A Deep Look into Cytokines and Septic Shock *

Bowen Chen[1], Sean La[2], Yue Liu[3,4], Babak Nasouri[4,5], Matthew Nguyen[1,6], Ka Mun Nip[7], Mingfeng Qiu[3,4], and Vasilii Triandafilidi[8]

[1] Department of Computer Science, Simon Fraser University, Burnaby, Canada
[2] Department of Mathematics, Simon Fraser University, Burnaby, Canada
[3] Department of Mathematics, University of British Columbia, Vancouver, Canada
[4] Institute of Applied Mathematics, University of British Columbia, Vancouver, Canada
[5] Department of Mechanical Engineering, University of British Columbia, Vancouver, Canada
[6] Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, Canada
[7] Department of Medicine, University of British Columbia, Vancouver, Canada
[8] Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, Canada

Abstract. Sepsis is the leading cause of death in the intensive care unit worldwide. Despite having many environmental factors, septic shock can be significantly attributed to cytokines which are specialized proteins that regulate inflammation in the body. Understanding the correlation between the cytokine levels, as well as genes that code for them, is crucial in reducing the rates of mortality. We performed genome-wide association studies (GWAS) to determine which single nucleotide polymorphisms (SNPs) are correlated with mortality of patients with septic shock. We also identified SNPs which are correlated with serum concentrations of various cytokines, determined the cytokines correlated with patient mortality, and trained various machine learning classifiers. We have identified several SNPs that correlate with mortality rates including one in the gene of a protein that is found to regulate cytokine levels.

1 Introduction

Sepsis is a serious medical condition usually caused by bacterial infection. It is the leading cause of death in the intensive care unit worldwide. On average, one of every 18 deaths in Canada occurs due to sepsis and septic shock [1]. Despite having many external environmental factors, septic shock can be greatly attributed to cytokines, which are special proteins in the body that regulate inflammation. Understanding the correlation between the cytokine levels as well as genes that code for them is crucial in developing therapy and reducing the rates of mortality.

* This report is prepared for PIMS data science workshop 2018, supported by Pacific Institute for the Mathematical Sciences. The authors are in alphabetical order.

Previous studies have shown correlations between various cytokine levels and death after 28 days in patients in septic shock [2]. However, the existence of a correlation does not imply that increased levels of cytokines cause death in these patients; there may be external factors which cause both. To rule out these hidden variables, one can perform genome-wide association studies (GWAS), where single nucleotide polymorphisms (SNPs) are found that correlate with both increased levels of particular cytokines and death. If a SNP is found to correlate with higher levels of a particular cytokine, we can infer that it is likely that the SNP causes increased levels of that cytokine. Of course, it is possible that such a SNP affects a different function, e.g. increased levels of a different protein which then causes an increased likelihood of death, but such an investigation is beyond the scope of this study and we will operate under the assumption that these SNPs modify no factors other than cytokines. With this assumption in mind, if it is also the case that the SNP is associated with increased likelihood of death, then the cytokines which correlate with both the SNP and death *cause* death, since the SNPs only modify cytokine levels. Figure 1 visually depicts this reasoning.

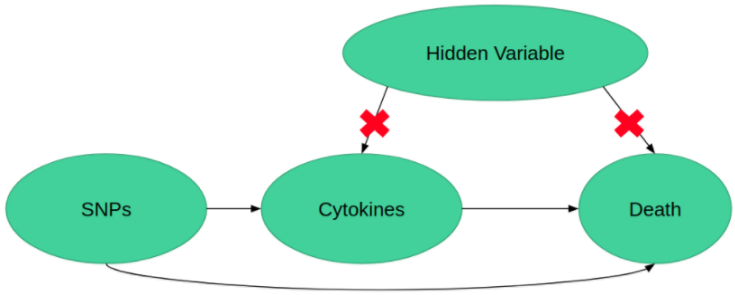

Fig. 1: A causal relationship between cytokines and death can be inferred if there is a correlation between certain SNPs and particular cytokines, as well as a correlation between these SNPs and death, since SNPs can only act through cytokines.

2 Method

Our dataset was from the VASST trial [3] at the St Pauls hospital in Vancouver, BC, Canada, on the effects of a certain vasopressin treatment on health outcomes, which included cytokine concentrations and 1.2 million SNP measurements from 330 patients with septic shock. We performed GWAS to determine which SNPs are correlated with mortality of the patients, as well as SNPs which are correlated with serum concentrations of various cytokines. We also determined which cytokines are correlated with mortality of the patients by using

statistical analysis and training various machine learning classifiers to predict whether a septic patient is likely to die within 28 days based on their cytokine concentrations.

The correlation results for SNPs with death and cytokine levels were intersected in order to obtain the SNPs that were correlated with both death and various cytokine levels. This was performed by eliminating all SNPs that did not meet a threshold p-value of 5×10^{-5} for both GWAS results. Only the cytokines that were found to significantly correlate with patient survival were considered. This enabled us to determine a list of SNPs that were related to the key cytokines correlating with death. The individual SNPs were then queried on the UC Santa Cruz Genome Browser to determine its location in the human genome. SNPs located in coding regions were further studied to reveal potential genomic mechanisms of any causal relationship found between certain cytokines and patient mortality.

3 Results

3.1 Correlation between cytokines and death

For each cytokine, we perform Student's t-test to determine whether the cytokine levels are significantly different in patients who have survived past 28 days, versus those who have not. The results are summarized in Table 1.

To further examine the correlation between cytokine level and survivability, for each cytokine, we use logistic regression to fit the probability of survival as a function of cytokine level, and use it to build a predictor for survivability. We split our dataset into 70% training and 30% testing. Since we have many outliers, we use L1 loss in order to build a most robust model. More weight is given to the data points where the patients have died, since we have fewer of them.

The accuracy of these models varies. The best two, using GRO (test accuracy 0.71429) and IL8_HW (test accuracy 0.69388) respectively, are given in Figure 2. The error produced by these models are one-sided, in the sense that it is more likely to predict "survival" for the patients who have died, than the other way around. It is also much more accurate on the regime where cytokine level is high.

Additionally, we attempt to create a more accurate predictor for survivability by using more sophisticated machine learning (ML) techniques, which use multiple cytokines simultaneously and allow us to explore the interaction between the cytokines. We have considered multivariable logistic regression, decision trees ranging from depth of 1 to 15, and neural networks with various architectures. Although we could obtain very high training accuracy, none of these models yields test accuracy of higher than 70%. This suggests that these models are suffering from overfitting. With a relatively small dataset, and considering the high amount of noise due to inherent randomness associated with human biology, this is as good as we can reasonably expect.

We would also like to visualize the possible interaction between cytokines. For each pair of cytokines, we produce a 2D plot displaying the location of each

Cytokines	t-value	p-value	Cytokines	t-value	p-value
GRO	-4.275651	0.000027	IL2	-1.40930	0.16003
IL8_HW	-4.04777	0.00007	CD40L	-1.40811	0.16038
IL7	-3.46039	0.00064	EGF	-1.34633	0.17945
EOTAXIN	-3.42037	0.00073	IL1A	-1.24536	0.21420
MIP1B	-3.33234	0.00100	IL1a_HW	-1.23291	0.21880
MIP1b_HW	-3.29874	0.00112	IL13	-1.18334	0.23783
MCP1	-3.29810	0.00112	Resistin_HW	1.15751	0.24820
IL8	-2.90489	0.00401	GMCSF	-1.13524	0.25739
IL4	-2.67821	0.00791	TNFA	-1.04183	0.29853
IL1B	-2.54473	0.01156	GranzymeB_HW	1.00870	0.31412
MIP1A	-2.41036	0.01668	IL12P70	-0.81889	0.41365
HSP70_HW	-2.36953	0.01859	Thrombospondin1_HW	0.80280	0.42287
TGFA	-2.30090	0.02225	Lactoferrin_HW	0.71628	0.47451
MCP3	-2.02820	0.04363	IFNG	0.70437	0.48188
IL17	-1.95551	0.05167	Elastase2_HW	0.69806	0.48581
GCSF	-1.95199	0.05209	IL9	-0.45252	0.65129
IL1RA	-1.94257	0.05322	IL5	-0.44668	0.65550
IL15	-1.88879	0.06011	IL12P40	-0.40331	0.68707
VEGF	-1.81033	0.07148	IL6	0.37954	0.70462
FGF2	-1.78716	0.07516	IP10	-0.36508	0.71537
FRACTALKINE	-1.73725	0.08361	IL10	0.35217	0.72501
MDC	1.73276	0.08441	IL2RA	-0.31630	0.75204
IFNA2	-1.63332	0.10370	MMP8_HW	0.24843	0.80401
IL3	-1.55221	0.12192	NGAL_HW	-0.14492	0.88490
MIP1a_HW	-1.45076	0.14814	TNFB	0.02158	0.98281
FLT3L	-1.44079	0.15093			

Table 1: Result of t-test on the correlation between cytokine level and death. There are 14 cytokines with $p < 0.05$.

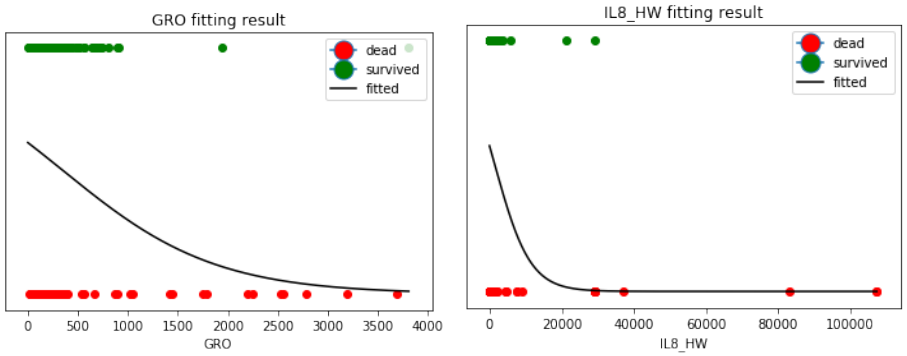

Fig. 2: Visualization of logistic regression result with (a) GRO and (b) IL8_HW respectively.

patient in the plane spanned by the two cytokines. The two interesting ones are displayed below. We also perform principal component analysis, keeping two dimensions, and plot the principle components of the data points. These plots (examples shown in Figure 4) seem to indicate that most data points cluster near zero cytokine level, and patients with high level of multiple cytokines are less likely to survive. However, we could not pick out any trend about interacting cytokines.

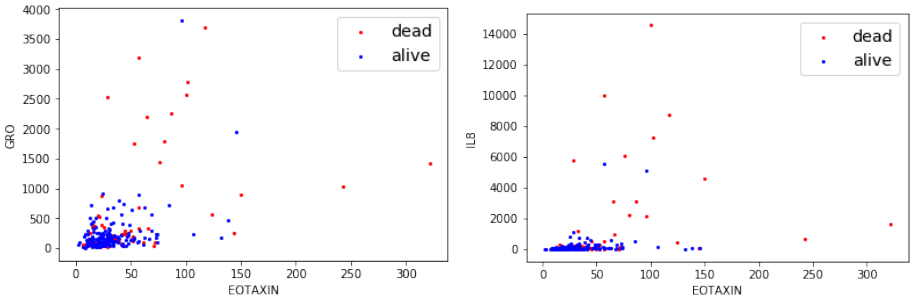

Fig. 3: 2D visualization of distribution of data points. Note that some red points are hidden underneath the blue cluster at bottom right.

We have also attempted unsupervised learning by clustering, as well as k-nearest-neighbor. However, neither gives us satisfactory results.

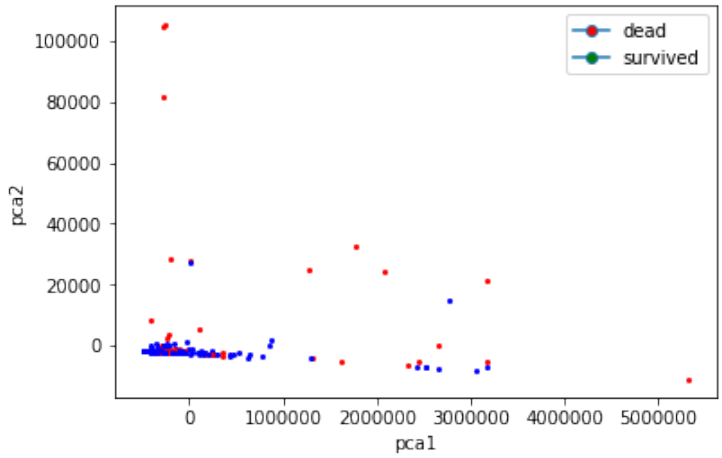

Fig. 4: Principle component analysis (PCA) of the data points. Only the 14 cytokines that were found to correlate strongly with survivability were used.

3.2 Statistical fitting and correlation of Cytokine levels to death rate

After 14 cytokines have been identified, we perform further analysis of the cytokine levels and their association to death. We plot the histogram of number of people who died on each day (Figure 5). As expected we see that number of people who survive the 28 day mark are approximately 67% of the whole pool of people.

For the purposes of this work, we have excluded the surviving category from consideration and analyzed only people who die within 28 day mark. For those people we analyze the histograms of the number of people dying each day versus the cytokine levels (Figure 6).

For each cytokine, the patient data is divided into two groups whose cytokine levels are within the 50% percentile and those whose cytokine levels are beyond that level. For both of these groups we have plotted the number of mortality-time histograms and fitted the data to exponential $e^{-\lambda x}$. The parameter λ (average death coefficient) is specific to each cytokine and identifies the likely-hood of a person to die. It is then plotted as a function of cytokine in Figure 7.

As could be seen from the graph, the 14 cytokines identified initially could be narrowed down to 2 specific cytokines, namely IL1B and MIP1A, which are more correlated to mortality outcomes. These cytokines could be used for further analysis of the association between the identified cytokines and mortality outcomes.

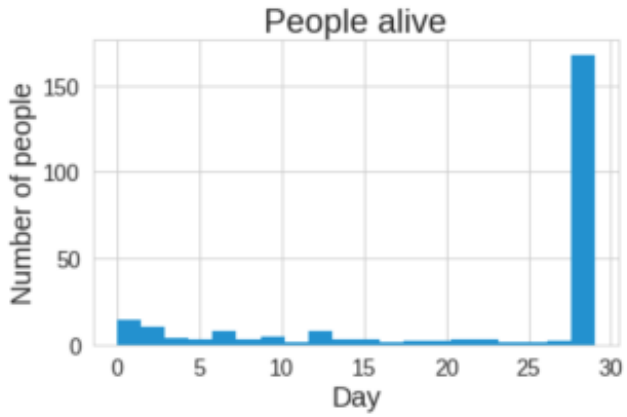

Fig. 5: Distribution of people who lived until the day shown.

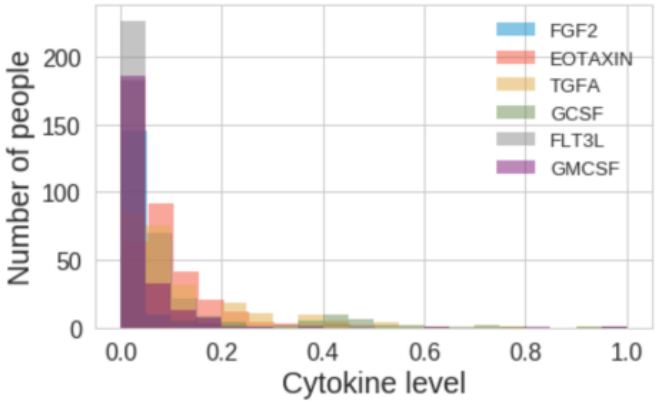

Fig. 6: Distribution of cytokine levels.

3.3 Correlation of SNP and death

We perform GWAS with the binary trait of patient being alive or dead within 28 days, using the software package PLINK [4]. The association between each SNP and patient survival rate iss analyzed using a χ^2 test. The result is shown in Figure 8, from which we can identify SNPs associated with survival with confidence level p smaller than any threshold.

3.4 Correlation of SNP and cytokine levels

We further perform GWAS with the quantitative trait of the cytokine concentration levels of the patients, also in PLINK. Least-square regression is used

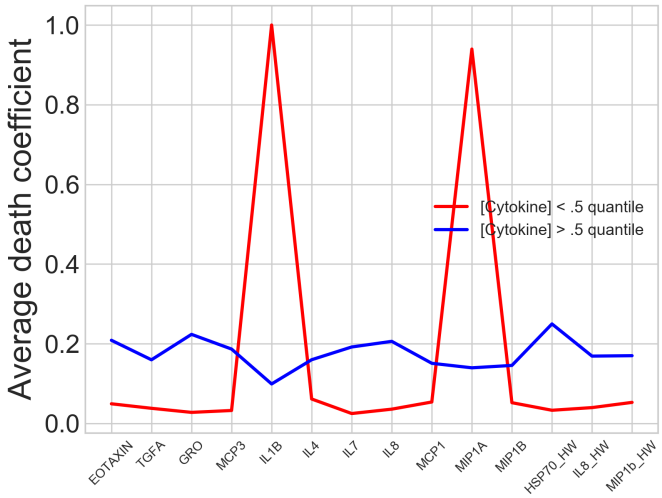

Fig. 7: The death coefficients for each cytokine.

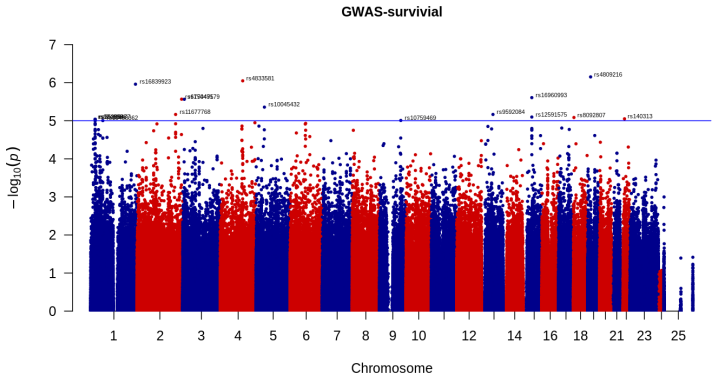

Fig. 8: GWAS of patient survival. The suggestive line is at $-\log_{10}(p) = 5$. As a demonstration the SNPs associated with patient survival with $p < 10^{-5}$ are labelled.

to associate each SNP to the cytokine level, and the analysis is done for each of the 50 different cytokines tested. Remark that this study is also performed with a binary trait of mean-split cytokine level. We find that using the binary trait of cytokine level gives results of lower quality due to increased noise from the imperfect splitting threshold. Therefore, we use here the results from the quantitative trait of continuous cytokine levels. An example of the results, for

the cytokine MCP3, is shown in Figure 9, where again we can identify the SNPs that are highly correlated to the cytokine levels.

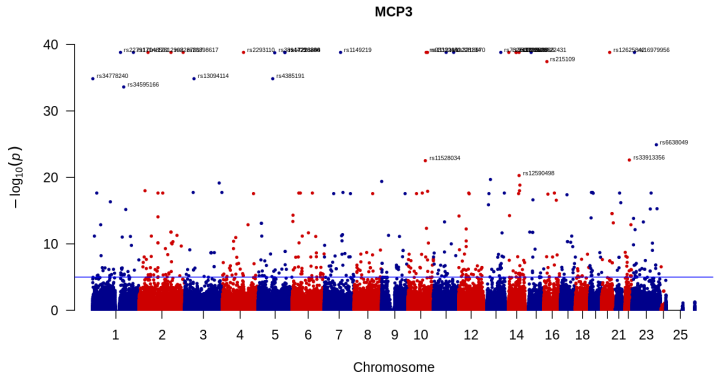

Fig. 9: GWAS of the level of the cytokine MCP3. The suggestive line is at $-\log_{10}(p) = 5$. As a demonstration the SNPs associated with MCP3 level with $p < 10^{-20}$ are labelled.

3.5 Assembling results

The three-way intersection allows us to find five cytokines that are correlated with a SNP located in a gene coding region of the human genome: GRO, MCP3, IL1B, MIP1A and MIP1B. The SNP is located in an intron of a gene coding the inhibitory subunit of a protein. Literature review reveals that the protein is involved in the increase of cytokine levels in the body. Therefore, it is possible that a SNP inhibiting the activity of the subunit will lead to the protein having normal activity and increasing cytokine levels, aggravating symptoms of sepsis.

4 Conclusion and Discussion

With the VASST dataset, we analyzed the correlation between the cytokine concentration levels and patient survival and identified cytokines that were highly related with patient death. To investigate this possible causal relationship, we further performed two GWAS on patient survival and cytokine levels to correlate these traits to SNP signatures of the genome. We have been able to identify several SNPs which are significantly correlated with both patient survival and at least one of the key cytokines related to patient mortality. It is therefore established that, certain cytokines cause patient death. We further clarified a potential genetic mechanism of this causal relationship through a particular

SNP which codes for a protein responsible for regulating cytokine levels. Due to time restriction, we were not able to dive more in depth into this topic. It is recommended that these identified SNPs and the associated key cytokines are studied in more detail for their functions and related signaling pathways, and they can be considered as potential drug targets for new therapy to treat sepsis.

Our dataset is small considering there are only 330 patients with both complete genome data and cytokine measurements. A much larger dataset is desirable to confirm our finding in this study, and for training proper machine learning models to make more accurate predictions. Investigating the connection of the key SNPs and cytokines with other inflammatory diseases than sepsis is also a valuable future study.

5 Supplementary Material

The Github repositories associated with this project are located at https://github.com/BigData2018ubc-stPaul, with three repos: GWAS-cytokines, data-visualization, Notebooks.

Acknowledgement

We sincerely thank Dr. Keith Walley from St. Pauls Hospital and UBC Department of Medicine, for the data provided to us and his valuable discussions. We also thank Dr. Brian Wetton, Aaron Berk, India Heisz and Matteo Lepur at UBC Department of Mathematics. We are grateful for funding and support from the Pacific Institute of Mathematical Sciences.

References

[1] Navaneelan, T., Alam, S., Peters, P. A., and Phillips, O. *Deaths involving sepsis in Canada*. Statistics Canada Catalogue no. 12-591-X, 2016. URL: https://www150.statcan.gc.ca/n1/pub/82-624-x/2016001/article/14308-eng.htm.

[2] Rautanen, A. et al. "Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study". In: *The Lancet Respiratory Medicine* 3.1 (2015), pp. 53–60. URL: https://www.sciencedirect.com/science/article/pii/S2213260014702905.

[3] Russell, J. A. et al. "Vasopressin versus norepinephrine infusion in patients with septic shock". In: *New England Journal of Medicine* 358.9 (2008), pp. 877–887. URL: https://www.nejm.org/doi/full/10.1056/NEJMoa067373.

[4] Purcell, S. et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American Journal of Human Genetics* 81.3 (2007), pp. 559–575. URL: https://www.sciencedirect.com/science/article/pii/S0002929707613524.