



UNIVERSITY OF NAIROBI

SEPSIS PREDICTION FROM COMPLETE BLOOD COUNT COMPONENTS.

PRINCIPAL INVESTIGATOR;

JAMES MWANGI WAWERU.

MSC. MEDICAL STATISTICS.

W62/12038/2018.

INSTITUTE OF TROPICAL AND INFECTIOUS DISEASES,

UNIVERSITY OF NAIROBI, 2020.

Research project submitted in part fulfillment of the requirements for the award of the degree of master of science in medical statistics of the University of Nairobi.

NOVEMBER 2020



DECLARATION

I declare that this research project is my own original work and to the best of my knowledge, it has not been presented anywhere else for consideration of publication or for the award of another degree.

Signature



Date26/11/2020.....

James Mwangi Waweru

Msc. Medical Statistic Year II.

W62/12038/2018.

UNITID, UoN.

SUPERVISORS.

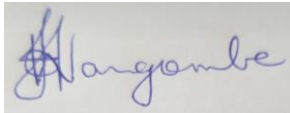
This research project has been submitted for examination with my approval as a university supervisor.

Dr. Anne Wango'mbe

Lecturer – school of mathematics, University of Nairobi.

P.O BOX 33507-00600, Nairobi.

Signed



Date.....1/12/2020.....

Dr. Anthony Karanja

Lecturer – school of computing sciences, Riara University.

P.O BOX 49940 – 00100, Nairobi.

Signed



Date1/12/2020.....

TABLE OF CONTENTS

TITLE PAGE.	i
DECLARATION	ii
TABLE OF CONTENTS	iii
ABBREVIATIONS	v
LIST OF FIGURES.	vi
LIST OF TABLES.	vii
ABSTRACT	viii
INTRODUCTION	2
2.1 Pathogenesis of sepsis	3
2.2 Immune response in sepsis	3
2.3 Blood cell changes in sepsis	3
2.4 Laboratory analysis of blood cells in sepsis.	4
2.5 Utility of complete blood count components as diagnostic tool	4
STUDY JUSTIFICATION	5
RESEARCH QUESTION.	6
STUDY OBJECTIVES.	6
METHODOLOGY	7
3.1 Study design	7
3.13 Statistical analysis.	8
RESULTS	11
Data cleaning and imputation (R software)	11
Test for multivariate assumptions (stata software).	16
Factor analysis.....	16

Linear discriminant analysis	18
Multivariate linear predictive model.....	19
DISCUSSION	21
CONCLUSION.....	21
APPENDIX 1: STUDY BUDGET.....	25
APPENDIX 2: TIME FRAME.	26
APPENDIX 3: DATA COLLECTION EXCEL FORM.....	27
APPENDIX 4: MIMIC III DATA ACCESS AND CONSENT WAIVER INFORMATION.....	28
APPENDIX 5: LETTER OF AUTHORITY TO ACCESS MIMIC III DATABASE.....	29
APPENDIX 6: ETHICS AND RESEARCH COMMITTEE LETTER.	Error!

Bookmark not defined.

ABBREVIATIONS.

BASO – Basophil

CBC – Complete Blood Count

ED – Emergency Department

EOS – Eosinophil

ERC – Ethical Research Committee

GRANU – Granulocytes

HB – Haemoglobin

HCT – Haematocrit

ICU – Intensive Care Unit

KNH – Kenyatta National Hospital.

LNR – Lymphocyte Neutrophil Ratio

MCH – Mean Cell Haemoglobin

MCHC – Mean Cell Haemoglobin Concentration

MCV – Mean Cell Volume

NEUT – Neutrophil

PLT – Platelets

RBC – Red Blood Cell

RDW – Red cell Distribution Width

SSC – Surviving Sepsis Campaign

TEMP - Temperature

UNITID - University of Nairobi Institute of Tropical and Infectious Diseases.

UoN – University of Nairobi.

WBC – White Blood Cell



LIST OF FIGURES.

Scatter plot matrix of variables.....	15
Box plot of variables.....	15
Scree plot of eigen values after factor analysis.....	18
Discriminant score plot on white cells.....	19
Variable loading plot on discriminant scores.....	19

LIST OF TABLES.

Table of generated factors after factor analysis.....	17
Table of variable loadings on generated factors.....	17
Table of obliquely rotated factors.....	17
Table of discriminant functions.....	18
Table of discriminant proportions of white cells.....	18
Table of multivariate predictive model coefficients.....	20

ABSTRACT.

Background. Sepsis is a clinical syndrome that defines human body reaction to infection by microbial pathogens. Sepsis is a major healthcare problem around the world with significant mortality. Timely and correct management improves outcome. Initial empirical therapy is started before definitive microbiological pathogen identification. Complete blood count is a major baseline laboratory investigation for a suspected case. Multiple parameters are analyzed and reported in the complete blood count analysis that estimate human body immune response to microbial infection. However, but for white cell count elevation, correlated information from majority of other measurements remain tangled up with minimal input as regards to diagnosis, stratification and management of sepsis. Statistical analysis of complete blood count in sepsis may reveal further clinical relevance of other components that is may be additive to diagnosis and management leading to improved outcome in sepsis.

Broad objectives; To measure the degree of relationship among complete blood count changes that occur in sepsis.

Study design and site; This study was a retrospective study using laboratory measurements recorded from patients treated for sepsis in Beth Israel hospital.

Materials and methods; This study used secondary data from medical information mart for intensive care (MIMIC III) database. This database contains de-identified clinical information of ICU patients admitted over ten years in Beth Israel hospital. Data cleaning and imputation was done using R software. Multivariate analysis was carried out using stata analytical software. Multiple linear Regression modelling; the main model had white cell count as response variable and differential counts and other complete blood count values as multivariate variables.

Data management; After approval, data was acquired from MIMIC III database management and stored in data files in the computer that was used for analysis. Data files were password protected and no manual coping was done.

Expected main outcome measure/ utility of the study; Main outcome measures were correlation of laboratory measurements from the initial complete blood count investigations. These correlated variables were used in prediction model for sepsis in addition to elevated white cell count.

Results; Mean cell volume, mean cell hemoglobin and red cell distribution width were the highest correlated variables with the principal factor explaining the variance in sepsis. They also loaded highly in discriminating elevated white cell count from low and normal counts. Multivariate predictive model using these variables had a cronbach's alpha of 0.4.

Conclusion; Mean cell volume, mean cell hemoglobin and red cell distribution width may be used in predictive diagnosis of sepsis in addition to white cell changes.

INTRODUCTION.

Sepsis is life threatening organ dysfunction caused by dysregulated host response to microbial infection(1) This current definition emphasizes on the presence of organ dysfunction as a pre-requisite for diagnosis of sepsis. Sepsis is a multiple organ clinical symptomatology that reflect physiologic, biologic, and biochemical abnormalities in the human body that results from microbial pathogen infection. Sepsis begins when infective microbial pathogen contaminate sterile areas of the body, proliferate and invade the body tissues resulting into suppuration in soft tissues, body cavities or other point of infection in any body part(2). The microorganisms then invade the bloodstream or may proliferate at a local body site and release large amounts of bacterial toxins into the bloodstream. Toxic products from the microbial pathogens, immunogenic structural components of the pathogens and reactive molecules from produced by the host act in concerted effort to trigger the dysregulated organ dysfunction.

Sepsis presents a colossal burden of disease worldwide with huge consumption of considerable amount of health- care resources. It is a common emergency disease condition, that is expensive to treat, associated with long term morbidity and frequently a common cause of death with a cause mortality approaching that due to acute myocardial infection(3). Recent studies report an increasing incidence in sepsis globally however, with reducing cause mortality(4). Sepsis is classified as a medical emergency and it is recommended that treatment and resuscitation begin immediately(5). At the emergency department, there is inadequate resource and time to definitively confirm sepsis and more so microbiology results are not available at initial diagnosis. Initial empirical therapy is usually started before definitive microbiological pathogen identification. Prompt recognition is particularly important to have a better outcome of sepsis.

Complete blood count is a major laboratory investigation widely available to clinicians at emergency department. In addition to clinical symptomatology, complete blood count laboratory measurement is the most resourceful investigation that helps to make a diagnosis of sepsis. Complete blood count avails multiple components of blood cell parameters as a measure of response of the human body to microbial infestation. Nearly all cellular elements in the blood are involved in the pathophysiology of sepsis. White cell alterations are common in majority of patients with sepsis with counts above $12 \times 10^9/L$ or less than $4.0 \times 10^9/L$ a major criterion in diagnosis of sepsis (5).

Complete blood count has been studied and utilized as a diagnostic, predictive and prognostic tool in heart and other diseases(6). Regression models has also been formulated to predict bacterial infection(7) Complete blood count is the single most laboratory investigation readily available at emergency department used for diagnosis of sepsis. However, but for white cell count changes, diagnostic utility of changes in other CBC components in sepsis is missing. Clinicians are un aware of the diagnostic utility of the other components of complete blood count and either are confounded by them or mistakenly err ignoring any input they may have in the diagnosis, stratification and prognosticating sepsis. The pattern of human body immune response in sepsis is represented in multiple components reported in CBC blood test. It is well known that all the CBC components are random measurements representing immune system in a single individual and therefore their diagnostic imputation cannot meaningfully be carried out separately. In-depth analysis of the attendant variation of all components of complete blood count in sepsis may improve to the quality of information and precision in diagnosis and management of sepsis leading to improved outcome.

LITERATURE REVIEW.

2.1 Pathogenesis of sepsis

Microbial infections have been described to afflict the human body as long as human history has been recorded. Several preventive and curative innovations have contributed great strides towards eliminating the menace of microbial diseases. However, to date, microbial infections remain a common cause of morbidity and mortality to human population. Commonest microbial pathogens consist of bacteria, fungi and viruses. Microbial pathogens transfer from environment or from specific sites of normal body commensals into sterile sites of the body where they invade proliferate and cause infection. Inoculation may be through ingestion, inhalation, skin breach or body site breakdown of immune barriers. The immune system in the body is responsible for protection of self from invading microbial pathogens. It is organised into local and systemic immune responses that are triggered in a stepwise structure regarding the degree of microbial invasion. Local immune system is located in the tissues exposed to pathogenic organisms. Microbial pathogens when inoculated at these sites, are recognised as foreign and destroyed. The host immune elements recognises the microbial pathogen structural components that triggers the immune reaction(8). Cytokines that are immune molecules produced at these sites circulate in the body and stimulate the systemic immune responses. If and when the bacterial load or the virulence of the infective pathogen overwhelms the local immune system, the pathogens gain entry into the blood system. The exaggerated inflammatory response rises to a level that physiologic alterations within the body occur with damage to body organs. The current definition of sepsis describes organ dysfunction due to dysregulated host response to the infecting pathogen(1).

2.2 Immune response in sepsis

The immune system consists of physiological barriers and biological components that protect the body from invading microbial pathogen. The biological components are developed and housed in the hematological system. The hematological system, among other functions carries the immune system of the body in two components; cellular and humoral components. Cellular elements include red blood cells, white cells, and platelets. Fluid-phase components include coagulation factors, natural anticoagulants, and proteins of the fibrinolytic system and the complement system (8). The cellular component is largely the first line of immune defense and triggers the activation of the humoral component of the immune system that largely consists of immunoglobulins. Due to the integrated nature of the hematologic system, almost every component is affected during immune response in sepsis.

2.3 Blood cell changes in sepsis

The main changes consist of elevated white cells and decreased red cells and platelets. White cell alterations are common in majority of patients with sepsis with counts above $12 \times 10^9/L$ or less than $4.0 \times 10^9/L$ a major criterion in diagnosis of sepsis(9). Increases above normal ranges reflect overproduction from the bone marrow source to counter the effect of invading pathogen. Low levels reflect over consumption of white cells in clearing the pathogens. Due to the wide reference range of the white cells, an increase or a decrease may still fall within normal range and therefore obscure diagnosis of sepsis. This informs the necessity of knowledge of blood cell changes in other cell lines other than white cell than may indicate a diagnosis of sepsis. Subsets of white cells that include neutrophils, lymphocytes, monocytes, eosinophils

and basophils are measured and reported. Changes in absolute values above normal may be suggestive of microbial infections other than other causes of elevated white cell counts. Ratios of neutrophils and lymphocytes changes may further suggest microbial infection. Analysis of the most consistent pattern of change in sepsis will help in interpreting these results and increasing the precision of sepsis diagnosis.

Hematopoietic progenitor cell from which immune cells develop give rise to more non immune cells with other functions in the hematologic system. As a result of this common source and interrelated development and maturation of hematologic cells, changes in one particular cell line has identifiable effects on other blood cell components. In sepsis, shift of development of immune cells through cell signaling molecules affect red cells and platelets due to competition for limited growth factors and space for maturation. Commonly low platelets have been associated with sepsis.

Blood cell size and content that are directed by hematopoietic growth factors and their changes are identifiable in sepsis effect on hematologic system. Red cell changes in size and content include hemoglobin content, red cell distribution width (RDW) mean cell volume (MCV) mean cell hematocrit (MCH) mean cell hemoglobin concentration (MCHC). Analysis in any present changes may be informative in diagnosis of sepsis.

2.4 Laboratory analysis of blood cells in sepsis.

The complete blood count measure (CBC) describes the number and morphology of different cell types in the hematological system. This is carried out in the laboratory by preparing a non-coagulated blood sample in a suspension liquid that is passed through the hematologic analyzer machine. The machine pulls a calculated volume fluid through a pipe system. Electric current is passed from one side of the fluid and changes due to the fluid characteristics are recorded across the opposite side. The number of pulses correlates to the number of particles. The height of the electrical pulse is proportional to the cell volume. Further individual white cells are classified according to their morphology into differential white cell counts. The hematologic analyzer machine is enabled to estimate the cell number and the cell size by detecting electrical changes as the cells are passed through it. These measurements are reported as absolute cell counts, white cell differential cell counts, cell contents and cell indices.

2.5 Utility of complete blood count components as diagnostic tool

Complete blood count changes have been studied and modelled as a diagnostic, prognostic and predictive tool in a number of disease conditions. Philips et al carried out multivariate logistic regression to predict bacterial infection from a microbiologically confirmed cases with healthy individuals (7). Their analysis mainly studied on presence or absence of sepsis regarding changes in white cell count differentials. They reported that positive cultures are associated with high band counts and high neutrophil counts.

Sedimentation rate and C-reactive protein are other laboratory investigation that have been studied for the purpose of early diagnosis of sepsis with poor correlation of sepsis(10). Disagreement exists regarding the clinical significance of many of these observations. Only lymphocytosis a few components of the differential count have been quantitatively evaluated to determine their diagnostic significance in sepsis. However, these observations have not been examined to determine their correlation in conjunction with other accompanying CBC observations. Barry et al (11) showed that high mean white cell count is associated

with infectious and inflammatory conditions but failed to show correlation between raised white cell count and percentage of granulocytes in the differential counts. We aim to tease out information on which of the numerous cell types and numbers in the commonly ordered CBC, are consistent with sepsis. Neutrophil – Lymphocyte count ratio has been studied previously as a diagnostic marker of sepsis with a normal value(12). Kim et al(13) found significantly higher values of the NLCR in patients with sepsis compared to patients without sepsis but on its own was a poor predictor of sepsis in ICU admitted patients. Factor analysis assesses correlation of multiple variables in explaining majority of the data variation (13). Selection of key variables representing largest variation may be used to predict sepsis in addition to elevated white cell count. Multivariate regression model with maximum likelihood estimation predictive model will be used to formulate a predictive model for sepsis.

STUDY JUSTIFICATION

There is need for improvement in timely diagnosis and management of sepsis to improve outcome. Majority of components reported in complete blood count have unknown clinical utility pertaining diagnosis and management of sepsis. Statistical analysis of these measurements may enlighten on their diagnostic utility and therefore improve outcome of sepsis.

RESEARCH QUESTION.

Is there correlation among complete blood count changes in sepsis?

STUDY OBJECTIVES.

Main objective; To measure the degree of relationship among complete blood count changes that occur in sepsis.

Specific objectives;

1. To determine the covariance of complete blood count components in patients with sepsis.
2. To determine the correlation among complete blood count components in patients with sepsis.
3. To fit a model of complete blood count changes that predicts sepsis.

METHODOLOGY.

3.1 Study design

The study was a retrospective analytical study. This study used secondary data from medical information mart for intensive care (MIMIC III v1.4) database. This database contains de-identified clinical information of ICU patients admitted over ten years in Beth Israel hospital. This clinical data was prospectively captured during routine hospital care using digital health record system in this hospital. Laboratory test results inclusive of complete blood count components were captured from laboratory health systems. Additional data was sourced from social security administration death master file.

3.2 Study area; Critical care units of a tertiary hospital, Beth Israel hospital.

3.3 Study population; patients treated for sepsis.

3.4 Inclusion criteria; patients admitted to ICU and with discharge or death diagnosis of sepsis.

3.5 Exclusion criteria; Patients with concomitant trauma, surgery and myocardial infarction or uncontrolled non communicable diseases e.g. diabetes and hypertension at the time of diagnosis of sepsis. This conditions induce systemic inflammatory reaction in the body that are confounding to expected complete blood count changes in sepsis

3.6 Study limitations; Retrospective data may have missing values that require imputation. The data source is of a different population and set up and therefore limited in generalization to our local population.

3.7 Sample size calculation;

Null hypothesis; There is no correlation among complete blood count changes in sepsis.

We will measure the degree of association of CBC changes with correlation coefficient parameter.

Ho; $r_1 = r_2 = \dots r_k = 0$

The study was designed to test for a correlation (r) equal to or greater than +0.4 or -0.4

Level of statistical significance was set at 0.05% and the power of the study set at 80% (14).

The correlation coefficient squared, represents the proportion of the variance that results from it linear association. Sample size for determining whether a correlation differs from zero as described by Hulley S.B et al was used(15)

r = expected correlation coefficient

$C = 0.5 \times \ln [(1 + r)/(1 - r)]$ fisher's correction for normalization of correlation coefficient

N = Total number of subjects required

Then

$N = [(Z\alpha + Z\beta) \div C]^2 + 3.$

$C = 0.5 \times \ln [(1 + 0.4)/(1 - 0.4)]$
 $= 0.5 \times \ln [1.4/0.6]$

$= 0.5 \times \ln [2.333]$

=0.424

$$N = [(1.96 + 0.84) \div 0.424]^2 + 3$$

$$N = [(2.8) \div 0.424]^2 + 3$$

$$N = [6.604]^2 + 3$$

$$N = 43.6 + 3$$

$$N = 46.6$$

$$N=47$$

3.8 Sampling procedure; consecutive sampling of data that met criteria was used.

3.9 Data management; Data collection was done through electronic retrieval from two different critical care information systems that were in use in the hospital (Phillips carevue clinical information systems and metavision ICU software). Data was stored as comma separated version of excel spread sheets.

3.10 Ethical approval; Institutional ethical review board authorization was obtained from UoN/KNH ERC before commencement of this study. This study used secondary data from medical information mart for intensive care (MIMIC III v1.4) database.

This database contains de- identified clinical information of ICU patients admitted over ten years in Beth Israel Deaconess Medical Center. The primary data collection was approved by the institutional review boards of Beth Israel Deaconess Medical Center and Massachusetts Institute of Technology. Subjects consenting has been waived for the secondary data use. Information on availability of data and waiver for subject consenting is provided in appendix 4 attached in this report.

3.11 Confidentiality maintenance; Complete de-identification of data was carried out in accordance with Health Insurance Portability and Accountability Act (HIPAA). Exclusive use of the data for the purposes of authorized research work in confidentiality was adhered. Password protection of the all documents was maintained at all level of data storage and transfer. No manual copying was done at any time. Authorization to use this data was granted through physioNet database managers (appendix 5). Exclusive use of the data for the purposes of authorized research work in confidentiality was adhered.

3.12 Main outcome measures; laboratory measurements from the initial complete blood count investigation carried out on admission. These were all metric measures of ratio scale.

3.13 Statistical analysis.

3.13.1 Computation of data; MIMIC-III is a relational database consisting of 26 tables linked by unique identifiers relating to in hospital admission. These were downloaded and imported into R software. No written copies of data were used.

3.13.2 Data imputation; Verification of data was carried out at all levels of data manipulation. Missing data and erroneous entries were counter checked by tracing back the unique identifier through all records for completeness. The extent of missing data will be assessed by tabulating the percentage of variables with missing data for each case and the number of cases with missing data for each variable. A limit of 10% overall missing data was set. An all available case imputation method was employed for scenarios below this limit, otherwise case substitution method was used(14).

3.13.2 Data coding; white cell counts were coded into three levels using reference ranges into high = 2, normal = 1 or low = 0.

3.13.3 Outlier Detection and Management; Box plot was used to assess for outliers in complete blood count measurements. A threshold of 4.0 standard deviations was used for outlier designation. Bivariate outlier assessment was carried out between white cell count and each of other CBC measurement using a scatter plot with 95% confidence ellipsoid. Mahalanobis D^2 measure was used to assess for multivariate outliers in complete blood count measurements (14). Identified significant outliers were treated by elimination from analysis.

3.13.4 Assessment for distribution of data; A normal probability plot was used to assess normality of individual variable distribution. Continuous variables were tested for normality with shapiro wilk's test.

3.13.5 Assessment for general association and linearity; a scatter plot matrix was used to assess for general relationship between pairs of complete blood count measurements and between white cell count and continuous baseline characteristics. Pearson correlation coefficient was calculated to test for the strength of linear relationships.

3.13.6 Descriptive statistics; mean and proportions were calculated as summary measures for continuous and categorical data respectively.

3.13.7 Multivariate analysis

Factor analysis was carried out on metric measurements of complete blood count. The objective was to analyze the underlying structure of interdependence among complete blood count measurements. Correlation matrix was calculated and presented in tabular format. Bartlett's test of sphericity for overall significance of correlation matrix was computed. Measure of sampling adequacy (MSA) was used to assess pattern of correlation between variables. **Factor extraction method;** Variables with significant correlation in the set of variables were selected for further analysis. Total variance explained by the components was calculated. Latent root criterion and scree plot analysis were used to select significant factors with Eigen values greater than one. Oblique factor rotation was used to summarize the factor matrix. Factor loadings equal to or greater than 0.75 were considered significant.

Multivariate discriminant analysis was carried out to discriminate changes in complete blood count among different levels of white cell count. Using the normal white cell count reference range ($4.0 - 11 \times 10^9/L$), white cell measurement was categorized into three levels, (high = 2, normal = 1 and low = 0). This analysis predicted the likelihood that a white cell count belonged to a particular category based on the pattern of changes in other complete blood count variables.

Multiple Regression predictive analysis was carried out with metric measurements of white cell count as response variable and metric measurements differential counts and other complete blood count values as predictor variables. This analysis predicted white cell count changes in response to changes in other complete blood count variables selected by factor and discriminant analysis. A model with weighted variate values was constructed and modified by addition of cross product and ratios of values. Likelihood ratio test for significance of model was carried out.

A summary of main variates of change were selected to form a summated scale of principle blood changes in sepsis. These were reported as the main changes in initial blood count measurement of patients suspected of sepsis. The information may help emergency doctors in early diagnosis of sepsis with more precision, which translates into prompt appropriate management with improved outcome.

All measurements were reported with 95% confidence interval and p value of 0.05 significance was used.

RESULTS

SEPSIS DATA ANALYSIS

JAMES

10/19/2020

{The data is from medical information mart for intensive care (MIMIC III v1.4) database. This database contains de-identified clinical information of ICU patients admitted over ten years in Beth Israel hospital. This clinical data was prospectively captured during routine hospital care using digital health record system in this hospital. Laboratory test results inclusive of complete blood count components were captured from laboratory health systems.

Data cleaning and imputation (R software)

1. loading admission data file with clinical diagnosis per admitted patient

```
admissions <- read.csv("C:/Users/HP/Desktop/Msc Project/MIMIC 3  
DATA/ADMISSIONS.csv")
```

2. loading diagnosis codes labels for clinical diagnosis.

```
codes <- read.csv("C:/Users/HP/Desktop/Msc Project/MIMIC 3  
DATA/DIAGNOSES_ICD.csv")
```

3. calling to console data imputation packages.

```
library(tidyverse)
```

```
library(dplyr)
```

4. filtering codes for patients with sepsis.

```
sepsiscodes <- filter(codes, ICD9_CODE == 99591)  
sepsiscodes2 <- filter(codes, ICD9_CODE == 99591 | ICD9_CODE == 99592)  
sepsiscodes3 <- filter(codes, ICD9_CODE == 99591 | ICD9_CODE == 99592)%>%  
arrange()  
sepsiscodes4 <- arrange(sepsiscodes2 , SUBJECT_ID)
```

5. loading laboratory data for all patients.

```
labdata <- read.csv("C:/Users/HP/Desktop/Msc Project/MIMIC 3  
DATA/LABEVENTS.csv")
```

6. pulling laboratory data specific to sepsis patients using above codes for sepsis and SUBJECT_ID.

```
xx <- pull(sepsiscodes4, SUBJECT_ID)  
````{  
sepsislab.xx <- labdata [xx,]
```

7. selecting initial measure of complete blood count component of each sepsis patient and binding all measures to one data file.

Selecting redcell component

```
redcells<- sepsislab.xx %>% filter(ITEMID ==51279) %>% group_by (SUBJECT_ID) %>% arrange
(CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID,
VALUE) %>% rename(REDCELL = VALUE)
```

##ADD PLATELETS

```
mydata1 <- sepsislab.xx %>% filter(ITEMID ==51265) %>% group_by (SUBJECT_ID) %>% arrange
(CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID,
VALUE) %>% rename(PLATELETS = VALUE) %>% merge(redcells,mydata1.by.x = "SUBJECT_ID" ,by.y
= "SUBJECT_ID", all=TRUE)
```

##ADD WHITE CELL

```
mydata2 <- sepsislab.xx %>% filter(ITEMID ==51301) %>% group_by (SUBJECT_ID) %>% arrange
(CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID,
VALUE) %>% rename(WBCCOUNT = VALUE) %>% merge(mydata1,mydata2.by.x = "SUBJECT_ID" ,by.y
= "SUBJECT_ID", all=TRUE)
```

##ADD RDW

```
mydata3 <- sepsislab.xx %>% filter(ITEMID ==51277) %>% group_by (SUBJECT_ID) %>% arrange
(CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID,
VALUE) %>% rename(RDW = VALUE) %>% merge(mydata2,mydata3.by.x = "SUBJECT_ID" ,by.y =
"SUBJECT_ID", all=TRUE)
```

##ADD NEUTROPHIL 51256

```
mydata4 <- sepsislab.xx %>% filter(ITEMID ==51256) %>% group_by (SUBJECT_ID) %>% arrange
(CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID,
VALUE) %>% rename(NEUTROPHIL = VALUE) %>% merge(mydata3,mydata4.by.x = "SUBJECT_ID"
,by.y = "SUBJECT_ID", all=TRUE)
```

##ADD MONOCYTES 51254

```
mydata5 <- sepsislab.xx %>% filter(ITEMID ==51254) %>% group_by (SUBJECT_ID) %>% arrange
(CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID,
VALUE) %>% rename(MONOCYTES = VALUE) %>% merge(mydata4,mydata5.by.x = "SUBJECT_ID"
,by.y = "SUBJECT_ID", all=TRUE)}
```

#ADD MCV 51250

```
mydata6 <- sepsislab.xx %>% filter(ITEMID ==51250) %>% group_by (SUBJECT_ID) %>% arrange
(CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID,
```

```
VALUE) %>% rename(MCV = VALUE) %>% merge(mydata5,mydata6.by.x = "SUBJECT_ID" ,by.y = "SUBJECT_ID", all=TRUE)
```

ADD MCH 51248

```
mydata7 <- sepsislab.xx %>% filter(ITEMID ==51248) %>% group_by (SUBJECT_ID) %>% arrange (CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID, VALUE) %>% rename(MCH = VALUE) %>% merge(mydata6,mydata7.by.x = "SUBJECT_ID" ,by.y = "SUBJECT_ID", all=TRUE)
```

##ADD MCHC 51249

```
mydata8 <- sepsislab.xx %>% filter(ITEMID ==51249) %>% group_by (SUBJECT_ID) %>% arrange (CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID, VALUE) %>% rename(MCHC = VALUE) %>% merge(mydata7,mydata8.by.x = "SUBJECT_ID" ,by.y = "SUBJECT_ID", all=TRUE)
```

ADD LYMPHOCYTES 51244

```
mydata9<- sepsislab.xx %>% filter(ITEMID ==51244) %>% group_by (SUBJECT_ID) %>% arrange (CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID, VALUE) %>% rename(LYMPHOCYTES = VALUE) %>% merge(mydata8,mydata9.by.x = "SUBJECT_ID" ,by.y = "SUBJECT_ID", all=TRUE)
```

ADD HEMOGLOBIN 51222

```
mydata10<- sepsislab.xx %>% filter(ITEMID ==51222) %>% group_by (SUBJECT_ID) %>% arrange (CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID, VALUE) %>% rename(HEMOGLOBIN = VALUE) %>% merge(mydata9,mydata10.by.x = "SUBJECT_ID" ,by.y = "SUBJECT_ID", all=TRUE)
```

ADD HEMATOCRIT 51221

```
mydata11<- sepsislab.xx %>% filter(ITEMID ==51221) %>% group_by (SUBJECT_ID) %>% arrange (CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID, VALUE) %>% rename(HEMATOCRIT = VALUE) %>% merge(mydata10,mydata11.by.x = "SUBJECT_ID" ,by.y = "SUBJECT_ID", all=TRUE)
```

ADD EOSINOPHIL 51200

```
mydata12<- sepsislab.xx %>% filter(ITEMID ==51200) %>% group_by (SUBJECT_ID) %>% arrange (CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID, VALUE) %>% rename(EOSINOPHIL = VALUE) %>% merge(mydata11,mydata12.by.x = "SUBJECT_ID" ,by.y = "SUBJECT_ID", all=TRUE)
```

ADD BASOPHIL 51200

```
mydata13<- sepsislab.xx %>% filter(ITEMID ==51200) %>% group_by (SUBJECT_ID) %>% arrange
(CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID,
VALUE) %>% rename(BASOPHIL = VALUE) %>% merge(mydata12,mydata13.by.x = "SUBJECT_ID"
,by.y = "SUBJECT_ID", all=TRUE)
```

```
ADD GRANULOCYTES 51218 mydata14<- sepsislab.xx %>% filter(ITEMID ==51218) %>% group_by
(SUBJECT_ID) %>% arrange (CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID)
%>% select(SUBJECT_ID, VALUE) %>% rename(GRANULOCYTE = VALUE) %>%
merge(mydata13,mydata14.by.x = "SUBJECT_ID" ,by.y = "SUBJECT_ID", all=TRUE)
```

```
ADD LYMPOCYTE PERCENT 51245
```

```
``}`
```

```
mydata15<- sepsislab.xx %>% filter(ITEMID ==51245) %>% group_by (SUBJECT_ID) %>% arrange
(CHARTTIME) %>% filter(row_number()==1) %>% arrange(SUBJECT_ID) %>% select(SUBJECT_ID,
VALUE) %>% rename(LYMPHOCYTE.PERCENT = VALUE) %>% merge(mydata14,mydata15.by.x =
"SUBJECT_ID" ,by.y = "SUBJECT_ID", all=TRUE)
```

8. selected complete blood count components.

```
mydata16 <- as_tibble(mydata15)
summary(mydata16)
```

9. omission of columns with more than 10% missing data.

```
mydata20 <- subset(mydata16,select = -
c(LYMPHOCYTE.PERCENT,GRANULOCYTE,BASOPHIL,EOSINOPHIL,LYMPHOCYTES,MONOCYTES,
NEUTROPHIL))
```

10.count the number of incomplete variables row-wise

```
mydata23 <- mutate(mydata20, na_count=rowSums(is.na(mydata20)))
```

11. selection of most complete subjects to n-size = 50

```
mydata24 <- filter(mydata23, na_count<=5)
```

12. imputation of missing values with column mean.

```
mydata24$HEMATOCRIT[is.na(mydata24$HEMATOCRIT)] <- 35.9
mydata24$HEMOGLOBIN[is.na(mydata24$HEMOGLOBIN)] <- 12.2)
mydata24$MCHC[is.na(mydata24$MCHC)] <- 33.9
mydata24$MCH[is.na(mydata24$MCH)] <- 29.8
mydata24$MCV[is.na(mydata24$MCV)] <- 84
mydata24$RDW[is.na(mydata24$RDW)] <- 12.9
mydata24$WBCCOUNT[is.na(mydata24$WBCCOUNT)] <- 3.6
mydata24$PLATELETS[is.na(mydata24$PLATELETS)] <- 87
mydata24$REDCELL[is.na(mydata24$REDCELL)] <-3.64
```

13. eliminating SUBJECT\_ID and na.count column from the dataframe.

```
mydata25<- subset(mydata24, select = -c(SUBJECT_ID, na_count))
```

14. multivariate normal distribution assesment.

```
library(car) plot1 <- scatterplotMatrix(mydata25[1:3]) plot2 <- scatterplotMatrix(mydata25[4:6]) plot3 <- scatterplotMatrix(mydata25[7:9])
```

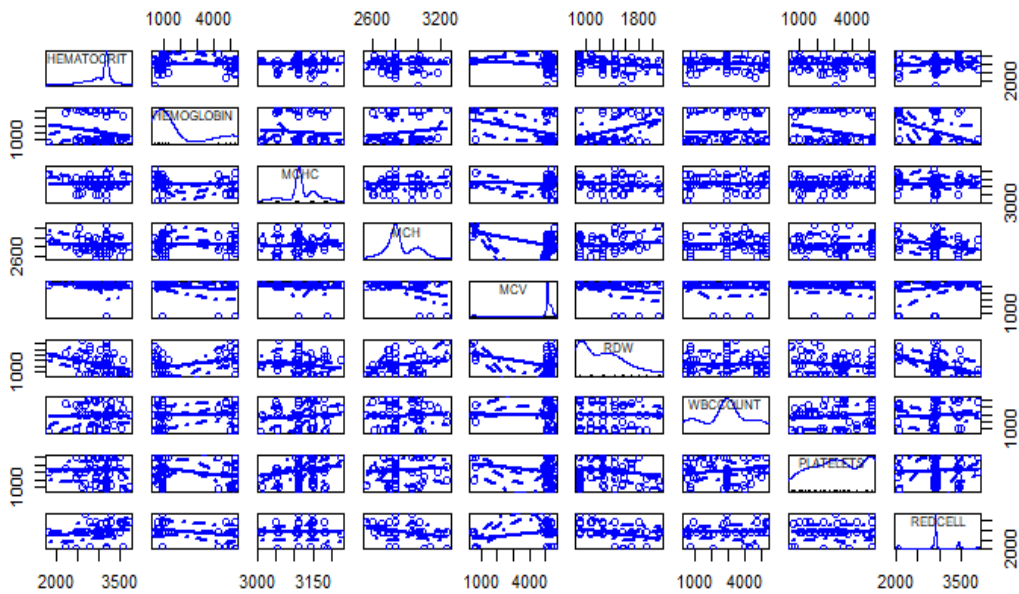


Diagram 1. Scatter plot matrix of variables

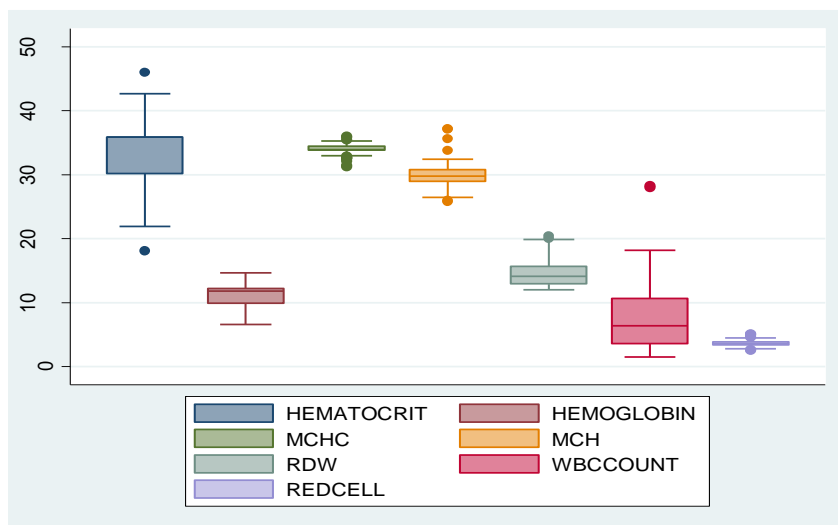


Diagram 2. Box plot matrix of variables.

## Test for multivariate assumptions (stata software).

### Analysis for multivariate equality of means.

```
import delimited "C:\Users\HP\Desktop\Msc Project\mmp.csv"
```

```
(10 vars, 64 obs)
```

```
. mvtest means hematocrit hemoglobin mchc mch mcv rdw wbccount platelets redcell
```

Test that all means are the same

```
Hotelling T2 = 66291.11
```

```
Hotelling F(8,56) = 7365.68
```

```
Prob > F = 0.0000
```

### Multivariate test of covariances among variables.

Test that covariance matrix is diagonal

```
Adjusted LR chi2(36) = 114.25
```

```
Prob > chi2 = 0.0000
```

### Factor analysis.

Table 1. Factor analysis/correlation	Number of obs =	64
Method: principal factors	Retained factors =	5
Rotation: (unrotated)	Number of params =	35

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	1.92169	1.00639	0.6905	0.6905
Factor2	0.91530	0.38385	0.3289	1.0194
Factor3	0.53145	0.46462	0.1910	1.2103
Factor4	0.06683	0.03868	0.0240	1.2343
Factor5	0.02815	0.08793	0.0101	1.2444
Factor6	-0.05977	0.03705	-0.0215	1.2230
Factor7	-0.09682	0.12085	-0.0348	1.1882
Factor8	-0.21767	0.08837	-0.0782	1.1100
Factor9	-0.30604	.	-0.1100	1.0000

LR test: independent vs. saturated:  $\chi^2(36) = 116.19$  Prob> $\chi^2 = 0.0000$

**Table 2. Factor loadings (pattern matrix) and unique variances**

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Uniqueness
hematocrit	-0.1579	0.3088	-0.1128	-0.1188	0.0172	0.8526
hemoglobin	-0.4967	0.5260	0.1644	0.0654	-0.0334	0.4442
mchc	0.0743	0.2236	-0.3030	0.0073	0.1141	0.8396
mch	0.6212	0.4736	-0.0298	0.0580	-0.0511	0.3830
mcv	0.7218	0.3184	0.2121	-0.0052	0.0153	0.3323
rdw	0.6253	-0.3576	0.1293	0.0996	0.0303	0.4535
wbccount	-0.1863	0.1395	0.4446	-0.0041	0.0959	0.7389
platelets	-0.3032	-0.1344	0.3510	-0.0177	-0.0197	0.7661
redcell	-0.4690	0.0464	-0.1283	0.1864	0.0186	0.7263

**Table 3. Rotated factor loadings (pattern matrix) and unique variances**

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Uniqueness
hematocrit	0.0374	0.3367	-0.1195	-0.1113	0.0773	0.8526
hemoglobin	-0.0379	0.7234	0.1590	0.0747	-0.0120	0.4442
mchc	0.1348	0.1147	-0.3114	0.0183	0.1783	0.8396
mch	0.7664	-0.0064	-0.1702	0.0256	-0.0022	0.3830
mcv	0.7897	-0.1895	0.0808	-0.0348	0.0186	0.3323
rdw	0.2977	-0.6651	0.0887	0.0801	-0.0348	0.4535
wbccount	0.0142	0.2259	0.4560	0.0082	0.0425	0.7389
platelets	-0.2545	0.0896	0.3924	-0.0108	-0.0835	0.7661
redcell	-0.3532	0.3184	-0.0694	0.2057	0.0203	0.7263



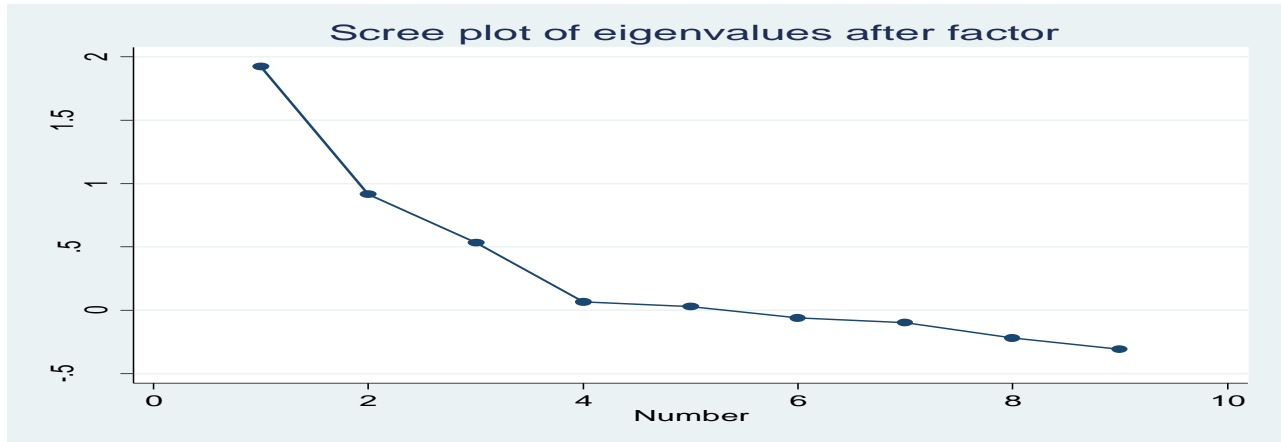


Diagram 3. Scree plot of eigen values after factor analysis.

### Linear discriminant analysis

Table 4. Standardized canonical discriminant function coefficients

	function1	function2
hematocrit	.1582713	-.4661949
hemoglobin	-.8421109	-.1855449
mchc	-.2560761	-.3884369
mcv	-.4867058	-.0245198
mch	.800188	.2489583
rdw	-.6442341	.4082755
platelets	-.5815287	-.0922558
redcell	.2944855	-.1197885

Table 5. Table of discriminant proportions of white cells

True Whitecell Level	Classified 0	1	2	
0	14 53.85	7 26.92	5 19.23	
1	9 39.13	7 30.43	7 30.43	
2	4 26.67	1 6.67	10 66.67	
				n= 64

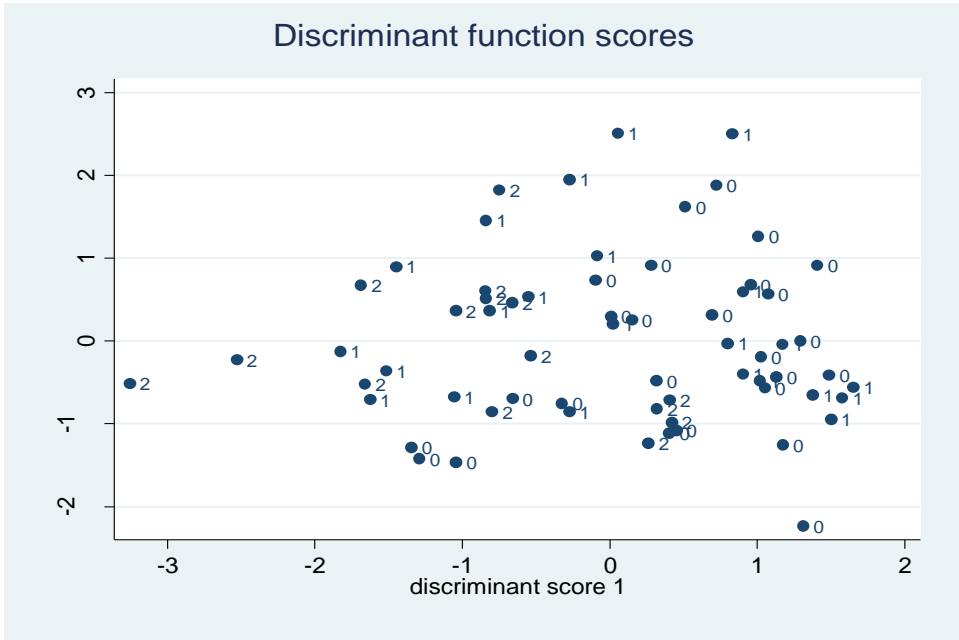


Diagram 4. Plot of discriminant scores on white cell levels

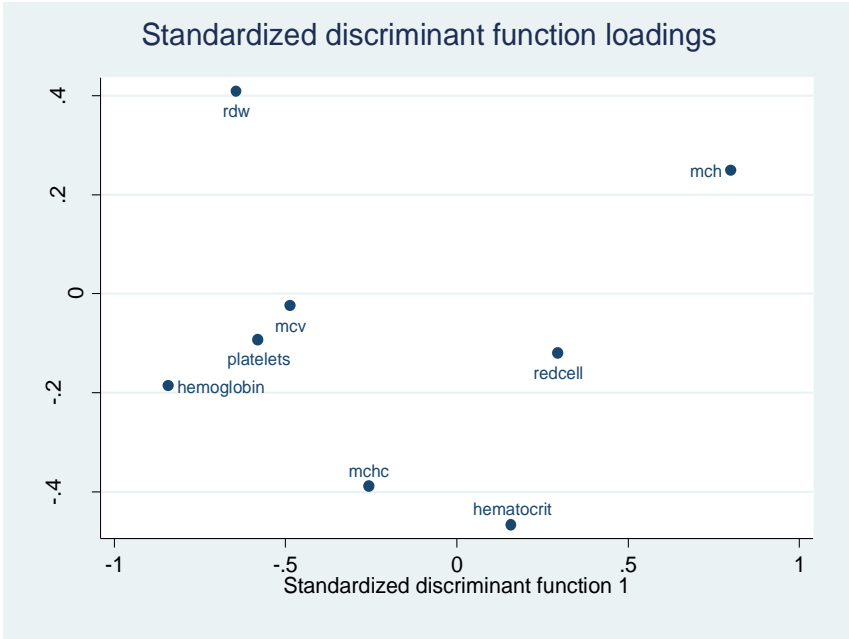


Diagram 5. Plot of variable loadings to discriminant function.

**Multivariate linear predictive model**

Multivariate regression on elevated white cell count and main variables of significance loadings.  
*keep if whitecelllevel ==2*

**Table 6 multivariate predictive coefficients.**

```

mvreg wbccount = mch mcv rdw
Equation Obs Parms RMSE "R-sq" F P

wbccount 15 4 4.766125 0.4657 3.196 0.0664

wbccount | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-----+-----
mch | .3031925 1.15475 0.26 0.798 -2.238395 2.84478
mcv | .6724049 .3146648 2.14 0.056 -.0201677 1.364977
rdw | -1.805831 .722948 -2.50 0.030 -3.397028 -.2146329
_cons | -26.95466 34.48205 -0.78 0.451 -102.8491 48.93983

-27 + 0.6mcv + 0.3mch -1.8rdw =>11 (sepsis)

```

**Cronbach's alpha;** mch mcv rdw

Test scale = mean(unstandardized items)  
Average interitem covariance: 2.073619  
Number of items in the scale: 3  
Scale reliability coefficient: 0.4277

## DISCUSSION

Our analysis reveals positive linear correlation among several complete blood count components. White cell count had largest correlation with platelets, while haemoglobin component had largest positive correlation with haematocrit. Other investigators who have looked into similar analysis included Barry et al (11) who showed that high mean white cell count is associated with infectious and inflammatory conditions but failed to show correlation between raised white cell count and percentage of granulocytes in the differential counts. Mean cell hemoglobin(MCH) and mean corpuscular volume(MCV) had positive correlation. Negative linear correlation was demonstrated between hemoglobin and redcell distribution width (RDW) component. Red cell changes in size and content include hemoglobin content, red cell distribution width (RDW) mean cell volume (MCV) mean cell hematocrit (MCH) mean cell hemoglobin concentration (MCHC).

Platelet counts were the most varied component with largest outliers. This may be explained by their diverse response to developing sepsis early in the disease process. In sepsis, shift of development of immune cells through cell signaling molecules affect red cells and platelets due to competition for limited growth factors and space for maturation. Commonly low platelets have been associated with sepsis.

There was sufficient statistical evidence to reject the null that all means were the same (Hotelling T square p .000). The blood components are in the same measurement scale with different summary measures therefore independent with some correlation.

The test that covariance matrix is diagonal is significant (chi2 p.000).

The first principle component had an eigen value of 1.92169 with cumulative variability Of 69.05% explained. The scree plot analysis showed selection of one factor with eigen value greater than one. MCV, MCH and RDW had positive coefficients while the rest had negative coefficient into the principle component. These three variables had highest correlation with the principal factor. On oblique rotation that aims to maximize the variance, high correlation with the principal factor was maintained in these three variables

In discriminant analysis, two thirds of elevated white cell count could be discriminated well using the linear discriminant functions. Not so much for low and normal white cell count. However, white cell count elevation is what is routinely used for diagnosis of sepsis and most important category to discriminate from other categories. Similar variables that explained largest variation (MCV, MCH and RDW) also loaded significantly in discriminant function to separate the categories. This informs on the choice of the three variables in predictive diagnostic model for sepsis.

Elevated white cell count in sepsis was regressed against the selected variables from factor and discriminant analysis. A value greater than eleven may be predictive of sepsis. This model may add to the utility of complete blood count components in diagnosis of sepsis. The cronbach's alpha is 0.4 which indicates a low reliability of our predictive model. The recommended reliability cut off is 0.5 (14). The direction may be correct to inform on future researchers in this field on such gaps to be filled.

## CONCLUSION

Mean cell volume, mean cell hemoglobin and red cell distribution width may be used in predictive diagnosis of sepsis in addition to white cell changes.

## Limitations.

Omitted variable bias; Missing data in some blood count components dropped in analysis may have some unknown significance.

### **Recommendations.**

Future studies in this field inclusive of all CBC measurements from local population may add strength to these findings and reveal further information towards variation in sepsis.

## REFERENCES

1. Hotchkiss, R. S., Moldawer, L. L., Opal, S. M., Reinhart, K., Turnbull, I. R., & Vincent, J. L. (2016). Sepsis and septic shock. *Nature reviews. Disease primers*, 2, 16045. <https://doi.org/10.1038/nrdp.2016.45>
2. Parrillo, J. E., Parker, M. M., Natanson, C., Suffredini, A. F., Danner, R. L., Cunnion, R. E., & Ognibene, F. P. (1990). Septic shock in humans. Advances in the understanding of pathogenesis, cardiovascular dysfunction, and therapy. *Annals of internal medicine*, 113(3), 227–242. <https://doi.org/10.7326/0003-4819-113-3-227>
3. Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., & Pinsky, M. R. (2001). Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Critical care medicine*, 29(7), 1303–1310. <https://doi.org/10.1097/00003246-200107000-00002>.
4. Martin, G. S., Mannino, D. M., Eaton, S., & Moss, M. (2003). The epidemiology of sepsis in the United States from 1979 through 2000. *The New England journal of medicine*, 348(16), 1546–1554. <https://doi.org/10.1056/NEJMoa022139>
5. Rhodes, A., Evans, L. E., Alhazzani, W., Levy, M. M., Antonelli, M., Ferrer, R., Kumar, A., Sevransky, J. E., Sprung, C. L., Nunnally, M. E., Rochwerf, B., Rubenfeld, G. D., Angus, D. C., Annane, D., Beale, R. J., Bellingham, G. J., Bernard, G. R., Chiche, J. D., Cooper-Smith, C., De Backer, D. P., ... Dellinger, R. P. (2017). Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. *Intensive care medicine*, 43(3), 304–377. <https://doi.org/10.1007/s00134-017-4683-6>
6. Lassale C, Curtis A, Abete I, Van Der Schouw YT, Verschuren WMM, Lu Y, et al. Elements of the complete blood count associated with cardiovascular disease incidence: Findings from the EPIC-NL cohort study. *Sci Rep [Internet]*. 2018;8(1):1–11.
7. Wile, M. J., Homer, L. D., Gaehler, S., Phillips, S., & Millan, J. (2001). Manual differential cell counts help predict bacterial infection. A multivariate analysis. *American journal of clinical pathology*, 115(5), 644–649. <https://doi.org/10.1309/J905-CKYW-4G7P-KUK8>
8. Stearns-Kurosawa, D. J., Osuchowski, M. F., Valentine, C., Kurosawa, S., & Remick, D. G. (2011). The pathogenesis of sepsis. *Annual review of pathology*, 6, 19–48. <https://doi.org/10.1146/annurev-pathol-011110-130327>
9. Goyette, R. E., Key, N. S., & Ely, E. W. (2004). Hematologic changes in sepsis and their therapeutic implications. *Seminars in respiratory and critical care medicine*, 25(6), 645–659. <https://doi.org/10.1055/s-2004-860979>
10. Al-Gwaiz, L. A., & Babay, H. H. (2007). The diagnostic value of absolute neutrophil count, band count and morphologic changes of neutrophils in predicting bacterial infections. *Medical principles and practice : international journal of the Kuwait University, Health Science Centre*, 16(5), 344–347. <https://doi.org/10.1159/000104806>
11. Wenz B, Gennis P, Canova C, Burns ER. The clinical utility of the leukocyte differential in emergency medicine. *Am J Clin Pathol*. 1986;86(3):298-303. doi:10.1093/ajcp/86.3.298
12. Zahorec R. (2001). Ratio of neutrophil to lymphocyte counts--rapid and simple parameter of systemic inflammation and stress in critically ill. *Bratislavske lekarske listy*, 102(1), 5–14.
13. Westerdijk, K., Simons, K. S., Zegers, M., Wever, P. C., Pickkers, P., & de Jager, C. (2019). The value of the neutrophil-lymphocyte count ratio in the diagnosis of sepsis in patients admitted to the Intensive Care Unit: A retrospective cohort study. *PloS one*, 14(2), e0212861. <https://doi.org/10.1371/journal.pone.0212861>
14. Beckett C, Eriksson L, Johansson E, Wikström C. Multivariate Data Analysis (MVDA). *Pharmaceutical Quality by Design: A Practical Approach*. 2017. 201–225 p.

15. Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., & Newman, T. B. (2013). *Designing Clinical Research: An Epidemiologic Approach* (2nd ed.). Philadelphia: Lippincott, Williams & Wilkins.

## APPENDIX 1: STUDY BUDGET.

Budget Item	Amount (K.shs.)
Research fee for KNH-ERC	2,500
Contingency fund	1,500
Total	4,000



## APPENDIX 2: TIME FRAME.

Time	Jan	March	May	November	November	December
Concept proposal						
Proposal presentation						
Ethical approval						
Statistical analysis & report writing						
Submission of report.						

### APPENDIX 3: DATA COLLECTION EXCEL FORM.

S/N	AGE	SEX	TEMP	SYST	WBC	RBC	HB	RDW	PLT	GRANU	NEUT	BASO	EOS	HCT	MCV	MCH	MCHC
1.																	
2.																	
3.																	
4.																	
.n=50																	

## **APPENDIX 4: MIMIC III DATA ACCESS AND CONSENT WAIVER INFORMATION.**

### **MIMIC-III Clinical Database**

Alistair Johnson, Tom Pollard, Roger Mark

Published: Sept. 4, 2016. Version: 1.4

MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (including post-hospital discharge). MIMIC supports a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement, and electronic tool development. It is notable for three factors: it is freely available to researchers worldwide; it encompasses a diverse and very large population of ICU patients; and it contains highly granular data, including vital signs, laboratory results, and medications.

Before data was incorporated into the MIMIC-III database, it was first deidentified in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting. The deidentification process for structured data required the removal of all eighteen of the identifying data elements listed in HIPAA, including fields such as patient name, telephone number, address, and dates. In particular, dates were shifted into the future by a random offset for each individual patient in a consistent manner to preserve intervals, resulting in stays which occur sometime between the years 2100 and 2200.

Protected health information was removed from free text fields, such as diagnostic reports and physician notes, using a rigorously evaluated deidentification system based on extensive dictionary look-ups and pattern-matching with regular expressions. The components of this deidentification system are continually expanded as new data is acquired.

The project was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived because the project did not impact clinical care and all protected health information was deidentified.

### **Acknowledgements**

This research and development was supported by grants NIH-R01-EB017205, NIH-R01-EB001659, and NIH-R01-GM104987 from the National Institutes of Health. The authors would also like to thank Philips Healthcare and staff at the Beth Israel Deaconess Medical Center, Boston, for supporting database development, and Ken Pierce for providing ongoing support for the MIMIC research community.

### **Conflicts of Interest**

The authors declare no competing financial interests.

## APPENDIX 5: LETTER OF AUTHORITY TO ACCESS MIMIC III DATABASE.

Your application for PhysioNet credentialing  
PhysioNet Automated System <noreply@physionet.org>

Mon, Apr  
13, 4:07  
PM

Dear James Mwangi,

Thank you for your interest in the PhysioNet Clinical Databases. We are pleased to say that your application for credentialed access has been approved. You are now able to access protected databases upon agreeing to the terms of usage. For example, you can access MIMIC-III by following the steps below:

- Go to the project page at <https://physionet.org/content/mimiciii/>
- Find the “Files” section in the project description
- Click “Sign the data use agreement” to agree to the terms of usage for this dataset

Regards,

The PhysioNet Team,  
MIT Laboratory for Computational Physiology,  
Institute for Medical Engineering and Science,  
MIT, E25-505 77 Massachusetts Ave. Cambridge, MA 02139

# Turnitin Originality Report

Document Viewer

Processed on: 24-Nov-2020 22:31 EAT  
ID: 1456322901  
Word Count: 6112  
Submitted: 1

Similarity Index	Similarity by Source
15%	Internet Sources: 15%
	Publications: 9%
	Student Papers: 10%



SEPSIS PREDICTION FROM COMPLETE BLOOD COUNT C... By James Waweru

[exclude quoted](#) [exclude bibliography](#) [exclude small matches](#)

mode: [quickview \(classic\) report](#)

[Change mode](#)

[print](#)

[refresh](#)

[download](#)

6% match (Internet from 01-Sep-2020)

<https://www.physionet.org/content/mimiciii/1.4/>

2% match (student papers from 22-Oct-2018)

Submitted to UT, Dallas on 2018-10-22

1% match (Internet from 27-Jul-2020)

[https://mafadoc.com/la-rd-dans-les-industries-culturelles-et-creatives\\_597704411723dd0928ed0aa6.html](https://mafadoc.com/la-rd-dans-les-industries-culturelles-et-creatives_597704411723dd0928ed0aa6.html)

1% match (Internet from 12-Jun-2020)

<https://www.researchsquare.com/article/rs-16270/v1>

1% match (Internet from 06-May-2010)

<http://www.lib.vmd.edu>

<1% match (Internet from 07-Nov-2019)

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0212861>

<1% match (Internet from 20-May-2019)

<http://bjcancer.org>

<1% match ()

<http://hvdra.hull.ac.uk>

<1% match (Internet from 22-Jan-2018)

<https://www.thieme-connect.de/products/ejournals/html/10.1055/s-3004-860979>

<1% match (Internet from 15-Feb-2019)

<https://www.stata.com/manuals14/mv.pdf>

<1% match (Internet from 09-May-2019)

<https://www.beckmancoulter.com/download/file/wsr-32913/4237523CB?type=pdf>

<1% match (Internet from 13-Oct-2020)

<https://www.ntnu.no/documents/10455/1290813005/EksamensoppgaveSOS1002H2019.pdf/274b2761-dfb7-105e-32e4-413093a70d1?i=1578929934502>

<1% match (Internet from 06-Nov-2020)

<https://www.physionet.org/galtin/getstarted/access/>

<1% match (student papers from 31-Aug-2010)

Submitted to University of Leicester on 2010-08-31

<1% match (student papers from 19-Apr-2018)

Submitted to CSU, San Jose State University on 2018-04-19

<1% match (Internet from 21-Sep-2020)

<https://www.researchsquare.com/article/rs-52103/v1>

<1% match (Internet from 09-Nov-2020)

<https://scforum.biomedcentral.com/articles/10.1186/s13054-018-1973-5>

<1% match (publications)

Dehghani, Mohammad Reza, Yousef Rezaei, Sanam Fakour, and Nasim Arjmand. "White Blood Cell Count to Mean Platelet Volume Ratio Is a Prognostic Factor in Patients with Non-ST Elevation Acute Coronary Syndrome with or without Metabolic Syndrome". *Korean Circulation Journal*. 2016.

<1% match (student papers from 17-Mar-2020)

Submitted to Vrije Universiteit Amsterdam on 2020-03-17

<1% match (Internet from 17-Apr-2015)

<http://www.aria.ru>

<1% match (publications)

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wai H. Lehman et al. "MIMIC-III, a freely accessible critical care database". *Scientific Data*. 2016

<1% match (student papers from 02-Aug-2017)

Submitted to Postgraduate Institute of Medicine on 2017-08-02

<1% match (Internet from 28-Oct-2020)

<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004947>

<1% match ()

<http://hdl.handle.net>

<1% match (Internet from 27-Mar-2020)

<https://physionet.org/content/?topic=mimic>

<1% match (Internet from 05-May-2013)

<http://www.heftopathology.com>

<1% match (Internet from 06-Feb-2019)

<http://iranjradiol.com>

<1% match (Internet from 29-Oct-2016)

<https://www.coursehero.com/file/14232216/mv/>

<1% match (publications)

MIT Critical Data. "Secondary Analysis of Electronic Health Records". *Springer Nature*. 2016

<1% match (publications)

U I. Dumaswala, M I. Wilson, Y L. Wu, J. Wykle, L. Zhuo, L M. Douglass, D L. Daleke. "Glutathione loading prevents free radical injury in red blood cells after storage". *Free Radical Research*. 2009