Testing the Efficiency of Statistical Methods in Pattern Recognition via Simulation

Maria Teresinha Arns Steiner *Reinaldo Castro Souza*

[1]DMAT, UFPR
Curitiba (PR)

[2]DEE
PUC-RIO

Abstract

This paper aims to verify the efficiency of statistical methods in pattern recognition via discriminant analysis and logistical regression, where sets of multivariate synthetic data are generated based on a set of known data. An application to preliminary medical diagnosis is discussed.

Keywords: Pattern recognition, statistical methods, simulation, medical diagnosis.

1 Introduction

The application of discriminant analysis techniques to pattern recognition has been the focus of special attention on the part of researchers. Several situations where such methods could possibly be employed have been investigated and promising results achieved. Particular applications such as the prediction of bank failures [8, Tam et al, 1992], preliminary medical diagnosis [5, Mangasarian et al, 1990], [1, Bennett et al, 1992], in the paper industry [3, 1993], and in many another areas [7, Sharda, 1994] may be mentioned.

This paper uses statistical discriminant analysis and logistical regression, as presented in section 3, applied to a preliminary medical diagnosis, described in section 2, via data simulation, as described in section 4. In section 5, the procedure for testing the efficiency of statistical methods in pattern recognition is presented. The conclusions are presented in section 6.

Real data from patients are statistically analyzed, resulting on what we will call "modulated data". We will show below how patients data may be simulated or generated, in case the available number is insufficient, based on the data already gathered

and maintaining the correlation structure of the characteristics observed in the existing patients. Three simulation alternatives are presented. The statistical methods were then applied to the modulated and simulated data.

2 The Medical Problem [2, Champion et al, 1983]

Icterus (from the Greek ikteros - yellowishness) is merely a symptom represented by yellowish skin and mucosae. The same symptom is also sometimes evident in secretions. It can be originated from a vast universe of diseases which the physician must sort into two major initial groups :

a) Cholestasis (chole = bile, stasis = stop)

(difficult or impaired flow of bile components from the liver to the intestine)

b) Other causes

This study involves only the cholestasis group. The physician usually bases his initial diagnosis on simple, routine tests that translate, in essence, the biochemical consequences of the obstruction to the flow of bile. Hence, the physician defines with reasonable safety which patients present cholestatic syndrome. This, however, is not enough and further split into two more groups is called for :

a1) Obstruction by gallstones

a2) Obstruction by cancer

It is generally possible to make this differential diagnosis with the data already available in combination with other tools such as ultrasound or even computerized axial tomography scans. About 16 to 22% of all patients are not classified, though, and the complementary scans mentioned present errors between 30 and 40% in the region of the main biliary duct. Even when gallstones are located through such tests, there is frequent data overlap and even concomitance of diseases. The case of gallstones and gall bladder cancer may be mentioned as an example. When combined with the foregoing, tests for the determination of the real morbidity provide a level of precision of over 95%. However, they are usually expensive and may entail serious, even lethal complications.

A simple, reliable and cheap technique is therefore necessary to help the physician at this stage of this diagnosis.

3 Statistical Techniques

Given the result of clinical tests, the icteric patient may be considered as a 14-dimension random vector, composed by the following 14 components : age (ag), sex, total bilirubin (tb), direct bilirubin (db), indirect bilirubin (ib), alkaline phosphatase (ap), sgot, sgpt, protrombine activity time (pat), albumin (alb), amylase (amy), creatinine (cr), leukocytes (leu) and vg.

These components, pointed out by experts in the field, were taken from 118 patients (35 cancer patients and 83 gallstone patients). Multivariate statistical techniques may be used to classify a random vector in one of the two populations. In this paper, a comparative study was carried out among the following methods: Fisher's Linear Discriminant Function [4, Johnson et al, 1988], the Logistic Regression Model [6, Nelder et al, 1972], and the nearest K'-Neighbors Method [8, Tam et al, 1988].A description of these methods follows.

3.1 The Methods

The problem consists in making a distinction between points of the R^n space belonging to two sets A and B, with cardinalities : $|\mathrm{A}| = \mathrm{m}$ and $|\mathrm{B}| = \mathrm{k}$, with a view to classifying new points. The three widely known statistical methods described below were employed for this purpose.

3.1.1 Logistic Regression Model

The sigmoidal (logistic) mathematical function is used to model a dichotomic response variable Y, i.e., a random variable which assumes only one of two values (0 and 1), explained by several covariables with entries of the vector $\underline{\mathrm{x}}' = [x_1, x_2, ..., x_n]$, which usually represent factors of interest.

$$\mathrm{P(Y = y) = f(\underline{x}) = (1 + e^{-\eta})^{-1}} \qquad \mathrm{y = 0, 1}$$

where $\eta = \mathrm{g}(\underline{\mathrm{x}})$ is obtained by linear adjustment. The quality of the adjustment [6, Nelder et al, 1972] is measured by the deviance function defined below in section 3.2.1.

3.1.2 Fisher's Linear Discriminant Function

Fisher's method transforms multivariate observations $\underline{x} \in \mathrm{R}^n$ of the A and B sets into corresponding univariate observations Y as distant as possible. The method creates the Y as linear combinations of the $\underline{\mathrm{x}}$, i.e., $\mathrm{Y} = \underline{c}'\underline{\mathrm{x}}$, onde $\underline{c} \in \mathrm{R}^n$. The best linear combination derives from the ratio between the square of the distance between the sample averages of the two sets, $\underline{\bar{x}}_A$ and $\underline{\bar{x}}_B$, and the variance of Y. The common variance is estimated by S_p, the combined (pooled) sample variance matrix. In this context, Fisher's linear discriminant function is given by :

$$\mathrm{Y} = (\ \underline{\bar{x}}_A \text{ - } \underline{\bar{x}}_B\ )' S_p^{-1} \underline{x}$$

where S_p^{-1} is the inverse matrix of the pooled sample covariance:

$$S_p = \frac{(m-1)S_A + (k-1)S_B}{m+k-2}$$

A new $\underline{x}_0 \in \mathrm{R}^n$ observation is classified in relation to an average value obtained from:

$$q = \tfrac{1}{2}(\underline{\bar{x}}_A \text{ - } \underline{\bar{x}}_B\ )' S_p^{-1} (\underline{\bar{x}}_A + \underline{\bar{x}}_B)\ ,$$

i.e., if $\underline{x}_0 \in$ A then $\mathrm{y}_0 = (\ \underline{\bar{x}}_A \text{ - } \underline{\bar{x}}_B\ )'\ S_p^{-1} \underline{x}_0 \geq$ q e se $\underline{x}_0 \in$ B then $\mathrm{y}_0 <$ q. For more details see [4, Johnson et al, 1988].

3.1.3 Nearest K'-Neighbors Method according to Mahalanobis's Distance

This method designates a given observation $\underline{x} \in \mathrm{R}^n$ to the A or B group to which the majority of its K' neighbors belong. The $\mathrm{d}(\underline{x}, \underline{x}_i)$ distance between two observations $\underline{x}$ and $\underline{x}_i \in \mathrm{R}^n$ may be defined by the Mahalanobis's distance, the expression of which may be obtained as an extension of Fisher's Linear Discriminant Function [4, Johnson et al, 1988]:

$$\mathrm{D}^2 = (\underline{x} \text{ - } \underline{x}_i\ )'\ S_p^{-1}\ (\underline{x} \text{ - } \underline{x}_i\ ), \qquad \mathrm{i} = 1, \ldots, (\mathrm{m+k})$$

where D^2 is Mahalanobis's quadratic distance and $\underline{x}_i \in \mathrm{R}^n$ is an observation belonging to either A or B. Among the (m+k) distances measured, the smallest K' are taken, i.e., the K' distances that show the K' $\underline{x}_i$ vectors nearest to $\underline{x}$. After the group to which most of these $\underline{x}_i$ belong to is identified, the $\underline{x}$ observation is designated to this group.

3.2 Statistical Analysis

A statistical analysis of the data preceded the application of the methods described. The patient data collected and organized into an $\mathrm{X} \in \mathrm{R}^{(m+k)xn}$ matrix were analyzed through a logistic adjustment [6, Nelder et al, 1972], possibly followed by a discard of points, thereby achieving better performance in all methods, empirically observed in tests carried out before this paper. These statistical techniques are described below:

3.2.1 Logistic Adjustment

When searching for a Multiple Linear Logistic Model to fit the dichotomic response variable and several covariables, the adequacy of the model is measured based on the deviation function. The deviance function is defined by :

$$\mathrm{s}_p = \text{-}2\{\ \mathrm{L}_p \text{ - } \mathrm{L}_{(m+k)})\ \}$$

where L_p is the maximum of the log-likelihood function for the model under investigation with p parameters and $L_{(m+k)}$ is the maximum of the log-likelihood function for the saturated model. A poorly fitted model has a large deviation and, obviously, a well-fitted model has a small deviation (zero, in the saturated model). The degrees of freedom associated to the deviation are defined by ν = (m+k) - p. The deviance function is a measure of the distance between the fitted values and those observed or, likewise, between the current and the saturated model. In general, an attempt is made to find models with moderate deviations. The likelihood ratio test may be used to choose the most adequate model. The test statistics is :

$$s_p = -2 \{Lp - Lp+1\} \sim \chi^2_v$$

3.2.2 Possible Discarding of Points

Pearson's residues for each observation can be calculated after fitting a Multiple Linear Logistic Model and is given by :

$$e_i = \frac{(y_i - \theta_i)}{\sqrt{(\theta_i(1-\theta_i)}}$$

where y_i is the value assumed by the variable in the saturated model and θ_i is an estimate of such value given by the model. A value of $|e_i| \geq 1$ indicates that the i observation is being erroneously classified by the model, i.e., the i observation is "displaced" in relation to its population, which means that this is an atypical observation. It is suggested that justifications be sought for these cases. If encountered, the atypical i observation may be discarded from the sample and, consequently, from the model. Note that the model estimates must be recalculated in this case.

3.3 Obtention of the Modulated Data

Hotelling's commonly used T^2 test was initially applied whose statistics is given by:

$$T^2 \frac{(m+k-n-1)}{(m+k-2)n}$$

where $T^2 = (\underline{\bar{x}}_A - \underline{\bar{x}}_B)' [(\frac{1}{m} + \frac{1}{k})S_p]^{-1} (\underline{\bar{x}}_A - \underline{\bar{x}}_B)$ and is compared with $F_{n,m+k-n-1}(0.95)$ from a F distribution to test the equality of the mean vectors of the two multivariate populations A and B. Such values for the original data matrix were : $4.84 > 1.78896 = F_{14,103}(0.95)$. It is thus possible to state with 95% probability of certainty that the populations A and B are distinct. Therefore, the cancer patient population is different from the icteric population as calculated by the variables studied.

The following covariables were defined in the logistic adjustment: ag, tb, db2 (=db.db), db, amy, 1nam (=loge amy), st2 (=st.st), st (=sgot/sgtp), ap, ap2n (=ap.ap/1000), vg, vg2 (vg.vg), tb2 (=tb.tb), that is, 13 variables. Note that the following variables were discarded : sex, ib, pat, alb, cr, leu. Covariables sgpt and

sgot entered the adjustment through the ratio placed in st. As the aim was the prognosis (1 or 0) with the smallest possible error, the relationship of the response variable with the 14 original covariables and other covariables derived from them were analyzed based on the deviation function. The covariable was incorporated to the adjusted or non-adjusted model, depending on the deviation function value being statistically significant or not. Some covariables were transformed to scale, in an attempt to better grasp their information.

In the process of discarding points, 7 points, considered atypical, were discarded (6% of the total). This was defined after discussion with specialists and identification of the causes. Thus 111 modulated data were obtained, each with 13 variables.

4 Simulations

In order to measure and compare the efficiency of the three methods, 500 synthetic multivariate observations were generated for each of the two groups (A and B) from the modulated data (simulations 1 and 2) or the original data (simulation 3).The observations were generated after the definition of the probability distribution of each random variable of the vector (quantitative test results), i.e., the probability distributions were tentatively modeled and the model adopted was the one resulting from the best adjustment indicated by the Chi-square and Kolmogorov-Smirnov tests. These tests are presented in appendices 1a and 1b together with the histograms for some of the variables used in simulation 3. The same procedure was applied to simulations 1 and 2.

A new $W_{1000x13}$ (simulations 1 and 2) and $W_{1000x14}$ (simulation 3) data matrix is generated taking $E(\underline{w}) = A^{-1}\underline{\mu}$ as mean for the random variables, where $\underline{\mu}$ is the vector of the means of the known X_{111x13} (simulations 1 and 2) and X_{118x14} (simulation 3) data matrix and A is the matrix of the $AW = \underline{Y}$ transformation.

These synthetic observations , $\underline{Y}$, were then built with the same covariance structure of the original data and centered around the same point, which is ensured by the following result.

4.1 Result 4.1

Considering the random sample $[\ \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{(m+k-7)}]$ of the random vector $\underline{x}_i \in R^p$ (simulations 1 and 2) (or random sample $[\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n]$ of the random vector $\underline{x}_i \in R^n$ for simulation 3) so that $\underline{x}_i \sim .\ (\underline{\mu}, \Sigma)$, and considering $\underline{w}_i$ as the random vector with a distribution $\underline{w}_i \sim .\ (A^{-1}\underline{\mu}, V)$, with a mean of $A^{-1}\underline{\mu}$,being $\underline{\mu}$ of dimension p,V as a covariance matrix of order p x p and $A = P\Lambda^{1/2}V^{-1/2}$ as a transformation matrix, where P is the eigenvectors matrix and Λ is the eigenvalues matrix of Σ. Then $\underline{y} = A\underline{w}$ has a distribution with mean $\underline{\mu}$ and covariance matrix Σ, i.e., $\underline{y} \sim .(\underline{\mu}, \Sigma)$.

The proof is straight forward and is omitted here.

The Minitab statistical package was used in the calculations needed to obtain the AW matrix and the correlation matrices mentioned ahead. The methods investigated were implemented in Pascal language, the statistical package GLIM (General Algebraic Modeling Systems) being preliminarly used to obtain the logistical regression model.

5 Procedure for Testing the Efficiency of Statistical Methods in Pattern Recognition

The numeric data available for the problem, shortly described in section 3, were used to verify the performance of the methods investigated concerning to the precision in the classification of new points.

5.1 Using the Real Data

Applying the methodology proposed by Bennett, 1992, the tests were performed in the following way: we randomly divided the set of points into two subsets. One of the subsets, the Training Set (Tr.S.) was used to train the program, and the other subset, the Testing Set (Tt.S.), was used to verify the trained program. The subset Tr.S. was also used to test the program as well. To perform the tests, we have got 3 Tr.S. with 106 points each one from the data original matrix and then we have got 3 Tt.S. with 12 points. These data were called Real Data 1. Furthermore, it was obtained 3 Tr.S. of 100 points each one, from the modelated data, with 3 Tt.S. of 11 points each one to perform the tests. These were called Real Data 2. The average of percentages of error for the three tests, in each case (Real Data 1 and Real Data 2) was calculated and it is showed in table 1.

5.2 Using the Simulated Data

First Simulation

The 111 points modulated with 13 variables were used to generate 1,000 points (500 for each group, cancer or gallstones). Appendix 2 shows the correlation matrices between X_{111x13} (known data matrix) and AW ($W_{1000x13}$ generated data matrix) for each group. These 111 points served to train the programs of each of the three methods focused. This training set (Tr.S.) was tested and the percentage of errors calculated. Afterwards, the trained program was used to verify the percentage of errors of the 1,000 generated points, the testing set (Tt.S.), contained in the AW matrix. These percentages are shown in table 1.

Second Simulation

The second simulation is a variation of the first. The set of 1,000 generated points was divided into two sub-sets : 600 of them formed the training set and the remaining

400 points, the testing set. The percentage of errors for the two sub-sets appears in table 1.

Third Simulation

In this third simulation the 118 original points, with 14 variables, were used to generate 1,000 points. The correlation matrices of X_{118x14} and AW, $W_{1000x14}$, for each group are presented in Appendix 2. This set of generated points was then divided into two sub-sets: 600 of these were submitted to logistic adjustment and discarding of points, which resulted in a sub-set of 579 points and 14 variables, the training set. The remaining 400 points formed the testing set. The percentages of errors are found in table 1. The Figure 1 shows the ways used to determine the subsets used to perform the tests.

Simulation	Log. Reg.	Log. Reg.	Fisher's	Fisher's	K' Neighbors	K' Neighbors
	Tr.S	Tt.S	Tr.S.	Tt.S.	Tr.S	Tt.S.
real data 1	16.67	19.44	18.87	19.44	22.95	25.00
real data 2	1.50	4.55	6.66	9.09	15.33	12.12
simulation 1	0	7.80	6.30	8.80	18.00	14.60
simulation 2	7.16	9.00	5.26	13.16	3.29	7.89
simulation 3	2.00	8.75	9.21	17.11	9.21	10.53

Table 1. Misclassification Percentual Rates, (P[(A|B)∪(B|A)]), where Tr.S. = Training Set, Tt.S. = Testing Set.

6 Conclusions

It may be observed in table 1 that the percentage of error in the Tr.S. is very low when the Logistic Regression is employed, except for simulation 2, where numbers above the Fisher's L.D.F. and K'-Neighbors methods are found. The percentage of points for the Tt.S. proves to be of a higher magnitude than those of the Tr.S., as expected, since they did not take part in the training. In this case, the Logistic Regression also evidenced a better performance than the other methods.

It is therefore possible to conclude that the Logistic Regression model, when used to differentiate two multivariate data groups subjected to the analysis described under 3.2 is more efficient than Fisher's L.D.F. and the K'-Neighbors method. The consistency of this statement is attested by the size of the samples generated.

During the statistical work developed in this study the result 4.1 was used in the generation of synthetic data with a covariance matrix Σ and a $\underline{\mu}$ mean vector, starting from a real multivariate data sample with such parameters. The validity of such result may be observed in the similarity between the correlation matrices of the original data and that of the generated data contained in Appendix 2.

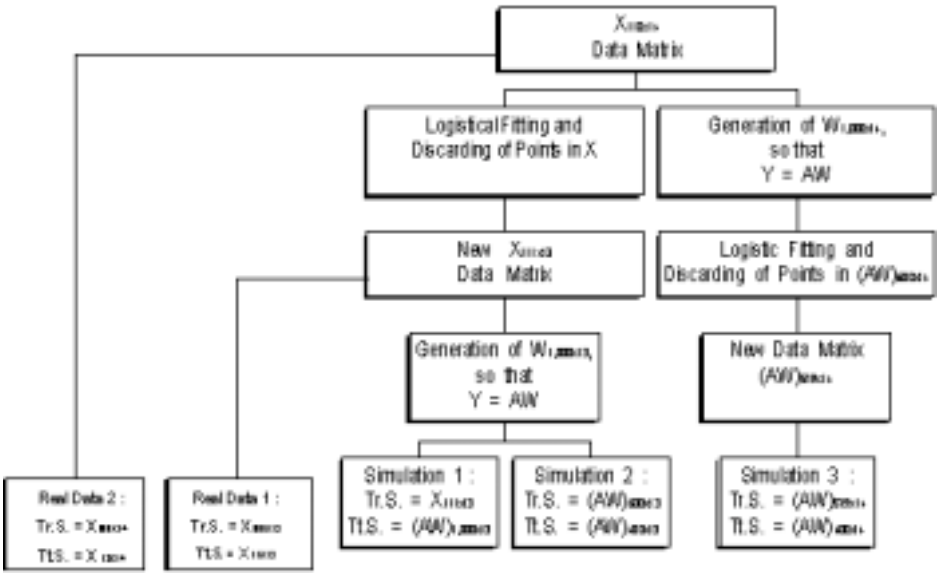


Fig. 1: The subsets for the tests, Tr.S. and Tt.S., using the real data and the simulated data.

References

[1] BENNETT, K. P. & MANGASARIAN, O. L. Robust Linear Programming Discrimination of Two Linearly Inseparable Sets, Optimization Methods and Software, 1992, vol.1, p. 23-34

[2] CHAMPION, H. R.; SACCO, W. J. and HUNT, T. K. , Trauma Severity Scoring to Predict Mortality, World J. Surg. 7, 1983, p. 4 - 11

[3] FADUM, O. Artificial Intelligence : expert systems, fuzzy logic and neural network applications in the paper industry, Pulp & Paper, 1993.

[4] JOHNSON, R. A. and WICHERN, D, W. Applied Multivariate Statistical Analysis, New Jersey, Prentice-Hall, inc., 1988.

[5] MANGASARIAN, O. L., SETIONO, R. & WOLBERG, W. H. Pattern Recognition via Linear Programming : Theory and Application to Medical Diagnosis, in : Large-Scale Numerical Optimization, Thomas F. Coleman and Yuying Li,

(Eds.), SIAM, Philadelphia 1990, p.22-30.

[6] NELDER, J. A. and WEDDERBURN, R. W. M. Generalized Linear Models, J. R. Statist. Soc., A, n.135, 1972, p. 370-384.

[7] SHARDA, R. Neural Networks for the MS/OR Analyst : An Application Bibliography, Interfaces, 1994, vol. 24, n. 2, p. 116-130.

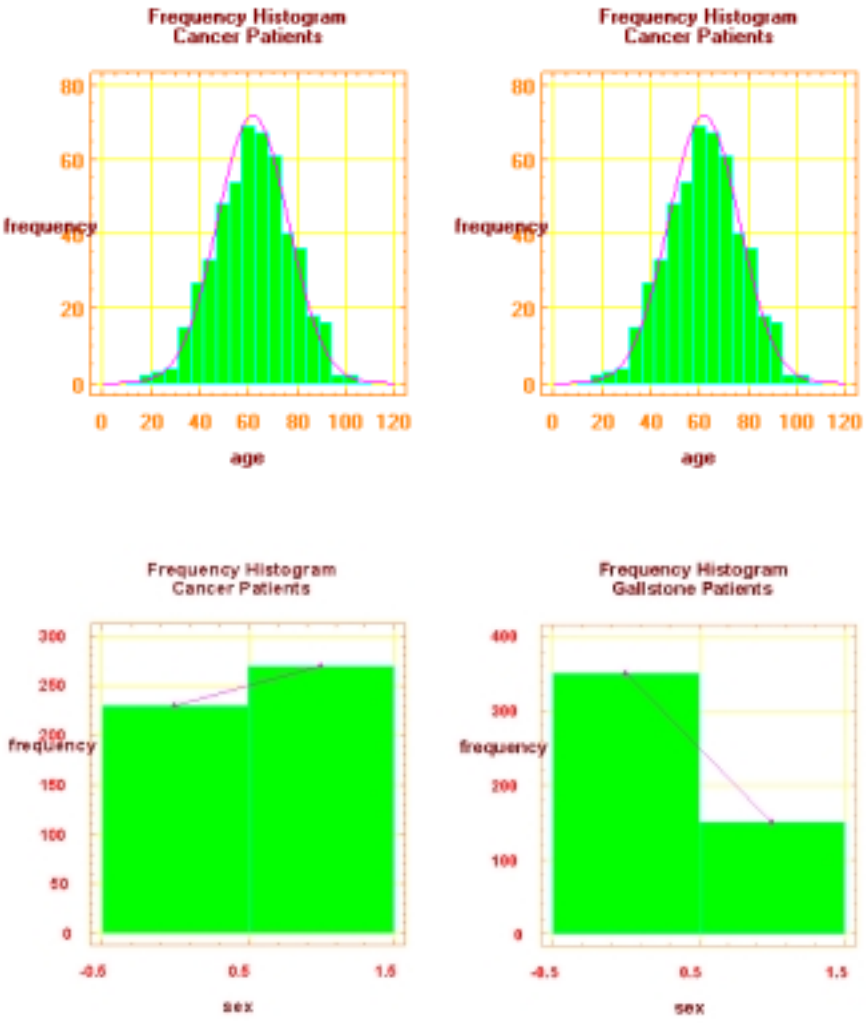
[8] TAM, K. Y. and KIANG, M. T. Managerial Applications of Neural Networks : The Case of Bank Failure Predictions, Management Sciences 39 n.7, 1992, p. 926 - 947.

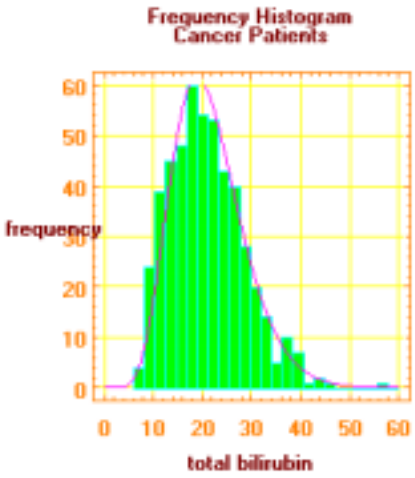
Appendix 1

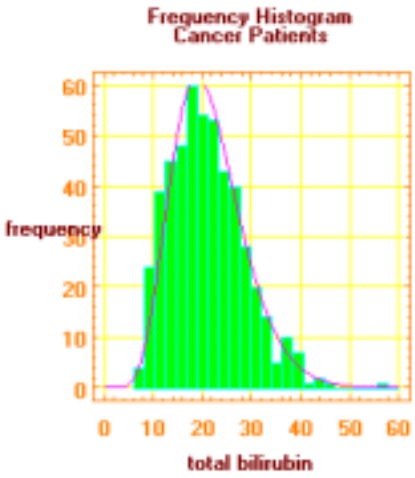
Appendix 1a. Tests for the variables under study (simulation 3)

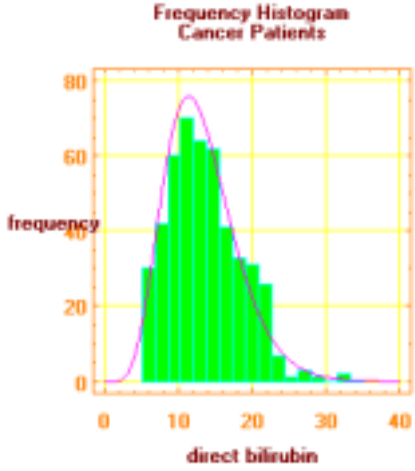
Variable	Origin	Dist.	Mean / Standard Deviation	KS or CSq (p)
Age	cancer	Normal	$\overline{x}$ = 57.1777 / s = 13.0607	0.995009
	gallstone		$\overline{x}$ = 47.4858 / s = 17.7342	0.657535
Sex	cancer	Bernoulli	$\overline{x}$ = 0.568 / s = 0.4954	
	gallstone		$\overline{x}$ = 0.300 / s = 0.4583	
Total bilirubin	cancer	Gama	$\overline{x}$ = 22.060 / s =7.802 ($\widehat{\alpha}$ = 7.994 / $\widehat{\beta}$ =0.362)	0.885345
	gallstone		$\overline{x}$ = 8.6072 / s =7.075 ($\widehat{\alpha}$ =1.479 $\widehat{\beta}$ = 0.171)	0.564497
Direct bilirubin	cancer	Gama	$\overline{x}$ = 12.8959 / s =4.504 ($\widehat{\alpha}$ = 8.195 $\widehat{\beta}$ = 0.635)	0.993016
	gallstone		$\overline{x}$ = 5.1492 / s = 4.504 ($\widehat{\alpha}$ = 1.306 $\widehat{\beta}$ = 0.253)	0.905743
Ind. bilirubin	cancer	Gama	$\overline{x}$ = 9.3517 / s = 4.355 ($\widehat{\alpha}$ = 4.610 $\widehat{\beta}$ = 0.492)	0.932499
	gallstone		$\overline{x}$ = 3.426 / α = 2.869 ($\widehat{\alpha}$ = 1.426 $\widehat{\beta}$ = 0.416)	0.900159
sgp	cancer	Gama	$\overline{x}$ = 81.918 / s =59.943 ($\widehat{\alpha}$ = 1.867 $\widehat{\beta}$ = 0.023)	0.680497
	gallstone		$\overline{x}$ = 86.1387 / s =83.158 ($\widehat{\alpha}$ =1.073, $\widehat{\beta}$ =0.012)	0.734534
sgot	cancer	Gama	$\overline{x}$ =97.3111 / s =49.210 ($\widehat{\alpha}$ =3.910 $\widehat{\beta}$ =0.040)	0.736189
	gallstone		$\overline{x}$ = 92.9211 / s=76.801 ($\widehat{\alpha}$ =1.464 $\widehat{\beta}$ =0.016)	0.942223
A. phosphatase	cancer	Gama	$\overline{x}$ = 383.194 / s =259.269 ($\widehat{\alpha}$ =2.184 $\widehat{\beta}$ =0.006)	0.383788
	gallstone		$\overline{x}$ = 226.044 / s =184.944 ($\widehat{\alpha}$ = 1.494 $\widehat{\beta}$ =0.007)	0.664853
Amylase	cancer	Gama	$\overline{x}$ = 103.003 / s =43.568 ($\widehat{\alpha}$ = 5.589 $\widehat{\beta}$ = 0.054)	0.587319
	gallstone		$\overline{x}$ = 198.970 / s = 167.519 ($\widehat{\alpha}$ = 1.41 $\widehat{\beta}$ = 0.007)	0.486539
tap	cancer	Gama	$\overline{x}$ = 14.2277 / s =1.5945 ($\widehat{\alpha}$ =79.619 $\widehat{\beta}$ = 5.596)	0.820555
	gallstone		$\overline{x}$ =13.9525 / s =1.464 ($\widehat{\alpha}$ =90.855 $\widehat{\beta}$ =6.512)	0.712707
Albumin	cancer	Normal	$\overline{x}$ =3.05277 / s =0.489265	0.903005
	gallstone		$\overline{x}$ =3.0141 / s = 0.68526	0.882127
Creatinine	cancer	Gama	$\overline{x}$ = 0.8582 / s =0.267 (α =10.306 $\widehat{\beta}$ =12.009)	0.981501
	gallstone		$\overline{x}$ = 0.9658 / s =0.415 (α =5.411 $\widehat{\beta}$ = 5.603)	0.900468
Leukocytes	cancer	Gama	$\overline{x}$ = 9.8722 / s =3.132 (α = 9.933 $\widehat{\beta}$ =1.006)	0.631138
	gallstone		$\overline{x}$ = 9.3099 / s =3.438(α =7.333 $\widehat{\beta}$ =0.787)	0.82377
Vg	cancer	Normal	$\overline{x}$ =36.4848 / s =6.03379	0.843828
	gallstone		$\overline{x}$ =39.0876 / s =6.50199	0.989614

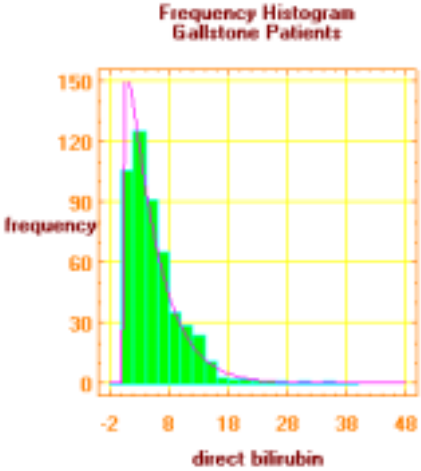
Appendix 1b. Histograms for the variables under study (simulation 3)

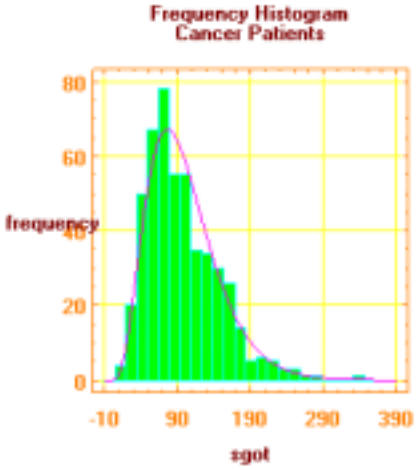


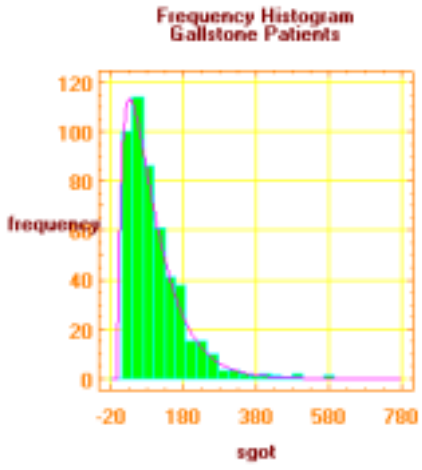


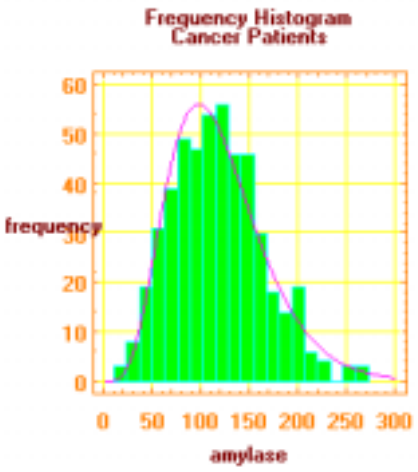


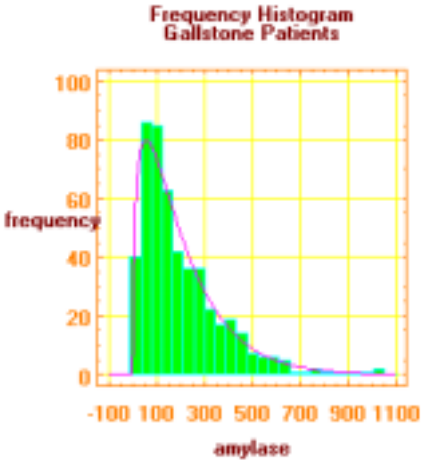


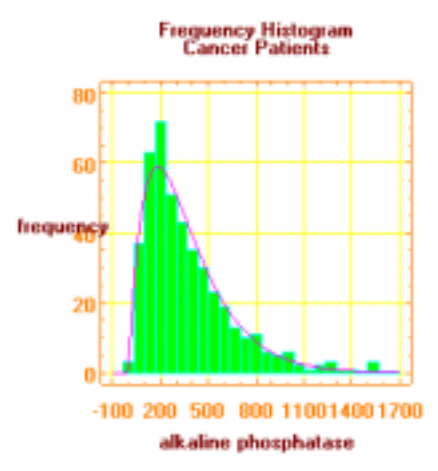


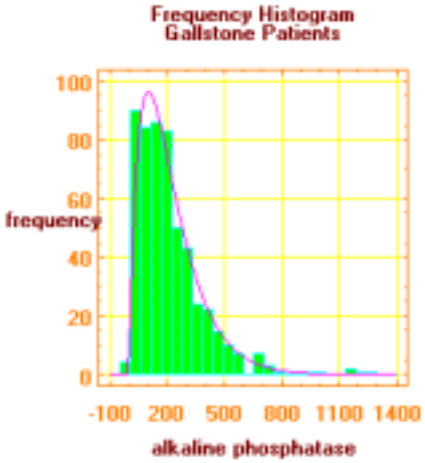


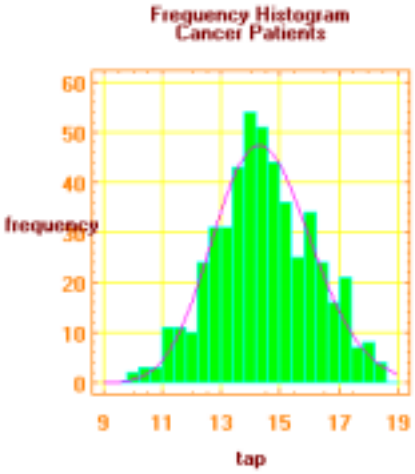


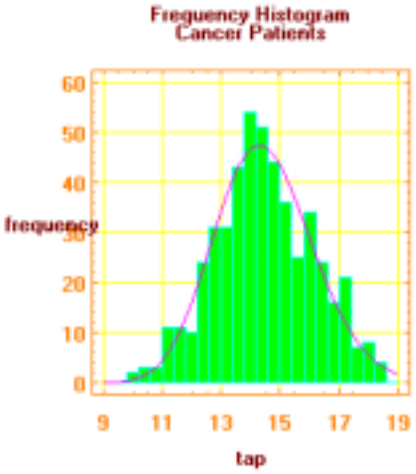


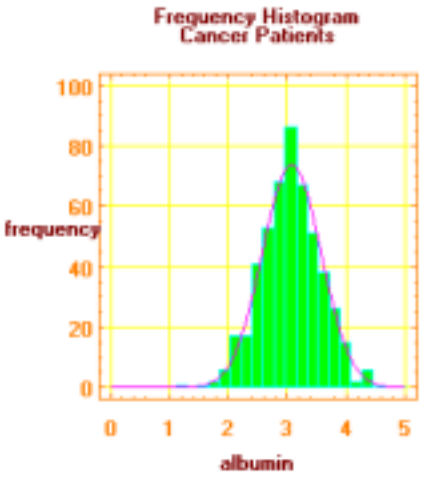


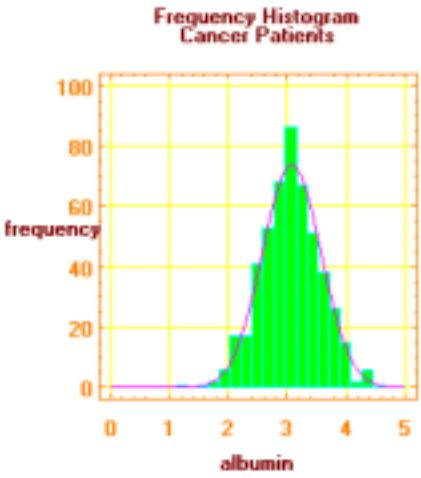


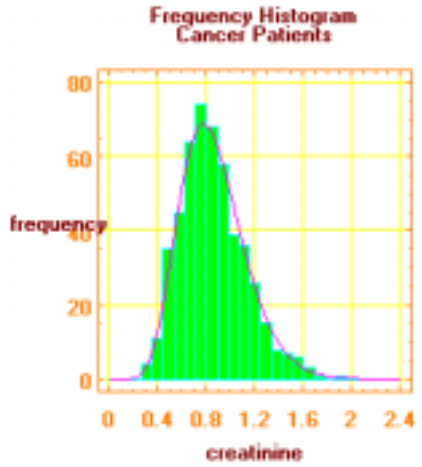


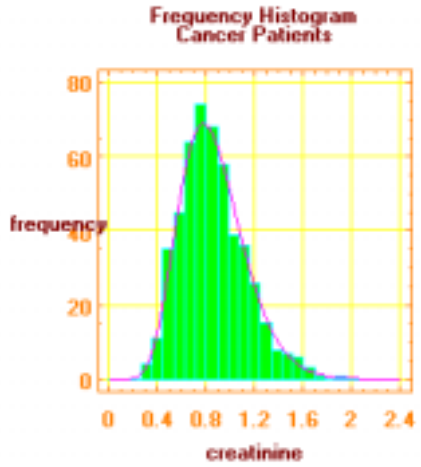


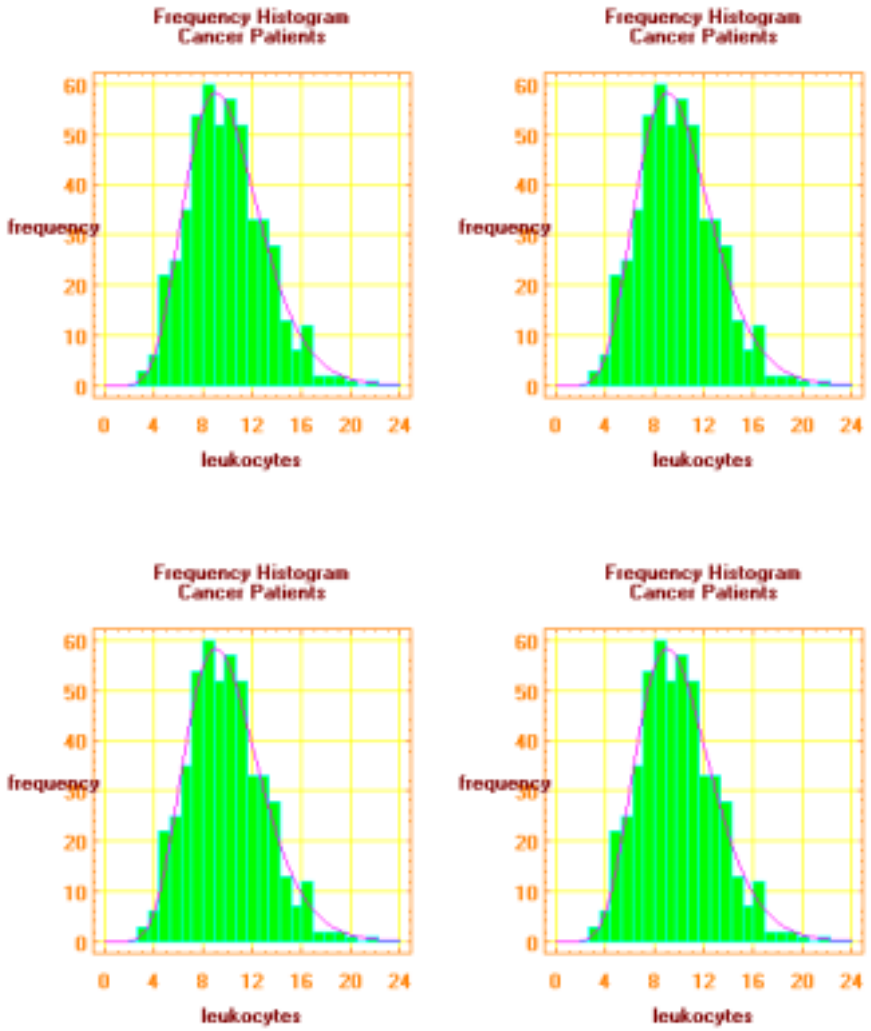












Appendix 2: Correlation Matrices

Correlation matrix for the 80 gallstone patients (simulations 1 and 2)

1.00000 0.09342 0.00072 0.04691 -0.10800 0.00006 0.21470 0.20604 -0.07229 -0.07065 0.14791 0.13309 0.08441

0.09342 1.00000 0.89506 0.97470 0.01060 -0.00350 -0.00135 -0.00388 0.33645 0.36554 0.04619 0.07173 0.95917

0.00072 0.89506 1.00000 0.93816 0.02847 0.04221 -0.06995 -0.06629 0.62870 0.66783 -0.05212 -0.03237 0.94199

0.04691 0.97470 0.93816 1.00000 0.02796 0.01981 -0.04046 -0.03355 0.44371 0.47606 0.02601 0.05369 0.93663

-0.1080 0.01060 0.02847 0.02796 1.00000 0.85529 -0.04263 -0.02239 0.13618 0.10653 -0.08344 -0.08942 -0.00969

0.00006 -0.00350 0.04221 0.01981 0.85529 1.00000 0.03194 0.06339 0.21100 0.17763 -0.17689 -0.18775 -0.00918

0.21470 -0.00135 -0.06995 -0.04046 -0.04263 0.03194 1.000 0.95146 -0.11472 -0.05758 -0.08710 -0.12184 -0.05335

0.20604 -0.00388 -0.06629 -0.03355 -0.02239 0.06339 0.95146 1.000 -0.12121 -0.04853 -0.08972 -0.12277 -0.06262

-0.07229 0.33645 0.62870 0.44371 0.13618 0.21100 -0.11472 -0.12121 1.00000 0.94524 -0.01488 -0.02437 0.41704

-0.07065 0.36554 0.66783 0.47606 0.10653 0.17763 -0.05758 -0.04853 0.94524 1.00000 -0.06098 -0.07619 0.44555

0.14791 0.04619 -0.05212 0.02601 -0.08344 -0.17689 -0.08710 -0.08972 -0.01488 -0.06098 1.0000 0.98286 0.02177

0.13309 0.07173 -0.03237 0.05369 -0.08942 -0.18775 -0.12184 -0.12277 -0.02437 -0.07619 0.98286 1.0000 0.04506

0.08441 0.95917 0.94199 0.93663 -0.00969 -0.00918 -0.05335 -0.06262 0.41704 0.44555 0.02177 0.04506 1.00000

Correlation matrix for the 500 gallstone "generated patients" (simulations 1 and 2)

1.00000 0.14894 0.00903 0.08249 -0.19764 -0.06610 0.23509 0.22291 -0.11568 -0.15251 0.10952 0.08566 0.12543

0.14894 1.00000 0.88068 0.97254 0.03098 0.01793 0.01295 -0.00961 0.27418 0.30636 0.05496 0.07783 0.95738

0.00903 0.88068 1.00000 0.92637 0.03157 0.03541 -0.06029 -0.08116 0.59684 0.64652 -0.04498 -0.02708 0.93456

0.08249 0.97254 0.92637 1.00000 0.03806 0.03165 -0.02839 -0.04291 0.38777 0.42299 0.03250 0.05834 0.93087

-0.19764 0.03098 0.03157 0.03806 1.00000 0.84880 -0.07075 -0.07004 0.11020 0.09229 -0.06393 -0.06945 0.00681

-0.06610 0.01793 0.03541 0.03165 0.84880 1.00000 0.03643 0.05501 0.18775 0.15402 -0.17763 -0.18303 -0.00335

0.23509 0.01295 -0.06029 -0.02839 -0.07075 0.03643 1.000 0.95199 -0.10917 -0.05646 -0.03405 -0.07142 -0.02554

0.22291 -0.00961 -0.08116 -0.04291 -0.07004 0.05501 0.95199 1.000-0.13128 -0.06757 -0.05118 -0.08395 -0.05909

-0.11568 0.27418 0.59684 0.38777 0.11020 0.18775 -0.10917 -0.13128 1.00000 0.93864 -0.01906 -0.03991 0.35930

-0.15251 0.30636 0.64652 0.42299 0.09229 0.15402 -0.05646 -0.06757 0.93864 1.00000 -0.07369 -0.10080 0.39597

0.10952 0.05496 -0.04498 0.03250 -0.06393 -0.17763 -0.03405 -0.05118-0.01906 -0.07369 1.0000 0.98016 0.02856

0.08566 0.07783 -0.02708 0.05834 -0.06945 -0.18303 -0.07142 -0.08395-0.03991 -0.10080 0.98016 1.0000 0.05207

0.12543 0.95738 0.93456 0.93087 0.00681 -0.00335 -0.02554 -0.05909 0.35930 0.39597 0.02856 0.05207 1.00000

Correlation matrix for the 31 cancer patients (simulations 1 and 2)

1.00000 -0.18564 -0.24347 -0.20191 0.05040 0.07228 -0.33392 -0.32821 -0.02802 -0.05652 0.22435 0.21851 -0.19517
-0.18564 1.00000 0.92176 0.93324 -0.31895 -0.25135 -0.06838 -0.06058 0.15862 0.07008 -0.25069 -0.24211 0.98106
-0.24347 0.92176 1.00000 0.98693 -0.34918 -0.24533 -0.00746 -0.00088 0.43249 0.34955 -0.23736 -0.21679 0.91129
-0.20191 0.93324 0.98693 1.00000 -0.36271 -0.27186 -0.05248 -0.03945 0.41738 0.33914 -0.21032 -0.18540 0.89660
0.05040 -0.31895 -0.34918 -0.36271 1.00000 0.87558 0.01070 -0.06825 -0.19724 -0.14265 0.18424 0.12636 -0.30905
0.07228 -0.25135 -0.24533 -0.27186 0.87558 1.00000 0.00409 -0.06436 -0.14766 -0.08062 0.37978 0.29615 -0.22481
-0.33392 -0.06838 -0.00746 -0.05248 0.01070 0.00409 1.00000 0.98334 0.10923 0.06158 -0.34308 -0.34457 -0.03978
-0.32821 -0.06058 -0.00088 -0.03945 -0.06825 -0.06436 0.98334 1.0000 0.11697 0.07382 -0.34884 -0.34540 -0.04016
-0.02802 0.15862 0.43249 0.41738 -0.19724 -0.14766 0.10923 0.11697 1.00000 0.95897 0.05304 0.10652 0.15591
-0.05652 0.07008 0.34955 0.33914 -0.14265 -0.08062 0.06158 0.07382 0.95897 1.00000 0.14306 0.18435 0.05343
0.22435 -0.25069 -0.23736 -0.21032 0.18424 0.37978 -0.34308 -0.34884 0.05304 0.14306 1.00000 0.99019 -0.25467
0.21851 -0.24211 -0.21679 -0.18540 0.12636 0.29615 -0.34457 -0.34540 0.10652 0.18435 0.99019 1.00000 -0.25373
-0.19517 0.98106 0.91129 0.89660 -0.30905 -0.22481 -0.03978 -0.04016 0.15591 0.05343 -0.25467 -0.25373 1.00000

Correlation matrix for the 500 cancer "generated patients" (simulations 1 and 2)

1.0000 -0.19668 -0.26483 -0.21796 0.07845 0.11599 -0.33890 -0.33424 -0.10965 -0.12217 0.29204 0.28265 -0.21779
-0.19668 1.00000 0.93046 0.94254 -0.40423 -0.31371 -0.10906 -0.10305 0.26374 0.16926 -0.24766 -0.23166 0.98290
-0.26483 0.93046 1.00000 0.98740 -0.41048 -0.28700 -0.03693 -0.03454 0.50496 0.41969 -0.22914 -0.20464 0.92473
-0.21796 0.94254 0.98740 1.00000 -0.43278 -0.32330 -0.08512 -0.07765 0.48765 0.40540 -0.20511 -0.17596 0.91172
0.07845 -0.40423 -0.41048 -0.43278 1.0000 0.87730 -0.03595 -0.11485 -0.26333 -0.22070 0.19682 0.14027 -0.38074
0.11599 -0.31371 -0.28700 -0.32330 0.87730 1.0000 -0.03939 -0.10467 -0.21171 -0.16157 0.38127 0.30176 -0.27331
-0.33890 -0.10906 -0.03693 -0.08512 -0.03595 -0.03939 1.0000 0.98437 0.09883 0.04721 -0.39521 -0.39591 -0.07166
-0.33424 -0.10305 -0.03454 -0.07765 -0.11485 -0.10467 0.98437 1.000 0.10713 0.06241 -0.39618 -0.39194 -0.07316
-0.10965 0.26374 0.50496 0.48765 -0.26333 -0.21171 0.09883 0.10713 1.00000 0.96251 0.04907 0.10797 0.26194
-0.12217 0.16926 0.41969 0.40540 -0.22070 -0.16157 0.04721 0.06241 0.96251 1.00000 0.13840 0.18897 0.15675
0.29204 -0.24766 -0.22914 -0.20511 0.19682 0.38127 -0.39521 -0.39618 0.04907 0.13840 1.00000 0.99133 -0.25064
0.28265 -0.23166 -0.20464 -0.17596 0.14027 0.30176 -0.39591 -0.39194 0.10797 0.18897 0.99133 1.00000 -0.24222
-0.21779 0.98290 0.92473 0.91172 -0.38074 -0.27331 -0.07166 -0.07316 0.26194 0.15675 -0.25064 -0.24222 1.00000

Correlation matrix for the 35 cancer patients (simulation 3)

1.0 0.08123 -0.02180 -0.02568 -0.01582 0.03701 0.08239 0.00781 0.02082 -0.18899 0.14619 0.08717 -0. 17175 0.18743
0.08123 1.0 -0.10376 -0.16597 -0.02286 0.36483 0.39938 0.12134 -0.04963 0.22627 -0.31123 0.13518 -0.27133 0.13748
-0.0218 -0.10376 1.0 0.9512 0.94189 -0.04772 -0.06372 0.11328 -0.26902 -0.13071 0.02625 -0.03458 0.30554 -0.28866
-0.02568 -0.16597 0.9512 1.0 0.79292 -0. 08499 -0.08489 0.31843 -0.29335 -0.08363 0.06710 0.0065 0.4049 -0.25339
-0.0158 -0.02286 0.94189 0.7929 1.0 -0.00937 -0.03745 -0.12085 -0.20526 -0.1566 -0.02996 -0.07945 0.16436 -0.29079
0.03701 0.36483 -0.04772 -0.08499 -0.00937 1.0 0.87381 0.22232 0.08214 0.02518 0.07059 0.10258 -0.25385 0.36858
0.08239 0.39938 -0.06372 -0.08489 -0.03745 0.8738 1.0 0.36728 -0.08438 0.03243 0.02473 0.28334 -0.16206 0.29897
0.00781 0.12134 0.11328 0.31843 -0.12085 0.22232 0.36728 1.0 -0.19879 0.06942 0.09141 0.27497 0.20004 0.08149
0.0208 -0.04963 -0.26902 0.29335 -0.20526 0.08214 -0.08438 -0.1987 1.0 0.2683 0.01769 -0.32896 -0.08444 0.23601
-0.18899 0.22627 -0.1307 0.08363 -0.15663 0.02518 0.03243 0.06942 0.2683 1.0 -0.29376 0.03967 0.33104 -0.13033
0.14619 -0.31123 0.02625 0.0671 -0.02996 0.07059 0.02473 0.0914 0.01769 -0.29376 1.0 -0.05669 -0.02186 0.4996
0.08717 0.13518 -0.03458 0.0065 -0.07945 0.10258 0.28334 0.27497 -0.32896 0.03967 -0.05669 1.0 0.27825 -0.1373
-0.17175 -0.27133 0.30554 0.4049 0.16436 -0.25385 -0.16206 0.200 -0.08444 0.33104 -0.02186 0.278254 1.0 -0.4159
0.18743 0.13748 -0.28866 0.25339 -0.29079 0.36858 0.29897 0.08149 0.2360 -0.13033 0.4996 -0.13732 -0.41590 1.0

Correlation matrix for the 500 cancer "generated patients" (simulation 3)

1.0 0.12750 -0.07963 -0.11188 -0.03711 0.06749 0.10203 -0.08604 0.06569 -0.16119 0.14367 0.10792 -0.16396 0.18270
0.12750 1.0 -0.11286 -0.16722 -0.03859 0.33030 0.36602 0.13437 -0.04670 0.19051 -0.26894 0.17920 -0.27283 0.18051
-0.07963 -0.11286 1.0 0.95619 0.94418 -0.11667 -0.14577 0.12462 -0.2835 -0.22394 0.04931 -0.07033 0.33807 -0.29242
-0.11188 -0.1672 0.95619 1.0 0.80705 -0.14626 -0.16034 0.31814 -0.29983 -0.16802 0.07687 -0.02233 0.42622 -0.25929
-0.03711 -0.03859 0.94418 0.80705 1.0 -0.07567 -0.11596 -0.10410 -0.22501 -0.25098 0.00368 -0.1236 0.2052 -0.29248
0.06749 0.33030 -0.11667 -0.14626 -0.07567 1.0 0.88313 0.24738 0.08402 0.06561 0.11671 0.1147 -0.22558 0.39071
0.10203 0.36602 -0.14577 -0.16034 -0.11596 0.88313 1.0 0.36887 -0.07807 0.08359 0.05059 0.26465 -0.13695 0.31954
-0.08604 0.13437 0.12462 0.31814 -0.10410 0.2473 0.36887 1.0 -0.17727 0.04344 0.06163 0.24089 0.17443 0.07248
0.06569 -0.0467 -0.2835 -0.29983 -0.22501 0.0840 -0.07807 -0.17727 1.0 0.30741 0.01611 -0.38540 -0.09472 0.24365
-0.16119 0.19051 -0.22394 -0.16802 -0.25098 0.0656 0.08359 0.04344 0.30741 1.0 -0.16620 0.01925 0.32817 -0.00142
0.14367 -0.26894 0.04931 0.07687 0.00368 0.1167 0.05059 0.06163 0.01611 -0.16620 1.0 -0.07944 -0.01330 0.48242
0.10792 0.17920 -0.07033 -0.02233 -0.12366 0.1147 0.26465 0.24089 -0.3854 0.01925 -0.07944 1.0 0.2181 -0.15506
-0.16396 -0.27283 0.33807 0.42622 0.20520 -0.2255 -0.13695 0.17443 -0.0947 0.32817 -0.01330 0.2181 1.0 -0.38482
0.1827 0.18051 -0.2924 -0.25929 -0.29248 0.3907 0.31954 0.07248 0.24365 -0.00142 0.48242 -0.15506 -0.38482 1.0

Correlation matrix for the 83 gallstone patients (simulation 3)

1.0 0.28388 0.08099 0.03753 0.14085 -0.31039 -0.27566 -0.07462 -0.10055 -0.13691 -0.21800 0.18535 0.05898 0.15410
0.28388 1.0 0.02203 -0.05894 0.14695 -0.26014 -0.14758 -0.08820 0.12389 -0.04386 0.00700 0.10493 -0.08359 0.07923
0.08099 0.02203 1.0 0.97491 0.93672 0.18099 0.17189 0.33354 0.00503 -0.00783 0.10468 -0.08874 0.06787 0.04339
0.03753 -0.05894 0.97491 1.0 0.83530 0.22031 0.20733 0.43672 0.02382 0.00208 0.02601 -0.09659 0.07286 0.02530
0.14085 0.14695 0.93672 0.83530 1.0 0.10120 0.09865 0.13599 -0.02561 -0.02278 0.21801 -0.06675 0.05295 0.06764
-0.31039 -0.26014 0.18099 0.22031 0.10120 1.0 0.80507 0.17647 -0.07643 0.00099 0.19464 -0.22740 -0.08631 0.20130
-0.27566 -0.14758 0.17189 0.20733 0.09865 0.80507 1.0 0.12641 -0.04735 0.01852 0.14049 -0.13878 -0.03011 0.17544
-0.07462 -0.08820 0.33354 0.43672 0.13599 0.17647 0.12641 1.0 0.13192 0.10293 -0.04564 -0.17897 0.07787 -0.01818
-0.10055 0.12389 0.00503 0.02382 -0.02561 -0.07643 -0.04735 0.13192 1.0 0.0645 -0.06608 -0.04701 0.07528 -0.07769
-0.13691 -0.04386 -0.00783 0.00208 -0.02278 0.0009 0.01852 0.102939 0.06450 1.0 0.05896 -0.12551 0.18460 -0.01450
-0.21800 0.00700 0.10468 0.02601 0.21801 0.19464 0.14049 -0.04564 -0.06608 0.05896 1.0 -0.21388 -0.12842 0.27432
0.18535 0.10493 -0.08874 -0.09659 -0.06675 -0.2274 -0.13878 -0.17897 -0.047 -0.1255 -0.21388 1.0 0.09994 -0.23996
0.05898 -0.08359 0.06787 0.07286 0.05295 -0.08631 -0.03011 0.07787 0.07528 0.18460 -0.12842 0.09994 1.0 -0.20417
0.15410 0.07923 0.04339 0.0253 0.06764 0.2013 0.17544 -0.01818 -0.07769 -0.01450 0.27432 -0.23996 -0.20417 1.0

Correlation matrix for the 500 gallstone " generated patients" (simulation 3)

1.0 0.30036 0.15648 0.10096 0.22663 -0.3268 -0.30513 -0.0972 -0.12136 -0.08179 -0.22129 0.22698 0.05557 0.17708
0.30036 1.0 0.09192 0.01092 0.20814 -0.2520 -0.14360 -0.03074 0.15239 -0.03455 -0.02654 0.12147 -0.0692 0.06156
0.15648 0.09192 1.0 0.97453 0.93760 0.12796 0.12849 0.36595 0.00629 0.04986 0.05831 -0.09625 0.10711 0.07217
0.10096 0.01092 0.97453 1.0 0.83574 0.17604 0.17043 0.47367 0.01794 0.05308 -0.02195 -0.09698 0.10799 0.04287
0.22663 0.20814 0.93760 0.83574 1.0 0. 04077 0.05021 0.16046 -0.01283 0.03972 0.17718 -0.08476 0.09464 0.11056
-0.3268 -0.25201 0.12796 0.17604 0.040771.0 0.79759 0.15696 -0.11808 -0.05224 0.13585 -0.21605 -0.10194 0.18621
0.30513 -0.14360 0.12849 0.17043 0.05021 0.79759 1.0 0.14920 -0.08979 0.01327 0.09271 -0.07891 -0.01181 0.17090
-0.09721 -0.03074 0.36595 0.47367 0.16046 0.15696 0.14920 1.0 0.07821 0.09739 -0.05805 -0.15322 0.05091-0.05857
-0.12136 0.15239 0.00629 0.01794 -0.01283 -0.1180 -0.08979 0.078218 1.0 0.02008 -0.09224 -0.0509 0.07870 -0.15777
-0.08179 -0.03455 0.04986 0.05308 0.0397 -0.0522 0.01327 0.097394 0.02008 1.0 0.04996 -0.09157 0.16964 -0.04139
-0.22129 -0.02654 0.0583 -0.02195 0.17718 0.1358 0.09271 -0.058055 -0.09224 0.04996 1.0 -0.2464 -0.12090 0.23320
0.227 0.12147 -0.09625 -0.09698 -0.08476 -0.21605 -0.0789 -0.15322 -0.05094 -0.09157 -0.24648 1.0 0.06704 -0.17704
0.05557 -0.06923 0.10711 0.10799 0.09464 -0.10194 -0.01181 0.05091 0.07870 0.16964 -0.12090 0.06704 1.0 -0.21590
0.17708 0.06156 0.07217 0.04287 0.11056 0.18621 0.17090 -0.05857 -0.15777 -0.04139 0.2332 -0.17704 -0.2159 1.0