CAN DEEP LEARNING PREDICT LABORATORY VALUES IN COVID-19?

NAZLIM AKTUĞ DEMİR[1], Onur Ural[1], Asli Ural[2], Sua Sumer[1], Hatice Esranur Kiratli[1], Lutfi Saltuk Demir[3], Ediz Uslu[2], Fikret Kanat[4], Ugur Arslan[1], Husamettin Vatansev[1], and Hakan Cebeci[1]

[1]Selcuk University Faculty of Medicine
[2]JotForm Yazilim A.S.
[3]Necmettin Erbakan University Faculty of Medicine
[4]Selcuk University, Faculty of Medicine

July 15, 2021

Abstract

Aims: Laboratory findings in COVID-19 patients vary according to the severity of the disease. This study aimed at defining a system of formulas that may predict the presence of thoracic CT involvement, the extent of such involvement and the need for intensive care stay on the basis of patient laboratory data using the Waikato Environment for Knowledge Analysis (WEKA) software. Methods: This study was conducted with 508 patients whose SARS-CoV-2 RT-PCR test was positive. These patients were divided into 2 groups, with and without thoracic CT involvement typical for COVID-19. Then, those patients who had signs of typical involvement for COVID-19 in their thoracic CT were divided into 3 groups depending on the extent of their lesions. J48 Decision Tree classification and Linear Regression methods were used on the WEKA software. The codes implemented in the Python programming language were used at the estimation, classification and testing stages. Results: Thoracic CT scans showed that lung involvement was absent in 93 of the patients, mild in 114, moderate in 115, and severe in 159. The success rates of WEKA Linear Regression Formulas calculated using laboratory values and demographic data, respectively 78.92%, 71.69% and 91%. The success rate of the J48 Decision Tree formula used to predict the presence of involvement in thoracic CT was found to be 95.95%. The success rate of the J48 Decision Tree, which was used to predict the degree of involvement in thoracic CT, was 84.39%. The success rate of the J48 Decision Tree used to predict the need for intensive care was found to be 93.06%. Conclusion: The results of this study will facilitate revealing the presence of lung involvement and identification of critical patients in the COVID-19 pandemic and particularly under circumstances and can be used effectively to ensure triage.

CAN DEEP LEARNING PREDICT LABORATORY VALUES IN COVID-19?LAB VALUES IN COVID-19 BY DEEP LEARNINGAims: Laboratory findings in COVID-19 patients vary according to the severity of the disease. This study aimed at defining a system of formulas that may predict the presence of thoracic CT involvement, the extent of such involvement and the need for intensive care stay on the basis of patient laboratory data using the Waikato Environment for Knowledge Analysis (WEKA) software.**Methods:** This study was conducted with 508 patients whose SARS-CoV-2 RT-PCR test was positive. These patients were divided into 2 groups, with and without thoracic CT involvement typical for COVID-19. Then, those patients who had signs of typical involvement for COVID-19 in their thoracic CT were divided into 3 groups depending on the extent of their lesions. J48 Decision Tree classification and Linear Regression methods were used on the WEKA software. The codes implemented in the Python programming language were used at the estimation, classification and testing stages.**Results:** Thoracic CT scans showed that lung involvement was absent in 93 of the patients, mild in 114, moderate in 115, and severe in 159. The success rates of WEKA Linear Regression Formulas calculated using laboratory values

and demographic data, respectively 78.92%, 71.69% and 91%. The success rate of the J48 Decision Tree formula used to predict the presence of involvement in thoracic CT was found to be 95.95%. The success rate of the J48 Decision Tree, which was used to predict the degree of involvement in thoracic CT, was 84.39%. The success rate of the J48 Decision Tree used to predict the need for intensive care was found to be 93.06%.**Conclusion:** The results of this study will facilitate revealing the presence of lung involvement and identification of critical patients in the COVID-19 pandemic and particularly under circumstances and can be used effectively to ensure triage.**Key words:** COVID-19; deep learning; laboratory; WEKA

What's known

1. COVID-19 may cause a variety of clinical conditions from an asymptomatic disease to severe viral pneumonia that can result in respiratory failure or death.
2. The gold standard diagnostic test in the diagnosis of COVID-19 is RT-PCR. Lung imaging is quite valuable, but having access to thoracic CT is not always possible.
3. Deep learning and artificial intelligence started to be used in many areas in medicine with a high diagnostic accuracy.

What's new

Formulas developed with WEKA software with Linear Regression and J48 Decision Tree algorithms can be used to determine the status of patients.

With these formulas, the presence and level of thoracic CT involvement and the need for intensive care can be predicted by using the laboratory data of COVID-19 patients.

INTRODUCTIONCoronavirus disease 2019 (COVID-19) is an alarming public health concern worldwide. This disease may cause a variety of clinical conditions from an asymptomatic disease to severe viral pneumonia that can result in respiratory failure or death.[1]Presence of accurate and fast diagnostic tests is of clinical importance in controlling the COVID-19 pneumonia. The gold standard diagnostic test in the diagnosis of COVID-19 is the reverse transcriptase polymerase chain reaction (RT-PCR). However, this test has several limitations such as potential false negative results, high cost, difficulties in collecting, storing and transport of sample materials. Serological tests have also attracted attention as alternative or complementary to RT-PCR and other nucleic acid tests in the diagnosis of acute infections.[2] However, they are not useful to diagnose acute cases as the IgM antibody response can be detected 6-15 days after the disease onset.[3] Studies have reported that the sensitivity of fast SARS-CoV-2 antigen diagnosis tests is between 45 and 97%.[4] Lack of common standardization among serological tests causes difficulties in the use of the test. As another diagnostic method in this disease, lung imaging is quite valuable, but having access to thoracic CT is not always possible. Moreover, early interpretation of this imaging, which is valuable for the diagnosis of COVID-19, requires an extra workload. This situation necessitates methods that can predict lung involvement and provide high diagnostic performance.[5]Another problem with this disease is that the patients experience unpredictable progression, giving rise to a need for intensive care. Although the intensive care need of COVID-19 patients varies by country and institution, it ranges between approximately 5 and 32%.[6] It is important in terms of disease morbidity and mortality to predict patients' intensive care needs and to transfer such patients quickly from healthcare institutions having no intensive care facilities.[7]There is a need for defining clinical and laboratory predictors enabling early prediction of progression to serious clinical form and intensive care need when combating this disease which puts severe stress on many healthcare systems across the world. Defining such predictors will enable risk classification so that patients having high risk of developing severe disease can be identified and necessary interventions initiated at an early stage. This will optimise human and technical resources during this ongoing pandemic.[8] It could be a good alternative in support of physicians to define, with the help of the deep learning technology, a system consisting of laboratory parameters to be used to predict presence of lung involvement, the extent of such involvement and intensive care need in COVID-19 patients. In recent years, technologies such as deep learning and artificial intelligence started to be used in many areas in medicine with a high diagnostic accuracy.[9,10] Although there are studies using these methods in the thoracic CT interpretation of COVID-19, we have not encountered any study on the use of these technologies for the assessment and interpretation of laboratory parameters in COVID-19. The present study aimed at defining a system of formulas that may predict the presence of thoracic CT involvement, the extent of such involvement and the need for intensive care stay on the basis of patient laboratory data using the Waikato Environment for Knowledge Analysis (WEKA) software, Linear Regression and J48 Decision Tree algorithms and that may in this way contribute to the process of diagnosing and treating the disease.

MATERIALS AND METHODS

Dataset

This study was conducted between 01.07.2020 - 15.02.2021 with 508 patients whose SARS-CoV-2 RT-PCR test was positive and who were being monitored and treated for COVID-19 diagnosis in XXXXXX University Medical School Hospital. Patients aged 18 years and older who were not pregnant and had no underlying lung disease were included in the study. These patients were divided into 2 groups, with and without thoracic CT involvement typical for COVID-19. Then, those patients who had signs of typical involvement for COVID-19 in their thoracic CT (ground glass opacities, consolidation or both) were divided into 3 groups depending on the extent of their lesions.

Mild involvement : Patients with 33% or less thoracic CT involvement

Moderate involvement: Patients with 34-66% thoracic CT involvement

Severe involvement: Patients with 67% or more thoracic CT involvement

The demographic characteristics of the patients and the laboratory values routinely required from the patients including hemogram, CRP, D-dimmer, ferritin, LDH, INR, fibrinogen, ALT, AST, troponin, CK, PCT and thoracic CTs were obtained from our hospital's database.

Ethical Concerns

This study was conducted in accordance with the Helsinki Declaration (2000) of the World Medical Association. Permissions from the Ethics Committee (2020/12) and the Ministry of Health were obtained for the study. This study was supported by the Scientific Research Projects Automation of XXXXXXX University (Project No: 20301005)

Objective

This study intends to estimate and formulate presence of thoracic CT involvement, extent of such involvement and intensive care need on the basis of patient laboratory data to contribute to the process of diagnosing and treating the disease. At the stage of formulation, the J48 Decision Tree classification and Linear Regression methods were used on the WEKA software. The codes implemented in the Python programming language were used at the estimation, classification and testing stages.

Data Mining

Data mining includes all the stages of extracting meaningful data from a large dataset, formulation, sensemaking, and purposeful use. The data used through the data mining stages are stored in large data warehouses and databases. The data used in this study were obtained from the databases of Selcuk University Medical School Hospital.

Preparation before Data Processing

A preliminary preparation work was conducted on the dataset first to be able to implement linear regression and J48 Decision Tree algorithms. Identity details such as name, surname and identity number were cleared off the dataset. At the data selection stage, the laboratory parameters intended to be used in COVID-19 diagnosis process were identified in the dataset with the help of doctors specialized in their fields.

Finally, the binary and nominal values in the dataset were converted into numeric data to be able to run the Linear Regression algorithm.

Absence of any missing values in the dataset used in the study provided advantage for the rest of the work.

WEKA (Waikato Environment for Knowledge Analysis)

Used in data mining, WEKA is an open-source software developed in Waikato University. This data mining instrument provides ready-made algorithms for data preparation and machine learning. It also provides tools for data and result visualization.

This study used the Nominal to Numeric and Binary to Numeric algorithms of WEKA at the dataset preparation stage. The Linear Regression and J48 Decision Tree algorithms were also run on WEKA.

Supervised Learning

In data mining, various algorithms are used to process large amounts of data. Designed to make estimations using data, these algorithms are divided into 2, Supervised and Unsupervised Learning. The Linear Regression and J48 algorithms used in this study are Supervised Learning algorithms. The main objective in Supervised Learning Algorithms is to infer, by training the dataset, a formula which has inputs and outputs and can be shown mathematically as y=f(x), In this formula x represents the input data, f the entire mathematical procedure applied to the data, and y estimated results from x.

Supervised Learning Algorithms are used in 2 areas, regression and classification. Both aim to form a model that can train itself using the dataset and make estimations.

Decision Tree

A decision tree is a supervised learning algorithm consisting of decision nodes based on classification, feature and target, and leaf nodes, all of which form a tree-like model. One of the most widely known decision tree algorithms is the Iterative Dichotomiser 3 (ID3). A substitute of this algorithm in WEKA is J48. The major difference of J48 compared to ID3 is its ability to make statistical normalization. For this reason, entropy is calculated in the J48 algorithm. This calculation is kept as a proportion. A pruning procedure is also carried out at the end to obtain the simplest form of the tree.

It chooses the features of data that divide the sample set to enriched subsets most effectively in any of classes for each node on the tree.

This algorithm is based on recursive classification of information in a dataset. Eventually, the information normalized most is chosen for decision making.

When the basic of algorithms are considered, they are thought to provide a correct solution path for a problem. First, when all the samples in a dataset belong to the same class, it forms a leaf node that tells the decision tree to choose that class. If the chosen feature does not provide any information to make a decision, a decision node is formed on the tree using the expected value. Finally, if the tree encounters a class sample it has not seen before, it forms a decision node at the top of the tree. A summary of the process is as follows: First, all algorithm steps are implemented with iterations. At each iteration, features in the dataset are reviewed. A value is calculated for each feature. This value is named as information gain. The best one of these values is assigned as decision and added to the tree. This process is continued by forming new decision routes from under the node added latest.

Application on WEKA

After opening the dataset on WEKA, the J48 algorithm was run. The dataset was divided as training and testing in this process. WEKA offers two options for this. The first is the cross-validation; the percentage of the dataset determined based on this value is allocated for testing. The second one, percentage split, tests to what extent the classification based on the percentage we assign continues successfully.

Interpretation of the Result and Visualization of the Tree

The tree obtained as a result of the J48 algorithm shows us according to what values the dataset of this algorithm was classified. The node at the very top of the tree is presented as the most important decision maker. Starting from the top, the tree moves down to lower nodes in terms of correct decision coefficient value.

The algorithm was run to predict thoracic CT involvement, extent of such involvement and need for intensive care. As a result of this, three separate decision trees were formed and illustrated.

After the dataset was compiled on WEKA, the J48 Decision Tree and linear regression algorithms were run in our study. In this process, the dataset was divided into training and testing for each algorithm. Cross-validation was checked as testing criterion of the algorithms. Based on this value, the percentage of the dataset at the specified value was assigned as testing. Percentage split was used as the second testing criterion. The classification performance rate was compared based on this.

The linear regression algorithm, which was chosen as the second step, is a statistical method enabling modelling and formulation of the relationship between the regression dependent and independent variables. When there are more than one value as input parameter in the regression method, the Linear Regression method is used. A formula is obtained as a result of this algorithm. When calculating the formula, the square of the difference between the real value and the estimated value is taken to get the absolute value. After doing this for each sample, the absolute values of the difference between the real and estimated values are added for all samples. Minimizing this total value gives the optimum correct formula.

RESULTS

The mean age of the 508 patients whose SARS-CoV-2 RT-PCR test turned out positive was 54 (19-99). Of these patients, 250 (49.2%) were female and 258 (50.8%) male.

Thoracic CT scans showed that lung involvement was absent in 93 of the patients, mild in 114, moderate in 115, and severe in 159 (Table 1).

The WEKA Linear Regression Formulas calculated using the laboratory values and demographic data (Formulas 1, 2, 3) and the J48 Decision Trees (Figures 1, 2, 3) are given below.

Formula 1: The formula constructed using the WEKA Linear Regression Formula to estimate presence of Thoracic CT involvement in patients

0.2537 * gender +0.0267 * age +-0.0001 * leukocyte +0.0001 * neutrophil +0.0001 * lymphocyte+0.0031 * CRP +-0.0029 * PCT +-0.0003 * ferritin +0.001 * fibrinogen +0.0013 * ALT +-0.0005 * AST +0.0004 * LDH+-0.0143 * creatine +-0.2005 * INR +0 * D-dimer +- 0.0868 *

The formula's rate of success in estimating the presence of thoracic CT involvement turned out to be 78.92%.

Formula 2: The formula constructed using the WEKA Linear Regression Formula to estimate the extent of Thoracic CT involvement in patients

0.2428 * gender +0.0265 * age +-0.0001 * leukocyte +0.0001 * neutrophil +0.0001 * lymphocyte+0,003 * CRP +-0,003 * PCT +-0.0003 * ferritin +0.001 * fibrinogen +0.0014 * ALT +-0.0006 * AST +0.0005 * LDH+-0.0143 * creatine +-0.1966 * INR +0 * D-dimer +-0.5318

The formula's rate of success in estimating the extent of thoracic CT involvement turned out to be 71.69%.

Formula 3: The formula constructed using the WEKA Linear Regression Formula to estimate whether or not intensive care would be needed for the patients with CT involvement

-0.0619 * CT +0.0033 * age +0.0001 * leukocyte +-0.0001 * neutrophil +0.0003 * lymphocyte+0.0009 * CRP +-0.0004 * fibrinogen +-0.0049 * APTT +0 * D-dimer +0.5923 * ex +-0.1209

The formula's rate of success in estimating whether or not intensive care would be needed for patients with CT involvement turned out to be 91%.

The J48 decision tree algorithms estimating presence of thoracic CT involvement, the extent of such involvement and whether or not intensive care would be needed for patients are given.

The success rate of the J48 Decision Tree formula used to predict the presence of involvement in thoracic CT was found to be 95.95%. The success rate of the J48 Decision Tree, which was used to predict the degree of involvement in thoracic CT, was 84.39%. The success rate of the J48 Decision Tree used to predict the need for intensive care was found to be 93.06%.

DISCUSSION

A literature search showed that there were no clinical studies on this subject. Our study is the first one conducted on this subject. The closest to our study was the one made by Alakuş et al.[11] where they developed clinical models predicting COVID-19 using a hypothetical deep learning model based on the laboratory data of COVID-19 patients. They tested 18 laboratory findings of 600 patients and found that their experimental estimation model detected the presence of COVID-19 with 86.6% accuracy, 86.7% specificity and 62.5% AUC.

Using the WEKA Linear Regression Formula, the formulas found in our study estimated the presence of involvement on thoracic CT at a rate of 78.92%, the extent of involvement on thoracic CT at 71.69%, and the need for intensive care in patients with thoracic CT involvement at 91%. Using the J48 Decision Tree Formula, the formulas found in the study estimated the presence of involvement on thoracic CT at a rate of 95.95%, the extent of involvement on thoracic CT at 84.39%, and the need for intensive care at 93.06%. A major limitation of this study is that the comorbidities of patients could not be assessed.

In the COVID-19 pandemic regions, it is important to diagnose the disease quickly and obtain medical resources. Limited medical resources have become a huge problem in pandemic regions. It is also quite important to determine the severity of the disease and identify priorities in treatment. The WEKA Linear Regression Formulas and J48 Decision Trees that were constructed with the help of the deep learning method in our study were able to estimate the presence of lung involvement, the extent of such involvement and the need for intensive care in COVID-19 patients at a high rate. We think that the results of this study will facilitate revealing the presence of lung involvement and identification of critical patients in the COVID-19 pandemic and particularly under circumstances where cases exponentially increase and resources are restricted, and can be used effectively to ensure triage.

REFERENCES

1. Jin Y, Yang H, Ji, W, et al. Virology, Epidemiology, Pathogenesis, and Control of COVID-19. *Viruses* . 2020;12:372.
2. Sidiq Z, Hanif M, Dwivedi KK, Chopra KK. Benefits and limitations of serological assays in COVID-19 infection. *Indian J Tuberc* . 2020;67:S163-S166.
3. Venter M, Richter K. Towards effective diagnostic assays for COVID-19: a review. *J Clin Pathol* . 2020;73:370-377.
4. Albert E, Torres I, Bueno F, et al. Field evaluation of a rapid antigen test (Panbio COVID-19 Ag Rapid Test Device) for COVID-19 diagnosis in primary healthcare centres. *Clin Microbiol Infect* . 2021;27: 472.e7-472.e10.
5. Farias LPG, Fonseca EKUN, Strabelli DG, et al. Imaging findings in COVID-19 pneumonia. *Clinics (Sao Paulo)* . 2020;75:e2027.
6. Halacli B, Kaya A, Topeli A. Critically-ill COVID-19 patient.*Turk J Med Sci* . 2020;50:585-591.
7. Phua J, Weng L, Ling L, et al. Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations.*Lancet Respir Med* . 2020;8:506-517.

8. Henry BM, de Oliveira MHS, Benoit S, Plebani M, Lippi G. Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): a meta-analysis. *Clin Chem Lab Med* . 2020;58:1021-1028.
9. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* 2018;18:500-510.
10. Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology.*Cancer Commun (Lond)* . 2020;40:154-166.
11. Alakus TB, Turkoglu I. Comparison of deep learning approaches to predict COVID-19 infection. *Chaos Solitons Fractals* . 2020;140:110120.

Table 1 . Patients' demographic data with respect to thoracic CT involvement

	No involvement (n: 93)	**Mild (n:141)**	**Moderate (n:115)**	**Severe (n:159)**	**P value**
Age	41 (19-62)	45 (19-76)	64 (31-90)	69 (23-99)	**0.001**
Female Male	45 (48.3%) 48 (51.7%)	67 (32.7%) 74 (35.2%)	65 (31.7%) 50 (23.8%)	73 (35.6%) 86 (41.0%)	0.191

Figure 1. J48 Decision Tree used to predict presence of involvement on thoracic CT

The formula's rate of success in estimation turned out to be 95.95%. It made the right decision in 166 of 173 data in the testing dataset. Cross-validation was checked as testing criterion.

Figure 2. J48 Decision Tree used to predict the extent of involvement on thoracic CT

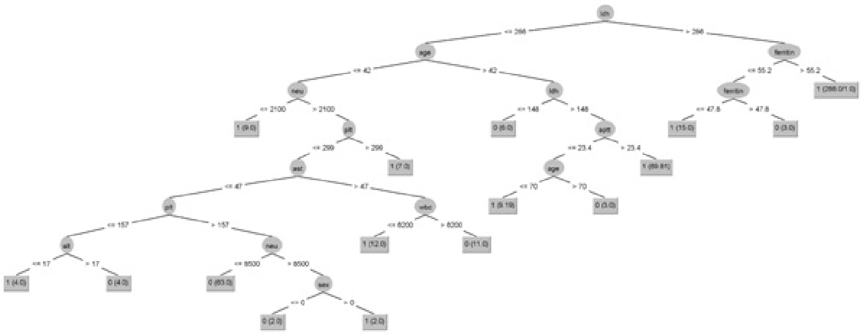
The formula's rate of success in estimation turned out to be 84.39%. Cross-validation was checked as testing criterion.

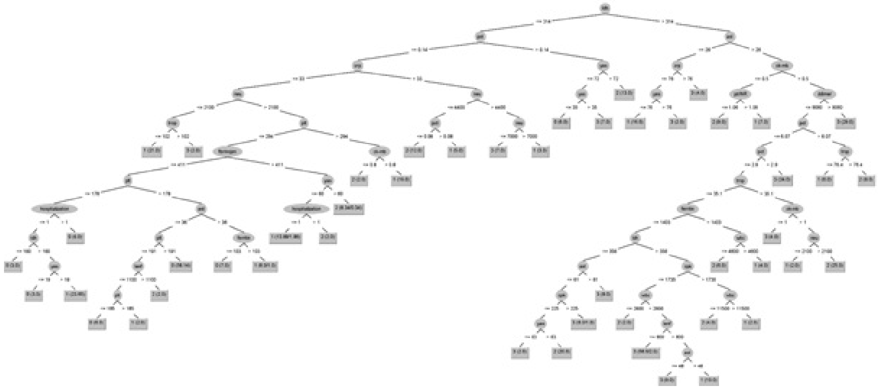
Figure 3. J48 Decision Tree used to predict intensive care need

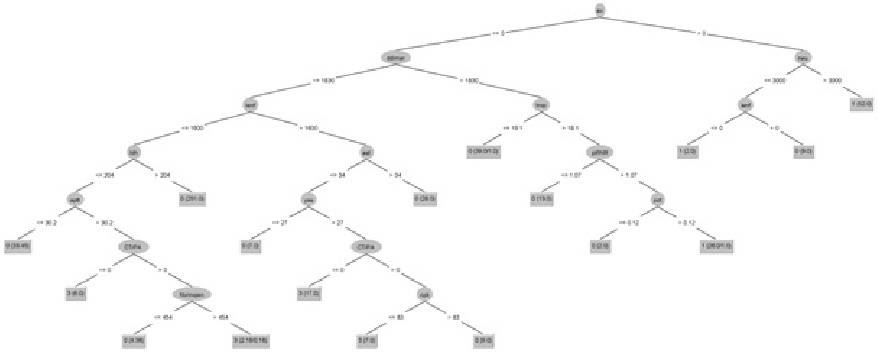
The formula's rate of success in estimation turned out to be 93.06%. Cross-validation was checked as testing criterion.

Hosted file

`Table 1.docx` available at https://authorea.com/users/425895/articles/530585-can-deep-learning-predict-laboratory-values-in-covid-19




Hospitalization
fibrinogen
Hospitalization


ddimer
<= 0
> 0
<= 1830
> 1830
<= 3000
> 3000
1 (52.0)
<= 1600
> 1600
<= 19.1
> 19.1
<= 0
> 0
0 (39.0/1.0)
1 (2.0)
0 (9.0)
<= 204
> 204
<= 34
> 34
<= 1.07
> 1.07
0 (251.0)
0 (28.0)
0 (15.0)
<= 30.2
> 30.2
<= 27
> 27
<= 0.12
> 0.12
0 (55.45)
CT/PA
0 (7.0)
CT/PA
0 (2.0)
1 (26.0/1.0)
<= 0
> 0
<= 0
> 0
3 (6.0)
fibrinogen
3 (17.0)
<= 454
> 454
<= 83
> 83
0 (4.96)
3 (2.18/0.18)
3 (7.0)
0 (9.0)