A machine learning approach for identification of gastrointestinal predictors for the risk of COVID-19 related hospitalization

Peter Lipták [Corresp., 1] , Peter Banovcin [1] , Róbert Rosoľanka [2] , Michal Prokopič [1] , Ivan Kocan [3] , Ivana Žiačiková [3] , Peter Uhrik [1] , Marian Grendar [4, 5] , Rudolf Hyrdel [1]

[1] Gastroenterology Clinic, University Hospital in Martin, Jessenius Faculty of Medicine in Martin, Comenius University in Bratislava, Martin, Slovak Republic

[2] Clinic of Infectology and Travel Medicine, University Hospital in Martin, Jessenius Faculty of Medicine in Martin, Comenius University in Bratislava, Martin, Slovak Republic

[3] Clinic of Pneumology and Phthisiology, University Hospital in Martin, Jessenius Faculty of Medicine in Martin, Comenius University in Bratislava, Martin, Slovak Republic

[4] Laboratory of bioinformatics and biostatistics, Biomedical Centre Martin, Jessenius Faculty of Medicine, Comenius University in Bratislava, Martin, Slovak Republic

[5] Laboratory of theoretical methods, Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia

Corresponding Author: Peter Lipták
Email address: peter.liptak@uniba.sk

Background and aim: COVID-19 can be presented with various gastrointestinal symptoms. Shortly after the pandemic outbreak several machine learning algorithms have been implemented to assess new diagnostic and therapeutic methods for this disease. Aim of this study is to assess gastrointestinal and liver related predictive factors for SARS-CoV-2 associated risk of hospitalization. **Methods:** Data collection was based on a questionnaire from the COVID-19 outpatient test center and from the emergency department at the University Hospital in combination with the data from internal hospital information system and from a mobile application used for telemedicine follow-up of patients. For statistical analysis SARS-CoV-2 negative patients were considered as controls in three different SARS-CoV-2 positive patient groups (divided based on severity of the disease). The data were visualized and analyzed in R ver. 4.0.5. The Chi-squared or Fisher test was applied to test the null hypothesis of independence between the factors followed, where appropriate, by the multiple comparisons with the Benjamini Hochberg adjustment. The null hypothesis of the equality of the population medians of a continuous variable was tested by the Kruskal Wallis test, followed by the Dunn multiple comparisons test. In order to assess predictive power of the gastrointestinal parameters and other measured variables for predicting an outcome of the patient group the Random Forest machine learning algorithm was trained on the data. The predictive ability was quantified by the ROC curve, constructed from the Out-of-Bag data. Matthews correlation coefficient was used as a one-number summary of the quality of binary classification. The importance of

the predictors was measured using the Variable Importance. A 2D representation of the data was obtained by means of Principal Component Analysis for mixed type of data. Findings with the p-value below 0.05 were considered statistically significant. **Results:** Total of 710 patients were enrolled in the study. The presence of diarrhea and nausea was significantly higher in the emergency department group than in the COVID-19 outpatient test center. Among liver enzymes only aspartate transaminase (AST) has been significantly elevated in the hospitalized group compared to patients discharged home. Based on the Random Forest algorithm, AST has been identified as the most important predictor followed by age or diabetes mellitus. Diarrhea and bloating have also predictive importance, although much lower than AST. **Conclusion:** SARS-CoV-2 positivity is connected with isolated AST elevation and the level is linked with the severity of the disease. Furthermore, using machine learning Random Forest algorithm, we have identified the elevated AST as the most important predictor for COVID-19 related hospitalizations.

1 **A machine learning approach for identification of gastrointestinal predictors**

2 **for the risk of COVID-19 related hospitalization**

4 Peter Lipták[1*], Peter Banovcin[1], Róbert Rosoľanka[2], Michal Prokopič[1], Ivan Kocan[3], Ivana

5 Žiačiková[3], Peter Uhrik[1], Marian Grendár[4,5], Rudolf Hyrdel[1]

6 1: Gastroenterology Clinic, University Hospital in Martin, Jessenius Faculty of Medicine in

7 Martin (JFM CU), Comenius University in Bratislava, Slovakia.

8 2: Clinic of Infectology and Travel Medicine, University Hospital in Martin, Jessenius Faculty of

9 Medicine in Martin (JFM CU), Comenius University in Bratislava, Slovakia.

10 3: Clinic of Pneumology and Phthisiology, University Hospital in Martin, Jessenius Faculty of

11 Medicine in Martin (JFM CU), Comenius University in Bratislava, Slovakia.

12 4: Laboratory of Bioinformatics and Biostatistics, Biomedical Centre Martin, Jessenius Faculty

13 of Medicine, Comenius University, Mala Hora 4C, 036 01 Martin, Slovakia

14 5: Laboratory of Theoretical Methods, Institute of Measurement Science, Slovak Academy of

15 Sciences, Dubravska cesta 9, 841 04 Bratislava, Slovakia

17 Corresponding Author:

18 Peter Liptak, MD, PhD

19 Gastroenterology Clinic, University Hospital in Martin, Kollarova 2, 03659, Martin, Slovakia

20 Email address: peter.liptak@gmail.com

22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39

Abstract

40

Background and aim: COVID-19 can be presented with various gastrointestinal symptoms.

41

Shortly after the pandemic outbreak several machine learning algorithms have been implemented

42

to assess new diagnostic and therapeutic methods for this disease. Aim of this study is to assess

43

gastrointestinal and liver related predictive factors for SARS-CoV-2 associated risk of

44

hospitalization.

45

Methods: Data collection was based on a questionnaire from the COVID-19 outpatient test center

46

and from the emergency department at the University Hospital in combination with the data from

47

internal hospital information system and from a mobile application used for telemedicine follow-

48

up of patients. For statistical analysis SARS-CoV-2 negative patients were considered as controls

49

in three different SARS-CoV-2 positive patient groups (divided based on severity of the disease).

50

The data were visualized and analyzed in R ver. 4.0.5. The Chi-squared or Fisher test was applied

51

to test the null hypothesis of independence between the factors followed, where appropriate, by

52

the multiple comparisons with the Benjamini Hochberg adjustment. The null hypothesis of the

53

equality of the population medians of a continuous variable was tested by the Kruskal Wallis test,

54

followed   by   the   Dunn   multiple   comparisons   test.

55

In order to assess predictive power of the gastrointestinal parameters and other measured variables

56

for predicting an outcome of the patient group the Random Forest machine learning algorithm was

57

trained on the data. The predictive ability was quantified by the ROC curve, constructed from the

58

Out-of-Bag data. Matthews correlation coefficient was used as a one-number summary of the

59

quality of binary classification. The importance of the predictors was measured using the Variable

60

Importance. A 2D representation of the data was obtained by means of Principal Component

61

Analysis for mixed type of data. Findings with the p-value below 0.05 were considered statistically

62

significant.

63

64

Results: Total of 710 patients were enrolled in the study. The presence of diarrhea and nausea

65

was significantly higher in the emergency department group than in the COVID-19 outpatient

66

test center. Among liver enzymes only aspartate transaminase (AST) has been significantly

67

elevated ($p<0.001$) in the hospitalized group compared to patients discharged home. Based on

68

the Random Forest algorithm, AST has been identified as the most important predictor followed

69

70 by age or diabetes mellitus. Diarrhea and bloating have also predictive importance, although

71 much lower than AST.

72 **Conclusion:** SARS-CoV-2 positivity is connected with isolated AST elevation and the level is

73 linked with the severity of the disease. Furthermore, using machine learning Random Forest

74 algorithm, we have identified the elevated AST as the most important predictor for COVID-19

75 related hospitalizations.

76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107

108 **Introduction**

109 Acute SARS-CoV-2 infection presents with variable symptoms associated with various organ

110 systems. Typical symptoms of COVID-19 are fever, cough, and in the case of a more severe course

111 of the disease, dyspnea with respiratory insufficiency occurs [1]. In addition, COVID-19 may be

112 presented with gastrointestinal symptoms, which include dominantly nausea, vomiting, diarrhea,

113 anorexia and abdominal pain with relatively wide range of prevalence among different published

114 studies [2]–[8]. Since COVID-19 pandemic is the cause of an immense world health crisis, new

115 diagnostic and therapeutic methods are rapidly emerging [9]. The use of artificial intelligence is

116 just one of them. Shortly after the COVID-19 outbreak, various machine learning algorithms

117 have been implemented [10]–[13]. Machine learning helps quickly identify patterns and trends of

118 the large volume of data, that are difficult for humans to recognize [14]. The availability of

119 objective stratification tools for the rapid assessment of a patient status and prognosis is of great

120 use for the frontline health-care providers [15].

121 The primary aim of this study is to assess the possible predictive factors for SARS-CoV-2 outcome

122 based on gastrointestinal symptoms and liver related laboratory results using machine learning

123 algorithms of the Random Forest [16], [17]. The secondary aim is to determinate the prevalence

124 of gastrointestinal symptoms among patients with COVID-19 within different groups based on the

125 severity of the disease.

126

127 **Materials & Methods**

128 The Study was performed from February through May 2021. Only subjects aged 18 years or older

129 were included in the study. All patients enrolled in this study had signed the informed consent.

130 This study was approved by the Ethical Committee of the University Hospital in Martin, decision

131 number: 14/2021.

132 Two distinct kinds of population were considered for this study. First group consists of patients

133 who underwent nasopharyngeal swab in the outpatient hospital testing center for COVID-19 in

134 order to determine whether they were SARS-CoV-2 positive. The method used for SARS-CoV-2

135 detection from nasopharyngeal swab was PCR (polymerase chain reaction). This group was then

136 subdivided based on their positivity. The negative group was thereafter used as a control group for

137 this study.

138 Second group consists of patients who attended COVID-19 emergency department (ED) in the

139 University Hospital. These patients were confirmed positive from nasopharyngeal swab either by

140 PCR or antigen method. Only patients with typical COVID-19 symptoms (fever, cough, dyspnoe)

141 were included in this study. Patients who were SARS-CoV-2 positive but, didn't present with

142 typical COVID-19 symptoms (e.g., patients who came to emergency room because of other

143 diagnoses, but simultaneously were SARS-CoV-2 positive) were excluded. Therefore, we

144 considered for this study only patients who were both tested positive and had at least one typical

145 COVID-19 symptom.

146 The second group was then divided based on further evaluation and course of the disease. First

147 subgroup consists of patients that didn't require admission to the hospital and were referred to the

148 outpatient care. Second subgroup of patients was admitted to the hospital. Consequently, this group

149 was observed until the end of hospitalization either because of their death or resolution of the

150 disease. This subgroup was also divided for analysis purposes to patients who required medical

151 care in standard hospital ward and those who needed intensive care unit (ICU).

152 Data was collected by using a questionnaire in the group from COVID-19 outpatient test center at

153 the University Hospital. Data from emergency room was obtained with the same questionnaire

154 which was combined with information from medical examination by an attending physician and

155 from the mobile application MEDAsistent used for telemedicine follow-up developed at the

156 Clinic of Pneumology and Phthisiology in the University Hospital in Martin. Further information

157 (including laboratory tests results, chest X-ray etc.) about patients who were hospitalized has been

158 obtained from hospital information system.

159 The questionnaire consists of questions related to the present health complaints typical for COVID-

160 19 and the spectrum of most common gastrointestinal symptoms which occurred in the last 5-7

161 days before examination. Patients were also allowed to write down other presented symptoms in

162 the case they were not in the original list. In order to include only new or worsened gastrointestinal

163 symptoms in the study the questionnaire also contained questions about chronic gastrointestinal

164 symptoms and their possible worsening in the last 5-7 days before examination.

165

166 **Data analysis**

167 The data was visualized and analyzed in R [18], ver. 4.0.5, with the aid of libraries gtsummary

168 [19], rstatix [20], DescTools [21], randomForestSRC [22], PCAmixdata [23] and ggpubr [24]. The

169 sample median and the lower and upper quartiles were used to summarize the data on continuous

170 variables (e.g., age); counts and percentages were used to summarize factors (e.g., gender). The

171   Chi-squared or Fisher test were applied to test the null hypothesis of independence between factors

172   (gender vs group; fever vs group; cough vs group; diarrhea vs group; constipation vs group;

173   bloating vs group; nausea vs group; heartburn vs group; abdominal pain vs group), followed, where

174   appropriate, by multiple comparisons with the Benjamini Hochberg adjustment. Using a

175   contingency table, an absence of trend was tested by Cochran Armitage test. The null hypothesis

176   of the equality of the population medians of the continuous variable: age, Oxygen (O2) saturation,

177   C-reactive protein (CRP), gamma glutamyltransferase (GMT), aspartate aminotrasferase (AST),

178   Bilirubin) was tested by the Kruskal Wallis test, followed by the Dunn multiple comparisons test

179   with the Benjamini Hochberg correction of p-values. Two-way ANOVA was used to model the

180   association between AST and group (discharged home, admitted to hospital) in interaction with

181   the recent ATB use. Another two-way ANOVA model was utilized to quantify the association

182   between AST and group (discharged home, admitted to hospital) in interaction with chronical liver

183   disease (yes, no). The AST values were log-transformed to bring data to normality. Normality of

184   residuals was assessed by the quantile-quantile plot with the 95% confidence band constructed by

185   bootstrap. Assumption of homogenity of variance was tested by the Levene test.

186   In order to assess the predictive power of the gastrointestinal parameters and other measured

187   variables (Gender, Age, No of Days of Symptoms, AST, alanine aminotrasferase /ALT/, Bilirubin,

188   Recent antibiotics /ATB/ Usage, Diabetes Mellitus, Arterial Hypertension, Chronic Liver Disease,

189   Fever, Cough, Diarrhea, Constipation, Bloating, Nausea, Heartburn, Abdominal Pain) for

190   predicting the outcome of the patient group the Random Forest machine learning algorithm was

191   trained on the data. The predictive ability was quantified by the ROC curve, constructed from the

192   Out-of-Bag data. Matthews correlation coefficient was used as a one-number summary of the

193   quality of binary classification. Importance of the predictors was measured by the Variable

194   Importance. A 2D representation of the data (the predictors used in Random Forest; i.e., Gender,

195   Age, Number of Days of Symptoms, AST, ALT, Bilirubin, Recent ATB Usage, Diabetes Mellitus,

196   Arterial Hypertension, Chronic Liver Disease, Fever, Cough, Diarrhea, Constipation, Bloating,

197   Nausea, Heartburn, Abdominal Pain) was obtained by Principal Component Analysis for a mixed

198   type of data. Findings with the p-value below 0.05 were considered statistically significant.

199

200

201

202

Results

204 Total of 710 patients were enrolled in the study. Thirty (30) patients were excluded from the
205 further analysis after primary screening. Participants (n=352) from the outpatient center who were
206 tested PCR negative for SARS-CoV-2 virus were considered- as the control group. SARS-CoV-2
207 positive group from outpatient center included 166 participants. One hundred and sixty-two
208 (n=162) patients from emergency department were enrolled. From this group 78 patient (48%)
209 were discharged home, 57 (35.3%) admitted to the hospital for standard care until discharged from
210 hospital. Twenty-seven (27) (16.7%) patients required intensive care unit. Based on age, the groups
211 from outpatient center had almost similar median of 42 and 41 years of age respectively.
212 Hospitalized patients were significantly older as shown in the Table 1. The presence of typical
213 COVID-19 symptoms such as fever and cough were significantly higher in the hospitalized groups
214 as opposed to outpatient participants. There were no significant differences between groups in the
215 men to women ratio.

216

Gastrointestinal symptoms occurrence and laboratory findings (Table 1 and 2)

218 The presence of diarrhea, constipation, bloating, nausea, heartburn and abdominal pain was
219 considered in this study. Presence of diarrhea and nausea was significantly higher in SARS-CoV-
220 2 positive patients than in SARS-CoV-2 negative controls. Comparing SARS-Cov-2 negative and
221 SARS-CoV-2 positive participants the cumulative presence of diarrhea is 21.3% (70/328) in the
222 positive group (combined outpatient center and emergency department) vs 6.2% (22/352) in the
223 negative group and for nausea it is 13.1% (43/328) in the positive group vs 3.4% (12/352) in the
224 negative group. This trend goes further considering ED patients and the severity of disease.

225 Among gastrointestinal symptoms, diarrhea and bloating were significantly more often manifested
226 in patients who were admitted to the hospital compared to those discharged home (40% for
227 diarrhea and 14% for bloating vs 18% and 2.6% respectively). Other symptoms such as abdominal
228 pain, heart burn, nausea, vomitus, anorexia, and constipation were not presented differently in
229 these groups in the meaning of statistical significance. C-reactive protein was also significantly
230 higher in hospitalized group. In case of alanin transaminase (ALT), aspartate transaminase (AST)
231 and bilirubin as markers of possible liver damage only AST (Figure 1) was significantly higher in
232 the hospitalized group. This difference is substantial. There is no statistically significant difference
233 in the levels of ALT (Figure 1) and Bilirubin when comparing different groups of patients.

234 **Predictors of hospitalization based on machine learning**

235 Based on the Random Forest algorithm with the data on demographic characteristics, symptoms
236 and gastrointestinal related laboratory findings in hospitalized and discharged patients, several
237 predictors for risk of hospitalization were identified. AST was pinpointed as the most important
238 predictor followed by age and diabetes mellitus. Diarrhea and bloating have also positive
239 importance, although much lower than AST. Gastrointestinal symptoms such as nausea, abdominal
240 pain or anorexia have none or negative predictive importance. The ROC curve for combined
241 factors is shown in the Figure 2 with AUC 0.76. The Matthews correlation coefficient was 0.48.
242 When using only liver enzymes (AST, ALT), gastrointestinal symptoms (diarrhea and bloating),
243 chronic liver disease, age and diabetes mellitus, the ROC curve (Figure 3) for this combination of
244 factors attained AUC 0.799 with AST as the strongest predictor for hospitalization (Table 3). The
245 Matthews correlation coefficient was 0.37.

246

247 **Principal components visualization of data**

248 Principal component analysis was used to get a two-dimensional visualization of the data, for
249 patients discharged home after ED examination and patients admitted to hospital. Data used for
250 the analysis consist of the data from Table 2, that means a combination of general patient
251 characteristics, typical COVID-19 symptoms and gastrointestinal symptoms and liver related
252 laboratory results. The PCA plot (Figure 4) is showing two distinct clusters which are partially
253 overlapping with tendencies to shift apart.

254

Discussion

256 Several studies and meta-analyses have pointed out the gastrointestinal involvement in the
257 SARS-CoV-2 infection [3], [4], [6], [8], [25]–[27]. The data from the pooled prevalence of
258 gastrointestinal symptoms are varying significantly from 10.5% to 53% between studies [3], [4],
259 [25], [28]. Based on comprehensive meta-analysis by Sultan et al. [4] the pooled prevalence of
260 diarrhea is 7.7%, nausea and vomiting 7.8% and abdominal pain 3.6%. In the presented study we
261 have focused on the presence of diarrhea, constipation, bloating, nausea, heart burn and
262 abdominal pain. Statistically significant differences have been found in the case of diarrhea and
263 nausea when comparing SARS-CoV-2 negative and positive patients. In the group of
264 hospitalized patients (with standard care) the diarrhea was presented in 40% patients and nausea

265   in 21%, which is higher compared to some meta-analysis mentioned, but consistent with the data

266   considering general presence of gastrointestinal symptoms and gut involvement. When

267   comparing only emergency department group the presence of bloating is significantly higher in

268   the hospitalized group than in those who were discharged home. Interestingly, bloating has

269   lower prevalence in the group of ICU patients than in patients with standard care management.

270   This could be explained by high subjectivity and interpersonal differences when reporting

271   symptom such as bloating. Considering differences between these two groups of patients, those

272   with more severe course of disease attach lower importance to less annoying symptoms such as

273   heart burn, bloating and nausea when compared to more manifested symptoms such as diarrhea,

274   abdominal pain or vomitus.

275   Focusing on the liver enzymes as markers of possible liver impairment resulting from SARS-

276   CoV-2 infection the AST, ALT and bilirubin were considered for the evaluation. The results are

277   showing that median level of liver enzymes was not elevated in the discharged group. Bilirubin

278   and ALT were also within normal range in the hospitalized group with no statistically significant

279   differences between these two groups. Only AST was elevated over the upper level of the

280   reference value in the hospitalized group with progressively higher values in patients who

281   required ICU. The differences between hospitalized and discharged patients are substantially

282   significant.

283   Several previously published data have shown an elevation in both transaminases and bilirubin to

284   a different extent ranging from 1% to 53% (mainly ALT and AST accompanied by slightly

285   increased bilirubin concentrations) [3]. In most published data, severe liver alterations were

286   uncommon [29] and the pooled prevalence of liver injury regarding severity was 12% based on

287   the meta-analysis by Mao et al. [3] More severe liver injury was also associated with worse

288   outcomes, including intensive care unit admission and mortality [30].

289   The pathophysiology of liver involvement in COVID-19 is still not completely understood. The

290   direct viral infection of the liver cells is proposed as one of potential causes of liver injury, but

291   the comprehensive studies are scarce. A study with pathological analysis of liver tissues from

292   dead victims of COVID-19 showed no viral inclusions in hepatocytes [31]. Another repeatedly

293   proposed and generally accepted mechanism of liver impairment could be drug toxicity [3]. In

294   order to determine the possible influence of recent ATB usage on the elevation of AST presented

295   in this paper, a two-way analysis of variance (two-way ANOVA) was performed. There are no

296   significant differences between the groups with or without recent antibiotics usage. Therefore,

297   we have concluded that ATB usage has no relevant influence on the elevated AST levels. The

298   two-way ANOVA was also performed to assess the relationship between the presence of chronic

299   liver disease and AST. There is no statistically relevant difference in AST levels in hospitalized

300   patients with and without chronic liver disease.

301   Another possible explanation of elevated transaminases is that it could be the result of a systemic

302   inflammation. ALT is an enzyme most commonly found in liver, with small levels in striated

303   muscle tissue and myocardium. On the other hand, AST could be found in liver, but also in

304   striated and myocardial muscle, kidneys, brain and red blood cells. AST had been used as a

305   marker for myocardial infarction for a long time before more sensitive markers were identified

306   and implemented into the routine clinical practice [32]. Based on the results of this study and

307   current knowledge of SARS-CoV-2 interaction in human organism it is possible that elevated

308   levels of AST in COVID-19 patients could be the result of a systemic inflammation with general

309   tissue hypoperfusion rather than a result of a direct influence of SARS-CoV-2 virus on the

310   hepatocytes or hepatotoxic drug use.

311   Further, we focused on identifying the possible predicting factors for hospitalization in COVID-

312   19 patients using the Random Forest (RF) machine learning algorithm.

313   Different types of machine learning are being used in an increased rate to determine the

314   predictors of outcome in various areas of clinical practice from brain trauma injuries [33],

315   radiology [34], oncology [35] to dermatology [36]. Since COVID-19 pandemic has been

316   affecting the global population for more than two years now and it is the cause of an immense

317   health crisis in most world countries new diagnostic tools – machine learning being one of them

318   – and therapeutic methods have been rapidly emerging [9]. Shortly after the COVID-19 outbreak

319   various machine learning techniques were used, including taxonomic classification of COVID-

320   19 genomes [10], determining the predictors of severe COVID-19 [11] and searching for new

321   potential drug candidates against SARS-CoV-2 viral infection [12]. Another example of a

322   successful implementation of artificial intelligence in COVID-19 diagnosis is the evaluation of

323   the CT scans detecting SARS-CoV-2 associated pneumonia and their differentiation from the

324   community acquired pneumonia and other similar conditions with specificity and sensitivity

325   higher than 90% [13].

326   So far several studies have been published using Random Forest Machine Algorithm for

327   identifying the predictors for COVID-19 outcome from a wide variety of symptoms,

328   socioeconomical factors [37] and laboratory results with various results [38], [39]. To our current

329   knowledge there are no studies specifically focused on the gastrointestinal symptoms and gut

330   related laboratory findings to this date.

331   In order to assess the predictive power of the gastrointestinal parameters and other measured

332   variables for predicting the need for hospitalization the Random Forest machine learning

333   algorithm was trained on the data from our study. Random Forest has become the Machine

334   Learning method of choice for several reasons: a) it usually appears among the top performing

335   classification algorithms; b) it has small number of tuning parameters; c) it does not overfit; d)

336   and last but not least, by its construction it provides a realistic estimate of the performance on a

337   future data via the Out-Of-Bag data. Moreover, Random Forest, at least as implemented in the R

338   library randomForestSRC, provides two different measures of importance of predictors. For

339   these reasons, we have selected RF algorithm to assess predictive power of the studied variables,

340   and to obtain their ranking.

341   Results were plotted as a ROC curve obtained from the Out-Of-Bag data. When considering the

342   general COVID-19 symptoms, gastrointestinal symptoms, age, sex, lasting of the symptoms and

343   comorbidities (diabetes mellitus, arterial hypertension and chronical liver diseases) the AUC is

344   0.76. The variable importance plot is shown in figure 5. When measuring the variable

345   importance, the most important predictor is AST followed by age and diabetes mellitus, which

346   are substantially less important. When using only liver enzymes (AST, ALT), gastrointestinal

347   symptoms (diarrhea and bloating), age and presence of chronic liver disease and diabetes

348   mellitus the AUC is 0.799 with AST as the strongest predictor for hospitalization. The variable

349   importance plot is shown in figure 6. Previously published studies, which used mostly the

350   methods of classical statistics, have identified the presence of gastrointestinal symptoms [40],

351   predominantly diarrhea [41], [42] and elevated liver enzymes [2] as predictors of hospitalization

352   associated with COVID-19. In our data, we have singled out aspartate transaminase (AST) as not

353   only the statistically significantly elevated liver enzyme in patients requiring hospitalization, but

354   using the artificial intelligence with the Random Forest algorithm the AST proved to be the most

355   important predictor of hospitalization. Finally, we performed the principal component analysis

356   for mixed type of data in order to obtain a two-dimensional representation of the data on patients

357   who were discharged home and those who were admitted to hospital. As could be seen on the

358   Plot 4 these two groups are partially overlapping, but with clear tendencies to shift apart, which

359   is in accordance with the predictive performance of the studied variables in the Random Forest

360   algorithm.

361

362   **Conclusions**

363   This study has identified elevated AST for being the most important predictor for COVID-19

364   related hospitalizations using machine learning Random Forest algorithm. We have also shown

365   that SARS-CoV-2 positivity is connected with isolated AST elevation and the level is linked

366   with the severity of the disease. Furthermore, the prevalence of diarrhea and nausea among

367   SARS-CoV-2 positive patients is significantly higher compared to SARS-CoV-2 negative

368   controls. Bloating is occurring significantly more frequently in COVID-19 patients who require

369   hospitalization than those who could be discharged to outpatient care.

370

371

372   **Abbreviations:**

373   PCR (polymerase chain reaction), ALT (alanin transaminase), AST (aspartate transaminase),

374   GMT (gamma glutamyltransferase), ICU (intensive care unit), ED (emergency department),

375   ROC (receiver operating characteristic), PCA (principal component analysis), RF (Random

376   Forest), ATB (antibiotics)

377

378

379

380

381

382

383

384

385

386

387

388

389

References

[1]   W. Guan *et al.*, "Clinical Characteristics of Coronavirus Disease 2019 in China," *N. Engl. J. Med.*, vol. 382, no. 18, pp. 1708–1720, Feb. 2020.

[2]   M. Aziz, H. Haghbin, W. Lee-Smith, H. Goyal, A. Nawras, and D. G. Adler, "Gastrointestinal predictors of severe COVID-19: systematic review and meta-analysis," *Ann. Gastroenterol.*, vol. 33, no. 6, p. 615, 2020.

[3]   R. Mao *et al.*, "Manifestations and prognosis of gastrointestinal and liver involvement in patients with COVID-19: a systematic review and meta-analysis," *Lancet Gastroenterol. Hepatol.*, vol. 5, no. 7, pp. 667–678, Jul. 2020.

[4]   S. Sultan *et al.*, "AGA Institute Rapid Review of the Gastrointestinal and Liver Manifestations of COVID-19, Meta-Analysis of International Data, and Recommendations for the Consultative Management of Patients with COVID-19," *Gastroenterology*, vol. 159, no. 1, pp. 320-334.e27, Jul. 2020.

[5]   K. P. Patel *et al.*, "Gastrointestinal, hepatobiliary, and pancreatic manifestations of COVID-19," *J. Clin. Virol.*, vol. 128, p. 104386, Jul. 2020.

[6]   F. D'Amico, D. C. Baumgart, S. Danese, and L. Peyrin-Biroulet, "Diarrhea During COVID-19 Infection: Pathogenesis, Epidemiology, Prevention, and Management," *Clinical Gastroenterology and Hepatology*, vol. 18, no. 8. W.B. Saunders, pp. 1663–1672, 01-Jul-2020.

[7]   X. Jin *et al.*, "Epidemiological, clinical and virological characteristics of 74 cases of coronavirus-infected disease 2019 (COVID-19) with gastrointestinal symptoms," *Gut*, vol. 69, no. 6, pp. 1002–1009, Jun. 2020.

[8]   F. Xiao, M. Tang, X. Zheng, Y. Liu, X. Li, and H. Shan, "Evidence for gastrointestinal infection of SARS-CoV-2," *Gastroenterology*, vol. 158, pp. 1831–1833, 2020.

[9]   A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial intelligence and machine learning to fight COVID-19," *https://doi.org/10.1152/physiolgenomics.00029.2020*, vol. 52, no. 4, pp. 200–202, Apr. 2020.

[10] G. S. Randhawa, M. P. M. Soltysiak, H. El Roz, C. P. E. de Souza, K. A. Hill, and L. Kari, "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study," *PLoS One*, vol. 15, no. 4, p. e0232391, Apr. 2020.

421 [11] L. Yan *et al.*, "A machine learning-based model for survival prediction in patients with
422   severe COVID-19 infection," *medRxiv*, p. 2020.02.27.20028027, Mar. 2020.

423 [12] Y. Ge *et al.*, "An integrative drug repositioning framework discovered a potential
424   therapeutic agent targeting COVID-19," *Signal Transduct. Target. Ther. 2021 61*, vol. 6,
425   no. 1, pp. 1–16, Apr. 2021.

426 [13] L. Li *et al.*, "Using Artificial Intelligence to Detect COVID-19 and Community-acquired
427   Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy," *Radiology*,
428   vol. 296, no. 2, pp. E65–E71, Aug. 2020.

429 [14] S. Kushwaha *et al.*, "Significant Applications of Machine Learning for COVID-19
430   Pandemic," *J. Ind. Integr. Manag.*, vol. 05, no. 04, pp. 453–479, Dec. 2020.

431 [15] P. Bachtiger, N. S. Peters, and S. L. Walsh, "Machine learning for COVID-19—asking the
432   right questions," *Lancet Digit. Heal.*, vol. 2, no. 8, pp. e391–e392, Aug. 2020.

433 [16] L. Breiman, "Bagging predictors," *Mach. Learn. 1996 242*, vol. 24, no. 2, pp. 123–140,
434   1996.

435 [17] L. Breiman, "Random Forests," *Mach. Learn. 2001 451*, vol. 45, no. 1, pp. 5–32, Oct.
436   2001.

437 [18] R. C. Team, "A language and environment for statistical computing. R Foundation for
438   Statistical Computing, Vienna, Austria." 2021.

439 [19] D. D. Sjoberg, M. Curry, M. Hannum, K. Whiting, and E. C. Zabor, "gtsummary:
440   Presentation-Ready Data Summary and Analytic Result Tables. R package version 1.3.5."
441   2020.

442 [20] A. Kassambara, "rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package
443   version 0.7.0." 2021.

444 [21] A. Signorell, "DescTools: Tools for descriptive statistics. R package version 0.99.41."
445   2021.

446 [22] H. Ishwaran and U. B. Kogalur, "Fast Unified Random Forests for Survival, Regression,
447   and Classification (RF-SRC), R package version 2.11.0." 2021.

448 [23] M. Chavent, V. Kuentz, A. Labenne, B. Liquet, and J. Saracco, "PCAmixdata:
449   Multivariate Analysis of Mixed Data. R package version 3.1." 2017.

450 [24] A. Kassambara, "ggpubr: 'ggplot2' Based Publication Ready Plots. R package version
451   0.4.0." 2020.

452 [25] L. Pan, M. Mu, P. Yang, and E. Al., "Clinical characteristics of COVID-19 patients with
453   digestive symptoms in Hubei, China: a descriptive, cross-sectional, multicenter study,"
454   *Am J Gastroenterol*, vol. 115, pp. 766–773, 2020.

455 [26] S. Villapol, "Gastrointestinal symptoms associated with COVID-19: impact on the gut
456   microbiome," *Translational Research*, vol. 226. Mosby Inc., pp. 57–69, 01-Dec-2020.

457 [27] M. Galanopoulos *et al.*, "COVID-19 pandemic: Pathophysiology and manifestations from
458   the gastrointestinal tract," *World Journal of Gastroenterology*, vol. 26, no. 31. Baishideng
459   Publishing Group Co, pp. 4579–4588, 21-Aug-2020.

460 [28] H. Ashktorab, A. Pizuorno, G. Oskroch, N. A. Fierro, Z. A. Sherif, and H. Brim,
461   "COVID-19 in Latin America: Symptoms, Morbidities, and Gastrointestinal
462   Manifestations," *Gastroenterology*, vol. 160, no. 3. W.B. Saunders, pp. 938–940, 01-Feb-
463   2021.

464 [29] G. Marasco *et al.*, "Implications of SARS-CoV-2 infection for neurogastroenterology,"
465   *Neurogastroenterology and Motility*, vol. 33, no. 3. Blackwell Publishing Ltd, p. e14104,
466   01-Mar-2021.

467 [30] M. M. Phipps *et al.*, "Acute Liver Injury in COVID-19: Prevalence and Association with
468   Clinical Outcomes in a Large U.S. Cohort," *Hepatology*, vol. 72, no. 3, pp. 807–817, Sep.
469   2020.

470 [31] C. Zhang, L. Shi, and F.-S. Wang, "Liver injury in COVID-19: management and
471   challenges," *Lancet Gastroenterol. Hepatol.*, vol. 5, no. 5, pp. 428–430, May 2020.

472 [32] G. Ndrepepa, "Aspartate aminotransferase and cardiovascular disease—a narrative
473   review," *J. Lab. Precis. Med.*, vol. 6, no. 0, pp. 6–6, Jan. 2021.

474 [33] M. Hanko *et al.*, "Random Forest-Based Prediction of Outcome and Mortality in Patients
475   with Traumatic Brain Injury Undergoing Primary Decompressive Craniectomy," *World
476   Neurosurg.*, vol. 148, pp. e450–e458, Apr. 2021.

477 [34] G. Choy *et al.*, "Current Applications and Future Impact of Machine Learning in
478   Radiology," *Radiology*, vol. 288, no. 2, pp. 318–328, Jun. 2018.

479 [35] J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in Cancer Prediction
480   and Prognosis:," *Cancer Inform.*, vol. 2, pp. 59–77, Oct. 2017.

481 [36] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *N. Engl. J. Med.*,
482   vol. 380, no. 14, pp. 1347–1358, Apr. 2019.

483   [37]   S. Wollenstein-Betech, A. A. B. Silva, J. L. Fleck, C. G. Cassandras, and I. C. Paschalidis,
484   "Physiological and socioeconomic characteristics predict COVID-19 mortality and
485   resource utilization in Brazil," *PLoS One*, vol. 15, no. 10, p. e0240346, Oct. 2020.

486   [38]   C. Iwendi *et al.*, "COVID-19 Patient Health Prediction Using Boosted Random Forest
487   Algorithm," *Front. Public Heal.*, vol. 0, p. 357, Jul. 2020.

488   [39]   J. Wang *et al.*, "A descriptive study of random forest algorithm for predicting COVID-19
489   patients outcome," *PeerJ*, vol. 8, p. e9945, Sep. 2020.

490   [40]   Y. Sun *et al.*, "Epidemiological and Clinical Predictors of COVID-19," *Clin. Infect. Dis.*,
491   vol. 71, no. 15, pp. 786–792, Jul. 2020.

492   [41]   N. Aumpan, P. Nunanan, and R. Vilaichone, "Gastrointestinal manifestation as clinical
493   predictor of severe COVID-19: A retrospective experience and literature review of
494   COVID-19 in Association of Southeast Asian Nations (ASEAN)," *JGH Open*, vol. 4, no.
495   6, pp. 1096–1101, Dec. 2020.

496   [42]   U. C. Ghoshal *et al.*, "The Spectrum of Gastrointestinal Symptoms in Patients With
497   Coronavirus Disease-19: Predictors, Relationship With Disease Severity, and Outcome,"
498   *Clin. Transl. Gastroenterol.*, vol. 11, no. 12, p. e00259, Dec. 2020.

499

Figure 1

Aspartate and alanine transaminase in hospitalized patients vs discharged home

AST: Aspartate transaminase; ALT: Alanine transaminase; p<0.001. AST activity is significantly higher in the hospitalized group compared to patients discharged home after visit to emergency department. There are no significant differences in ALT activity between this groups of patients.

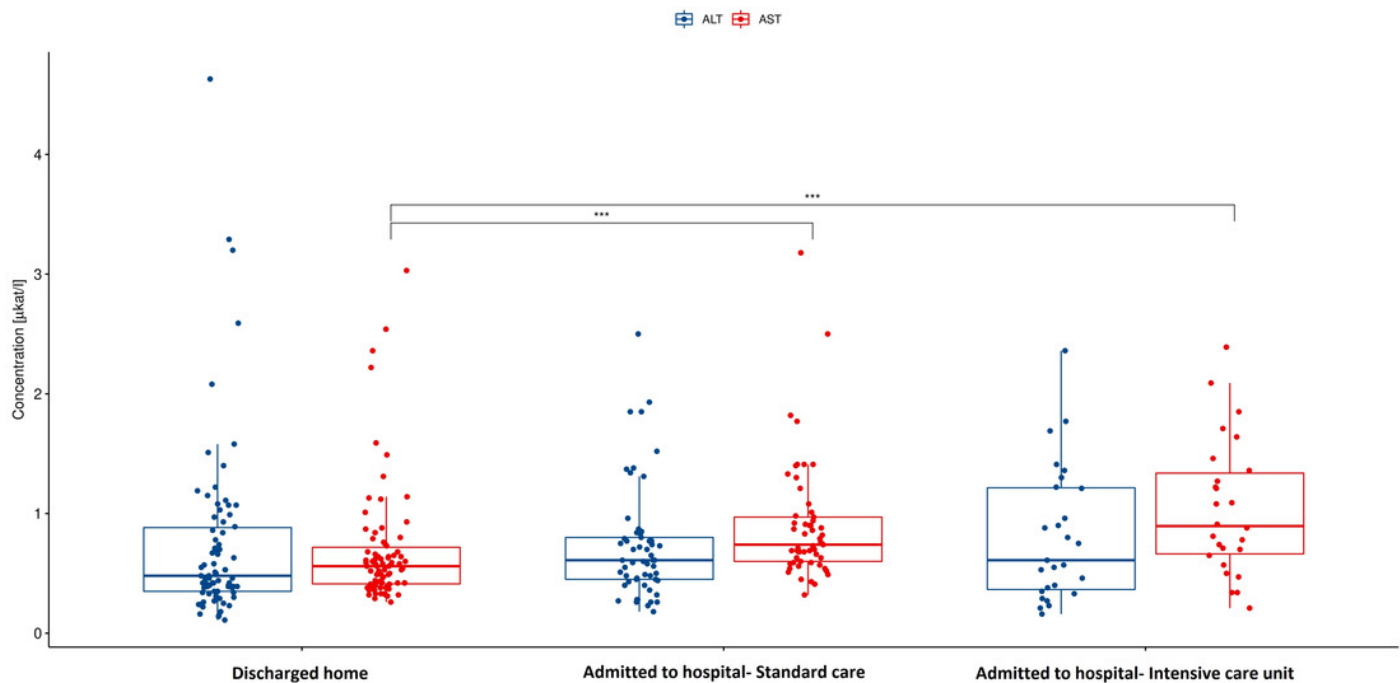



Figure 2

The ROC curve for general COVID-19 and gastrointestinal symptoms and other measurable data in general clinical settings

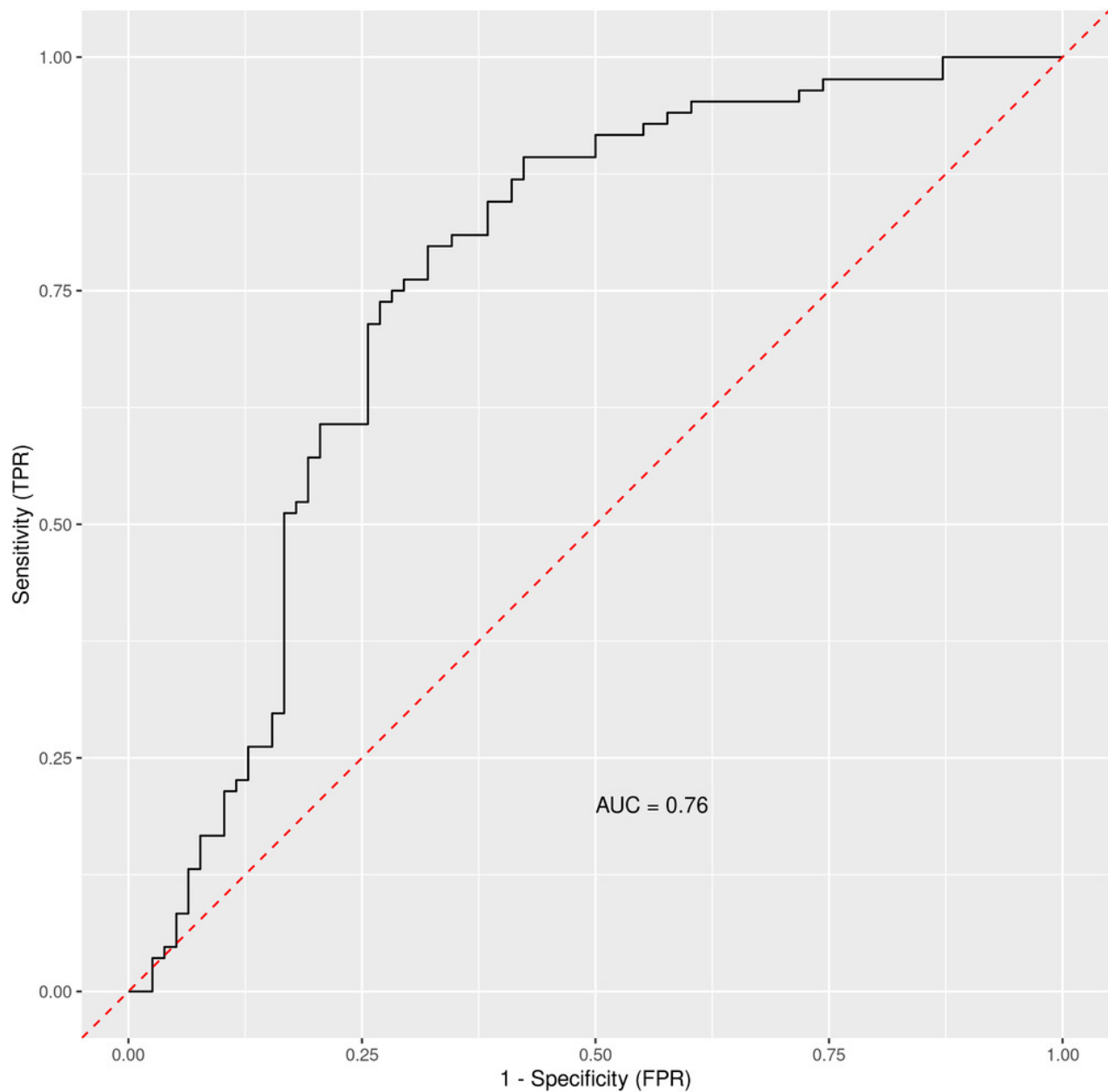
Out-of-bag receiver operating characteristic curve with calculated area under the curve (AUC)= 0.76. The Matthews correlation coefficient was 0.48. For analysis were considered: general COVID-19 symptoms, gastrointestinal symptoms, age, sex, lasting of the symptoms and comorbidities (Diabetes Mellitus, Arterial hypertension and Chronical liver diseases).

AUC = 0.76

Figure 3

The ROC curve for selected parameters

Out-of-bag receiver operating characteristic curve with calculated area under the curve (AUC)= 0.799. The Matthews correlation coefficient was 0.37. For analysis were considered selected parameters (clinically easily measurable): liver enzymes (AST, ALT), gastrointestinal symptoms (diarrhea and bloating), chronic liver disease, age and Diabetes Mellitus.

AUC = 0.799

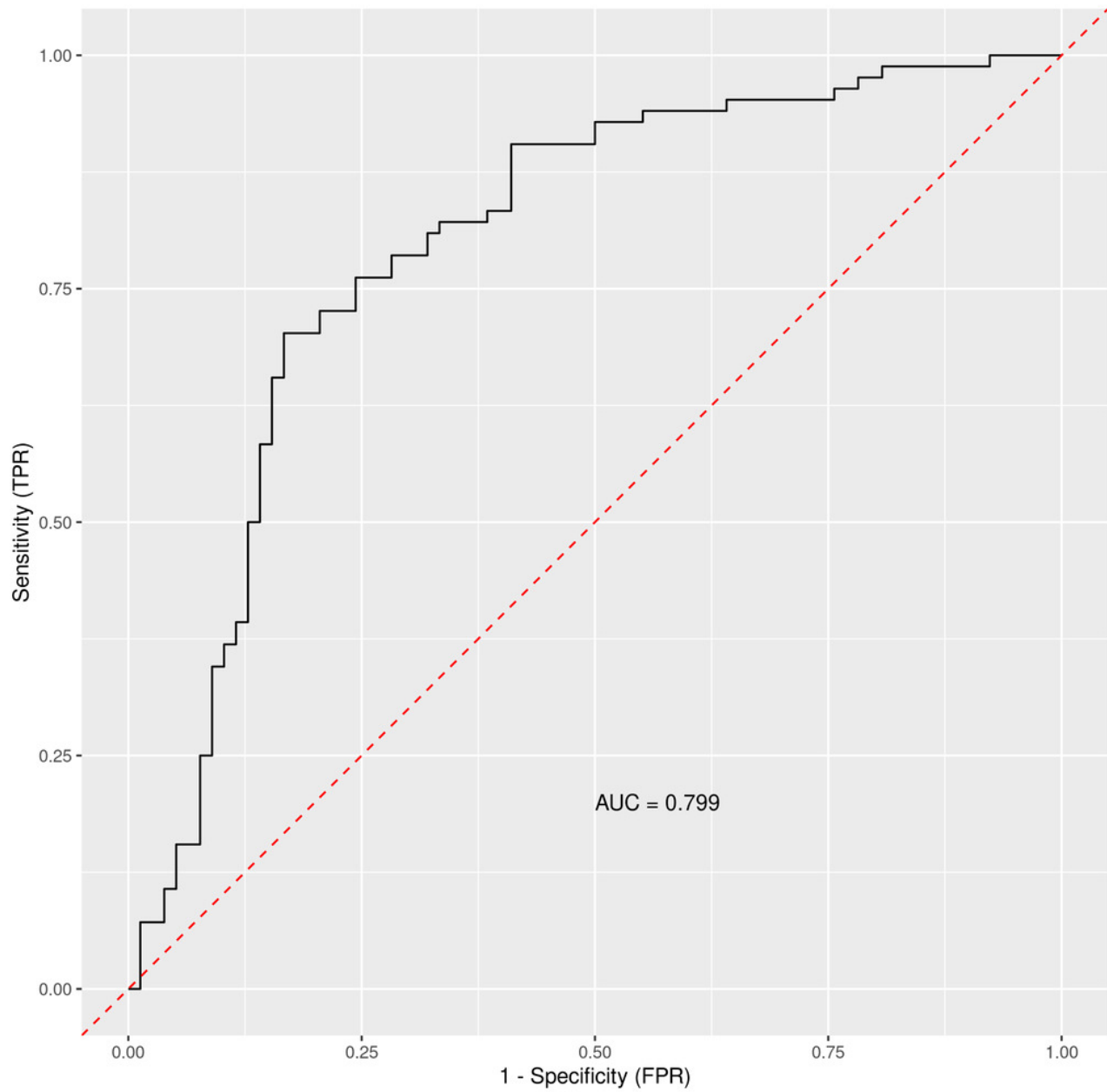



Figure 4

Principal component analysis for mixed type of data to obtain two-dimensional representation of the data

Patients who were discharged home are marked as black dots and those who were admitted to the hospital marked as red dots. The first principal component (x axis) explains 14.04% of the variability; the second principal component (y axis) explains 10.44% of the variability in data. The two groups cannot be completely separated, as there is some overlap of the observations but there is a clear tendency to shift apart of the clusters.

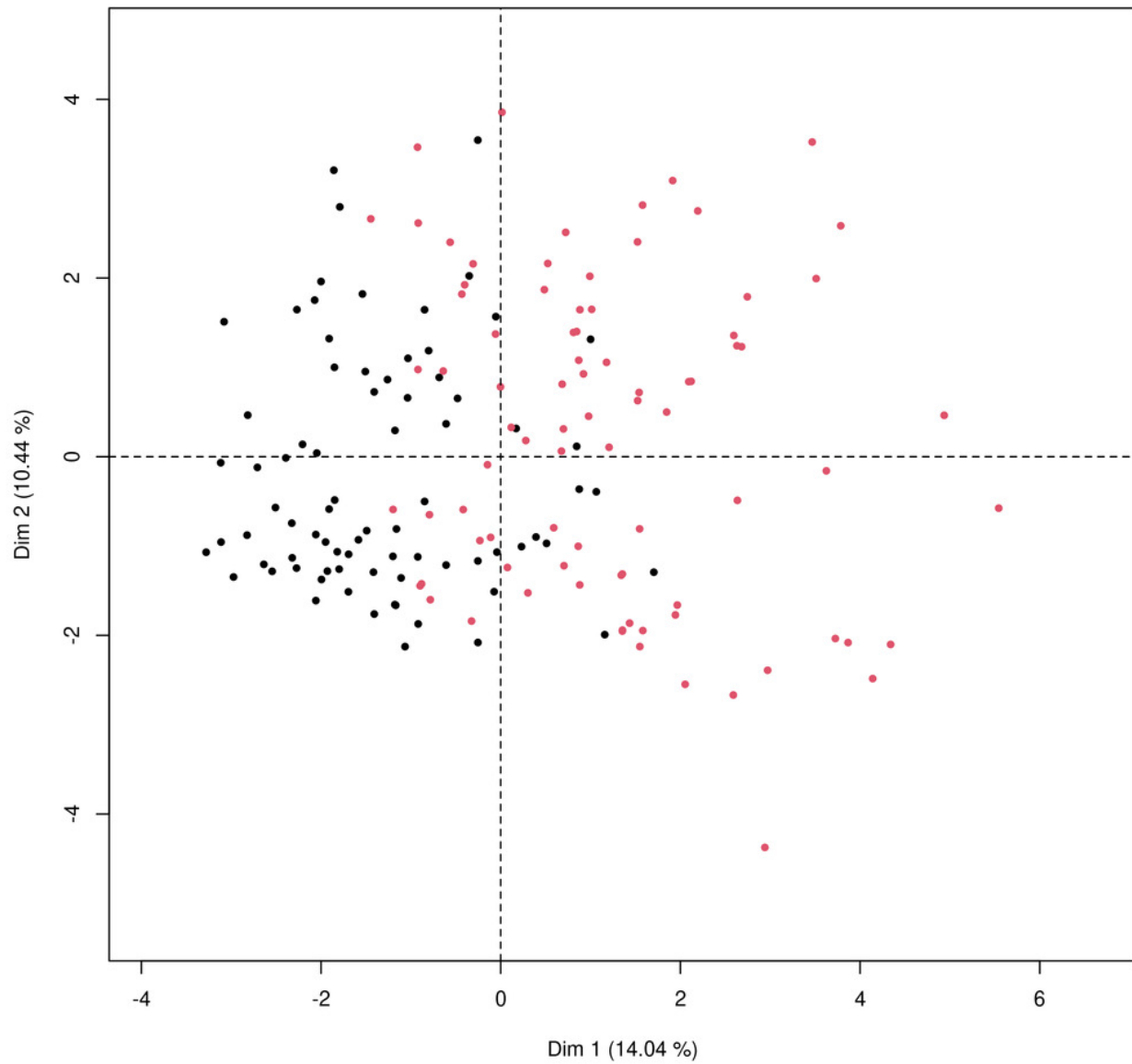



Figure 5

Variable importance plot for all measured factors

Variable importance plot for all measured factors. The positive value of importance of a predictor represents a positive factor for the predictive accurancy of random forrest algorithm. The negative value of importance of a predictor indicates that omitting the predictor increases the predictive accuracy of random forrest algorithm.

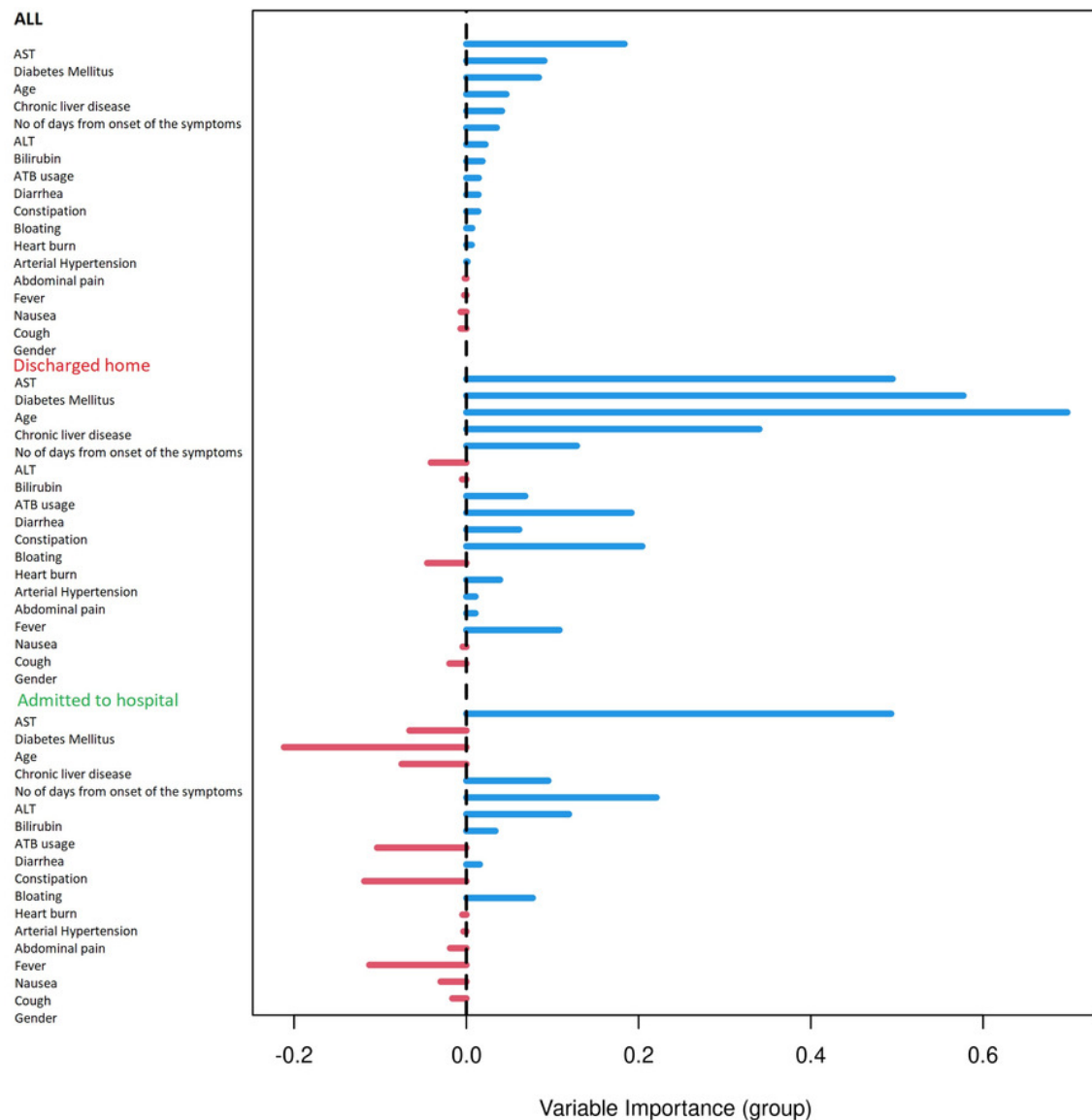



Figure 6

Variable importance plot for selected factors

Variable importance plot for selected factors that are fast and easy to measure in the emergency department setting (liver enzymes: AST and ALT, gastrointestinal symptoms /diarrhea and bloating/, age and presence of chronic liver disease and diabetes mellitus). The positive value of importance of a predictor represents a positive factor for the predictive accurancy of random forrest algorithm. The negative value of importance of a predictor indicates that omitting the predictor increases the predictive accuracy of random forrest algorithm.

all

AST
Age
Diabetes Mellitus
Chronic liver disease
ALT
Diarrhea
Bloating

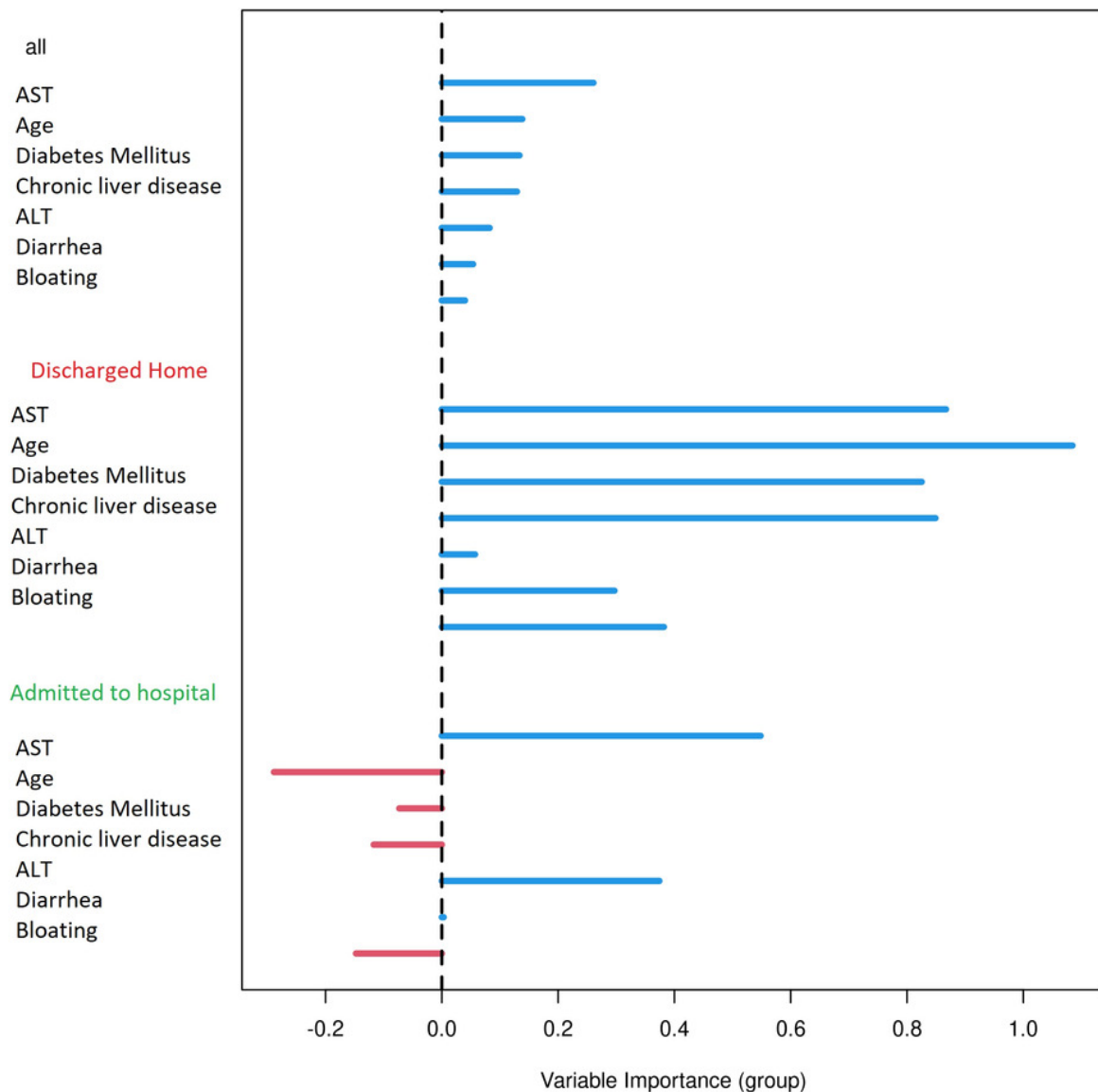
Discharged Home

AST
Age
Diabetes Mellitus
Chronic liver disease
ALT
Diarrhea
Bloating

Admitted to hospital

AST
Age
Diabetes Mellitus
Chronic liver disease
ALT
Diarrhea
Bloating

-0.2   0.0   0.2   0.4   0.6   0.8   1.0

Variable Importance (group)

Patients characteristics

[1]Statistics presented: n (%); Median (IQR). [2] Statistical tests performed: Kruskal-Wallis test; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

1 **Table 1**

		Outpatient test center (SARS-CoV-2 negative) (n=352)[1]	Outpatient test center (SARS-CoV-2 positive) (n=166)[1]	Discharged home (n=78)[1]	Admitted to hospital-standard care (n=57)[1]	Admitted to hospital-ICU (n=27)[1]	p value[2]
Age		42 (31, 51)	41 (32, 52)	48 (42, 58)	55 (46, 67)	65 (55, 72)	<0.001
Gender							0.07
	female	206 (59%)	104 (63%)	41 (53%)	27 (47%)	11 (41%)	
	male	141 (40%)	62 (37%)	37 (47%)	30 (53%)	16 (59%)	
	NA	5 (1%)					
Fever							<0.001
	Not-presented	340 (97%)	146 (88%)	48 (62%)	25 (44%)	15 (56%)	
	Presented	12 (3.4%)	20 (12%)	30 (38%)	32 (56%)	12 (44%)	
Cough							<0.001
	Not-presented	342 (97%)	144 (87%)	48 (62%)	24 (42%)	14 (52%)	
	Presented	10 (2.8%)	22 (13%)	30 (38%)	33 (58%)	13 (48%)	
Diarrhea							<0.001
	Not-presented	330 (94%)	141 (85%)	64 (82%)	34 (60%)	19 (70%)	
	Presented	22 (6.2%)	25 (15%)	14 (18%)	23 (40%)	8 (30%)	
Constipation							0.37
	Not-presented	347 (99%)	164 (99%)	76 (97%)	55 (96%)	26 (96%)	
	Presented	5 (1.4%)	2 (1.2%)	2 (2.6%)	2 (3.5%)	1 (3.7%)	
Bloating							0.13
	Not-presented	341 (97%)	161 (97%)	76 (97%)	51 (89%)	26 (96%)	
	Presented	11 (3.1%)	5 (3.0%)	2 (2.6%)	6 (11%)	1 (3.7%)	
Nausea							<0.001
	Not-presented	340 (97%)	147 (89%)	70 (90%)	45 (79%)	23 (85%)	
	Presented	12 (3.4%)	19 (11%)	8 (10%)	12 (21%)	4 (15%)	
Heart burn							0.86
	Not-presented	341 (97%)	163 (98%)	76 (97%)	56 (98%)	26 (96%)	
	Presented	11 (3.1%)	3 (1.8%)	2 (2.6%)	1 (1.8%)	1 (3.7%)	
Abdominal pain							0.07
	Not-presented	334 (95%)	150 (90%)	71 (91%)	49 (86%)	26 (96%)	
	Presented	18 (5.1%)	16 (9.6%)	7 (9.0%)	8 (14%)	1 (3.7%)	

2 [1]Statistics presented: n (%); Median (IQR).
3 [2] Statistical tests performed: Kruskal-Wallis test; Fisher's Exact Test for Count Data with simulated p-value (based
4 on 2000 replicates)

39

40

41

42

43

44

45

46

47

48

49

50

Gastrointestinal symptoms occurrence and laboratory findings

[1]Statistics presented: n (%); Median (IQR). [2] Statistical tests performed: chi-square test of independence; Kruskal-Wallis test; Fisher's exact testCRP: C- reactive protein; AST: aspartate transaminase; ALT: alanine transaminase; ATB: antiobiotics

1 **Table 2**

		Discharged home (n=78)[1]	Admitted to hospital-standard care (n=57)[1]	Admitted to hospital-ICU (n=27)[1]	p value[2]
Gender					0.55
	female	41 (53%)	27 (47%)	11 (41%)	
	male	37 (47%)	30 (53%)	16 (59%)	
No of days with symptoms		7 (5,10)	7 (5.10)	5 (2,7)	0.03
Age		48 (42, 58)	55 (46, 67)	65 (55, 72)	<0.001
Diarrhea					0.02
	Not-presented	64 (82%)	34 (60%)	19 (70%)	
	Presented	14 (18%)	23 (40%)	8 (30%)	
Abdominal pain					0.23
	Not-presented	71 (91%)	48 (84%)	26 (96%)	
	Presented	7 (9.0%)	9 (16%)	1 (3.7%)	
Vomitus					0.24
	Not-presented	65 (83%)	53 (93%)	23 (85%)	
	Presented	13 (17%)	4 (7.0%)	4 (15%)	
Nausea					0.22
	Not-presented	70 (90%)	45 (79%)	23 (85%)	
	Presented	8 (10%)	12 (21%)	4 (15%)	
Heart burn					0.83
	Not-presented	76 (97%)	56 (98%)	26 (96%)	
	Presented	2 (2.6%)	1 (1.8%)	1 (3.7%)	
Bloating					0.03
	Not-presented	76 (97%)	49 (86%)	26 (96%)	
	Presented	2 (2.6%)	8 (14%)	1 (3.7%)	
Constipation					>0.9
	Not-presented	76 (97%)	55 (96%)	26 (96%)	
	Presented	2 (2.6%)	2 (3.5%)	1 (3.7%)	
O2 saturation [%]		96 (95, 98)	93 (89, 96)	87 (83, 91)	<0.001
CRP [norm < 5 mg/l]		29 (8, 74)	99 (65, 140)	102 (56, 166)	<0.001
AST [norm < 0.6 ukat/l]		0.56 (0.41, 0.72)	0.74 (0.60, 0.97)	0.91 (0.68, 1.41)	<0.001
ALT [norm < 0.6 ukat/l]		0.48 (0.35, 0.88)	0.61 (0.45, 0.80)	0.61 (0.36, 1.21)	0.25
Bilirubin [norm <21 ukat/l]		8.8 (7.1, 11.1)	9.2 (8.3, 13.9)	9.5 (7.8, 12.5)	0.25
Recent ATB usage					0.09
	No	45 (58%)	22 (39%)	13 (48%)	
	Yes	33 (42%)	35 (61%)	14 (52%)	
Diabetes Mellitus					<0.001
	No	72 (92%)	45 (79%)	15 (56%)	

	Yes	6 (7.7%)	12 (21%)	12 (44%)	
Arterial Hypertension					0.003
	No	49 (63%)	26 (46%)	7 (26%)	
	Yes	29 (37%)	31 (54%)	20 (74%)	
Chronic liver disease	No	72 (92%)	47 (82%)	22 (81%)	
	Yes	6 (8%)	10 (18%)	5 (19%)	

3 [1]Statistics presented: n (%); Median (IQR).

4 [2] Statistical tests performed: chi-square test of independence; Kruskal-Wallis test; Fisher's exact test

5 CRP: C- reactive protein; AST: aspartate transaminase; ALT: alanine transaminase; ATB: antiobiotics

Relative predictors values

AST: aspartate transaminase; ALT: alanine transaminase

1   **Table 3**

	Variable importance		
	All	Discharged home	Admitted to hospital
AST	0.1451	0.5729	0.2217
Diabetes Mellitus	0.0248	0.1080	0.0288
Chronic liver disease	0.0169	0.0882	0.0061
ALT	0.0110	0.1025	-0.0379
Diarrhea	0.0068	0.0087	0.0272
Age	0.0139	0.0561	0.0197
Bloating	0.0011	0.0052	0.0006

2   AST: aspartate transaminase; ALT: alanine transaminase