Universidad del País Vasco
Euskal Herriko Unibertsitatea

ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

FACULTY OF SCIENCE AND TECHNOLOGY. LEIOA

BACHELOR'S FINAL WORK
BIOTECHNOLOGY

Assessing the urinary lithogenic risk by multivariate data analysis using analytical results and historic archives

Student: Alvira Larizgoitia, José Ignacio
Date: June 2019

Director
Dr. Pedro Castaño Sánchez

Co-Director
Dr. Federico Mijangos Antón

Academic year
2018/19

INDEX

1. INTRODUCTION 1

2. OBJECTIVES 4

3. METHODS 5

3.1. PATIENTS 5

3.2. VARIABLES 5

3.3. PRE-PROCESSING 6

3.4. PRINCIPAL COMPONENT ANALYSIS (PCA) 6

3.5. χ^2 TEST AND CRAMER'S V 8

3.6. ODDS RATIO 8

3.7. RANDOM FOREST CLASSIFIER 9

4. RESULTS AND DISCUSSION 9

5. CONCLUSIONS 19

6. ACKNOWLEDGEMENTS 20

7. REFERENCES 20

8. APPENDIX I 25

1. INTRODUCTION

Renal lithiasis (RL) is a chronic disease with huge impact in life quality (Tefekli and Cezayirli, 2013). It is characterized by the formation of calculus with different chemical composition in the urinary apparatus due to the commence of a crystallization process. As the pathological origin is the crystallization process attributable to supersaturation, there are substances that can act as promoters, enhancing and favouring the crystallization process, thus increasing the risk to form a kidney stone, and as inhibitors if the crystallization process is difficulted (Basavaraj et al., 2007). The incidence of RL is increasing in modern, developed, countries as a result of lifestyle changes (Ferreira and Dias Sarti, 2019). It is more prevalent in white race and less in black race. In Hispanic and Asian people, the prevalence is intermediate. However, globalization is homogenizing these racial and geographical differences (Curhan, 2007). The lifetime risk of suffering from RL is about 10 – 15% in developed world and 20 – 25% in the middle east. Nephrolithiasis is largely a recurrent disease with a relapse rate of 50% in 5 – 10 years and 75% in 20 years. Once recurrent, the subsequent relapse risk is raised ant the interval between recurrences is shortened. This recurrent nature underscores the importance of prevention (Moe, 2006).

The physiopathological process or RL relies on urine saturation, supersaturation, nucleation (primary and secondary), crystal growth, aggregation, retention and calculus formation. The stone formed may be small enough to be excreted normally when urinating without causing symptoms or colics. Nevertheless, when the calculus is big, it can cause the obstruction of the renal system (Dardamanis, 2013).

Kidney stones are composed of inorganic and organic crystals amalgamated with proteins. Calcareous stones are by far the most common nephroliths (Pak et al., 2003), accounting for more than 80% of stones. Uric acid stones represent about 5 – 10% trailed by cystine, struvite (also called Staghorn stone) and ammonium acid urate stones. Miscellaneous types of highly uncommon stones are associated with xanthine, 2,8-dihydroxyadenine, protein matrix, and drugs such as indinavir and triamterene (Moe, 2006).

The clinical diagnosis is done nowadays by ultrasonography and computed tomography (Smith-Bindman et al., 2014).

There are several risk factors for urolithiasis: (i) genetic factors, associated with a polygenic inheritance where the environmental factors are very influential; (ii) race, as it has been exposed, kidney stones are yet more prevalent on white race; (iii) sex, it is generally more prevalent in men; (iv) age, kidney stones affect all ages, but it is less prevalent in children and teenagers; (v) anatomic alterations such as polycystic kidney or horseshoe kidney; (vi) endocrinological factors, its high incidence in men can be explained by high levels of testosterone; (vii) pregnancy, due to the increase in the glomerular filtration rates; (viii) lifestyle, stress, drugs intake, physical activity and diet; and (ix) other pathologies such as diabetes, obesity or dyslipidemia (Romero et al., 2010). Some of them are modifiable (dietary habits) and others are not (genetic factors), but the different interaction among them may cause different results (Vitale et al., 2008). Therefore, the analysis of RL as a complex, multivariate, process is of paramount importance to have a better understanding of the pathology and build better diagnostic tools.

But, apart from the multivariate view of the problem, RL cannot be only seen under the biomedical lens. The knowledge existent in the biochemical engineering field helps gain an insightful comprehension of the process taking the kidney as a filtration unit and studying the formation of stones as a crystallization process. Also, a biochemical engineering perspective on RL as a dynamical system can be very helpful. This view parts of the base that the human body is a collection of networked processing units (reactors), which exchange mass and energy with the environment. This structure tends to keep the steady state when possible and when it is not, little disturbances towards the correct, critical state are applied, trying also to maintain homeostasis (Androulakis, 2014). Thus, the proper modelling and study of the steady state and its stability, as a biochemical engineer would analyse complex reaction networks, can help study the RL and improve the predictive tools.

Then, under this lens, RL is just a crystallization process, which heavily depends on the factors above exposed, where supersaturation is the first one. Supersaturation refers to a solution that contains more of the dissolved material than could be dissolved by the solvent under normal circumstances. It is a key parameter in nucleation and normal urine is supersaturated with respect to crystalline components. Nucleation is the formation of a solid crystal phase in a solution and can be homogeneous or

heterogeneous, depending on the purity of the solution, and primary or secondary, depending on the presence or absence of precursor crystals, respectively (Briuglia, 2018). Urine is not a pure solution, so only heterogeneous nucleation occurs. The nucleation process often occurs over an existing surface or an alternative structure: epithelial cells, red blood cells, cell debris, urinary casts, other crystals and bacteria. After the nucleation, aggregation occurs: crystal in solution stick together and form a larger particle. These particles can be retained an accumulated in the kidney and this may be caused by association of crystals with the epithelial cells lining the renal tubules (Basavaraj et al., 2007).

The proper assessment of the effect of all biomedical and chemical variables affecting the RL process is challenging due to the immense possible interactions of the correlated and uncorrelated variables. In this context, PCA is a valuable statistical tool that can aid categorizing the variables by replacing a large set of parameters with a smaller set of new ones, termed principal components (PCs), that describe the observed patterns (Alvira et al., 2018). Categorizing and grouping variables is the main aim of PCA and it has been widely used in biomedical and chemical studies (De Perrot et al., 2019; Sparkes et al., 2019; Fanning et al., 2019; Welch, 2019; and Aguado et al., 2014). PCA is among the disruptive methodologies which can lead to new and powerful diagnostic tools, together with artificial neural networks. This last approach is fundamentally distinct to the empiric one that uses principal component equations to describe observations or predict new cases, namely the principal component regression (PCR) or the principal component analysis and multiple linear regression (PCA-MLR). The partial least squared regression (PLS) uses PCA methodology to select the most influential variables and, subsequently, calculate the empirical functions that describe the results with less error (Alvira et al., 2018).

However, the RL process is not still as well understood as it should be to construct a proper, powerful, predictive model based on the mathematical functions of biochemical characteristics. At this point, artificial intelligence (AI) tools can provide the predictive power based on empirical data to fill this gap. The most used AI methods are machine and deep learning. Both methods rely on empirical data to construct a mathematical, abstract, model that can predict an outcome based on the observations, internal patterns of the data and inference, without explicit instructions. Machine

learning is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task, usually prediction, effectively. There are many machine learning algorithms: K-nearest neighbours, support vector machines, Bayesian networks, decision trees, random forests, etc. (Kannan and Vasanthi, 2019; and Pedregosa et al., 2011). Deep learning models relies on artificial neural networks, which have emerged as a result of simulation of biological nervous system, such as the brain on a computer. Neural networks are represented as a set of nodes called neurons connected among them (Kumar and Abhishek, 2012). The predictive power of ANNs heavily depends on the architecture and training process, but it is generally more accurate than machine learning models. In RL context, new methods of kidney stones detection through these methodologies, applied to image analysis have been described: principal component analysis (PCA) plus a prediction/analysis algorithm of machine and deep learning (Attia et al., 2015; Aldoukhi et al., 2019).

The application of these methodologies, together with statistical analyses that help gain more knowledge on the studied process, can give access to powerful, accurate and biomedical models that can predict the presence or absence of kidney stones in a given patient.

2. OBJECTIVES

- Exploratory Data Analysis (EDA): explore the database from an univariate perspective to analyse the distribution of the data.
- Shed light on the nephrolithiasis process by studying it from a multivariate, interdisciplinary perspective, analysing the properties and characteristics of the variables collectively together with their importance through PCA.
- Calculate and interpret the correlations and interactions between every variable and grouped in clusters.
- Use the correlations and interactions among variables to train an AI-based model for the clinical diagnosis and prevention of urolithiasis that can be implemented in hospitals and primary attention clinics.

- Make the AI-based model capable of predicting the probability that a problem patient has kidney stones and predict which type.

3. METHODS

3.1. PATIENTS

The historic and clinical archives used in this work gather 936 patients and were collected from 2012 to 2016 in the Razi Hospital (Tehran, Iran) as part of a previous study approved by the ethics committee of the Gulan University of Medical Sciences. This study was published by Kazemi and Mirroshandel (2017) and the whole dataset was kindly provided by the authors. All data is anonymized and was obtained with the informed consent of all patients.

3.2. VARIABLES

The complete dataset includes 42 features, however, only 21 were considered relevant and thus, kept in this study.

The 21 variables are: A*ge*, age of the patient; *Alcohol*, usual consumer of alcoholic drinks; *BP*, blood pressure; *Bun*, urea levels in blood; *C.C*, complaint or patient's general condition upon arrival at the hospital; *CA*, calcium levels in blood; *Cr*, creatinine levels in urine; *Disease Diagnosis*, medical diagnosis of the pathology suffered by the patient; *DM*, diabetic; *fever*, patient with fever; *FH*, kidney stone histories in the family; *GH*, blood in urine; *Hb*, haemoglobin levels in blood; *Marital Status*, patient married; *Nausea*, patient with nauseas; *NH*, own nephrolithiasic history; *Opium*, usual consumer of opiate-derived substances; *Sex*, male or female; *Smoking*, usual smoker; *Type of Stone*, calcium, aciduric or Staghorn stone; and *uric acid*, uric acid levels in urine.

Purely analytical variables where divided into three categories, ranging from 0 to 2, where 0 indicates low levels of the compound, 1, normal levels and 2, high levels. Normal, low and high references were based on the World Health Organisation (WHO) standards.

Alcohol, *Smoking*, *DM*, f*ever*, *FH*, *GH*, *Marital status*, *Nausea*, *NH*, and *Opium* are binary variables, where 0 indicates absence, no consumption or not married and 1, presence, consumption or married. *Sex* is also a binary variable, but 0 stands for male and 1 for female.

C.C ranges from 0 to 14: 0, flank pain; 1, urinary tract infection; 2, blood in urine; 3, weakness and lethargy; 4, consciousness reduction; 5, abdominal pain; 6, frequent urination; 7, shortness of breath; 8, replacement of ureteral stent; 9, urinary obstruction; 10, coughing; 11, vomiting; 12, headache; 13, melena; 14, wound infection. *Disease_Diagnosis* ranges from 0 to 12: 0, kidney medulla stone; 1, kidney calyx stone; 2, kidney pelvis stone; 3, ureteral stone; 4, bladder stone; 5, hydronephrosis of the kidney; 6, polycystic kidney disease; 7, non-functional kidney; 8, pyelonephritis kidney disease; 9, hematoma surgical site; 10, BPH (Benign Prostatic Hyperplasia); 11, chronic kidney disease; 12, end-stage kidney disease. *Type of Stone* ranges from 0 to 2: 0, calcium stone; 1, aciduric stone; 2, Staghorn stone (struvite). *Age* represents the age of the patient, from 4 to 91.

3.3. PRE-PROCESSING

The raw-data was pre-processed as proposed by Kazemi and Mirroshandel (2017) to discretize the continuous variables (*Hb*, *Cr* and *Bun*), to eliminate redundant variables and to fill in the gaps of the clinical archives. This way, all data is in the same format and can be taken by the designed PCA and prediction algorithms. 30 patients were eliminated from the study due to the absence of enough information in their archives, thus reducing the data matrix to 906 patients and 21 variables.

3.4. PRINCIPAL COMPONENT ANALYSIS (PCA)

An updated routine of that proposed by Alvira et al. (2018) has been developed in MATLAB^{TM} (Version 2015a, Mathworks, Natick, MA, USA, 2015) for the multivariate analysis of the experimental results. The dataset X was configured as a matrix consisting of n columns (for n variables) and m rows (for m number of independent experiments).

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & & & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \quad (1)$$

The data has been normalized with the command *normc*. Then, the data matrix X has been decomposed into a score vector t_i and a loading vector p_i with a residual error E, as:

$$X = t_1 p_1^T + t_2 p_2^T + \dots + t_k p_k^T + E \quad (2)$$

The correlations of the variables were obtained by first calculating the covariance (*cov* command) and then transforming it to a correlation matrix (*corrcov* command). The covariance matrix can also be expressed as:

$$R = \sum_{j=1}^{n} p_j \lambda_j p_j^T = P \Lambda P^T \quad (3)$$

Where P is a *mxn* matrix of eigenvectors and Λ is a matrix of eigenvalues associated to those eigenvectors, as:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \lambda_n \end{bmatrix} \quad (4)$$

The eigenvalues and eigenvectors (calculated with *eig*) are used for the calculation of the principal components of the correlation matrix. To compute the loadings of each PC, the eigenvectors were multiplied by the square root of the eigenvalues diagonal matrix (Equation 4). The eigenvalues also allow for the calculation of the portion of the total variance associated with the *j*-th eigenvalue:

$$\nu_j = \frac{\lambda_j}{\sum_j \lambda_j} \quad (5)$$

In addition, the Varimax rotation was applied to facilitate the interpretation of the results and was calculated using the command *rotatefactors* over the loading matrix.

The command *pca* was used to obtain the score matrix and *zscore* to obtain the z-score matrix.

All these calculations and transformation are nested in a for-loop that saves the results and the images of the calculated PCAs for given values of PCs in each iteration.

Additionally, all PCAs were performed in IBM SPSS Statistics (v23.0) and R (64-bit version) to check the reproducibility of the results.

3.5. χ^2 TEST AND CRAMER'S V

The χ^2 statistic is a non-parametric (distribution free) tool designed to analyse group differences when the dependent variable is measured at a nominal level. Like all non-parametric statistics, the χ^2 is robust with respect to the distribution of the data. Specifically, it does not require equality of variances among the study groups (homoscedasticity in the data). It permits evaluation of both dichotomous independent variables, and of multiple group studies. Unlike many other non-parametric and some parametric statistics, the calculations needed to compute the χ^2 provide considerable information about how each of the groups performed in the study. This richness of detail allows the researcher to understand the results and thus to derive more detailed information from this statistic than from many others (McHugh, 2013). High values of the statistic thus indicate that the variables are not independent and Cramer's V (values from 0 to 1) measure the correlation between the variables: 1 for fully correlated variables and 0 for the opposite (Prentice and Andersen, 2007).

3.6. ODDS RATIO

The odds ratio is a widely used tool in medical reports. They provide an estimate for the relationship between two binary variables, they enable researchers to examine the effects of other variables on that relationship, using logistic regression and, finally, they have a special and very convenient interpretation in case-control studies (Bland and Altman, 2000).

For this data, the odds ratio has been used to calculate how much prevalent are the kidney stones in men than in women. To do this, the next equation has been applied for each type of stone:

$$OR = \frac{\text{Number of men with stone/Number of men without that stone}}{\text{Number of women with stone/Number of women without that stone}} \quad (6)$$

3.7. RANDOM FOREST CLASSIFIER (RFC)

Random forests, or random decision forests, are an ensemble learning method for classification and regression that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of individual trees. Random forests correct the habit of decision trees of overfitting to their training set (Trevor et al., 2009).

To build the RFC model, the python library sci-kit learn (sklearn) was used (Pedregosa et al., 2011). The data was randomly split into training data (77%) and testing data (33%). Both models use 1000 estimators to get the predictions.

The results were obtained in the form of a confusion matrix and classification reports.

4. RESULTS AND DISCUSSION

All the results and figures of the exploratory data analysis, prior to the PCA analysis are shown in **Appendix I**. After this preliminary analysis, an individual idea of each variable has been inferred and some questions, such as what the influence of sex is in the stone formation, arose. However, the behaviour of the system cannot be analysed only by means of univariate data analysis techniques, that is why the PCA methodology was used to extract information considering all variables at once. A first PCA was done considering all variables as enlisted in **Section 3.2**.

The PCA interpretation made in this study relies on clustering and distribution of the patients in the score plot and of variables in the Varimax-rotated plot. In the score plot, associations and clustering patterns between patients are represented as close points and inversely correlated patients appear on the contrary diagonal. Therefore, patients with statistically similar characteristics or symptoms, as well as outliers or patterns can be identified. The relative position of each patient in a PC_i-PC_j axis also determines how much those PCs affect that patient. Since PCs are represented by associations of variables, the coordinates of the patient in that plane determine what variables affect

more and in which way (directly or indirectly proportional). To assess the importance of the variables in each PC, a Varimax rotation is performed. This rotation forces the PCs to be orthogonal, thus compelling the variables to associate more with one PC, ideally obtaining a cross-like pattern. The distribution of the variables in the Varimax plot indicates the correlation of the variables both individually and as a cluster (Abdi, 2003). Variables that appear close to others are highly, positively, correlated, whereas clusters of variables that appear on the opposite diagonal from another cluster are highly, negatively, correlated. Like this, a comprehension of the system as a whole can be obtained. Also, since PCs are ordered based on their importance, i.e., the percentage of the variance explained, variables with higher loadings (coordinate in the axis) in the first PCs are more important in the studied process. Note that the sign of the coordinate does not necessarily mean that the variable is affecting positively or negatively, but its effect is antagonist of other variables that appear on the opposite side (changed signs).

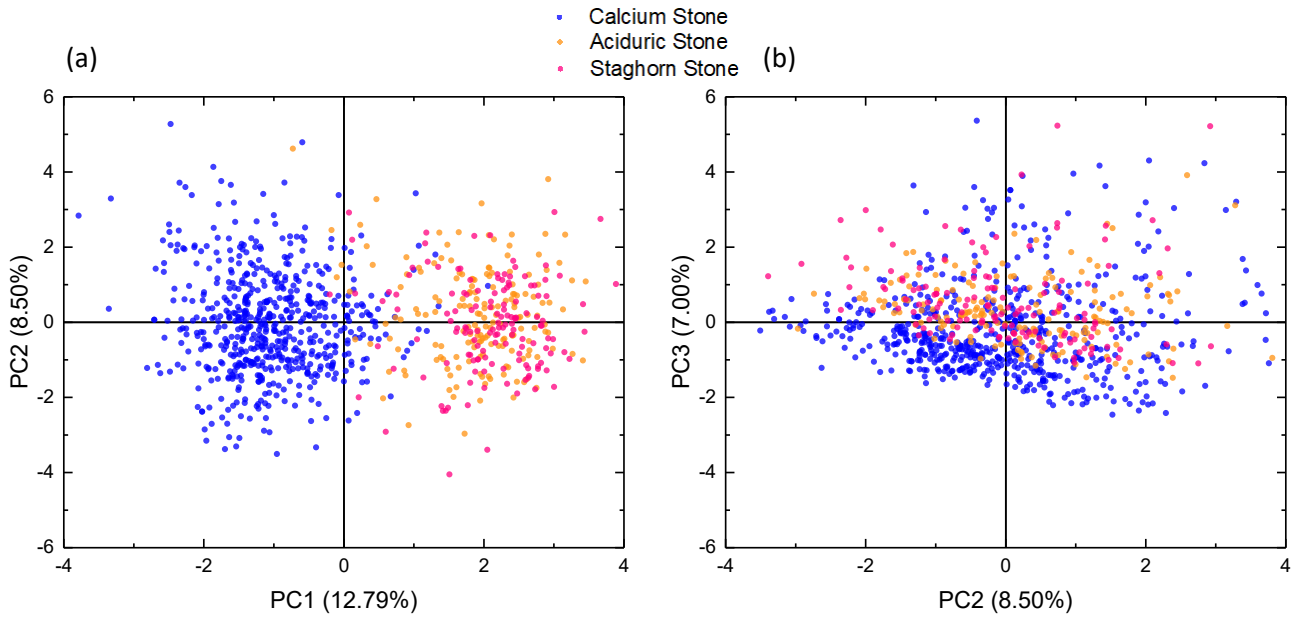


Figure 1. Score plot of the first PCA. All 906 patients are represented as points in a **(a)** PC1-PC2 plane; and **(b)** PC2-PC3 plane.

From the clinical data consisting of 906 uncorrelated patients and 21 measured variables, 8 PCs were extracted and the sum of them characterize 57.54% of the total variance and therefore are discussed below. **Figure 1** shows the score plot of these PCs, which is a representation of the values derived from the score vectors computed

according to **Equation 2** and provide insights into potential deviations in the data or data that do not fit the PCA at a confidence level of 95% (Alvira et al., 2018). The clinical data is widely spread along PC1. Two clusters can be seen in **Figure 1 (a)**, patients with calcium stones at the left side of the graph and patients with aciduric or Staghorn stones at the right. Basically, there is a split between patients having calcium stones and those who have another type of stone (aciduric or Staghorn). In **Figure 1 (b)**, the separation between the three groups is not obvious, thus indicating that the effect of the variables corresponding to PC2 and PC3 do not present relevant differences between the patients, meaning that the clinic and symptoms suffered are statistically similar just considering the variables associated with both PCs. Since the distribution of patients along PC1 is that good, the variables affecting PC1 can be considered as good predictors for the type of stone a problem patient may have, but this will be discussed further in the analysis of calcium stones and the construction of the prediction algorithm.

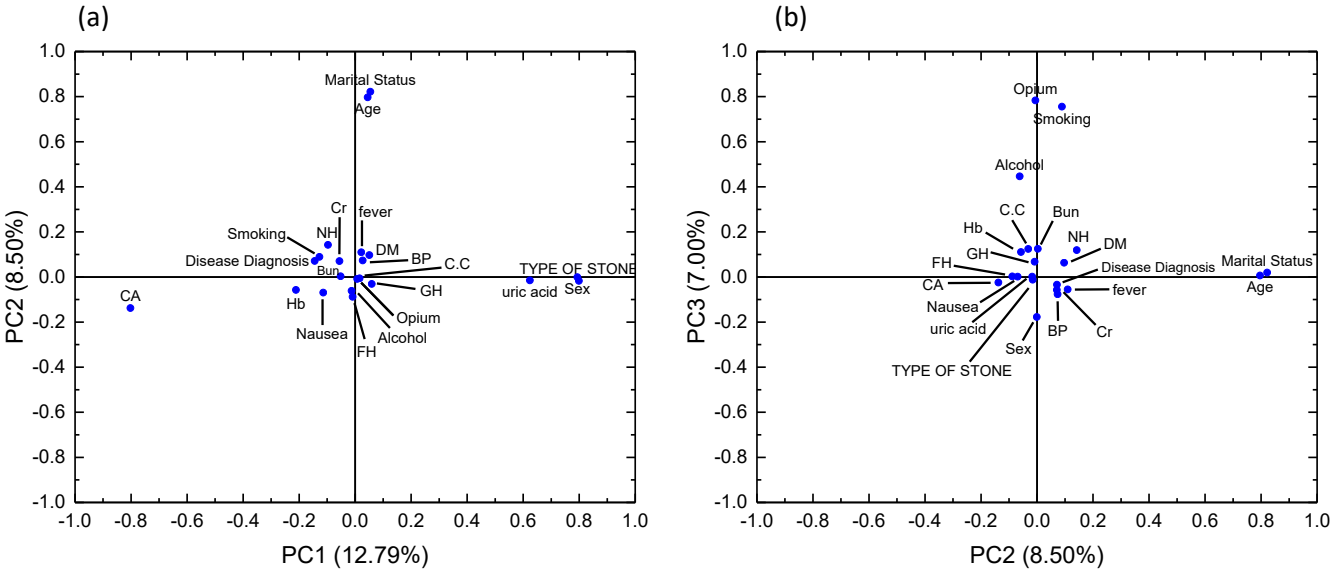


Figure 2. Plot of the Varimax rotated loadings. Each variable is represented as a point in a **(a)** PC1-PC2 plane; and **(b)** PC2-PC3 plane.

Figure 2 displays the Varimax-rotated loadings for this PCA. A cross-like pattern is observable in **Figure 2 (a and b)**. Starting with **Figure 2 (a)**, the variables with the higher Varimax-rotated loadings are: *Sex*, *Type of Stone*, *uric acid* and *CA*, meaning that these four variables are of paramount importance in the nephrolithiasic process. Also, *Sex*, *Type of Stone* and *uric acid* appear close together, which means that they are highly, positively, correlated. *CA* appears on the opposite side, thus meaning that

it is highly, inversely, correlated. The high correlation between the kidney stones and sex indicates that the formation of stones may be prevalent in one sex over the other. To check this association, a χ^2-test of independence was run and the results are discussed ahead. The value of the statistic $\chi^2 = 464.65$ (>> 3.8415, the reference value for this system where the degrees of freedom are 2) indicates that the independence hypothesis must be discarded, this is, sex and kidney stones are tightly associated. The Odds Ratio was calculated for this data to check the results of the independence test and $OR_{calcium} = 44.1201$, thus meaning that the probability of forming a calcium kidney stone is much higher in men than in women, confirming that sex and stone formation are tightly associated. $OR_{aciduric} = 0.0468$ and $OR_{Staghorn} = 0.0566$; these results indicate that aciduric and Staghorn stones are more prevalent in women, according to this data. Be as that it may, the sex is very correlated with kidney stone formation: calcium stones are more prevalent in men and aciduric and Staghorn stones, in women. However, considering that the most prevalent stone in overall is the calcium stone (Basavaraj et al., 2007), it can be said that men have more risk of forming a kidney stone than women. Back to **Figure 2 (a)**, *uric acid* is also correlated with *Type of Stone*, which makes sense because one possible type of stone is the aciduric one. This means that patients with altered levels of uric acid in urine are probably forming an aciduric stone in the kidney. At the opposite side of this cluster there is *CA*, the calcium levels affect a lot the process, but not in the same way as uric acid does. This is because the calcium concentration is being measured in blood, whereas the concentration of uric acid is measured in urine. Having less calcium in the blood may indicate that it is being accumulated in the renal system in the form of a stone. This is why *CA* appears on the opposite side of *Type of Stone*, but is still highly correlated with it.

The variables that affect most to the process after those discussed above are those related to PC2: *Age* and *Marital Status*, explaining 14.77% of the total variance. They appear very close, meaning that they are highly, positively, correlated. This is obvious because average age of marriage in Iran is 21.9 ± 4.1 years old (Mahdaviazad et al., 2019), so most people older than 25 years old are married in Iran. Of course, the marital status does not have anything to do with the process of stone formation itself, but rather indicates that most people suffering from kidney stones are married. This is probably due to a change in the nutritional habits or a more sedentary lifestyle, which may lead

to the formation of stones (Manfredini et al., 2019). Moreover, *Marital Status* has proven to be valuable enough as a predictor of the nephrolithiasic process and hence this variable has been conserved throughout the study and the construction of a prediction model. Also, a simple relationship such as the high correlation between *Age* and *Marital Status* is so obvious and expectable that the fact that they appear close together gives more reliability to the model. This is important to gain the trust of a customer or committee when deploying a prediction algorithm or present the results of the study (Siau and Wang, 2018),

The PC3 explains 12.17% of the total variance, very close to PC2, thus meaning that the variables associated with PC3 are just slightly less important than those related with PC2. In **Figure 2 (b)**, the three variables most affecting PC3 appear in a similar cross-like pattern: *Opium*, *Smoking* and *Alcohol*. *Opium* and *Smoking* are very close and hence are very positively correlated. *Alcohol* is also correlated with them, but with a little weaker correlation. These results indicate that the drug-consuming habits are important in the nephrolithiasic process, specially the consumption of opium-related substances and smoking. Alcohol intake may also affect the nephrolithiasic process, but not as much as *Opium* or *Smoking*. The high, positive, correlation between this two can be explained by the fact that opiate substances can be smoked and most opium consumers would also be, thereby, smokers. However, the reverse is not necessarily true, since there can be cigarette smokers that do not consume opiate substances, thus separating a little bit *Smoking* from the *Opium* in **Figure 2 (b)**. Alcohol intake appear slightly separated from the *Opium-Smoking* cluster and this might be explained by two factors: first, it affects the stone-formation process, but not as much as *Opium* and *Smoking*; or, second, patients lied in the clinical interview, alleging that they do not consume alcohol when they do. This is possible because in Iran, alcohol intake is not permitted by law and, hence, is not socially accepted and would be a shame to admit the intake of alcoholic substances (Mehrabi et al., 2019). For these results, it is impossible to know if the effect of *Alcohol* in the process is less important than *Opium* or *Smoking* or it is closely related, but the patients would not admit its intake. Nevertheless, some authors (Littlejohns et al. (2019), Ferraro et al. (2013)) demonstrate that high intakes of fluids, including alcohol, may reduce the probability to form a kidney stone due to its diuretic effect. However, since in this analysis *Alcohol*

is positively and not negatively associated with *Opium* and *Smoking*, the fact that these results may be biased cannot be discarded and another clinical interview should be done with emphasis in the alcohol intake. Ketabchi et al. (2012) demonstrated that the probability of forming kidney stones was about 7.42 times higher in opium addicts. They also demonstrated that 56.68% of urolithiasis patients, who had a history of more than 15 renal colics, related to stone forming frequencies, were addicts to opium for more than 10 years. These results are heavily supported by this PCA, adding that the effect of smoking may be as bad as opiate consumption. Tamadon et al. (2013) demonstrated that smoking also increases the probability to form a kidney stone independently of other factors. Hence, the consumption of opiate substances and smoking together may heavily increase the probability to form a kidney stone and, as the results of this analysis suggest, this is the case in most opium consumers.

Table 1. Varimax-rotated loadings for the eight PCs extracted. The variables that are associated with each PC are highlighted.

Variables	Principal Component (PC)							
	1	2	3	4	5	6	7	8
Bun	-0.052	0.003	0.125	0.001	0.783	-0.070	-0.193	-0.066
Cr	-0.056	0.071	-0.057	0.009	0.781	0.121	0.170	0.087
Hb	-0.211	-0.057	0.111	-0.111	-0.322	-0.097	-0.377	0.224
Sex	0.794	-0.001	-0.177	0.184	-0.094	-0.044	0.071	-0.069
Age	0.045	0.797	0.006	-0.020	0.192	0.154	-0.056	-0.037
Marital Status	0.055	0.822	0.019	0.034	-0.093	-0.099	0.118	-0.012
C.C	0.060	-0.031	0.124	-0.261	0.117	0.560	-0.085	0.309
fever	0.022	0.109	-0.055	-0.143	0.024	0.723	0.097	0.047
Nausea	-0.114	-0.069	0.001	0.203	-0.022	0.540	0.186	-0.077
GH	0.007	-0.009	0.068	0.105	-0.027	0.113	-0.088	0.720
DM	0.051	0.097	0.063	0.311	0.013	0.413	-0.379	-0.454
NH	-0.097	0.142	0.119	-0.094	0.078	-0.008	0.481	-0.032
Smoking	-0.127	0.089	0.755	-0.018	-0.056	-0.036	-0.087	0.116
Alcohol	-0.013	-0.062	0.446	-0.012	0.066	0.029	0.095	0.284
Opium	0.016	-0.006	0.783	-0.022	0.039	0.033	0.103	-0.193
BP	0.027	0.074	-0.077	0.707	0.053	-0.027	-0.145	0.006
FH	-0.009	-0.088	0.003	-0.003	-0.127	0.167	0.674	0.029
Disease Diagnosis	-0.144	0.071	-0.034	-0.791	0.020	0.047	-0.059	-0.058
uric acid	0.624	-0.015	-0.012	-0.178	-0.037	0.021	-0.060	-0.282
CA	-0.802	-0.138	-0.024	0.032	-0.047	-0.035	0.078	-0.113
TYPE OF STONE	0.799	-0.017	0.002	0.326	-0.014	-0.074	-0.027	0.175
% of explained variance	12.79	8.50	7.00	6.94	6.33	5.69	5.26	5.02

So, in this analysis, the multivariate problem of nephrolithiasic process has successfully been reduced to eight factors, from which three stand out: the first and most important is the analytical and gender one (PC1: *Sex*, *uric acid*, *CA* and *Type of Stone*), the second is A*ge* (PC2) and the third is the dietary habits. The eight factors, along with their relative importance and the variables associated with them are shown in **Table 1**.

After the univariate and multivariate analysis of this database, it can be said that its main drawback is that it contains no controls. Therefore, for further analyses and for a proper construction of a prediction algorithm, we transformed the patients with aciduric and Staghorn stones to calcium-stone controls, since they do not present that type of stone. The decision of keeping calcium stones while converting the rest to controls was done on the basis of several researchers (Basavaraj et al., 2007; Kumar and Abhishek, 2012; and Pozdzik et al., 2019) that claim that calcium stones are the most widely spread and most prevalent. Another PCA was applied to this modified database and the results are shown in **Figure 3 and 4** and in **Table 2**.

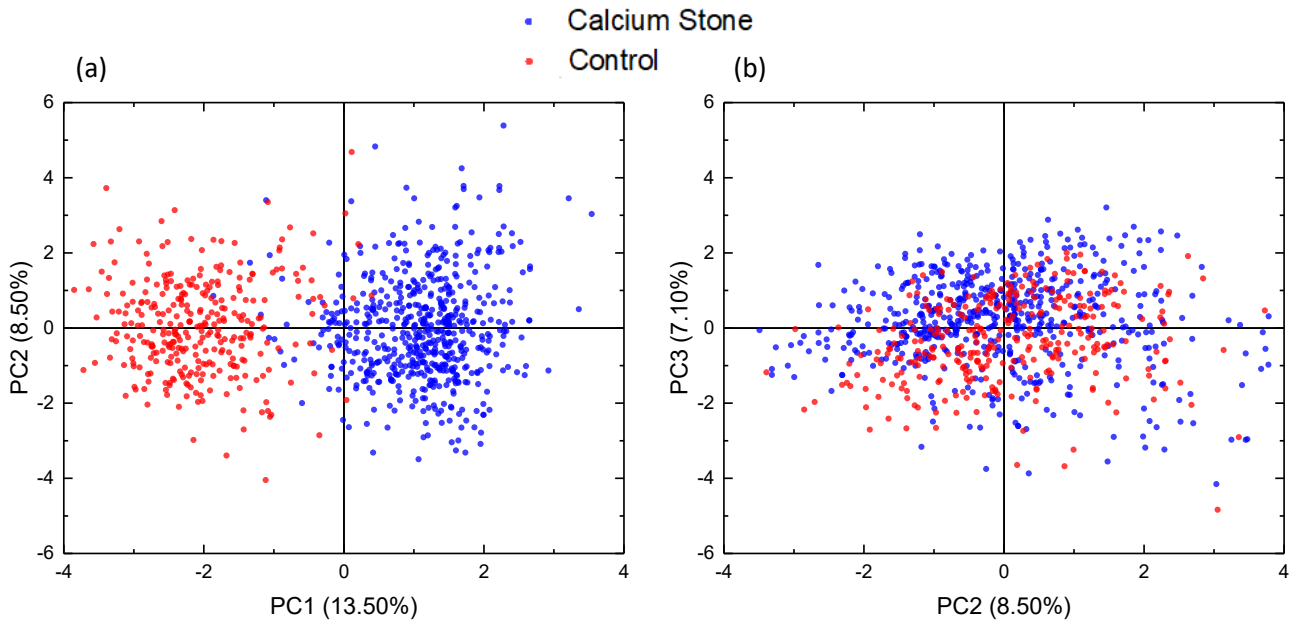


Figure 3. Score plot of the second PCA, where only calcium stones were considered. All the same 906 patients are represented in the plane formed by the new PCs: **(a)** PC1-PC2; and **(b)** PC2-PC3.

In this new analysis, also eight PCs have been extracted, but the sum of them explain 58.19% of the total variance while in the previous analysis, the eight PCs characterized

57.54%. Although the difference among them is not paramount, it can be said that this second analysis explains better the data and thus the prediction algorithm will work better.

In **Figure 3 (a)**, two differentiable clusters can be distinguished with a good distribution between the relative centres for the presence/absence of calcium stones along PC1. This is very important because it would facilitate the classificatory task of the prediction algorithm. Also, there are not outliers in this dataset.

In **Figure 3 (b)**, a homogeneous distribution of the patients can be viewed along the PC2-PC3 axis. This means that, as in the previous case, patients suffering from calcium stones cannot be distinguished from those who do not have that type of stone, thus making the differentiation impossible if only these two PCs are considered. However, the fact that classification cannot be properly done on the PC2-PC3 basis does not imply that these PCs and their associated variables are not important in the nephrolithiasic process.

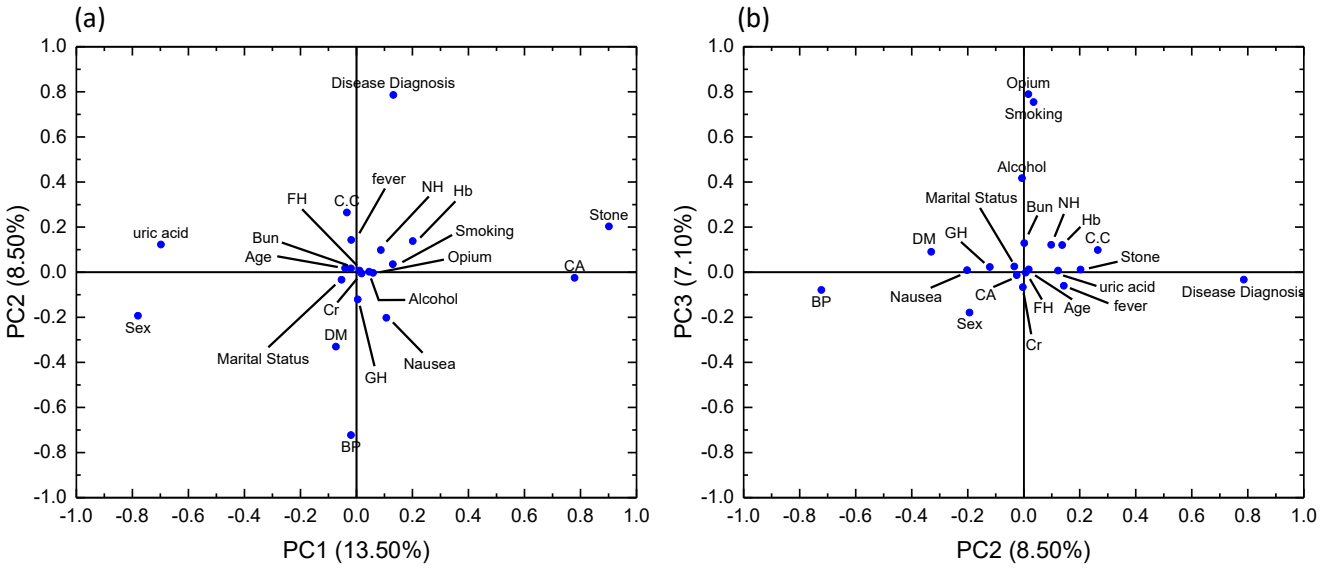


Figure 4. Varimax-rotated loadings of the PCA analysis applied to the modified database. The same variables are represented in the plane formed by the new PCs: **(a)** PC1-PC2; and **(b)** PC2-PC3.

In the Varimax-rotated loadings represented in **Figure 4 (a)**, a cross-like pattern is also achieved, but with important differences from the previous case (**Figure 2 (a)**). The same variables are associated with PC1: *Sex*, *uric acid*, *CA* and *Stone*, but the relationships among them have changed. Now *CA* is highly, positively, correlated with

Stone, whereas *uric acid* is highly, negatively, correlated. That is because the other stones have been eliminated for this study. This way, altered levels of uric acid, which may lead to the formation of an aciduric stone, will not help the formation of a calcium stone, thus negatively affecting the calcium-nephrolithiasic process. For this same reason, the relationship between *CA* and *Stone* has changed to a positive correlation, meaning that altered levels of calcium in blood may lead to a calcium stone formation. *Sex* is still very related to stone formation and the fact that it appears inversely correlated is just because of the election of the binary values for *Sex*, this is, *Sex* = 0 means that the patient is a male, whereas *Sex* = 1 stands for a female patient. So, as it has been previously demonstrated, men have higher probability of forming a calcium stone (*Stone* = 1) and thus, the relationships between these variables is inverse, while being tightly associated, because a male patient with calcium stone in a *Sex-Stone* plane would have coordinates (0,1). More substantial differences appear in PC2, where *Age* and *Marital Status* lose importance towards *BP* and *Disease Diagnosis*. This means that in the specific formation of calcium stones, the blood pressure and the clinical diagnosis of the pathology have much more importance than the age of the patients. This is supported by Boyd et al. (2019), who demonstrated that older age was associated with decreased calcium excretion and therefore it is less probable to form a calcium stone. High blood pressure is very associated with the urinary calcium excretion and, therefore, it acts as a promoter for calcium stone formation (Krishna et al., 2012). Also, *BP* appears to be highly, negatively, correlated with *Disease Diagnosis*. This could mean that *BP* has a high positive effect in the calcium stone formation, but not in where that calcium stone is located (*Disease Diagnosis* = 0, 1, 2, 3 or 4) or the renal pathology suffered by the patient (*Disease Diagnosis* = 5, 6, 7, 8, 9, 10, 11 or 12). *DM* is, to a certain extent, associated with PC2, thus meaning that having diabetes affects the calcium nephrolithiasic process. It is associated with more oxalate excretions, thus leading to formation of calcium oxalate stones (Boyd et al., 2019). The results of this analysis support this statement because *DM* is positively correlated with *BP* and both help the calcium stone formation. However, although having diabetes is a risk factor for the formation of calcium stones, it is not as important as *BP*, according to our results.

As in the previous analysis, the variables mostly affecting PC3 are *Opium*, *Smoking* and *Alcohol*, shown in **Figure 4 (b)**. There are no significant differences between this PC3 and the previously calculated, except for *Smoking*, that appears slightly closer to *Opium*, implying that opiate consumption and smoking are even more positively correlated. However, the fact that the PC3 remains virtually the same, means that the effects of these variables, discussed above, do not lose any importance and equally affect the process of calcium stones formation.

Table 2. Varimax-rotated loadings for the eight new PCs extracted. The variables that are associated with each PC are highlighted.

Variables	Principal Component (PC)							
	1	2	3	4	5	6	7	8
Bun	0.045	0.002	0.129	0.787	-0.001	-0.067	-0.187	-0.061
Cr	0.059	-0.003	-0.067	0.778	0.069	0.120	0.173	0.087
Hb	0.201	0.138	0.120	-0.320	-0.076	-0.092	-0.386	0.158
Sex	-0.780	-0.193	-0.179	-0.094	0.007	-0.042	0.075	-0.053
Age	-0.040	0.018	0.012	0.191	0.796	0.155	-0.053	-0.038
Marital Status	-0.053	-0.034	0.025	-0.095	0.821	-0.102	0.118	-0.012
C.C	-0.034	0.265	0.098	0.107	-0.023	0.566	-0.090	0.304
fever	-0.018	0.143	-0.060	0.022	0.110	0.724	0.099	0.042
Nausea	0.107	-0.202	0.009	-0.020	-0.077	0.538	0.190	-0.085
GH	0.005	-0.121	0.023	-0.034	-0.005	0.094	-0.091	0.777
DM	-0.073	-0.330	0.090	0.027	0.097	0.412	-0.363	-0.417
NH	0.087	0.098	0.121	0.080	0.135	-0.013	0.485	-0.019
Smoking	0.130	0.035	0.754	-0.054	0.080	-0.029	-0.088	0.122
Alcohol	0.018	-0.006	0.417	0.067	-0.054	0.017	0.101	0.371
Opium	-0.020	0.016	0.789	0.047	-0.008	0.038	0.112	-0.138
BP	-0.020	-0.723	-0.079	0.051	0.079	-0.035	-0.142	0.033
FH	0.011	0.007	-0.002	-0.133	-0.092	0.164	0.671	0.028
Disease Diagnosis	0.132	0.786	-0.033	0.023	0.074	0.053	-0.060	-0.060
uric acid	-0.698	0.123	0.007	-0.019	-0.019	0.009	-0.042	-0.171
CA	0.779	-0.025	-0.013	-0.043	-0.145	-0.041	0.077	-0.129
Stone	0.902	0.203	0.012	0.015	0.024	0.071	0.020	-0.068
% of explained variance	13.50	8.50	7.10	6.85	6.27	5.74	5.25	4.97

After the two PCAs, whose results for the variables are summarized in **Table 1** and **Table 2**, it can be concluded that the clusters of variables that affect most the process are, in order of importance: the analytical and sexual (*Sex*, *uric acid*, *CA* and *Stone*),

clinical (*BP* and *Disease Diagnosis*), dietary habits (*Smoking*, *Opium* and *Alcohol*), urinary (*Bun* and *Cr*), the age and the associated status (*Age*, *Marital Status*), symptomatic (*C.C*, *fever* and *Nausea*), haemoglobin levels and presence of historical nephrolithiasic cases (*Hb*, *NH* and *FH*) and other harmful or pathological processes (*GH* and *DM*).

All the variables in this study has proven valuable as kidney-stone-type and calcium stones predictors, especially for the last, so a random forest classifier model (RFC) has been trained for detecting the stone type of a patient or the presence of calcium stones. The results of the classification reports for stone type and calcium stone are summarized in **Table 3** and **Table 4**, respectively. The accuracy of these models is higher than that of those proposed by other authors (Kumar and Abhishek, 2012; and Kazemi and Mirroshandel, 2018).

Table 3. Results of the RFC for predicting the type of stone.

Value	Precision	Recall	F1 score	Support
0	0.97	0.99	0.98	192
1	0.95	0.95	0.95	63
2	1.00	0.93	0.96	44
average/total	0.97	0.97	0.97	299

Table 4. Results of the RFC for predicting the presence/absence of calcium stones.

Value	Precision	Recall	F1 score	Support
0	0.97	0.95	0.96	107
1	0.97	0.98	0.98	192
average/total	0.97	0.97	0.97	299

The RFC models trained are also capable of calculating the probability that a given patient has calcium, aciduric or Staghorn stones with a confidence level of 97%.

5. CONCLUSIONS

- The PCA methodology has proven very valuable for the study of a complex pathology as urolithiasis. New insights in the form of interaction between variables and correlation among clusters of variables have been exposed.

- The study demonstrated the key role that sex, uric acid and calcium levels, blood pressure, diagnosed pathologies and dietary habits (smoking, opiate and alcohol intake) play in the formation of calcium, aciduric and Staghorn stones. Thus, these variables should be controlled in risk patients to help prevent new cases.
- All the variables included in this study have also demonstrated to be good predictors, collectively, for the presence of kidney stones.
- The RFC models developed to predict the type of stone in patients with a kidney stone and presence of calcium stone in not diagnosed patients can predict both features with an accuracy of 97%, thus making these models reliable enough to be implemented in hospitals and primary attention clinics.
- The RFC models can also predict the probability that a patient has a kidney stone and which type it is. Thus, making it potentially useful for the prevention of RL in risk patients.

6. ACKNOWLEDGMENTS

We want to thank Dr. Mirroshandel for kindly providing the database used in this study.

We also want to thank the researchers involved in the ALIRE (analysis of lithogenic risk) project: Cruces Hospital: Antonio Arruza Echevarría, Francisco Javier Gainza de los Ríos, Marta Antón Juanilla, José Antonio Quintanar Lartundo, María Begoña Bralo Berastegui, Laura López Bueno, Lorena Salazar Ibáñez, Susana del Corral Navarro, Aitor García de Vicuña Meléndez, Ana Santorcuato Bilbao, Magdalena Carreras Aja and Irma Arrieta Artieda; Galdakao Hospital: José Antonio Gallego Sánchez and Verónica Pando Olaso; Biocruces Research Institute: Eunate Arana Arri, Ariane Imaz Ayo and Natale Imaz Ayo.

7. REFERENCES

Abdi, H. 2003. Factor rotations in factor analyses. Encyclopedia for Research Methods for the Social Sciences. Sage: Thousand Oaks, CA, 792-795.

Aguado, R., Elordi, G., Arrizabalaga, A., Artetxe, M., Bilbao, J., Olazar, M. 2014. Principal component analysis for kinetic scheme proposal in the thermal pyrolysis of waste HDPE plastics. Chemical Engineering Journal, 254, 357-364.

Aldoukhi, A. H., Law, H., Black, K. M., Roberts, W. W., Deng, J., Ghani, K. R. 2019. PD04-06 Deep learning computer vision algorithm for detecting kidney stone composition: towards an automated future. The Journal of Urology, 201, e75-e76.

Alvira, J., Hita, I., Rodríguez, E., Arandes, J., Castaño, P. 2018. A Data-Driven Reaction Network for the Fluid Catalytic Cracking of Waste Feeds. Processes, 6, 243.

Androulakis, I. P. 2014. A chemical engineer's perspective on health and disease. Computers & chemical engineering, 71, 665-671.

Attia, M. W., Abou-Chadi, F. E. Z., Moustafa, H. E. D., Mekky, N. 2015. Classification of ultrasound kidney images using PCA and neural networks. International Journal of Advanced Computer Science and Applications, 6, 53-57.

Basavaraj, D. R., Biyani, C. S., Browning, A. J., Cartledge, J. J. 2007. The role of urinary kidney stone inhibitors and promoters in the pathogenesis of calcium containing renal stones. EAU-EBU update series, 5, 126-136.

Bland, J. M., Altman, D. G. 2000. The odds ratio. Bmj, 320, 1468.

Boyd, C., Whitaker, D., Ashorobi, O., Poore, W., Oster, R., Knight, J., Holmes, R., Assimos, D., Wood, K. 2019. MP08-05 Impact of demographic factors and systemic disease on urinary stone risk parameters amongst stone formers. The Journal of Urology, 201, e100-e100.

Briuglia, M. 2018. Primary and Secondary Crystal Nucleation. New approaches to measure nucleation rates. https://www.crystallizationsystems.com/news/june-2018/primary-and-secondary-crystal-nucleationnew-approaches-to-measure-nucleation-rates. Last checked: 13/06/2019.

Curhan, G. C. 2007. Epidemiology of stone disease. Urologic Clinics of North America, 34, 287-293.

Dardamanis, M. 2013. Pathomechanisms of nephrolithiasis. Hippokratia, 17, 100.

De Perrot, T., Hofmeister, J., Burgermeister, S., Martin, S. P., Feutry, G., Klein, J., Montet, X. 2019. Differentiating kidney stones from phleboliths in unenhanced low-dose computed tomography using radiomics and machine learning. European radiology, 1-7.

Fanning, J., Rejeski, W. J., Chen, S. H., Nicklas, B. J., Walkup, M. P., Axtell, R. S., Fielding, R. A., Glynn, N. W., King, A. C., Manini, T. M., Newman, A. B., Pahor, M., Tudor-Locke, C., Miller, M. E., McDermott, M. M. 2019. A Case for Promoting Movement Medicine: Preventing Disability in the LIFE Randomized Controlled Trial. The Journals of Gerontology: Series A.

Ferraro, P. M., Taylor, E. N., Gambaro, G., Curhan, G. C. 2013. Soda and other beverages and the risk of kidney stones. Clinical Journal of the American Society of Nephrology, 8, 1389-1395.

Ferreira Fontenelle, L., Dias Sarti, T. 2019. Kidney Stones: Treatment and Prevention. American family physician, 99.

Kannan, R., Vasanthi, V. 2019. Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease. In Soft Computing and Medical Bioinformatics (pp. 63-72). Springer, Singapore.

Kaur, G., Sidhu, E. B. K. 2014. Proposing Efficient Neural Network Training Model for Thyroid Disease Diagnosis. International Journal for Technological Research in Engineering, 1.

Kazemi, Y., Mirroshandel, S. A. 2018. A novel method for predicting kidney stone type using ensemble learning. Artificial intelligence in medicine, 84, 117-126.

Ketabchi, A. A., Ebad-Zadeh, M. R., Parvaresh, S., Moshtaghi-Kashanian, G. R. 2012. Opium dependency in recurrent painful renal lithiasis colic. Addiction & health, 4, 73.

Krishna, K., Rayavarapu, A., Vadlapudi, V. 2012. Statistical and data mining aspects on kidney stones: a systematic review and meta-analysis. Open Access Scientific Reports, 1.

Kumar, K., Abhishek, B. 2012. Artificial neural networks for diagnosis of kidney stones disease. GRIN Verlag.

Littlejohns, T. J., Neal, N. L., Bradbury, K. E., Heers, H., Allen, N. E., Turney, B. W. 2019. Fluid Intake and Dietary Factors and the Risk of Incident Kidney Stones in UK Biobank: A Population-based Prospective Cohort Study. European urology focus.

Mahdaviazad, H., Malekmakan, L., Sayadi, M., Tadayon, T. 2019. Trends in age at first marriage in women and related factors: A population-based study in southwestern Iran. Women & health, 59, 171-180.

Manfredini, R., Cappadona, R., De Giorgi, A., Fabbian, F. 2019. To Marry or Not. American Journal of Cardiology, 123, 1185.

McHugh, M. L. 2013. The chi-square test of independence. Biochemia medica: Biochemia medica, 23, 143-149.

Mehrabi, M., Hajebi, A., Mohebbi, E., Baneshi, M. R., Khodadost, M., Haghdoost, A. A., Sharifi, H., Noroozi, A. 2019. Prevalence and Correlates of Lifetime Alcohol Use among Adult Urban Populations in Iran: A Knowledge, Attitude, and Practice Study. Journal of psychoactive drugs, 1-8.

Moe, O. W. 2006. Kidney stones: pathophysiology and medical management. The lancet, 367, 333-344.

Pak, C. Y., Poindexter, J. R., Adams-Huet, B., Pearle, M. S. 2003. Predictive value of kidney stone composition in the detection of metabolic abnormalities. The American journal of medicine, 115, 26-32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Vanderplas, J. 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12, 2825-2830.

Pozdzik, A., Maalouf, N., Letavernier, E., Brocheriou, I., Body, J. J., Vervaet, B., Haute, C. V., Noels, J., Gadisseur, R., Castiglione, V., Gambaro, G., Daudon, M., Sakhaee, K., Cotton, F. 2019. Meeting report of the “Symposium on kidney stones and mineral metabolism: calcium kidney stones in 2017”. Journal of nephrology, 1-18.

Prentice, R., Andersen, V. 2007. Interpreting heritage essentialisms: Familiarity and felt history. Tourism Management, 28, 661-676.

Romero, V., Akpinar, H., Assimos, D. G. 2010. Kidney stones: a global picture of prevalence, incidence, and associated risk factors. Reviews in urology, 12, e86.

Siau, K., Wang, W. 2018. Building trust in artificial intelligence, machine learning, and robotics. Cutter Business Technology Journal, 31, 47-53.

Smith-Bindman, R., Aubin, C., Bailitz, J., Bengiamin, R. N., Camargo Jr, C. A., Corbo, J., Kang, T. L. 2014. Ultrasonography versus computed tomography for suspected nephrolithiasis. New England Journal of Medicine, 371, 1100-1110.

Sparkes, S. P., Atun, R., Bärnighausen, T. 2019. The impact of the Family Medicine Model on patient satisfaction in Turkey: Panel analysis with province fixed effects. PloS one, 14, e0210563.

Tamadon, M. R., Nassaji, M., Ghorbani, R. 2013. Cigarette smoking and nephrolitiasis in adult individuals. Nephro-urology monthly, 5, 702.

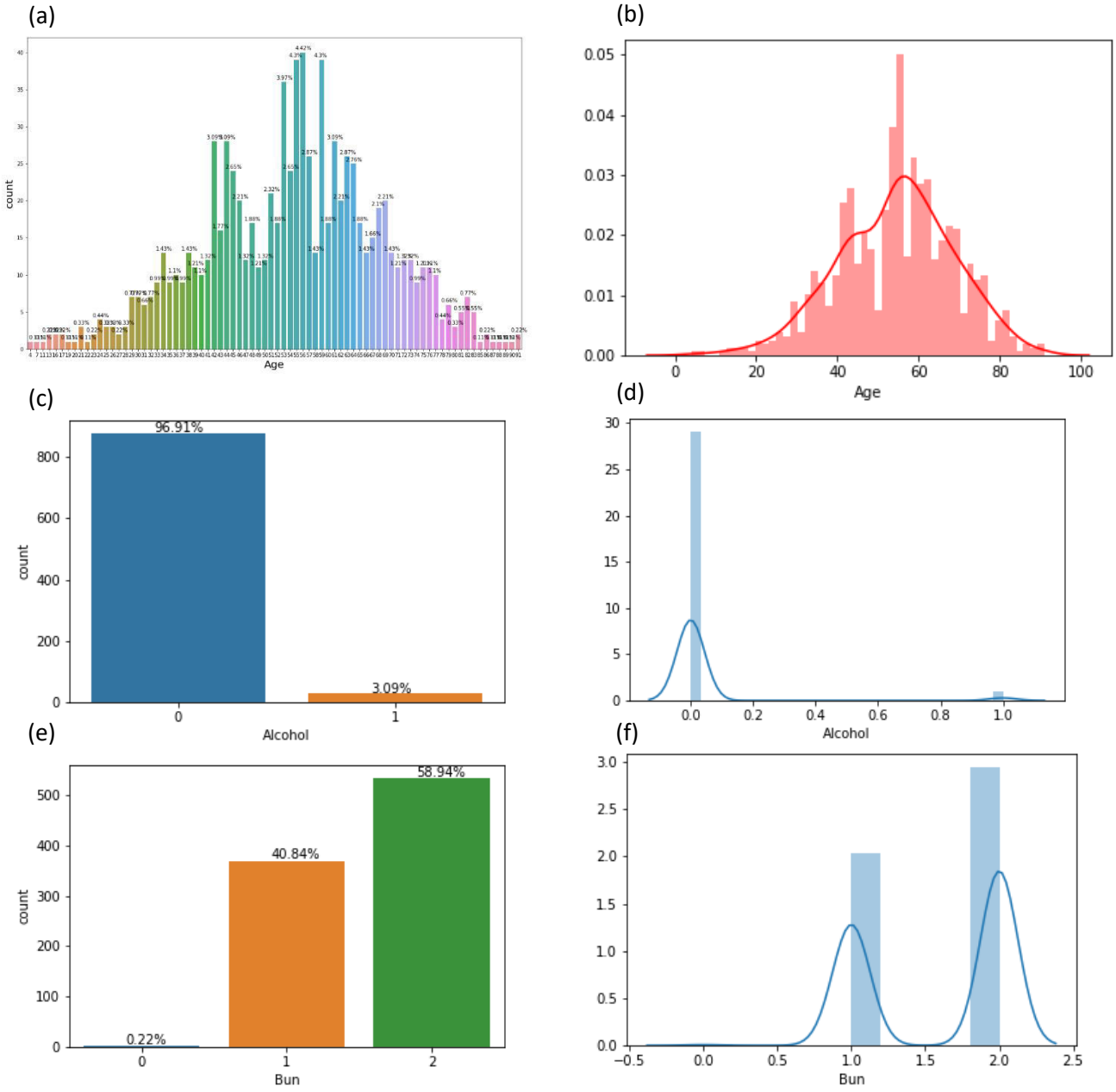
Tefekli, A., Cezayirli, F. 2013. The history of urinary stones: in parallel with civilization. The Scientific World Journal, 2013.

Trevor, H., Robert, T., JH, F. 2009. The elements of statistical learning: data mining, inference, and prediction.

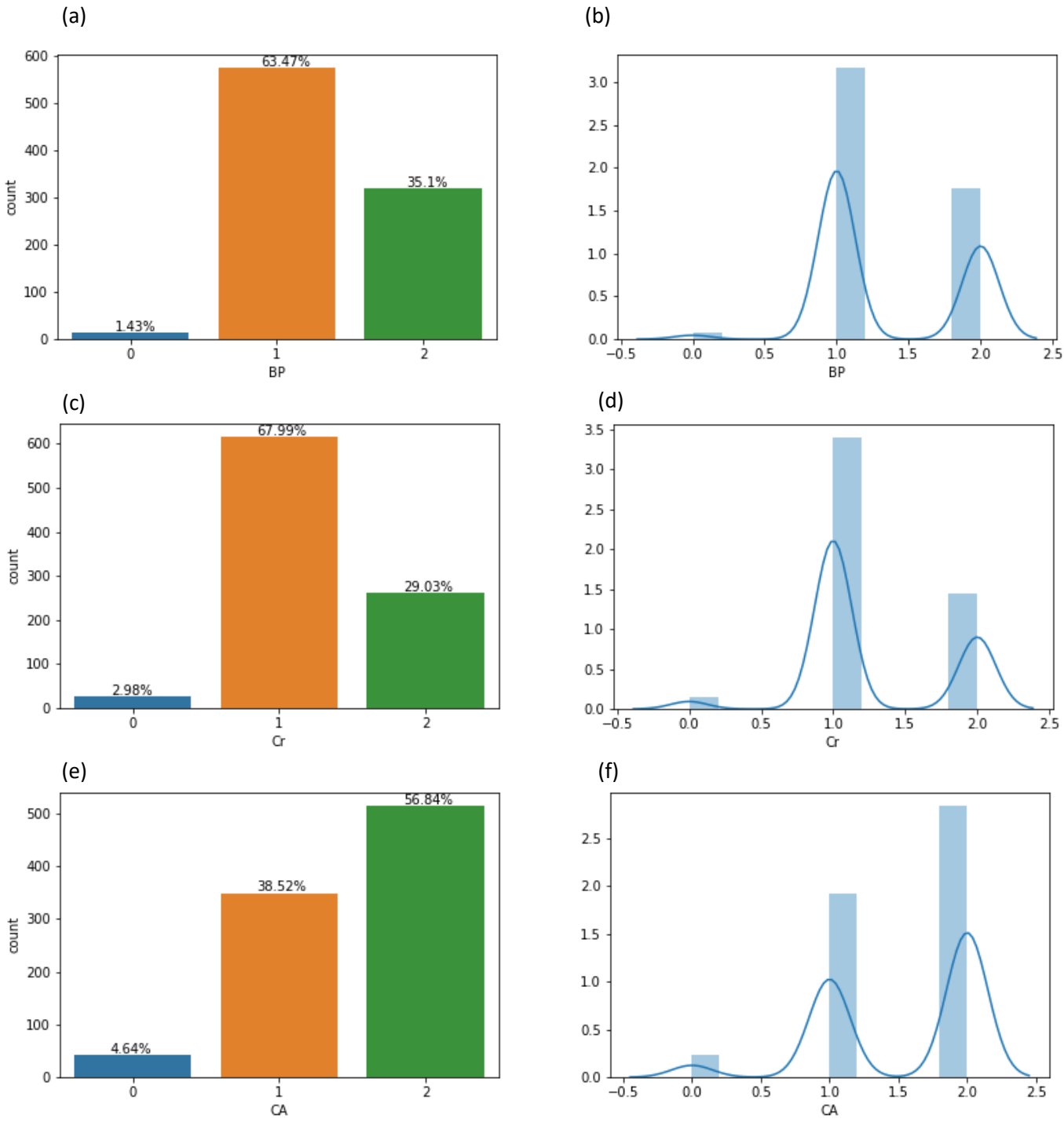
Vitale, C., Croppi, E., Marangella, M. 2008. Biochemical evaluation in renal stone disease. Clinical cases in mineral and bone metabolism, 5, 127.

Welch, N., Richter, C., Franklyn-Miller, A., Moran, K. 2019. Principal Component Analysis of the Biomechanical Factors Associated with Performance During Cutting. Journal of strength and conditioning research.

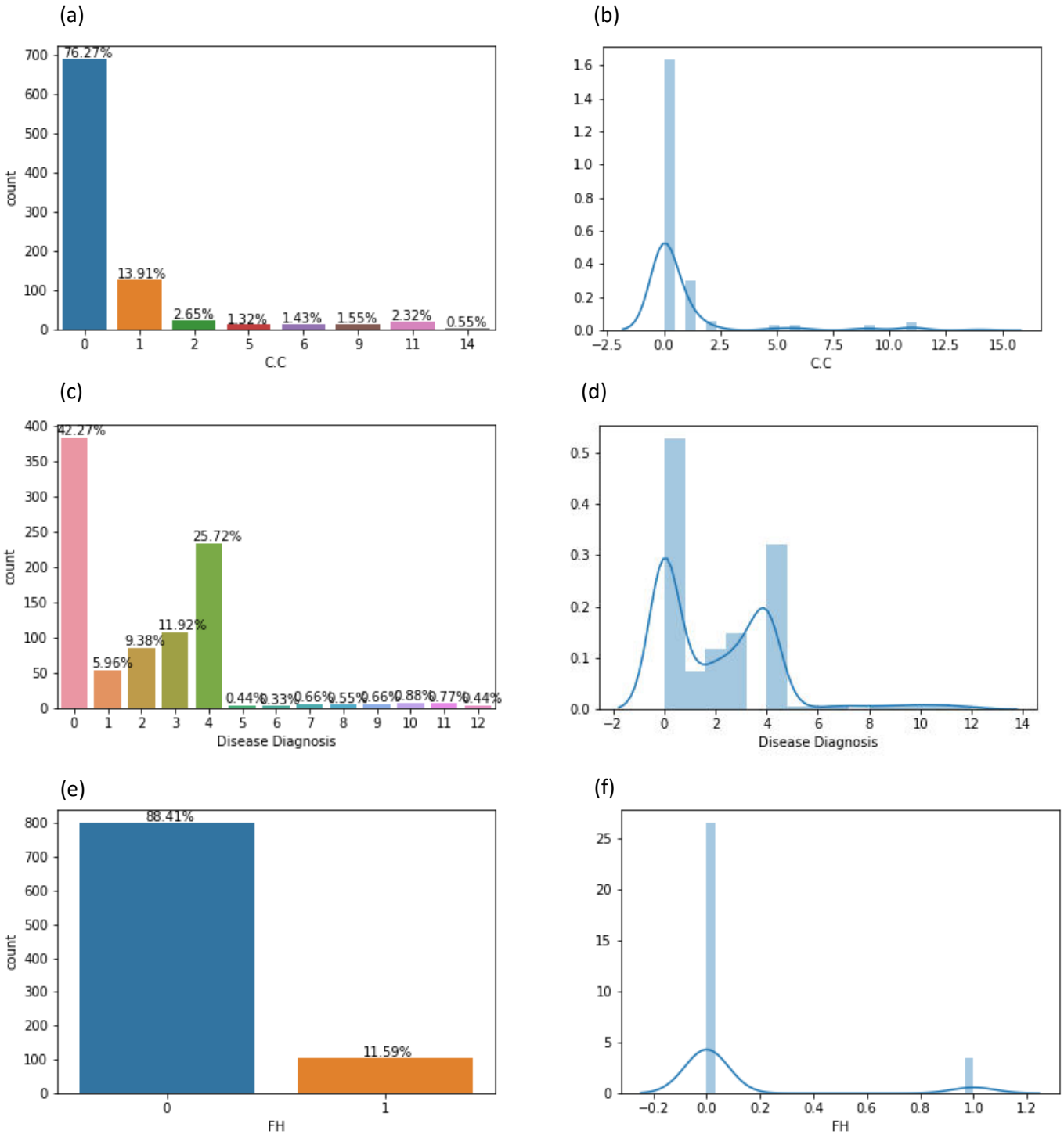
APPENDIX I



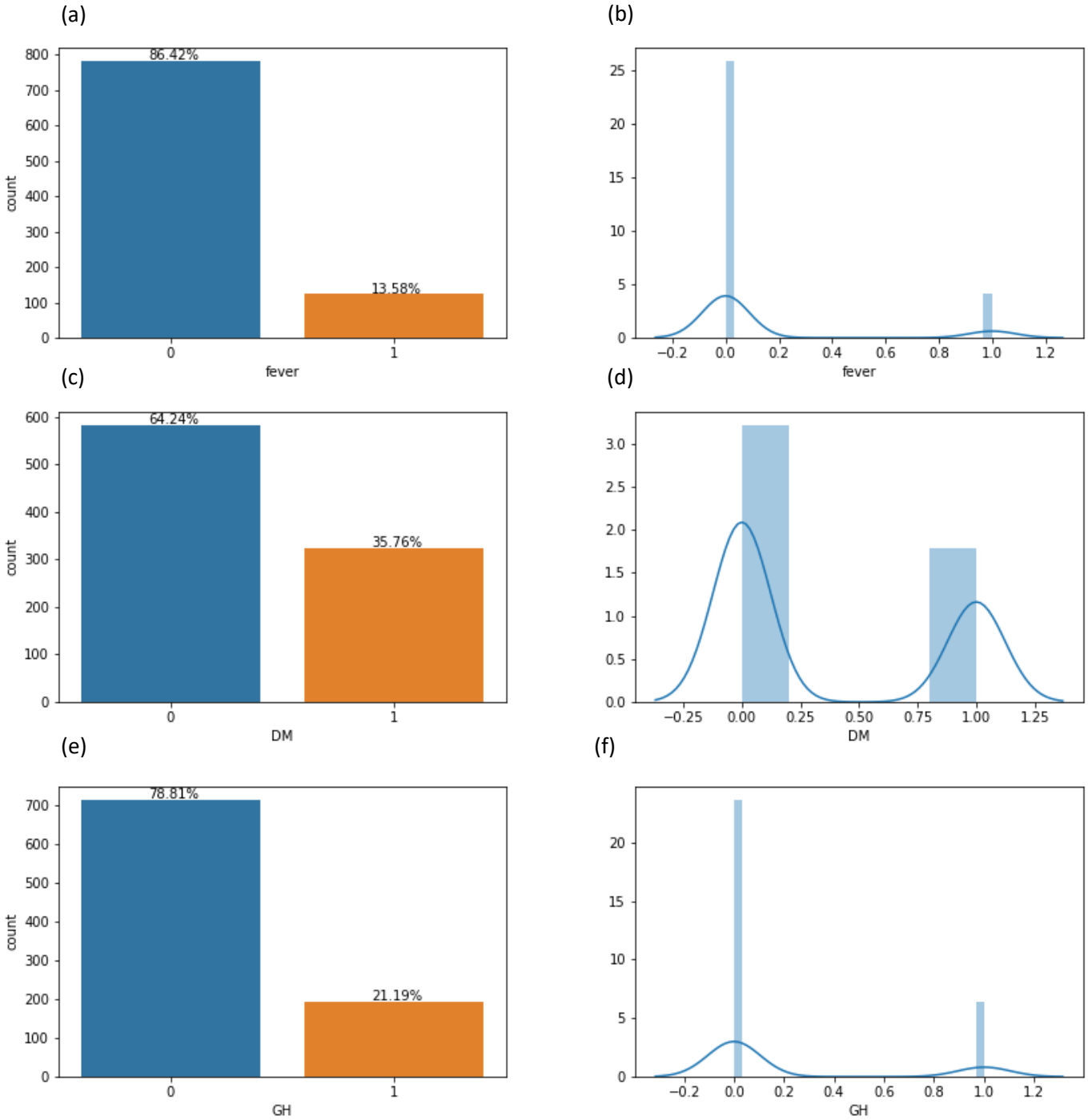
Supplementary Figure 1. Countplots and distribution plots, respectively, for **(a)** and **(b)**, Age; **(c)** and **(d)**, Alcohol; and **(e)** and **(f)**, Bun.



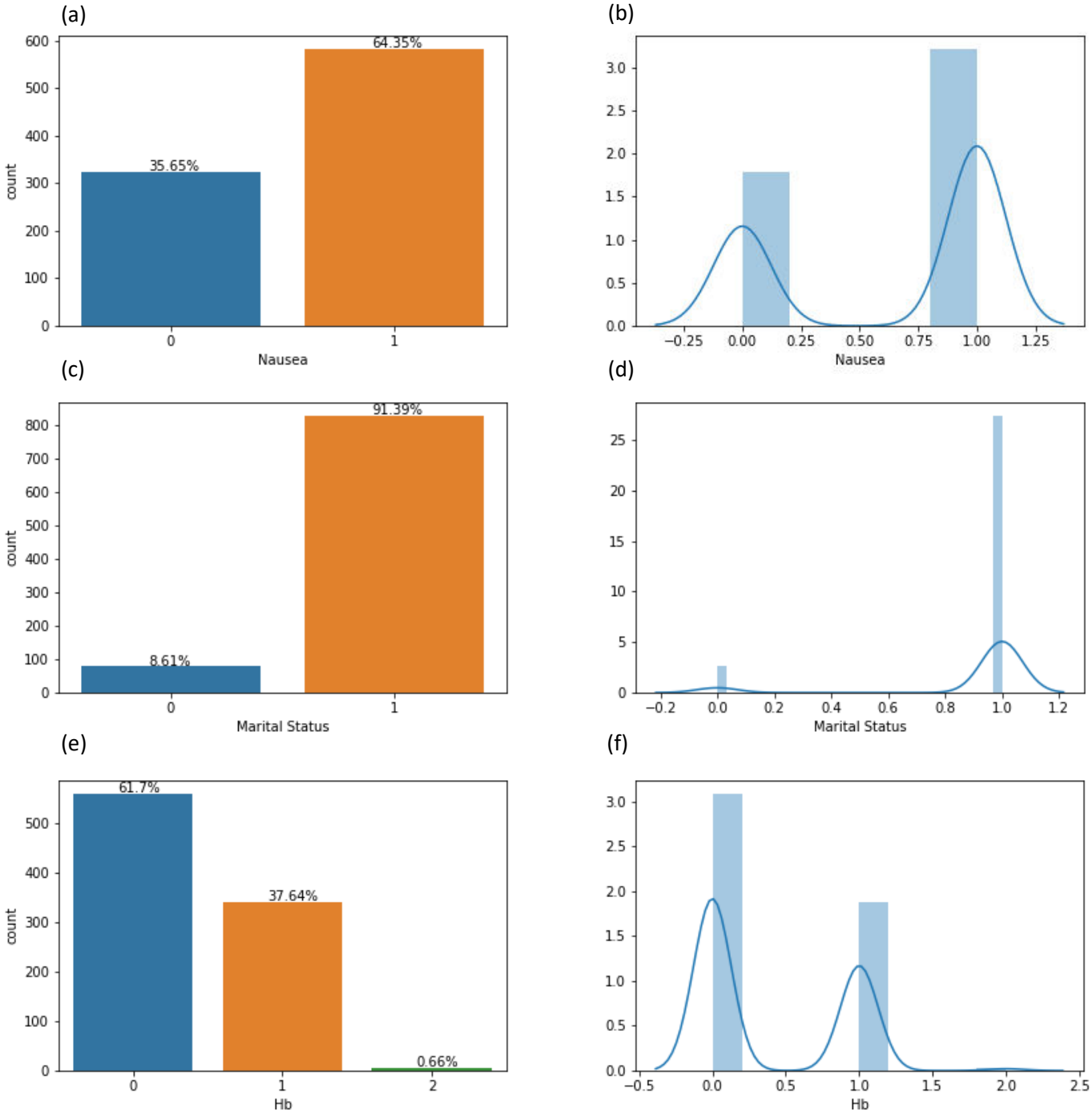
Supplementary Figure 2. Countplots and distribution plots, respectively, for **(a)** and **(b)**, BP; **(c)** and **(d)**, Cr; and **(e)** and **(f)**, CA.



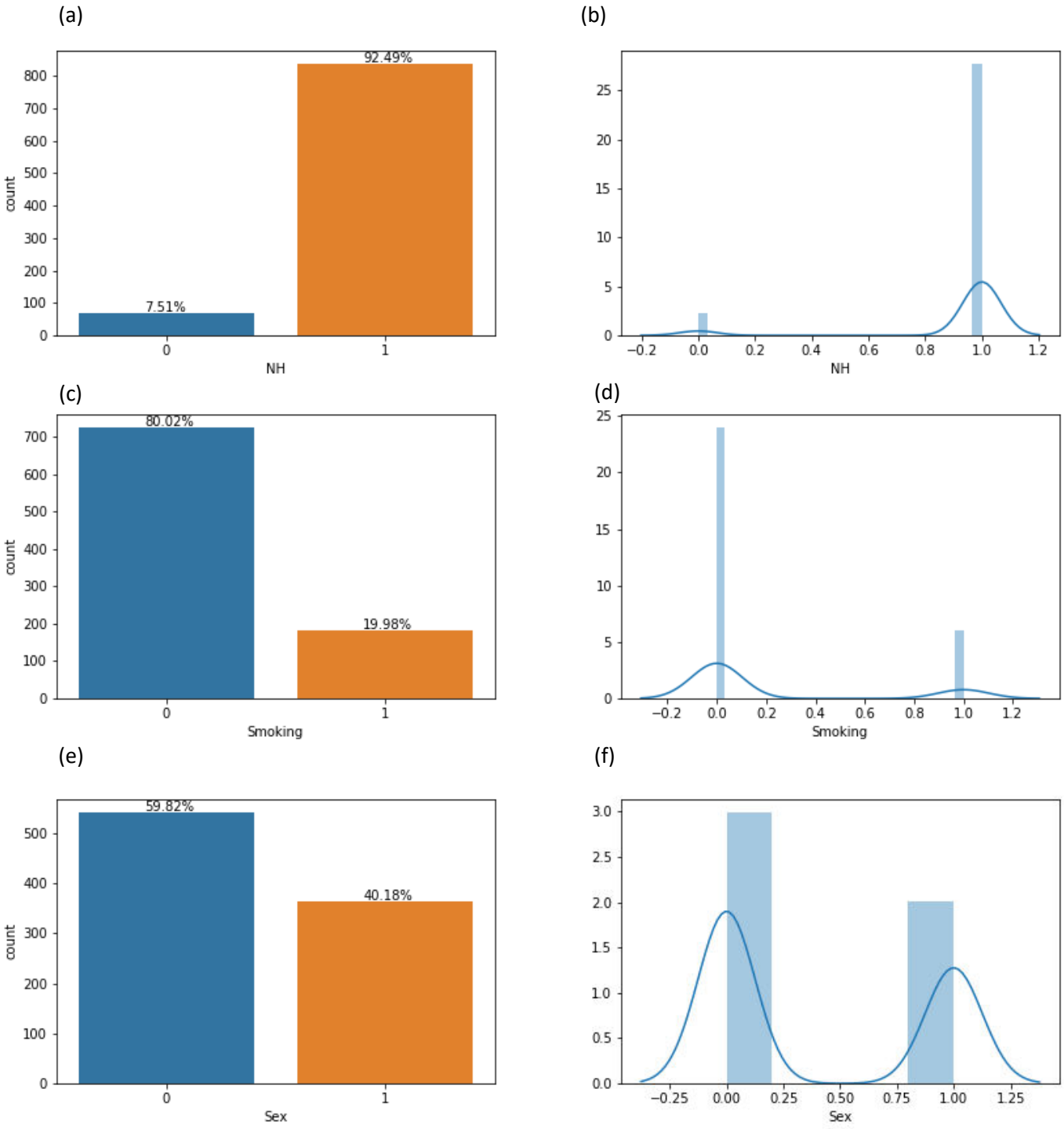
Supplementary Figure 3. Countplots and distribution plots, respectively, for **(a)** and **(b)**, C.C; **(c)** and **(d)**, Disease Diagnosis; and **(e)** and **(f)**, FH.



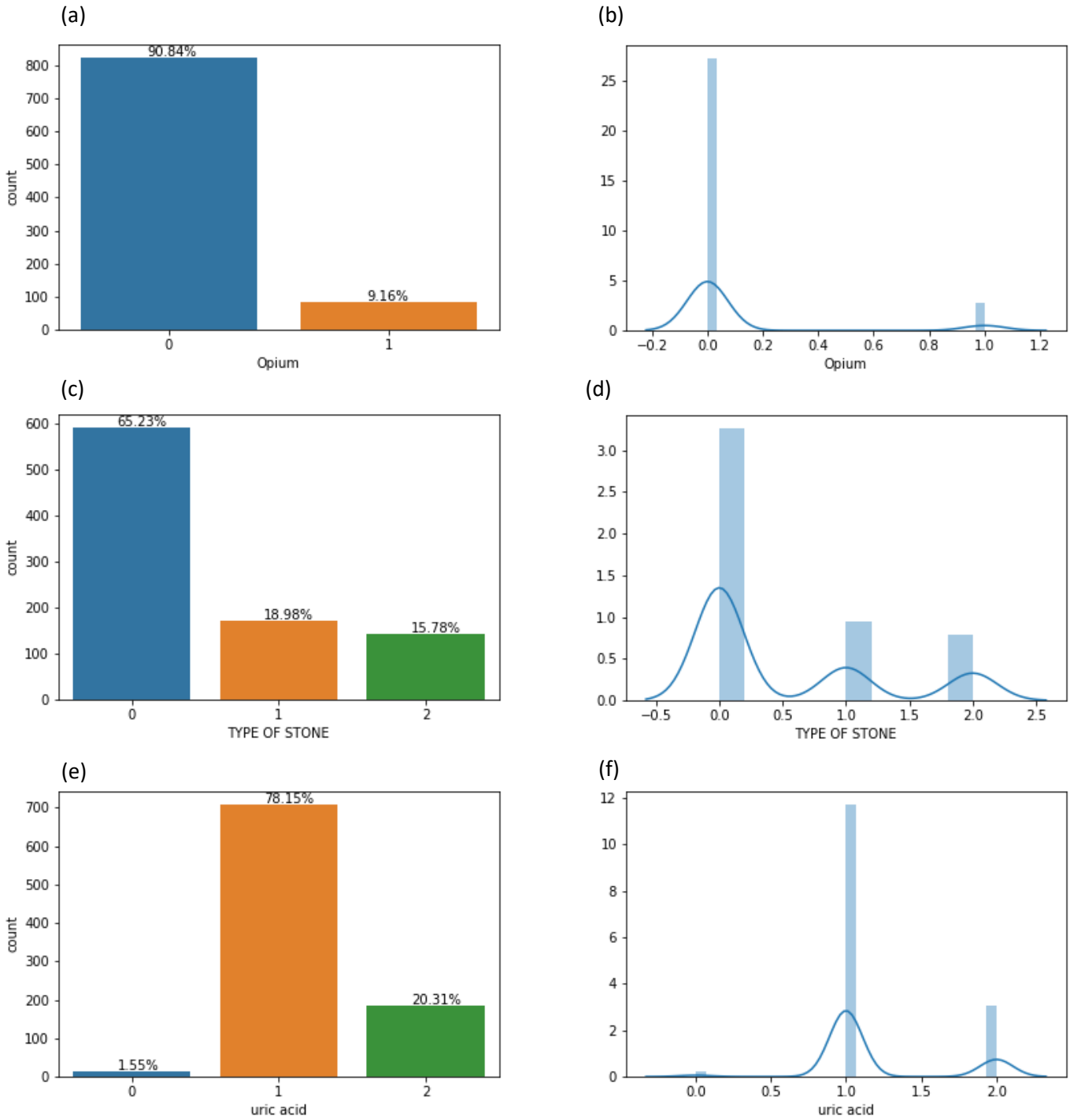
Supplementary Figure 4. Countplots and distribution plots, respectively, for **(a)** and **(b)**, fever; **(c)** and **(d)**, DM; and **(e)** and **(f)**, GH.



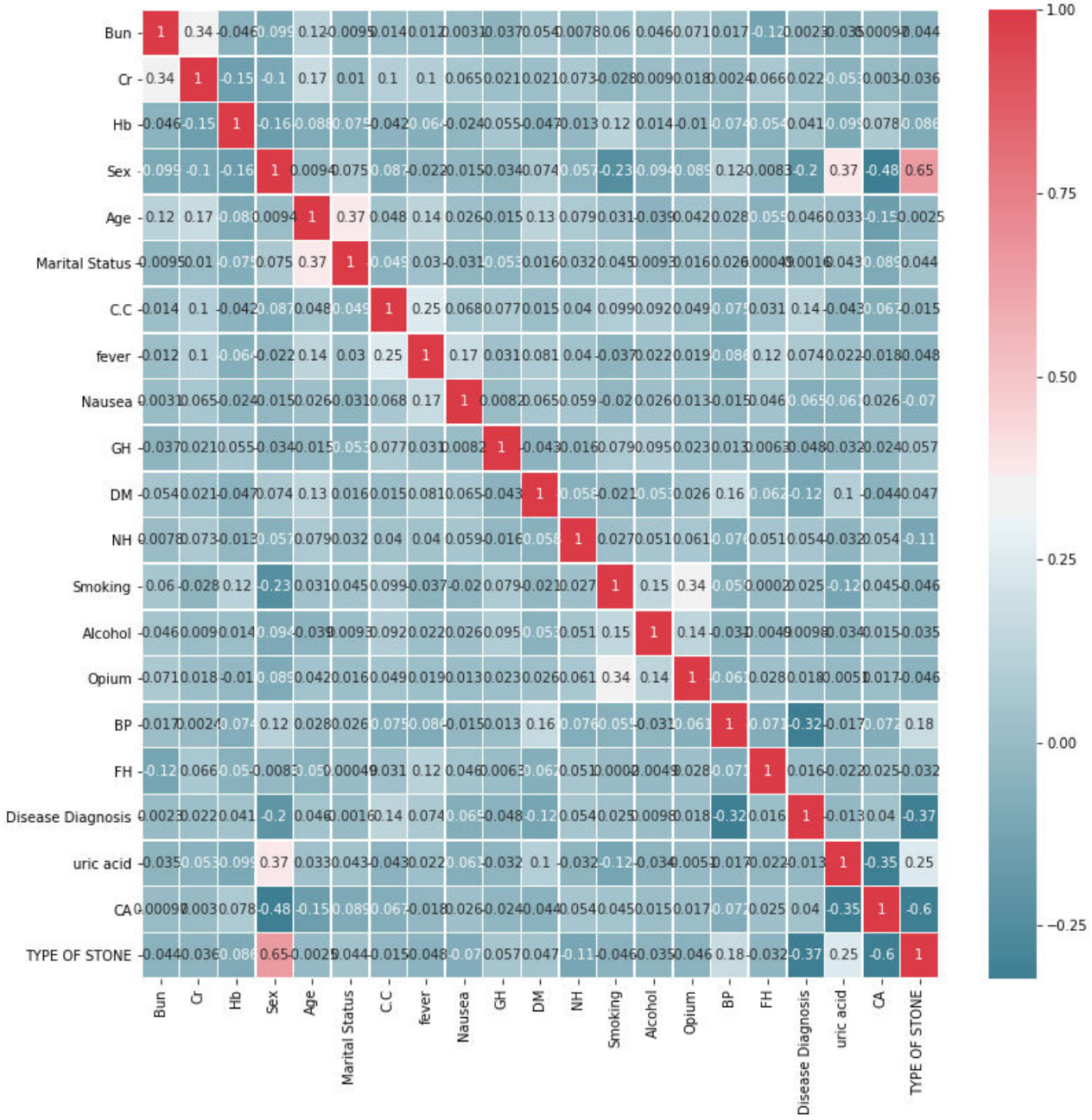
Supplementary Figure 5. Countplots and distribution plots, respectively, for **(a)** and **(b)**, Nausea; **(c)** and **(d)**, Marital Status; and **(e)** and **(f)**, Hb.



Supplementary Figure 6. Countplots and distribution plots, respectively, for **(a)** and **(b)**, NH; **(c)** and **(d)**, Smoking; and **(e)** and **(f)**, Sex.



Supplementary Figure 7. Countplots and distribution plots, respectively, for **(a)** and **(b)**, Opium; **(c)** and **(d)**, Type of stone; and **(e)** and **(f)**, uric acid.



Supplementary Figure 8. Correlation plot. The number in each square is the R-Pearson coefficient for every two variables.