Statistical Analyses for a Clinical Sepsis Study

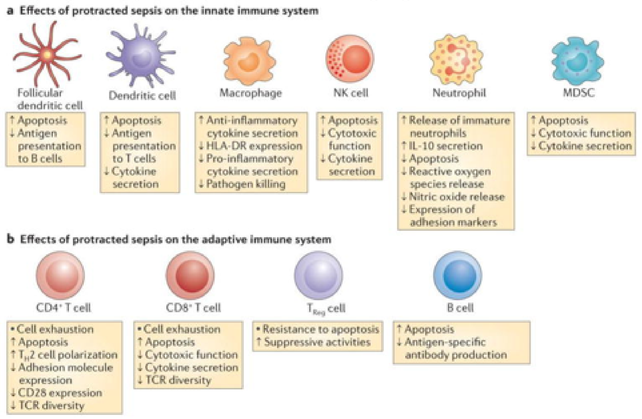
Gansen Deng *supervised by* Dr. Wenqing He
Department of Statistical and Actuarial Sciences, University of Western Ontario
gdeng7@uwo.ca

Abstract

Sepsis is a life-threatening syndrome caused by the immune system working overtime to fight infection. A sepsis clinical study has been conducted to investigate the effect of sepsis in humans on MAIT cells in the blood by looking at their surface markers and their function (production of cytokines), and we provide statistical analysis to the team. The difference between different cohorts are compared using t-tests and Mann-Whitney U tests, and GEE (Generalized Estimating Equations) is utilized to identify variables with significant longitudinal trends. Correlation analysis is also involved to check if variables are significantly correlated with each other. The key conclusion is that MAIT cells decrease in abundance in septic patients comparing with healthy control, but surprisingly that their function progressively declines as the patient gets better. *MAIT % of total* and *CD14 HLADR* are found to be two important indicators of sepsis.

Background

Sepsis is a life-threatening syndrome typically associated with early hyperinflammation, immunosuppression in its protracted phase, and a continuum of organ dysfunction abnormalities [1]. It is usually caused by the body's immune system working overtime to fight infection. In this case, a large number of chemicals released into the blood during this process triggers widespread inflammation, which can lead to organ damage [2]. We can see how sepsis affects our immune system from the following figure [3]:



The study objective of this project is to **identify the phenotype and functional state of MAIT cells throughout sepsis**. We want to determine the effect of sepsis in humans on MAIT cells in the blood by looking at their surface markers (to determine whether they are activated or anergized), and their function (production of cytokines).

Data collection

The data about MAIT cells was collected by drawing blood from three different cohorts at six different timepoints using flow cytometry. The patient data was collected from patients that were admitted to the ICU exhibiting hypotension and being put on vasopressors.

Timepoint	
A	Within 24 hours of admission
B	Off vasopressors
C	24 hours after vasopressor is off
D	72 hours after vasopressor is off
E	At ICU discharge
F	At hospital discharge

Cohort	
septic	Patients with sepsis
non_septic	Patients without sepsis
healthy	Healthy volunteers

After collecting the data, information is measured using MFI (Mean Fluorescence Intensity), and the percentage of MAIT cells (the proportion of MAIT cells in all the cells).

Comparison

The algorithm for t-test:

- Given two samples Y_1, Y_2:

1. Perform the Shapiro-Wilk test to check if Y_1 and Y_2 are normally distributed
2. If both samples are normally distributed, then perform two sample t-test. Otherwise, go to step 3
3. Apply the Box-Cox transformation with the parameter λ on the shifted sample $Y_1 + 1$ and $Y_2 + 1$, such that

$$Y_i^* = \begin{cases} log(Y_i + 1) & \text{if } \lambda = 0 \\ ((Y_i + 1)^\lambda - 1)/\lambda & \text{otherwise} \end{cases} \quad (1)$$

where the initial value of λ is set at -5

4. Use the Shapiro-Wilk test to check the normality of Y_1^* and Y_2^*. If at least one of them is not normally distributed, then set $\lambda = \lambda + 0.25$ (Stop the algorithm when $\lambda > 5$) and go back to step 3. Otherwise, go to step 5
5. Perform t-test to compare the transformed sample Y_1^* and Y_2^*. The test statistic is

$$t = \frac{\bar{Y}_1^* - \bar{Y}_2^*}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2)$$

where $s_p = \sqrt{\frac{(n_1-1)s_{Y_1^*}^2 + (n_2-1)s_{Y_2^*}^2}{n_1+n_2-2}}$

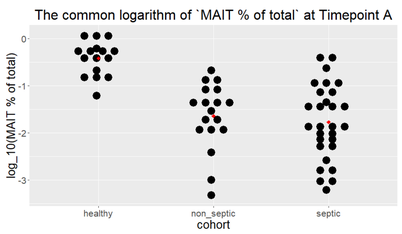
Mann-Whitney U test:

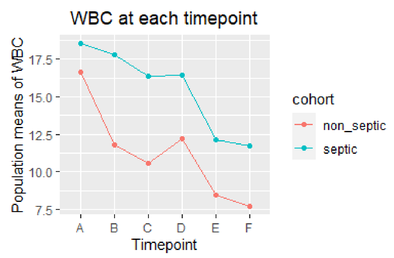
When using the Mann-Whitney U test, two samples are compared directly using test statistic:

$$U = min\{R_1 - \frac{n_1(n_1+1)}{2}, R_2 - \frac{n_2(n_2+1)}{2}\} \quad (3)$$

Under the null hypothesis that there is no difference between two groups, U follows a distribution with mean $n_1n_2/2$ and variance $n_1n_2(n_1+n_2+1)/12$. [4] The distribution is tabulated for small samples, but when the sample size is above 20, the normal distribution can be used to approximate the distribution.

Two variables that are significantly different in different cohorts:





Longitudinal analysis

Generalized Estimating Equations:

The GEE model for this problem is:

$$g(\mu_M) = \beta_0 + \beta_B x_B + \beta_C x_C + \beta_D x_D + \beta_E x_E + \beta_F x_F + \epsilon, \ M = A, ..., F \quad (4)$$

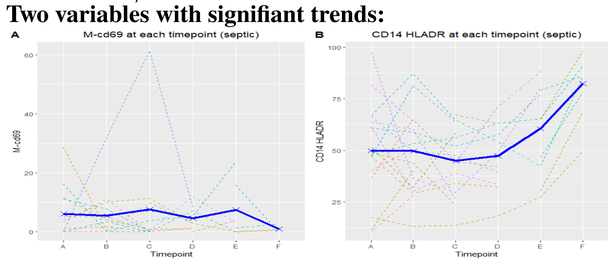
where $\mu_A, ..., \mu_F$ are the population means at timepoint A,..., F, respectively. $x_B, ..., x_F$ are dummy variables such that for $M = B, C, ..., F$, $x_M = 1$ at timepoint M and $x_M = 0$ otherwise. $\beta_0, \beta_B, ..., \beta_F$ are the coefficients to be estimated.

The coefficients are estimated by solving the equations:

$$\sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) = 0 \quad (5)$$

where $D_i = \frac{\partial \mu_i}{\partial \beta^T}$ and $V_i = cov(Y_i)$ is the working covariance matrix [5].

Two variables with signifiant trends:



Correlation analysis

Correlation coefficients:

Variable type	Correlation coefficient
Two continuous variables	Pearson
Two categorical variables	Phi
One continuous & one categorical	Point-Biserial

Logistic regression:

- **Goal:** To investigate the association between the binary variable *Survival* and multiple other variables.
- **Steps:**

1. Fit a logistic regression model with *Survival* as the response and the other variables as the predictors
2. Fit a reduced model with *Survival* as the response and selected predictors.
3. Use likelihood ratio test to compare the fitted model with the reduced model. The test statistic is $\Lambda = -2\,ln(L_0/L_1) \sim \chi_p^2$ [6], where L_0, L_1 are the maximum likelihood of the reduced model and the fitted model, respectively, and p is the number of predictors reduced from the previous model.
4. If the fitted model is not significantly better than the reduced model, then we claim that the association between *Survival* and reduced variables is not significant.
5. Repeat step 1 to 4 until we cannot reduce the model. The variables in the final model will have significant association with *Survival* response.

t-test VS Mann-Whitney U test

Result:

# of comparisons	467
# that t-test has a smaller p-value	303
# that t-test is significant but M-W test is not	18
# that M-W test is significant but t-test is not	18

Findings:

- The p-values of t-tests tend to be smaller
- Two tests give consistent results in most cases

Conclusion:

The Mann-Whitney U test can be used as an alternative of t-test when the normal assumption of t-test is not satisfied. It's pretty useful for ordinal data where data transformation cannot be used. However, for continuous variables, t-test and Mann-Whitney U test give consistent results.

Key Conclusion

- MAIT cells decrease in abundance in septic patients versus healthy control, but their function (production of cytokines) progressively declines as the patient gets better.
- *MAIT % of total* is an important indicator of the severity of sepsis. A patient with lower *MAIT % of total* value might have severer sepsis.
- *CD14 HLADR* is another important indicator of sepsis. The *CD14 HLADR* values of different cohorts are significantly different, and the trend of it is significant in the *sepsis* cohort. Besides, it is also correlated with many stimulated variables.

Future work

- Use random effect model to investigate variables at the individual level
- Handle missing values using imputation or adjusted statistical methods before analyzing the data
- Try different tests and collect more data to verify the conclusions.

References

[1] Peter A Szabo, Ram V Anantha, Christopher R Shaler, John K McCormick, and SM Haeryfar. Cd1d-and mr1-restricted t cells in sepsis. *Frontiers in immunology*, 6:401, 2015.

[2] Minesh Khatri. Sepsis (blood infection): Symptoms, causes and treatments. https://www.webmd.com/a-to-z-guides/sepsis-septicemia-blood-infection#1, May 2019.

[3] Fuhua Li and Jianhai Xiang. Signaling pathways regulating innate immune responses in shrimp. *Fish & shellfish immunology*, 34(4):973–980, 2013.

[4] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.

[5] G.Y. Yi. *Lecture Notes of "SS9878/CS9878 - Analysis of High Dimensional Noisy Data"*. University of Western Ontario, 2020.

[6] Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 14: Logistic regression. *Critical care*, 9(1):112, 2005.

Acknowledgment

I would like to thank Dr. Wenqing He for guiding me to complete the project, and thank the team of Joshua Choi for providing the data.