t0 | pH (in blood air). The acidity or alkalinity of blood | Keep | 6.06 | |
| 126 | cls_km_ph_t6 | pH at 6 hours after hospital admission | Remove | 68.48 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 127 | cls_km_ph_t30 | pH at 30 hours after hospital admission | Remove | 61.82 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 128 | cls_km_ph_t54 | pH at 54 hours after hospital admission | Remove | 74.55 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | be inappropriate if we keep |
| 129 | cls_km_paco2_t0 | paCo2 partial pressure of carbon dioxide in arterial blood | Keep | 6.67 | |
| 130 | cls_km_paco2_t6 | paCo2 at 6 hours after hospital admission | Remove | 68.48 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 131 | cls_km_paco2_t30 | paCo2 at 30 hours after hospital admission | Remove | 62.42 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 132 | cls_km_paco2_t54 | paCo2 at 54 hours after hospital admission | Remove | 74.55 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 133 | cls_km_paco2_t72 | paCo2 at 72 hours after hospital admission | Remove | 75.76 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 134 | cls_km_pao2_t0 | PaO2, is a measurement of oxygen pressure in arterial blood. It reflects how well oxygen is able to move from the lungs to the blood, and it is often altered by severe illnesse | Keep | 6.67 | |
| 135 | cls_km_pao2_t6 | pao2 at 6 hours after hospital admission | Remove | 67.88 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 136 | cls_km_pao2_t30 | pao2 at 30 hours after hospital admission | Remove | 62.42 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 137 | cls_km_pao2_t54 | pao2 at 54 hours after hospital admission | Remove | 74.55 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | be inappropriate if we keep |
| 138 | cls_km_pao2_t72 | pao2 at 72 hours after hospital admission | Remove | 75.76 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 139 | cls_km_hco3_t0 | Bicarbonate. Blood brings bicarbonate to your lungs, and then it is exhaled as carbon dioxide. Normal range 22 to 28 milliequivalents per liter (mEq/L) | Keep | 7.88 | |
| 140 | cls_km_hco3_t6 | subclinical examination - HCO3-(in blood air) measured at 6 hours after the hospital admission | Remove | 69.09 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 141 | cls_km_hco3_t30 | subclinical examination - HCO3-(in blood air) measured at 30 hours after the hospital admission | Remove | 61.82 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 142 | cls_km_hco3_t54 | subclinical examination - HCO3-(in blood air) measured at 54 hours after the hospital admission | Remove | 74.55 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 143 | cls_km_hco3_t72 | subclinical examination - HCO3-(in blood air) measured at 72 hours after the hospital admission | Remove | 75.76 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 144 | cls_km_be_t0 | Base Excess and base deficit refer to an excess or deficit, respectively, in the amount of base present in the blood. used to determine the pH of the blood, or how acidic it i | Keep | 10.30 | |
| 145 | cls_km_be_t6 | Base Excess at 6 hours after admission to the hospital | Remove | 69.70 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 146 | cls_km_be_t30 | Base Excess at 30 hours after admission to the hospital | Remove | 63.64 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | analysis.Imputation will be inappropriate if we keep |
| 147 | cls_km_be_t54 | Base Excess at 54 hours after admission to the hospital | Remove | 75.15 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 148 | cls_km_be_t72 | Base Excess at 72 hours after admission to the hospital | Remove | 77.58 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 149 | cls_km_pf_t0 | PaO2 measurement shows the oxygen pressure in the blood. PaO2 normal range is 80–100 mmHg. This variable is calculated by Pao2/Oxygen percentage | Keep | 40.61 | |
| 150 | cls_km_pf_t6 | P/F at 6 hours after hospital admission | Remove | 84.24 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 151 | cls_km_pf_t30 | P/F at 30 hours after hospital admission | Remove | 75.76 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 152 | cls_km_pf_t54 | P/F at 54 hours after hospital admission | Remove | 80.00 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 153 | cls_km_pf_t72 | P/F at 72 hours after hospital admission | Remove | 84.85 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 154 | cls_km_lac_t0 | Normal blood lactate level is 0.5-1 mmol/L. A high lactate level in the blood means that the disease or condition a person has is causing lactate to accumulate. Greater increase in lactate means a greater severity of the condition. When associated with lack of oxygen, an increase in lactate can indicate that organs are not functioning properly | Keep | 10.91 | |

| | | | | | |
|---|---|---|---|---|---|
| 155 | cls_km_lac_t6 | Lactate level in blood measured at 6 hours after hospital admission | Remove | 70.91 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 156 | cls_km_lac_t30 | Lactate level in blood measured at 30 hours after hospital admission | Remove | 66.06 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 157 | cls_km_lac_t54 | Lactate level in blood measured at 54 hours after hospital admission | Remove | 74.55 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 158 | cls_km_lac_t72 | Lactate level in blood measured at 72 hours after hospital admission | Remove | 76.97 | Just a record of this parameter in the patient at 6,30,54 and 72 hours. Unimportant in the analysis.Imputation will be inappropriate if we keep |
| 159 | dt_dich_vao_t24 | A record of treatment fluid when the patient is in hospital at 24 hours from admisison | Keep | 10.30 | |
| 160 | dt_dich_vao_t48 | A record of treatment fluid when the patient is in hospital at 48 hours from admisison | Keep | 13.94 | |
| 161 | dt_dich_vao_t72 | A record of treatment fluid when the patient is in hospital at 72 hours from admisison | Keep | 23.64 | |
| 162 | dt_dich_ra_t24 | A record of treatment fluid when the patient is in hospital at 24 hours from admisison | Keep | 9.09 | |
| 163 | dt_dich_ra_t48 | A record of treatment fluid when the patient is in hospital at 48 hours from admisison | Keep | 12.73 | |
| 164 | dt_dich_ra_t72 | A record of treatment fluid when the patient is in hospital at 72 hours from admisison | Keep | 23.64 | |
| 165 | dt_dich_biln_t24 | Fluid in and out record at 24 hours from hospital admission | Keep | 10.91 | |
| 166 | dt_dich_biln_t48 | Fluid in and out record at 48 hours from hospital admission | Keep | 15.76 | |
| 167 | dt_dich_biln_t72 | Fluid in and out record at 72 hours from hospital admission | Keep | 26.06 | |
| 168 | dt_nhin_ngay | Fasting treatment to the patient. No of days a patient was on fasting treatment that is without food. | Keep | 6.06 | |
| 169 | dt_pex_ngaybenh | Probably after how many days after a patient was referred for PEX treatment. | Keep | 50.91 | |
| 170 | dt_pex_lan | Number of PEX treatments a patient received | Keep | 49.70 | |
| 171 | dt_pex_sauvv | The variable 169 means number of days after a patient was referred for PEX. This variable could probably mean - the time in hours in | Remove | 49.09 | Above variable also tells us after how many days the pex treatment was |

| | | | | | |
|---|---|---|---|---|---|
| | | addition to the number of days, a patient was referred for PEX treatment. | | | started. This variable is redundant |
| 172 | DT_PE0 | Empty variable | Remove | 100.00 | Empty variable |
| 173 | dt_pex_tri_t_lan1 | triglycerid before first time of PEX | Keep | 50.30 | |
| 174 | dt_pex_tri_s_lan1 | triglycerid after first time of PEX | Keep | 51.52 | |
| 175 | dt_pex_chol_t_lan1 | cholesterol before first time of PEX | Remove | 56.36 | Missing percentage is more. Cannot impute appropriately and make sense of the values in this variable. Also, the record of this variable before and after PEX treatment may not significantly impact the model |
| 176 | dt_pex_chol_s_lan1 | cholesterol after first time PEX | Remove | 72.73 | Missing percentage is more. Cannot impute appropriately and make sense of the values in this variable. Also, the record of this variable before and after PEX treatment may not significantly impact the model |
| 177 | dt_pex_ldl_t_lan1 | LDL before first time of PEX | Remove | 73.33 | Missing percentage is more. Cannot impute appropriately and make sense of the values in this variable. Also, the record of this variable before and after PEX treatment may not significantly impact the model |
| 178 | dt_pex_ldl_s_lan1 | LDL after first time of PEX | Remove | 90.30 | Missing percentage is more. Cannot impute appropriately and make sense of the values in this variable. Also, the record of this variable before and after PEX treatment may not significantly impact the model |
| 179 | dt_pex_hdl_t_lan1 | HDL - before first time PEX | Remove | 72.73 | Missing percentage is more. Cannot impute appropriately and make sense of the values in this variable. Also, the record of this variable before and after PEX treatment may not significantly impact the model |
| 180 | dt_pex_apache_t_lan1 | APAche 2 score before first time PEX | Keep | 50.30 | |
| 181 | dt_pex_apache_s_lan1 | APAche 2 score after first time PEX | Keep | 51.52 | |

| | | | | | |
|---|---|---|---|---|---|
| 182 | dt_pex_ranson_t_lan1 | ranson score before first time PEX | Remove | 49.70 | High missing values in the record of this score after the first time treatment. Therefore removing both before and after record |
| 183 | dt_pex_ranson_s_lan1 | ranson score after first time PEX | Remove | 98.79 | High missing values in the record of this score after the first time treatment. Therefore removing both before and after record |
| 184 | dt_pex_imrie_t_lan1 | Imre score before first time of PEX | Keep | 50.30 | |
| 185 | dt_pex_imrie_s_lan1 | Imre score after first time of PEX | Keep | 52.12 | |
| 186 | dt_pex_balthazar_t_lan1 | balthazar score (with computer tomography) before first time PEX | Remove | 60.61 | High missing values in the record of this score after the first time treatment. Therefore removing both before and after record |
| 187 | dt_pex_balthazar_s_lan1 | balthazar score (with computer tomography) after first time PEX | Remove | 95.76 | High missing values in the record of this score after the first time treatment. Therefore removing both before and after record |
| 188 | dt_pex_sofa_t_lan1 | sofa score before first time of PEX | Remove | 49.70 | High missing values in the record of this score after the first time treatment. Therefore removing both before and after record |
| 189 | dt_pex_sofa_s_lan1 | sofa score after first time of PEX | Remove | 53.94 | High missing values in the record of this score after the first time treatment. Therefore removing both before and after record |
| 190 | dt_pex_alob_t_lan1 | Abdominal pressure before first time of PEX | Remove | 66.06 | High missing values in the record of this score after the first time treatment. Therefore removing both before and after record |
| 191 | dt_pex_alob_s_lan1 | Abdominal pressure after first time of PEX | Remove | 72.73 | High missing values in the record of this score after the first time treatment. Therefore removing both before and after record |
| 192 | kq | Weather a patient is Dead or Alive | Keep | 24.24 | |
| 193 | bcxa | Often, some complications can be reported in patients subjected to PEX treatments. Weather any such complications are seen in the patient or not | Keep | 52.12 | |

| | | | | | |
|---|---|---|---|---|---|
| 194 | pex | Was the patient subjected to PEX treatment or not. If not, then the patient was subjected to the conventional treatment. | Keep | 1.82 | |

**Accuracy Check of Variables:**

Based on statistical requirements for missing values and domain knowledge of the variables, a total of 79 variables are found to be important for the analysis of PEX treatment for Acute Pancreatitis. A subset was created for only these important variables and accuracy check was performed in SAS. Out of these 79 variables, 17 variables had visible accuracy issues and these variables would require appropriate treatment of the inaccuracy during the imputation step. In addition to the below mentioned issues, the categorical variables should also be encoded to numerical values, as well as the units of variables must be verified during modelling.
The descriptions of the accuracy issues can be found from the below table:

**Table 2: Study of PEX treatment for Acute Pancreatitis - Accuracy Issues in the variables**

| Var No. | Name | Description | Meaning of the Variable | Status | Accuracy Problem |
|---|---|---|---|---|---|
| 1 | Age | Patient's age | Patient's age | Keep | |
| 2 | Gender | Patient's gender | Patient's gender | Keep | |
| 3 | vv_reason_1 | Main reason to admit hospital | Major reason of the patient to visit the hospital | Keep | |
| 4 | rv_ngaydt | Duration of staying in hospitals | No of days a patient stayed in the hospital. Min =1, Max = 18 in this dataset | Keep | |
| 5 | ts_giadinh | Hereditary information | some patients could get AP as hereditary. By this meaning, the variable could mean that is the AP of the patient is hereditary or not. | Keep | |
| 6 | ts_benhmat | Gallbladder problem | Gall stones causes about 40% of the cases of AP. These gallstones pass in the duct connected to pancreas and causes severe pain leading to AP. The variable means that weather the current AP is due to this cause or not | Keep | |
| 7 | ts_ruou | Drinking problem | Alcohol causes about 30% of the cases of AP. The risk of developing AP increases with increasing amounts of alcohol (4 to 7 drinks/day in men and 3 or more drinks/day in women) However less than 10% people who drink alcohol are reported to develop AP. The variable means weather alcohol is the cause for AP in the patient. | Keep | |

| | | | | | |
|---|---|---|---|---|---|
| 8 | ts_ruou_nam | A breakdown of drinking problem | The whole number values like 8,10,20 probably indicates how many drinks the patient has consumed in a day | Keep | |
| 9 | ts_dtd | Diabetes problem | one meaning is severe AP can cause improper development of insulin in a patient and in severer cases of AP, a patient can develop diabetes due to the existing AP. Second meaning is a patient who admitted in the hospital with AP has a previous diabetes problem or not | Keep | 3' value present for 1 observation. Should replace this to Nan and Impute appropriately |
| 10 | ts_vtc | Historical cholecystitis problem | cholesystitis means swelling of gallbladder due to the gallstones blocking the tube, or due to bile juice getting trapped in the gallbladder. The variable means weather the patient has previously reported to cholesystis or not | Keep | 3' and '5' values present for 2 observations. Should replace this to Nan and Impute appropriately |
| 11 | non | Vomitting | Vomitting or nausea is another most common symptom of AP. The variable indicated weather the person is reporting to being nauseous or vomitting or not | Keep | |
| 12 | ls_cn_bidaitien | Clinical symptoms of defecation | Pale yellow, greasy, foul-smelling stool: malabsorption of fat due to pancreatic insufficiency can be seen in patients with AP. Variable indicates weather these symptoms present in the patient or not | Keep | some values have 't0' and some has 'T0' both means the same (time at point of admisison/ 0th Hour). Some values are 't30' and T30 |
| 13 | ls_cn_ialong | Clinical symptoms of Diarrhea | Both acute and chronic pancreatitis can cause the pancreas to produce fewer enzymes that are needed to process nutrients causing diarrhea. Variable indicates weather patient has diarrhea symptom or not | Keep | values are same but in upper and lower case. Ex: t0 and T0. Should be altered for uniformity |
| 14 | ls_tht_bungchuong | Clinical symptoms of Abdominal distension | Severe abdominal pain is a predominant symptom. A person may also develop some swelling in the upper abdomen. T0 means time at which patient was admitted. | Keep | values are same but in upper and lower case. Ex: t0 and T0. Should be altered for uniformity |
| 15 | ls_tt_alob_t0 | Clinical symptoms | Using the Atlanta criteria, acute pancreatitis is diagnosed when a patient presents with two of three findings, including abdominal pain suggestive of pancreatitis, serum amylase and/or lipase levels at least three times the normal level. Based on the values, this variable should be Lipase level in Blood. normal range <67 U/l | Keep | |
| 16 | ls_tt_bmi_t0 | Clinical symptoms | Should be amylase level aka serum amylase in blood. Normal range 22-80 U/l | Keep | |
| 17 | ls_tn_mach_t0 | Clinical symptoms | Should be systolic blood pressure | Keep | |

| | | | | | |
|---|---|---|---|---|---|
| 18 | ls_tn_nhiet_t0 | Body temperature | Body temperature | Keep | Min & max values are 3.7 and 366 respectively. Unrealistic values found in this variable |
| 19 | ls_tn_ha_t6 | Blood pressure | Blood pressure | Keep | values seperated by a /. Should create two variables - systolic pressure and diastolic pressure |
| 20 | ls_tn_spo2_t0 | Saturation of peripheral oxygen | Oxygen saturation in patient. Denoted by Spo2. lesser oxygen saturation could be a measure of damaged lungs | Keep | |
| 21 | ls_diem_apache_t0 | apache 2 score at the points of admitting hospitals | APACHE II is a severity-of-disease scoring systems. Applied within 24 hours of admission of a patient to an ICU. Range = 0 to 71. Higher scores correspond to more severe disease and a higher risk of death | Keep | |
| 22 | ls_diem_ranson_t0 | ranson score at the points of admitting hospitals | Ranson criteria isused to predict the severity and mortality of acute pancreatitis. Five parameters are assessed on admission, and the other six are assessed at 48 hours post-admission. One point is given for each positive parameter for a maximum score of 11. The modified criteria have a max score of 10. Five parameters are assessed on admission and the other 5 at the 48-hour mark | Keep | |
| 23 | ls_diem_ct_t0 | CTSI score at the points of admitting hospitals | CT severity index sums two scores, Balthazar score and grading the extent of pancreatic necrosis. Grading of balthazar - 0 to 4, pancreatic necrosis values are 0,2,4,6. Max ct score can be 10 | Keep | |
| 24 | ls_diem_imrie_t0 | imre score at the points of admitting hospitals | Glasgow imrie score is a modification of Ranson's criteria for AP. Measured at 48 hours after hospital admission. Max score is 8 | Keep | |
| 25 | ls_diem_sofa_t0 | sofa score at the points of admitting hospitals | The Sequential Organ Failure Assessment (SOFA) score is a simple and objective score that allows for calculation of both the number and the severity of organ dysfunction in six organ systems (respiratory, coagulatory, liver, cardiovascular, renal, and neurologic). ranges from 0 to 24 | Keep | |

| | | | | | |
|---|---|---|---|---|---|
| 26 | cls_sa_dichob_t0 | subclinical examination - (Abdominal fluid) ultrasound at the points of admitting hospitals | Ultrasound finding - Weather patient had abdominal fluid. The pancreas secretes pancreatic fluid into the first part of the small intestine (duodenum). This contains digestive enzymes that help digest food. If a gallstone becomes stuck (where the pancreatic duct empties into the duodenum), the fluid stops flowing. Usually, the blockage is temporary and causes limited damage, which is soon repaired. But if the blockage remains, the enzymes collect in the pancreas and begin to digest the cells of the pancreas, causing severe inflammation | Keep | |
| 27 | cls_ct_dichob_lan1 | subclinical examination - (Abdominal fluid) computer tomography | Weather or not abdominal fluid is present, during the computer tomography | Keep | Has values other than 'have' and 'No'. Other values should be altered |
| 28 | cls_ct_balthazar_lan1 | subclinical examination - balthazar score (with computer tomography) | Balthazar score calculated from computer tomography. Grading from 0,1,2,3,4 also represented by A,B,C,D,E. | Keep | Upper and lower case values for the same value. Ex: a and A both mean same |
| 29 | cls_ct_ctscore_lan1 | subclinical examination - CTSI score (with computer tomography) | CTSI calculated from computer tomography sums two scores, Balthazar score and grading the extent of pancreatic necrosis. Grading of balthazar - 0 to 4, pancreatic necrosis values are 0,2,4,6. Max ct score can be 10 | Keep | value 'e' present for 1 observation. Must be altered |
| 30 | cls_hh_bc_t0 | subclinical examination - white blood cell; t0: at the points of admitting hospitals, t6: after 6h of admitting hospitals... | WBC level in the blood at the point of admisison to the hospital | Keep | |
| 31 | cls_hh_tc_t0 | subclinical examination - | TC count or total cell count of WBCs. A high white blood cell count may indicate that the immune system is working to destroy an infection | Keep | |
| 32 | cls_hh_hct_t0 | subclinical examination - Hematocrit | Hematocrit is the percentage by volume of red blood cells in the blood. Normal levels for men is 41% to 50%. For women is 36% to 48% | Keep | |
| 33 | cls_hh_hc_t0 | red blood cell | RBC level in the blood. Normal range for men – 4.7 to 6.1 million cells per microlitre (cells/mcL), women – 4.2 to 5.4 million cells/mcL | Keep | |
| 34 | cls_hh_pt_t0 | prothrombin | Prothrombin is a protein made by the liver. A substance known as clotting (coagulation) factors. | Keep | |

| | | | | | |
|---|---|---|---|---|---|
| 35 | cls_hh_aptt _t0 | APTT | aPTT is a screening test that helps evaluate a person's ability to appropriately form blood clots. It measures the number of seconds it takes for a clot to form in a sample of blood after reagents are added | Keep | |
| 36 | cls_hh_fib_t 0 | subclinical examination - Fibrinogen | Fibrinogen in Blood | Keep | |
| 37 | cls_sh_ure_t 0 | subclinical examination - ure | Urea in Blood | Keep | |
| 38 | cls_sh_cre_t 0 | subclinical examination - creatinin | Creatinine in Blood | Keep | |
| 39 | cls_sh_glu_t 0 | subclinical examination - glucose | Glucose in Blood | Keep | |
| 40 | cls_sh_bil_t 0 | subclinical examination - bilirubin total | Bilirubin level in Blood | Keep | values seperated by a /. Should create two variables for individual values. Also has 1 value as 'duc/duc' |
| 41 | cls_sh_gan_ t0 | AST, ALT (liver funtion) | AST - aspartate aminotransferase, ALT - alanine transaminase. High AST and ALT levels can indicate heart problems, liver damage or pancreatitis. Normal range of AST - 5 to 40 units per liter of serum, Normal range of ALT - 7 to 56 units per liter of serum | Keep | values seperated by a /. Should create two variables for individual values of AST and ALT.' |
| 42 | cls_sh_ck_t 0 | subclinical examination - | Creatinine Kinase in blood. normal range is 22 to 198 U/L (units per liter) | Keep | |
| 43 | cls_sh_chol _t0 | cholesterol | Cholesterol level in Blood | Keep | Min value is 3.9 too low for this variable. Need to check |
| 44 | cls_sh_tri_t 0 | triglycerid | Triglycerid level in Blood. Normal range- less than 1.7 millimoles per liter (mmol/L), Borderline high -(1.8 to 2.2 mmol/L), High -(2.3 to 5.6 mmol/L) | Keep | |
| 45 | cls_sh_amy _t0 | subclinical examination - amylase | Amylase level in Blood. Normal range is 40 to 140 units per liter (U/L) or 0.38 to 1.42 microkat/L (µkat/L) | Keep | |
| 46 | cls_sh_lip_t 0 | subclinical examination - lipase | Lipase level in Blood. The normal range for adults younger than 60 is 10 to 140 U/L. Normal results for adults ages 60 and older is 24 to 151 U/L. Higher than normal levels of lipase mean problems with pancreas. If the blood has 3 to 10 times the normal level of lipase, then it's likely that you have acute pancreatitis | Keep | |

| | | | | | |
|---|---|---|---|---|---|
| 47 | cls_sh_pro_t0 | subclinical examination - protein | Total protein level in blood. A high total protein level could indicate dehydration or a certain type of cancer, such as multiple myeloma, that causes protein to accumulate abnormally. further tests will be needed to identify which proteins are too high or too low. The normal range is 60 to 83 g/L | Keep | |
| 48 | cls_sh_alb_t0 | subclinical examination - albumin | Albumin level in blood. | Keep | |
| 49 | cls_sh_na_t0 | subclinical examination – natri (sodium) | Sodium level in blood | Keep | |
| 50 | cls_sh_ka_t0 | subclinical examination - potasium | Potassium level in blood | Keep | |
| 51 | cls_sh_ca_t0 | subclinical examination - calci total | Calcium level in blood | Keep | values seperated by a /. Some values are decimal |
| 52 | cls_km_ph_t0 | subclinical examination - pH (in blood air) | pH (in blood air). The acidity or alkalinity of blood | Keep | |
| 53 | cls_km_paco2_t0 | subclinical examination - paCo2(in blood air) | paCo2 partial pressure of carbon dioxide in arterial blood | Keep | |
| 54 | cls_km_pao2_t0 | subclinical examination - pa Oxy (in blood air) | PaO2, is a measurement of oxygen pressure in arterial blood. It reflects how well oxygen is able to move from the lungs to the blood, and it is often altered by severe illnesse | Keep | |
| 55 | cls_km_hco3_t0 | subclinical examination - HCO3-(in blood air) | Bicarbonate. Blood brings bicarbonate to your lungs, and then it is exhaled as carbon dioxide. Normal range 22 to 28 milliequivalents per liter (mEq/L) | Keep | negative values found in the variable. Cannot be negative |
| 56 | cls_km_be_t0 | BE (in blood air) | Base Excess and base deficit refer to an excess or deficit, respectively, in the amount of base present in the blood. used to determine the pH of the blood, or how acidic it i | Keep | negative values found in the variable. Cannot be negative |
| 57 | cls_km_pf_t0 | p/f (paO2/%O2) | PaO2 measurement shows the oxygen pressure in the blood. PaO2 normal range is 80–100 mmHg. This variable is calculated by Pao2/Oxygen percentage | Keep | |
| 58 | cls_km_lac_t0 | lactatr (in blood air) | Normal blood lactate level is 0.5-1 mmol/L. A high lactate level in the blood means that the disease or condition a person has is causing lactate to accumulate. Greater increase in lactate means a greater severity of the condition. When associated with lack of oxygen, an increase in lactate can indicate that organs are not functioning properly | Keep | |
| 59 | dt_dich_vao_t24 | treatment - fluide intake | A record of treatment fluid when the patient is in hospital at 24 hours from admisison | Keep | |

| | | | | | |
|---|---|---|---|---|---|
| 60 | dt_dich_vao_t48 | treatment - fluide intake | A record of treatment fluid when the patient is in hospital at 48 hours from admisison | Keep | |
| 61 | dt_dich_vao_t72 | treatment - fluide intake | A record of treatment fluid when the patient is in hospital at 72 hours from admisison | Keep | |
| 62 | dt_dich_ra_t24 | treatment - fluide output | A record of treatment fluid when the patient is in hospital at 24 hours from admisison | Keep | |
| 63 | dt_dich_ra_t48 | treatment - fluide output | A record of treatment fluid when the patient is in hospital at 48 hours from admisison | Keep | |
| 64 | dt_dich_ra_t72 | treatment - fluide output | A record of treatment fluid when the patient is in hospital at 72 hours from admisison | Keep | |
| 65 | dt_dich_biland_t24 | treatment - balance fluid in and out | Fluid in and out record at 24 hours from hospital admission | Keep | |
| 66 | dt_dich_biland_t48 | treatment - balance fluid in and out | Fluid in and out record at 48 hours from hospital admission | Keep | |
| 67 | dt_dich_biland_t72 | treatment - balance fluid in and out | Fluid in and out record at 72 hours from hospital admission | Keep | |
| 68 | dt_nhin_ngay | treatment - day without food intake | Fasting treatment to the patient. No of days a patient was on fasting treatment that is without food. | Keep | |
| 69 | dt_pex_ngaybenh | treatment - PEX treatment of which day of the diagnosis | Probably after how many days after a patient was referred for PEX treatment. | Keep | |
| 70 | dt_pex_lan | treatment - number of PEX treatment | Number of PEX treatments a patient received | Keep | |
| 71 | dt_pex_tri_t_lan1 | treatment - triglycerid before first time of PEX | triglycerid before first time of PEX | Keep | |
| 72 | dt_pex_tri_s_lan1 | treatment - triglycerid after first time of PEX | triglycerid after first time of PEX | Keep | |
| 73 | dt_pex_apache_t_lan1 | treatment - APAche 2 score before first time PEX | APAche 2 score before first time PEX | Keep | |
| 74 | dt_pex_apache_s_lan1 | treatment - APAche 2 score after first time PEX | APAche 2 score after first time PEX | Keep | |
| 75 | dt_pex_imrie_t_lan1 | treatment - Imre score before first time of PEX | Imre score before first time of PEX | Keep | |
| 76 | dt_pex_imrie_s_lan1 | treatment - Imre score after first time of PEX | Imre score after first time of PEX | Keep | |
| 77 | kq | Result - dead or alive | Weather a patient is Dead or Alive | Keep | one value is Alive another is 0. Must change for uniformity |
| 78 | bcxa | Potential complication | Often times, some complications can be reported in patients subjected to PEX treatments. Weather any such complications are seen in the patient or not | Keep | |

| | | | | | |
|---|---|---|---|---|---|
| 79 | pex | Patient with PEX or without PEX | Was the patient subjected to PEX treatment or not. If not, then the patient was subjected to the conventional treatment. | Keep | |

***Additional accuracy issues if found during the data cleaning will be reported and dealt with later.**

Extra: Screenshot of SAS output to determine out of range scores.

| Label | N | N Miss | Mean | Median | Minimum | Maximum | Range |
|---|---|---|---|---|---|---|---|
| Age | 165 | 0 | 41.1090909 | 41.0000000 | 21.0000000 | 77.0000000 | 56.0000000 |
| Duration of staying in hospitals | 165 | 0 | 6.7333333 | 6.0000000 | 1.0000000 | 18.0000000 | 17.0000000 |
| A breakdown of drinking problem...13 | 77 | 88 | 10.9610390 | 10.0000000 | 0 | 30.0000000 | 30.0000000 |
| Vomitting | 88 | 77 | 0.7045455 | 1.0000000 | 0 | 1.0000000 | 1.0000000 |
| Clinical symptoms...24 | 97 | 68 | 19.1443299 | 18.0000000 | 2.0000000 | 46.0000000 | 44.0000000 |
| Clinical symptoms...25 | 162 | 3 | 24.1122222 | 23.0950000 | 15.6300000 | 258.0000000 | 242.3700000 |
| Clinical symptoms...26 | 160 | 5 | 107.3250000 | 107.5000000 | 68.0000000 | 158.0000000 | 90.0000000 |
| Body temperature | 158 | 7 | 39.0132911 | 37.0000000 | 3.7000000 | 366.0000000 | 362.3000000 |
| Saturation of peripheral oxygen | 159 | 6 | 97.0943396 | 98.0000000 | 90.0000000 | 100.0000000 | 10.0000000 |
| apache 2 score at the points of admitting hospitals | 140 | 25 | 3.4714286 | 3.0000000 | 0 | 16.0000000 | 16.0000000 |
| ranson score at the points of admitting hospitals | 140 | 25 | 1.4928571 | 2.0000000 | 0 | 5.0000000 | 5.0000000 |
| CTSI score at the points of admitting hospitals | 113 | 52 | 3.7345133 | 4.0000000 | 0 | 10.0000000 | 10.0000000 |
| imre score at the points of admitting hospitals | 141 | 24 | 1.3758865 | 1.0000000 | 0 | 4.0000000 | 4.0000000 |
| sofa score at the points of admitting hospitals | 140 | 25 | 1.4714286 | 1.0000000 | 0 | 7.0000000 | 7.0000000 |
| subclinical examination - white blood cell; t0: at the points of admitting hospitals, t6: after 6h of admitting hospitals... | 160 | 5 | 10.5843500 | 10.2150000 | 1.6700000 | 22.5900000 | 20.9200000 |
| subclinical examination -...51 | 161 | 4 | 193.2807453 | 177.0000000 | 14.6000000 | 422.0000000 | 407.4000000 |
| subclinical examination - Hematocrit | 161 | 4 | 0.4680745 | 0.3900000 | 0.2260000 | 10.4300000 | 10.2040000 |
| red blood cell | 146 | 19 | 4.4743082 | 4.4600000 | 2.7000000 | 6.8800000 | 4.1800000 |
| prothrombin | 155 | 10 | 95.1541935 | 87.0000000 | 36.0000000 | 879.0000000 | 843.0000000 |
| APTT | 154 | 11 | 21.9937662 | 1.0550000 | 0.7600000 | 3212.00 | 3211.24 |
| subclinical examination - Fibrinogen | 154 | 11 | 5.8643052 | 5.5495000 | 1.2540000 | 45.0000000 | 43.7460000 |
| subclinical examination - ure | 161 | 4 | 6.4987578 | 3.8000000 | 1.2000000 | 193.0000000 | 191.8000000 |
| subclinical examination - creatinin | 159 | 6 | 83.6067925 | 68.0000000 | 1.8000000 | 727.0000000 | 725.2000000 |
| subclinical examination - glucose | 130 | 35 | 11.2892308 | 9.6750000 | 3.2000000 | 66.0000000 | 62.8000000 |
| subclinical examination -...97 | 100 | 65 | 435.6608000 | 270.0000000 | 9.7800000 | 3546.00 | 3536.22 |
| cholesterol | 136 | 29 | 15.3570588 | 13.1850000 | 3.9100000 | 99.0000000 | 95.0900000 |
| triglycerid | 160 | 5 | 27.4805000 | 18.5250000 | 11.2100000 | 131.5500000 | 120.3400000 |
| subclinical examination - amylase | 103 | 62 | 414.3060194 | 320.0000000 | 6.6700000 | 1519.80 | 1513.13 |
| subclinical examination - lipase | 86 | 79 | 513.6627907 | 414.0000000 | 6.0000000 | 1728.10 | 1722.10 |
| subclinical examination - protein | 125 | 40 | 59.1856000 | 59.8000000 | 32.1000000 | 84.3000000 | 52.2000000 |
| subclinical examination - albumin | 136 | 29 | 30.9713235 | 30.4000000 | 13.6000000 | 56.2000000 | 42.6000000 |
| subclinical examination - sodium | 157 | 8 | 131.0458599 | 132.0000000 | 4.2000000 | 154.0000000 | 149.8000000 |
| subclinical examination - potasium | 154 | 11 | 3.6787013 | 3.6000000 | 2.6000000 | 5.3000000 | 2.7000000 |
| subclinical examination - pH (in blood air) | 155 | 10 | 12.1230452 | 7.3900000 | 7.1000000 | 741.0000000 | 733.9000000 |
| subclinical examination - paco2(in blood air) | 154 | 11 | 31.8850649 | 31.7500000 | 9.0000000 | 97.0000000 | 88.0000000 |
| subclinical examination - pa Oxy (in blood air) | 154 | 11 | 92.3012987 | 88.3500000 | 32.0000000 | 251.0000000 | 219.0000000 |
| subclinical examination - HCO3-(in blood air) | 152 | 13 | 30.6890132 | 19.7500000 | -18.6000000 | 1708.00 | 1726.60 |
| BE (in blood air) | 148 | 17 | -4.8736486 | -4.5500000 | -24.7000000 | 16.0000000 | 40.7000000 |
| p/f (paO2/%O2) | 98 | 67 | 334.8857143 | 341.4000000 | 3.8000000 | 562.0000000 | 558.2000000 |
| lactatr (in blood air) | 147 | 18 | 2.1251020 | 1.6000000 | 0.4000000 | 9.0000000 | 8.6000000 |
| treatment - fluide intake...159 | 148 | 17 | 4617.16 | 4275.00 | 60.0000000 | 9650.00 | 9590.00 |
| treatment - fluide intake...160 | 142 | 23 | 4234.79 | 4000.00 | 1000.00 | 9500.00 | 8500.00 |
| treatment - fluide intake...161 | 126 | 39 | 3758.35 | 3500.00 | 1200.00 | 8500.00 | 7300.00 |
| treatment - fluide output...162 | 150 | 15 | 2572.93 | 2360.00 | 950.0000000 | 6900.00 | 5950.00 |
| treatment - fluide output...163 | 144 | 21 | 3154.31 | 2650.00 | 620.0000000 | 10760.00 | 10140.00 |
| treatment - fluide output...164 | 126 | 39 | 3035.75 | 2605.00 | 270.0000000 | 8020.00 | 7750.00 |
| treatment - balance fluid in and out...165 | 147 | 18 | 2329.12 | 1980.00 | -2100.00 | 23200.00 | 25300.00 |
| treatment - balance fluid in and out...166 | 139 | 26 | 1142.73 | 1120.00 | -4550.00 | 6830.00 | 11380.00 |
| treatment - balance fluid in and out...167 | 122 | 43 | 745.3032787 | 820.0000000 | -3780.00 | 2650.00 | 6430.00 |
| treatment - day without food intake | 155 | 10 | 1.6258065 | 1.0000000 | 0 | 12.0000000 | 12.0000000 |
| treatment - PEX treatment of which day of the diagnosis | 81 | 84 | 2.6419753 | 3.0000000 | 1.0000000 | 7.0000000 | 6.0000000 |
| treatment - number of PEX treatment | 83 | 82 | 1.1566265 | 1.0000000 | 1.0000000 | 3.0000000 | 2.0000000 |
| treatment - triglycerid before first time of PEX | 82 | 83 | 31.9035366 | 22.6800000 | 2.4100000 | 131.5500000 | 129.1400000 |
| treatment - triglycerid after first time of PEX | 80 | 85 | 9.7173750 | 6.2250000 | 1.0100000 | 76.9400000 | 75.9300000 |
| treatment - APAche 2 sHavere before first time PEX | 82 | 83 | 3.9709756 | 4.0000000 | 0 | 16.0000000 | 16.0000000 |
| treatment - APAche 2 sHavere after first time PEX | 80 | 85 | 1.9625000 | 2.0000000 | 0 | 9.0000000 | 9.0000000 |
| treatment - Imre sHavere before first time of PEX | 82 | 83 | 1.5000000 | 2.0000000 | 0 | 4.0000000 | 4.0000000 |
| treatment - Imre sHavere after first time of PEX | 79 | 86 | 0.7468354 | 1.0000000 | 0 | 3.0000000 | 3.0000000 |
| Potential Complication | 79 | 86 | 1.0000000 | 1.0000000 | 1.0000000 | 1.0000000 | 0 |
| Patient with PEX or without PEX | 162 | 3 | 0.5000000 | 0.5000000 | 0 | 1.0000000 | 1.0000000 |

**Selection of Features:**

A subset of 79 variables is selected to continue the analysis further. The choice of keeping these variables was on the basis that they describe key characteristics of the patient condition suffering with AP and would also play a significant role in studying the effect of PEX treatment and determining a difference in the two treatments.

The variables removed are on the basis that they had a large number of missing values and importance of the variable in determining the response. Also, many subclinical observations like – values at $24^{th}$ hour, $30^{th}$ hour, $54^{th}$ hour and $72^{nd}$ hour are found to be less important as they do not play an important role in determining our response variables. For such subclinical observations, only the T0 value – that is the value at the point of hospital admission has been retained in the dataset.

Moreover, these variables had many missing values – which indicate that these observations were not recorded for the patient. If these variables were retained, the imputation of the missing values could not be performed appropriately. The nature of the missing values will be discussed in the $5^{th}$ question of this report.

The 79 variables that are retained are found to have a meaningful impact on the study and the response, and the missing values in the retained variables can be imputed appropriately during further analysis. Certain variables like the ones starting with 'treatment' were retained as they had very a smaller number of missing values. The unimportant ones among these, will be removed based on statistical analysis, feature selection techniques or can be reduced with Principal component analysis.

**Note:** More details and reasons on the removal of each variable can be found in 'Reason to Remove' column in **Table 1: Study of PEX treatment for Acute Pancreatitis – Meaning of Each Variable.**

The list of the 79 Retained variables is as following:

[1] "Age"
[2] "Gender"
[3] "Main reason to admit hospital"
[4] "Duration of staying in hospitals"
[5] "Hereditary information"
[6] "Gallbladder problem"
[7] "Drinking problem"
[8] "A breakdown of drinking problem...13"
[9] "Diabetes problem"
[10] "Historical cholecystitis problem"
[11] "Vomitting"
[12] "Clinical symptoms of defecation"

[13] "Clinical symptoms of Diarrhea"

[14] "Clinical symptoms of Abdominal distension"

[15] "Clinical symptoms...24"

[16] "Clinical symptoms...25"

[17] "Clinical symptoms...26"

[18] "Body temperature"

[19] "Blood pressure"

[20] "Saturation of peripheral oxygen"

[21] "apache 2 score at the points of admitting hospitals"

[22] "ranson score at the points of admitting hospitals"

[23] "CTSI score at the points of admitting hospitals"

[24] "imre score at the points of admitting hospitals"

[25] "sofa score at the points of admitting hospitals"

[26] "subclinical examination - (Abdominal fluid) ultrasound at the points of admitting hospitals"

[27] "subclinical examination - (Abdominal fluid) computer tomography"

[28] "subclinical examination - balthazar sHavere (with computer tomography)"

[29] "subclinical examination - CTSI score (with computer tomography)"

[30] "subclinical examination - white blood cell; t0: at the points of admitting hospitals, t6: after 6h of admitting hospitals..."

[31] "subclinical examination -...51"

[32] "subclinical examination - Hematocrit"

[33] "red blood cell"

[34] "prothrombin"

[35] "APTT"

[36] "subclinical examination - Fibrinogen"

[37] "subclinical examination - ure"

[38] "subclinical examination - creatinin"

[39] "subclinical examination - glucose"

[40] "subclinical examination - bilirubin total"

[41] "AST, ALT (liver funtion)"

[42] "subclinical examination -...97"

[43] "cholesterol"

[44] "triglycerid"

[45] "subclinical examination - amylase"

[46] "subclinical examination - lipase"

[47] "subclinical examination - protein"

[48] "subclinical examination - albumin"

[49] "subclinical examination - sodium"

[50] "subclinical examination - potasium"

[51] "subclinical examination - calci total"

[52] "subclinical examination - pH (in blood air)"

[53] "subclinical examination - paco2(in blood air)"

[54] "subclinical examination - pa Oxy (in blood air)"

[55] "subclinical examination - HCO3-(in blood air)"

[56] "BE (in blood air)"

[57] "p/f (paO2/%O2)"

[58] "lactatr (in blood air)"

[59] "treatment - fluide intake...159"

[60] "treatment - fluide intake...160"

[61] "treatment - fluide intake...161"

[62] "treatment - fluide output...162"

[63] "treatment - fluide output...163"

[64] "treatment - fluide output...164"

[65] "treatment - balance fluid in and out...165"

[66] "treatment - balance fluid in and out...166"

[67] "treatment - balance fluid in and out...167"

[68] "treatment - day without food intake"

[69] "treatment - PEX treatment of which day of the diagnosis"

[70] "treatment - number of PEX treatment"

[71] "treatment - triglycerid before first time of PEX"

[72] "treatment - triglycerid after first time of PEX"

[73] "treatment - APAche 2 score before first time PEX"

[74] "treatment - APAche 2 score after first time PEX"

[75] "treatment - Imre score before first time of PEX"

[76] "treatment - Imre score after first time of PEX"

[77] "Result - dead or alive"

[78] "Potential Complication"

[79] "Patient with PEX or without PEX"

**Missing Values Visualization:**

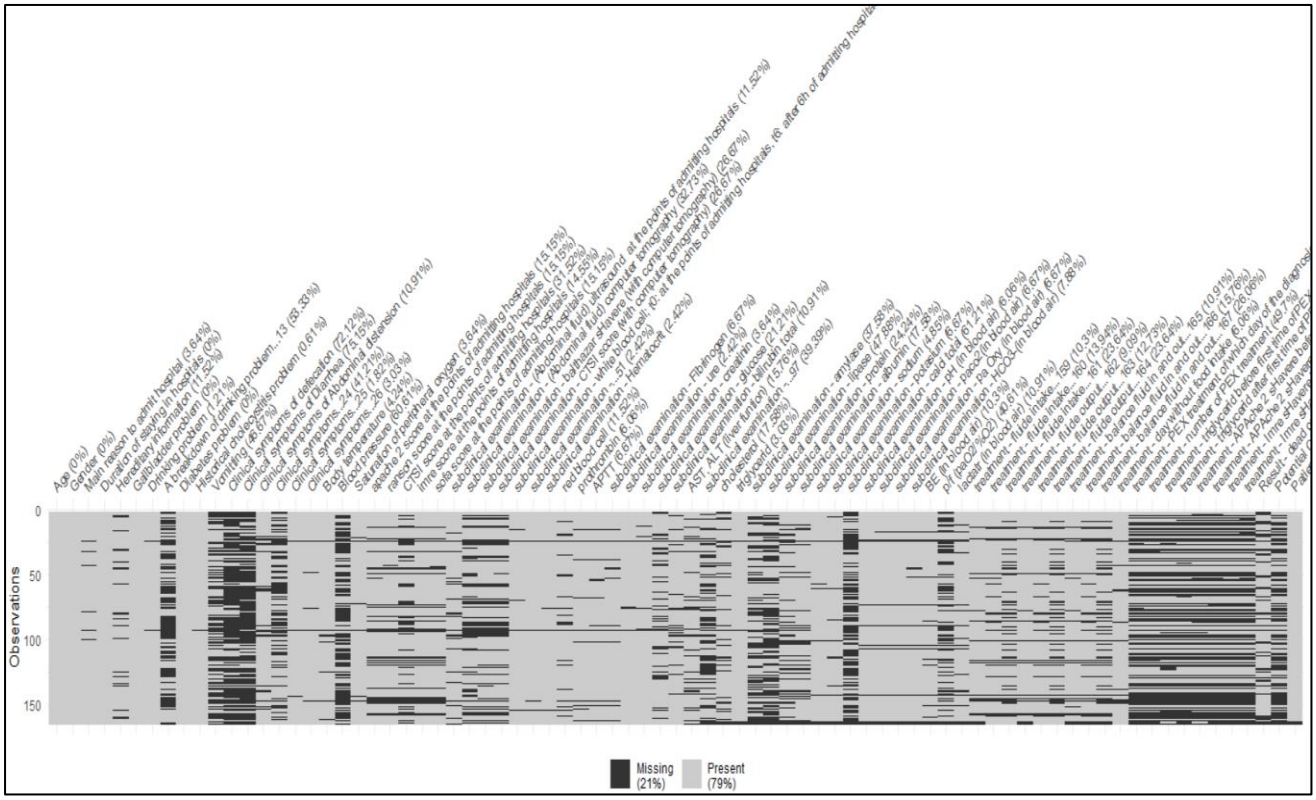


Fig 1: General overview of distribution of missing values.

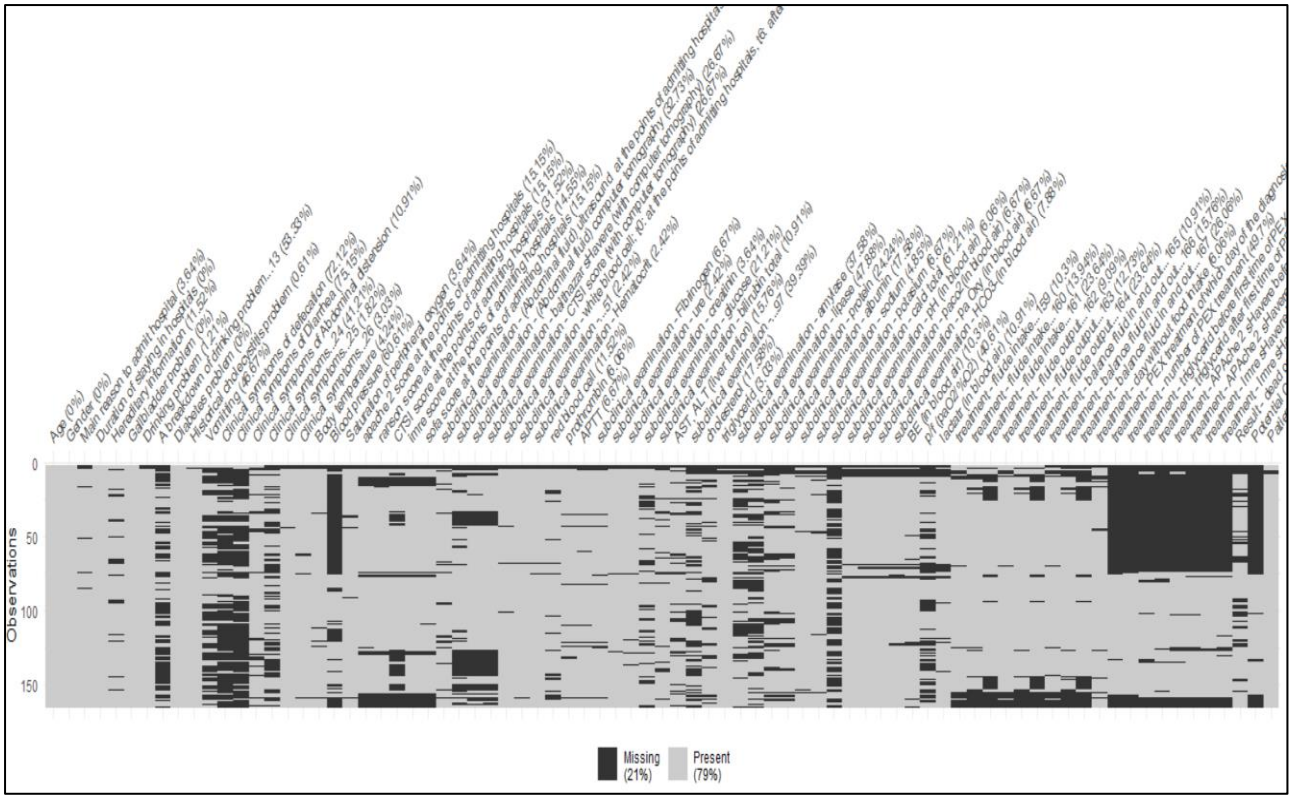


Fig 2: Clustered overview of distribution of missing values.

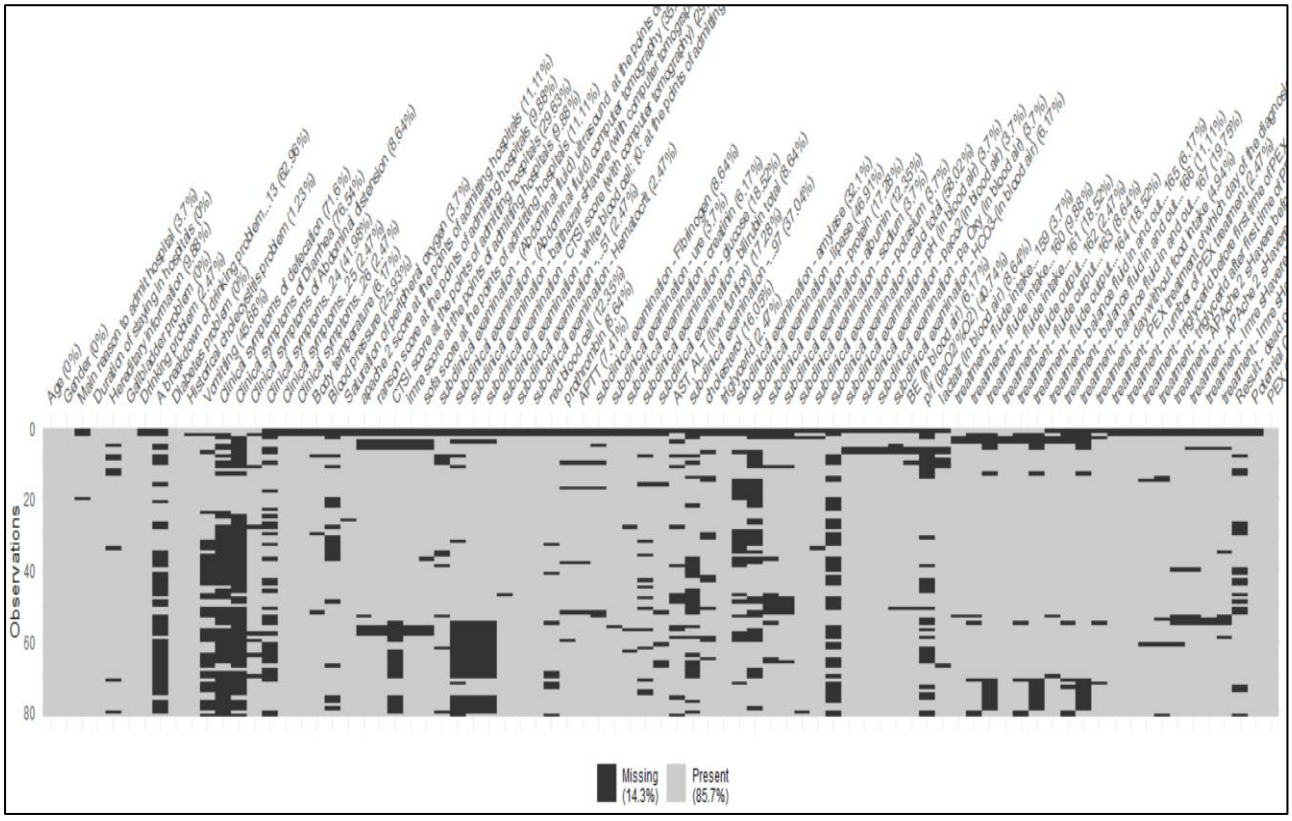


Fig 3: Heatmap of Missing values – For patients who underwent PEX treatment

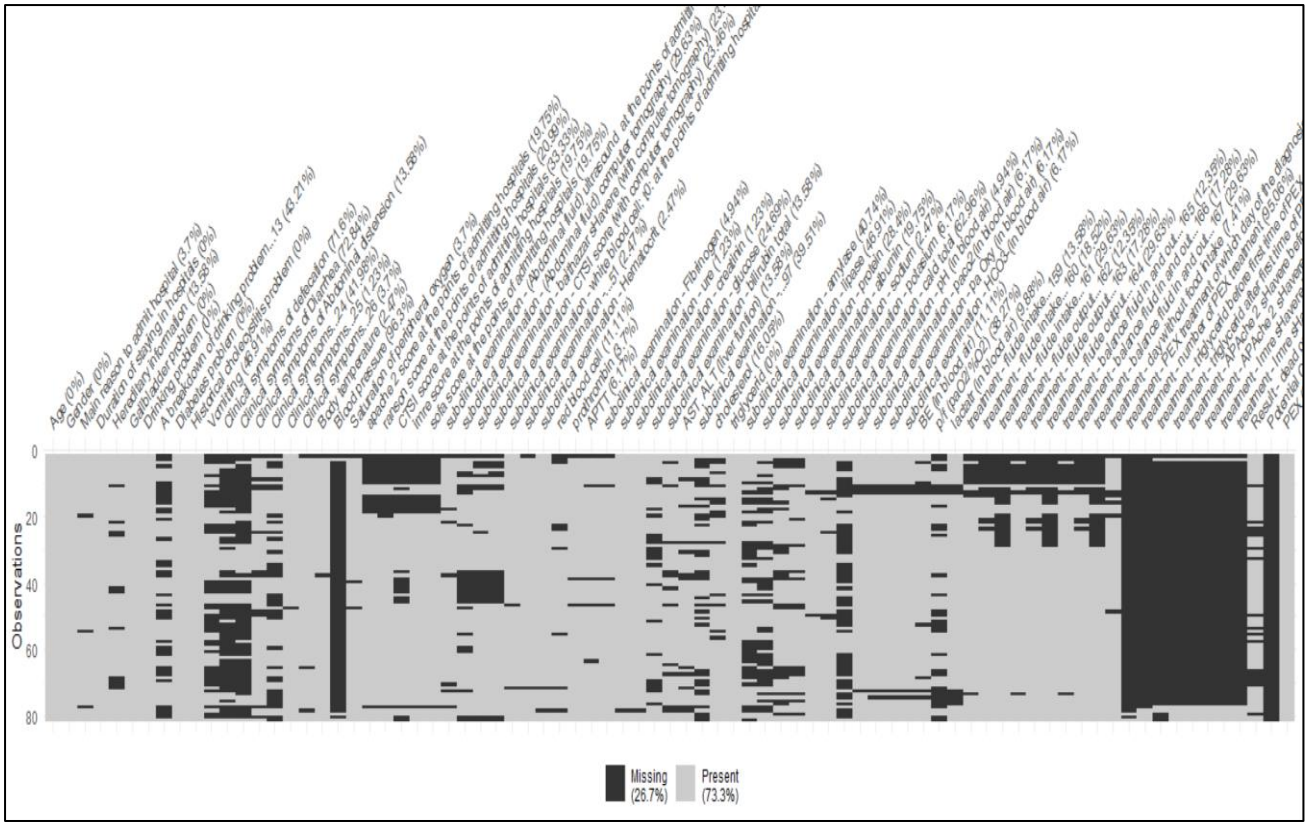


Fig 4: Heatmap of Missing values – For patients without PEX treatment

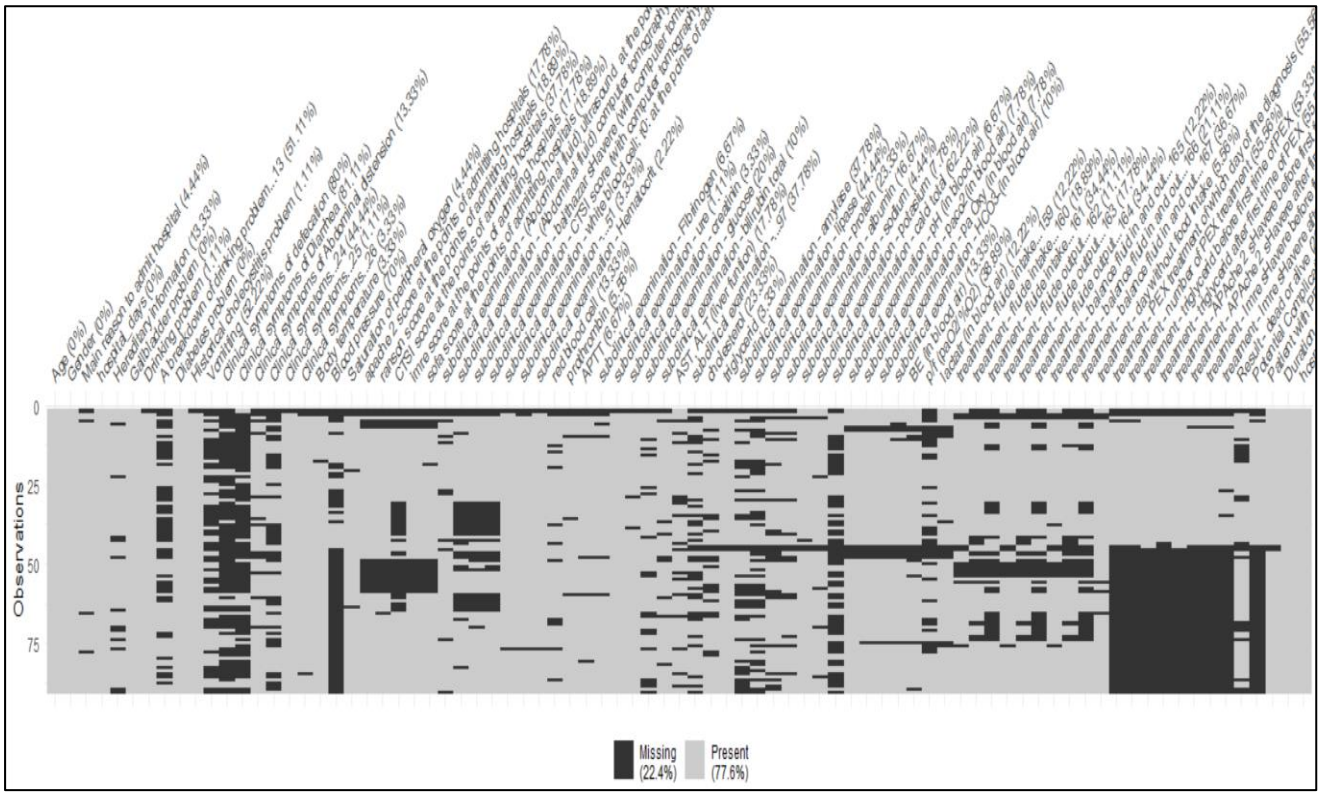


Fig 5: Missing values distribution – For patients who had a shorted stay in hospital (less than 6 days)

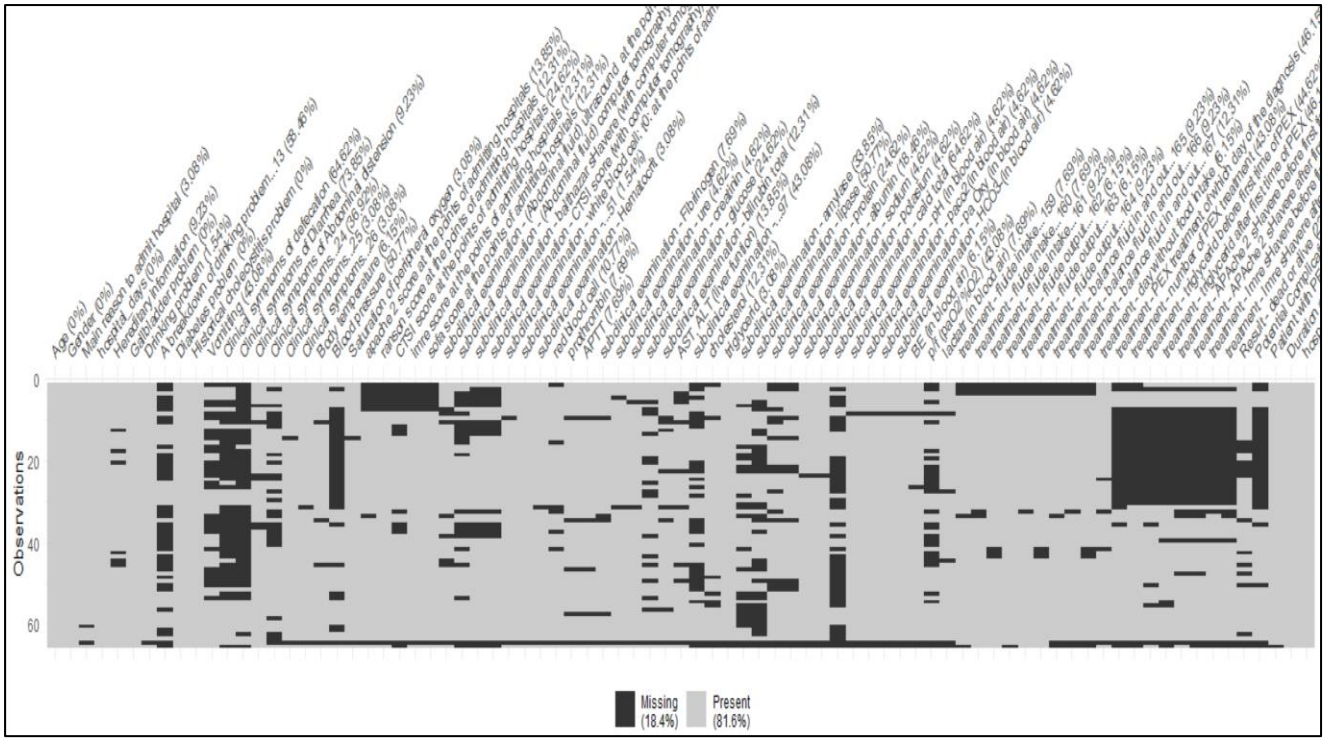


Fig 6: Missing values distribution – For patients who had a medium stay in hospital (Between 6 days to 12 days)

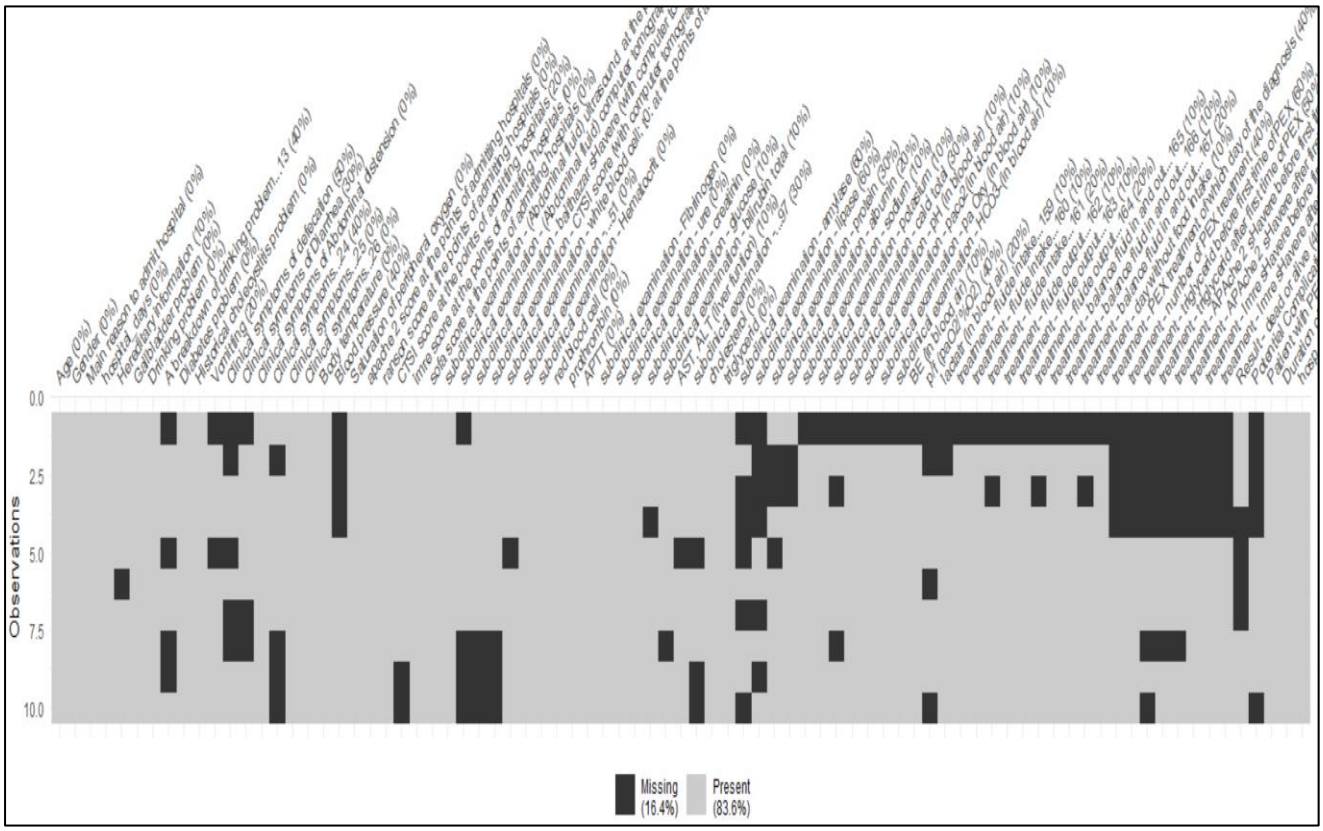


Fig 7: Missing values distribution – For patients who had a Longer stay in hospital (Between 12 days to 18 days)

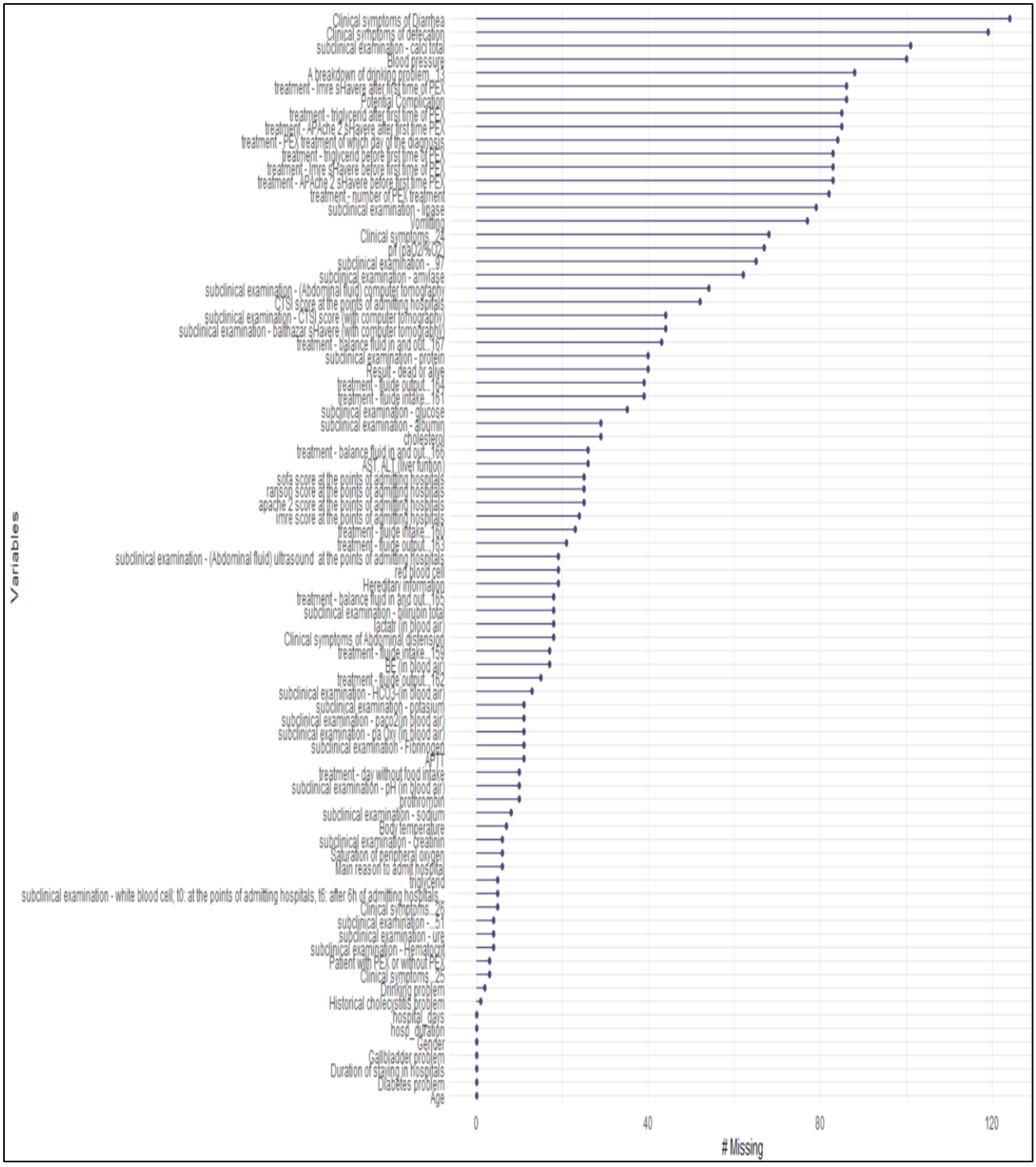


Fig 8: No. of missing cases (out of 165 observations) in each of the variable.

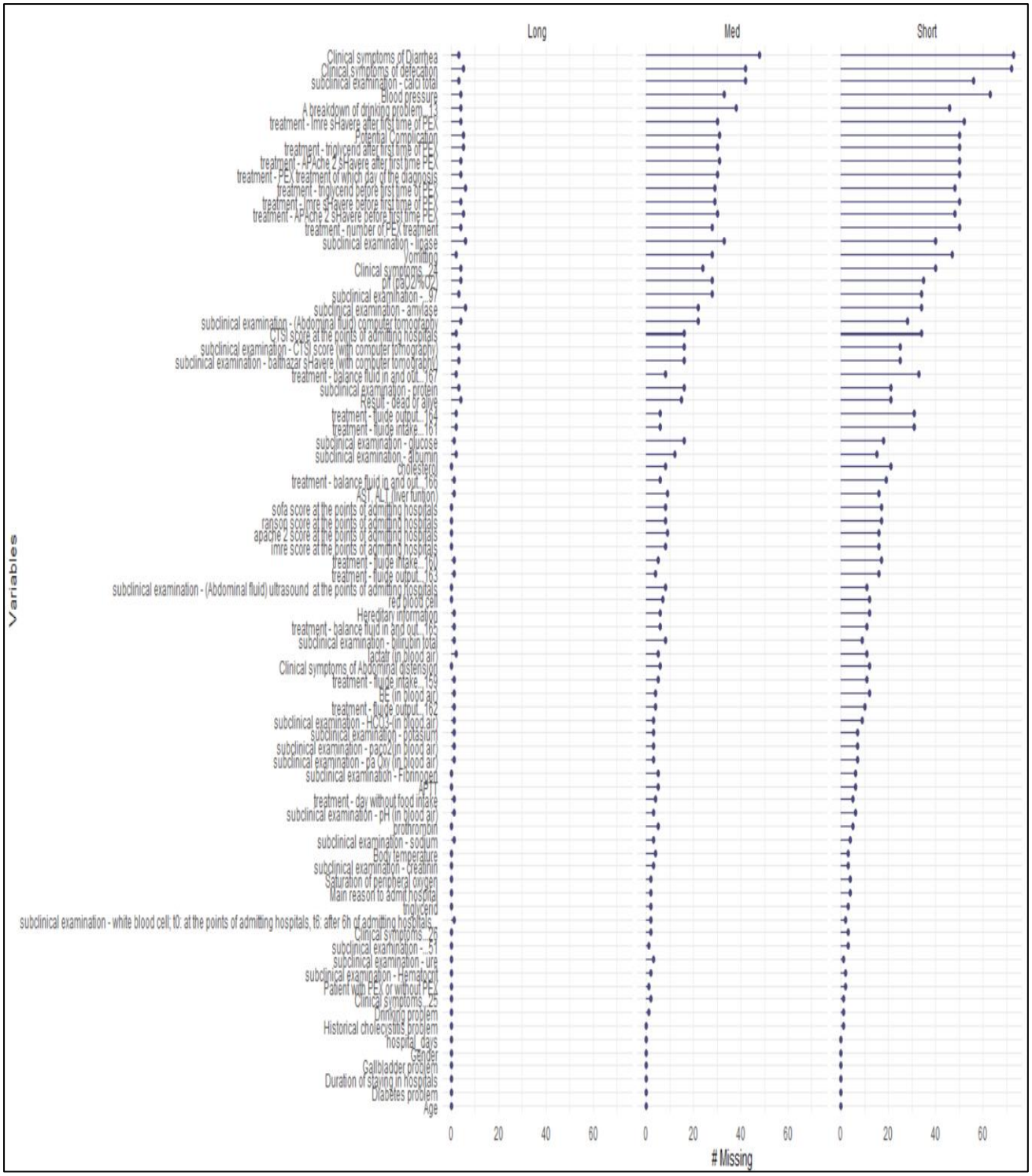


Fig 8: No. of missing cases (out of 165 observations) in each of the variable – For Short, Medium and Longer duration of hospital stays.

**Characterization of Missingness:**

Interpretations of the graphs:

From the figures, we can observe the following patterns of missingness in the variables –

1. **A breakdown of drinking problems has 53% missing values:** A similar variable exist in the dataset 'Drinking Problem' which has about 1% missing values, 43% 'Have' and 55% 'No' values. We classify this pattern as **'MAR'** as the missingness of the breakdown of the drinking problem variable can be attributed to the values in the variable 'Drinking Problem'. The 53% missing values in the breakdown variable are due to the patient not having a drinking problem.

2. **Vomiting has 46% missing, Clinical symptoms of defecation and diarrhea has about 75% missing values:** This pattern could be classified as **MNAR**. Though Vomiting, defecation and diarrhea are most important symptoms of AP, we have a large amount of data missing in these variables. We cannot attribute the missingness to any observed factor in the dataset, as well as we cannot say that the data has been missed completely at random. Perhaps the data in these variables was missed intentionally, or it was missed to capture due to reasons not attributable in the dataset. The variable vomiting has values 0 and 1. The missingness can be because for some patients, the symptom of vomiting was either not checked or was not entered into the dataset.

   Similarly, the values in the variables, defecation, and diarrhoea, are 'no' for some observations and 't6' or t30' for a few observations. Even accounting for all the available observed information, the reason for observations being missing still depends on the unseen observations.

3. **Blood pressure has 60% missing values that is more than half of the observations:** This pattern could also be classified as **MNAR**. After evaluating all the graphs, we still cannot find a reason for the missingness of values in blood pressure. One theory for missing values in the blood pressure could be, may be the patient stayed for a very short time in the hospital and hence this variable was not captured, but we can clearly observe, that for all durations of stay, that is short, medium, and long, the missing values in the variable blood pressure exists in the range 40 -70% respectively. Therefore, either the blood pressure was intentionally not measured for 60% of the patients or was not entered into this dataset. Whatever, the case may be, the missingness is neither completely at random, nor at random.

4. **Patient's severity of disease classification criteria like apache 2 score, ranson score, CTSI score (31% missing), imre score and sofa score has about 15% missing values:** We can observe from the graphs that, nearly the same 11% missing values in these scores exists for patients who underwent PEX treatment, as well as the patients who underwent the conventional treatment. Moreover, the 11% missing values also exists for patients who stayed for less than 6 days in the hospital (short stay) as well as the patients who stayed between 6 to 12 days (medium stay). However, from the graph (Fig 7 – Longer stay in Hospital), we can observe that, the missing percentage of apache 2 score, ranson score, imre score and sofa score was 0% (that is all values are present), and missing percentage of CTSI score was 20%. Therefore, we can relate the missingness with the duration of stay of patients in the hospital. We can also assume that these tests were performed on only the patients who had a longer hospital stay, which means the patients who had severe complications of AP. The tests were perhaps not performed on the patients who discharged from the hospital early and therefore the missingness in these variables. This pattern since attributable to the subsets within the data, is classified as **'Missing at Random' MAR**.

5. **Abdominal fluid ultrasound has about 11% missing values, abdominal fluid with computer tomography, balthazar score and CTSI score has about 30% missing values:** From the graphs, we can observe that missingness of these variables exist for all subsets of data. That is, we can observe missing values for all durations of stay and for patients undergoing PEX treatment and without PEX treatment. We can assume that patients generally perform ultrasound test to get the condition of their abdomen. Certain patients undergo all of the tests, but certain patients would choose to perform only one ultrasound test and skip other tests. As this assumption cannot be verified by the data, we have available, there is no other pattern observed within this dataset that would explain the missingness of these variables. Therefore, either the data was not captured intentionally as the patients did not perform the test, or we are unable to relate the missingness of this data to any other attribute of the dataset, and the missing is not at random, we could classify the missingness of these variables as **MNAR**.

6. **Treatment variables like – pex treatment on which day of diagnosis, number of pex tratments, triglyceride levels, apache 2 scores and Imre scores before and after pex treatments:** For these variables, we can observe a pattern from the graphs (Fig 3 – Patients

who underwent PEX and Fig 4 – Patients who did not have PEX treatment) that, the missing percentages of these variables in the case where the patients did not have a PEX treatment is over 90% whereas for the patients with PEX treatment, the missingness is about 2 – 5%. We can clearly infer that these treatment variables are captured only for patients with PEX treatment and the missingness can be attributed to the reason that these values are not applicable to patients undergoing conventional treatment and therefore the missingness. Though some observations who has missingness in both the cases are classified as **MCAR**, but majority of the observations are **MAR** as they are attributed to the cause within the dataset.

7. **Potential complications:** This variable is Binary. The variable has 49 '1' value indicating that a potential complication was observed in the patient and the rest of the values are blanks indicating '0' – No complications observed, or the status of complication not entered. We can also observe a pattern from the graphs that, patients with PEX have about 2% missing values which means 98% values are '1'. Whereas patients without PEX have 100% missing values meaning '0' – No complication. If we assume that the blanks are to be imputed as '0', we determine that majority of the patients (over 98%) who underwent PEX treatment, were reported of having potential complications after the treatment. However, for the conventional treatment, no patient was observed to have reported complications. Since the missingness of this variable can be directly attributed to the cause within our dataset, we can classify this missingness as **'MAR'.**