2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871716

Towards Remote Continuous Monitoring of Cytokine Release Syndrome

Michael J. Pettinati*, Arad Lajevardi-Khosh, Kuldeep Singh Rajput, Maulik Majmudar, and Nandakumar Selvaraj

Abstract— **Cytokine release syndrome (CRS) is a noninfectious systemic inflammatory response syndrome condition and a principle severe adverse event common in oncology patients treated with immunotherapies. Accurate monitoring and timely prediction of CRS severity remain a challenge. This study presents an XGBoost-based machine learning algorithm for forecasting CRS severity (no CRS, mild– and severe–CRS classes) in the 24 hours following the time of prediction utilizing the common vital signs and Glasgow coma scale (GCS) questionnaire inputs. The CRS algorithm was developed and evaluated on a cohort of patients (n=1,139) surgically treated for neoplasm with no ICD9 codes for infection or sepsis during a collective 9,892 patient-days of monitoring in ICU settings. Different models were trained with unique feature sets to mimic practical monitoring environments where different types of data availability will exist. The CRS models that incorporated all time series features up to the prediction time showcased a micro-average area under curve (AUC) statistic for the receiver operating characteristic curve (ROC) of 0.94 for the 3 classes of CRS grades. Models developed on a second cohort requiring data within the 24 hours preceding prediction time showcased a relatively lower 0.88 micro-average AUROC as these models did not benefit from implicit information in the data availability. Systematic removal of blood pressure and/or GCS inputs revealed significant decreases ($p<0.05$) in model performances that confirm the importance of such features for CRS prediction. Accurate CRS prediction and timely intervention can reverse CRS adverse events and maximize the benefit of immunotherapies in oncology patients.**

I. INTRODUCTION

Infection and injury [1], [2] as well as certain immunotherapies [3] can trigger amplified inflammatory responses of the human body known as Systemic Inflammatory Response Syndrome (SIRS). SIRS is associated with common symptoms and signs including: fever or hypothermia, tachycardia, tachypnea, and leucocytosis or leucopoenia [4]. The definition for SIRS is overly sensitive in acute-care settings [2].

Cytokine Release Syndrome (CRS) is a serious noninfectious SIRS condition that develops rapidly following certain types of immunotherapy and cancer treatments [3], [5]. It is critical to detect the development and severity of CRS and prompt for timely clinical attention and intervention. The healthcare costs and likelihood of clinical complications multiply as patients deteriorate to severe stages of CRS [6]. Almost half of patients in early CAR-T cell treatment trials required intensive care management following infusion [7]. Time is of the essence when CRS can become severe. This necessitates a patient monitoring system and automated artificial intelligence algorithms that can detect CRS onset or early stages of CRS well in advance and allow for early intervention to produce better clinical and therapeutic outcomes with lower healthcare delivery costs.

Machine learning-based solutions can offer earlier insights into SIRS conditions. We have previously showcased how early prediction of sepsis [8], an infection-related SIRS condition, solutions need to consider the environments in which they will be deployed, what data will be available in these environments and what populations they will be applied on. The solutions in this work allow for monitoring across different patient environments including remote settings.

There is a paucity of research regarding the remote continuous monitoring of CRS patients. The task is made harder by the ambiguous application of CRS definition [9]. The malaise, fever, hypoxia, and hypotension that define CRS are common to many conditions, and it often falls on clinicians to make a determination that the condition is CRS. Previous works have used clinical tests to separate noninfectious and infectious inflammatory responses (e.g. [2]).

CRS is often studied as a side effect of cancer treatments, especially immunotherapies [3], [5]–[7], [10]. It is a noninfectious systemic inflammatory response in oncology patients that fit the grading definitions for CRS [9]. This study presents machine learning algorithms to predict the CRS grades in the 24 hours following the prediction time using common vital signs and questionnaires obtained in ICU patients receiving treatments for neoplasms and having no ICD9 codes indicative of infectious disorders or sepsis.

II. METHODOLOGY

A. Dataset and Patient Cohort Extraction

MIMIC3 [11] is an extensive dataset consisting of electronic health records (EHRs) from tens of thousands of intensive care unit (ICU) patient stays. These records include: patients' demographic information, ICD9 discharge diagnoses, events during the patients' stays such as manually recorded vitals, labs, and treatments.

To begin CRS model development, we extracted a cohort of oncology patients from the MIMIC3 dataset that did not have indications of infection nor sepsis to explain an immune response. The designation of a patient as an oncology patient without infection or sepsis was made via ICD9 codes. Rassek et al. [12] grouped neoplasm ICD9 codes, which were used to define oncology patients (ICD9 codes: 140-239). Patients were excluded if they had an infection indicated by having an ICD9 code for an infectious disorder as identified in Rassek et al. [12]. Sepsis patients were also excluded; these patients had the following ICD9 codes: 995.91 (sepsis),

Authors are with Biofourmis Inc, Boston, MA 02110, USA.
(*correspondence e-mail: michael.p@biofourmis.com)

995.92 (severe sepsis), and/or 785.52 (septic shock). For patients with multiple stays in the dataset, only the first was included. Additionally, we excluded patients who were under the age of 18 and patients who died during their admission as these patients are graded differently [9], [13]. After filtering for these inclusion and exclusion criteria, n= 1,139 patients with 9,892 days of patient data were extracted.

Close examination of a subset of patient medical records revealed that a lot of patients were receiving surgical interventions for their cancers. These surgical interventions can result in immune responses, hypoxia and hypotension stemming from a host of issues, including infection, that can be hard to disentangle or conclusively rule out. The shortcomings of this population and future directions are further discussed in the paper's concluding section.

B. Patient Day Labelling

Subsequent to the patient extraction, we labelled the individual patient days in the ICU. The goal of our classifier was to predict the CRS grade of a patient during the following 24 hours. The mildest CRS grades are defined by fever and managed hypoxia and hypotension; the most severe grades require life-saving intervention, e.g. mechanical ventilation (hypoxia) or vasoporessors (hypotension) [9], [13].

The time of a patient's admission was considered time 0. Each 24 hour period following the time of the admission until the patient was discharged for which there was data in the EHR was labelled with an outcome variable. If a patient was treated for a severe condition, with a vasopressor, mechanical ventilation or FiO2 greater than 40%, then the day was labelled as a severe CRS day. If the patient had mild symptoms, fever and hypotension, or less serious treatments, such as O2 supplement with inspired oxygen at less than 40%, but did not receive the more serious treatments, then the day was labelled as mild CRS day. If the patient had none of the treatments or symptoms indicated, then the day was labelled as a no CRS day. This process is summarized in Fig. 1. After checking for these conditions, the 9,892 extracted patients days consisted of 5,215 no CRS days, 1,931 mild CRS days, and 2,746 severe CRS days.

C. Patient Day Cohort Description

There were 9892 patient days extracted via the procedure described above. The majority of these days were demarcated as no CRS days (5215 days). A manual examination of the data revealed almost 80% of the no CRS days had no patient monitoring data in the preceding 24 hours compared to approximately 2% of mild CRS days and approximately 3% of the severe CRS days. We developed a second cohort of patient days which required vitals and/or GCS data within the 24 hours before the prediction. The second cohort consisted of the same 1,139 patients but with 1,134 no CRS days, 1,906 mild CRS days, and 2,667 severe CRS days. See Table 1.

D. Predictive Features

The features fall into two general groups, vitals and glascow coma score (GCS) features. The vitals consisted of heart rate, respiration rate, body temperature, SpO2 levels, systolic and diastolic blood pressure. The GCS features are the verbal, motor and eye opening response scores. The features derived from these measurements include the extreme values from the data in the last 24 hours as well as all days preceding the last 24 hours. These were the only features used in the set of models for the first cohort. The models for the cohort requiring data within 24 hours of prediction contained additional statistical moments and trends from these data. In both groups, various combinations of features were evaluated to determine the resulting performance when only a subset of the vitals or GCS data was used.

TABLE I
SUMMARY OF THE UNIQUE PATIENTS' DAYS EXTRACTED USING THE DESCRIBED PROCEDURE

Patient Cohorts	Unique Patient Stays	No CRS Days	Mild CRS Days	Severe CRS Day	Total Days
Cohort 1	1139	5215	1931	2746	9892
Cohort 2	1139	1134	1906	2667	5707

E. XGBoost Models

The output of the models was the probability of the patient being no CRS, mild CRS, or severe CRS during the 24 hours following the prediction time using all of the patient's data up until the time of prediction. The probabilities sum to one. The models were XGBoost models[1]; scripts to train and test the models were written in Python.

The basis of XGBoost models are decision trees. Decision trees split training data into subgroups with single classes being over represented based on informative feature values. For a given decision tree, a test point will follow a path along the tree based on its features' values and the learned splits to a leaf node; the probability of that test sample belonging to a certain class depends on the proportion of examples of that class at that leaf node. XGBoost devises decision tree models using the training data. Each decision tree gets training examples wrong. Additional trees can be added that accurately predict examples previously missed. Trees can be added and weighted depending upon their performances on the training examples. The prediction for a test point is the weighted sum of the predictions of the trees.

F. Model Evaluation

There were six models developed and tested as a part of this work, two sets of three models each. The first set of three models were the models that did not require vitals or GCS data within the 24 hours leading up to the time of prediction. The second set of three models did require vitals and/or GCS data within the 24 hours leading up to the time of prediction. Each of the sets had one model that incorporated features from all nine data types discussed above (vitals and GCS), a second model that incorporated only the vital sign data types discussed above, and a final model that incorporated

[1]https://xgboost.readthedocs.io/en/stable/python/index.html

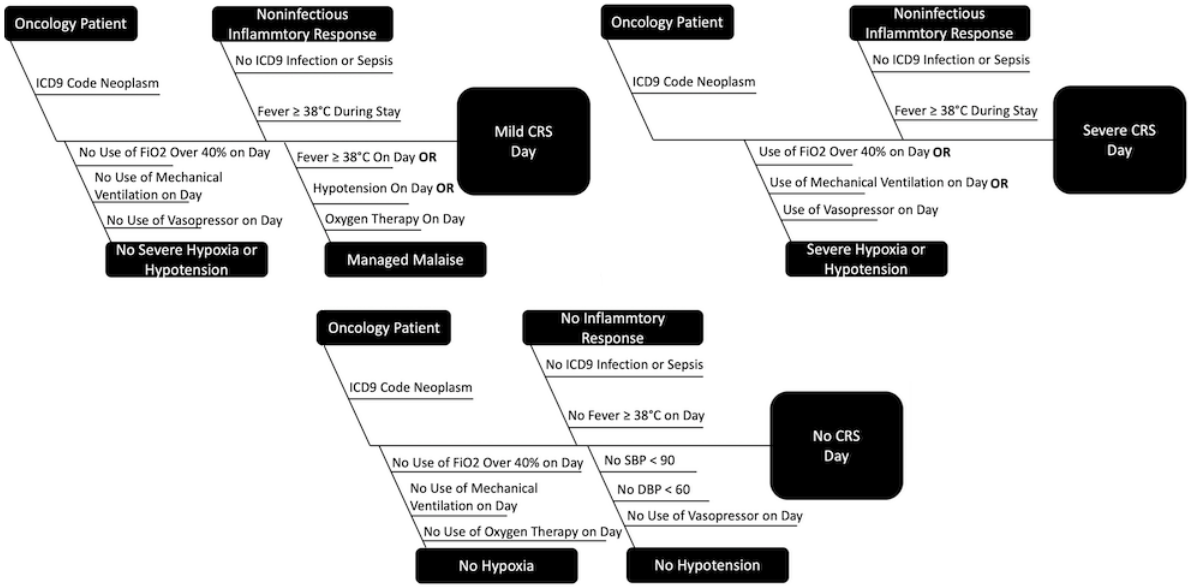

Fig. 1. Summary of Rules Used to Extract Patients' Days From MIMIC3 to Include in Our CRS Cohort

TABLE II
AUROC STATISTICS - EACH MODEL SEPARATING EACH CRS GRADE

Patient Cohort	Features Groups Included	AUROC No CRS (± SD)	AUROC Mild CRS (± SD)	AUROC Severe CRS (± SD)
Cohort 1	All Features	0.96 (0.00)	0.89 (0.01)	0.93 (0.01)
	All Vitals	0.96 (0.00)	0.85 (0.01)	0.89 (0.00)
	HR, RR, SpO2, Temp	0.95 (0.00)	0.83 (0.02)	0.87 (0.01)
Cohort 2	All Features	0.89 (0.01)	0.80 (0.02)	0.89 (0.02)
	All Vitals	0.85 (0.01)	0.72 (0.02)	0.79 (0.01)
	HR, RR, SpO2, Temp	0.82 (0.01)	0.68 (0.02)	0.76 (0.02)

only heart rate, respiration rate, SpO2 and body temperature features. These models became increasingly more suited for remote and continuous patient monitoring. The data types are accurately measurable with common wearable devices.

Five-fold cross validation was used to validate each of the six models. Days were split randomly amongst the five folds while balancing no CRS, mild CRS, and severe CRS grades. The splits were not even because patient stays were contained to a single fold to ensure there was not any data leakage. A single patient's day data was not incorporated for both training and testing. Performance metrics were calculated as the mean $\pm$ standard deviation of the AUROC.

III. RESULTS

The results for our six models are summarized in Table 2. There is clear evidence that the no, mild, and severe CRS classes are well separated regardless of the data used when using the patient day cohort that does not require vitals or GCS data within the 24hrs before the time of prediction. The models separating severe CRS from non-severe CRS (no CRS and mild CRS) had a minimum AUROC of 0.87 and separating no CRS from CRS (mild or severe) had a minimum AUROC of 0.95. The low standard deviations in performance across the five folds for all models shows the consistency with which these classes are separable. The ROC for our model using all features is shown in Fig 2.

There is evidence of GCS scores and BP data being predictive of the CRS grades of patients in the following 24 hours. We tested for statistically significant ($p < .05$) differences between the average AUROCs when using all features, when removing GCS, and when removing GCS and BP features for models separating severe CRS from non-severe CRS, mild CRS from the other two conditions, and no CRS from CRS. If assumptions of normality held, we used a Repeated Analysis of Variance (ANOVA) with paired posthoc t tests if significant difference was found in the ANOVA. If normality did not hold, we used a Friedman test with Nemenyi post hoc testing. In all three model types, there was a significant difference between groups. All group pairs were significantly different when separating severe CRS from non-severe CRS and no CRS from CRS.

To address the issue of missing data being predictive, we consider the models that predicted CRS grades on days for which there was vitals and/or GCS data in the preceding 24 hours. These models are intended to be more practical as they attempt to not exploit the different monitoring levels.

These models show predictive value in discriminating the three classes in the 24 hours following prediction with a minimum AUROC of 0.68. We note an AUROC of 0.76 separating severe CRS and non-severe CRS using only 4 vital signs. AUROC statistics are highest for no CRS vs. CRS. See

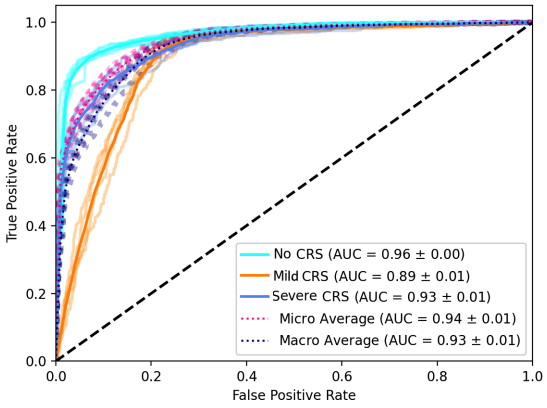

Fig. 2. All Feature ROC Curves - Cohort 1

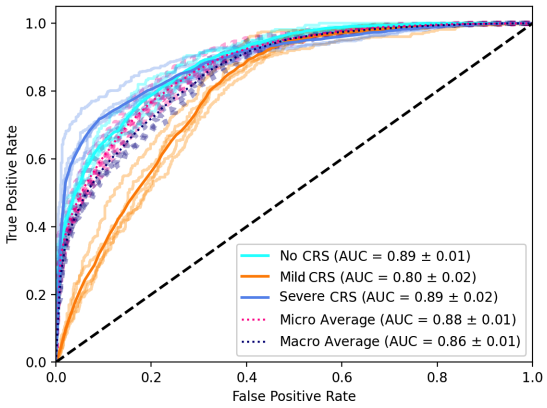

Fig. 3. All Feature ROC Curves - Cohort 2

Table 2. ROCs for the All Features model are shown in Fig. 3. Again performance is consistent across the five folds.

With these three model types, we looked for statistically significant differences between the models incorporating different feature groups using the same procedure as above. There was a statistically significant difference ($p < .05$) between all pairings of groups for all model types. The all feature model was the highest performing. The model using all vital signs performed better than the model using HR, RR, SpO2 and body temperature. The importance of GCS and BP is supported by looking at the average feature importance from the five-fold cross validation for the all feature model and the vitals-only model. The top ten features are shown in Figs 4 and 5. Features derived from GCS verbal and eye scores appear in Fig 4. SBP-related features appear in Fig. 5. These data sources are important predictors of CRS grades in the following 24 hours.

IV. Discussion

This paper showcases the feasibility of developing predictive models that allow for the continuous and remote monitoring of patients at risk for CRS and assessing CRS

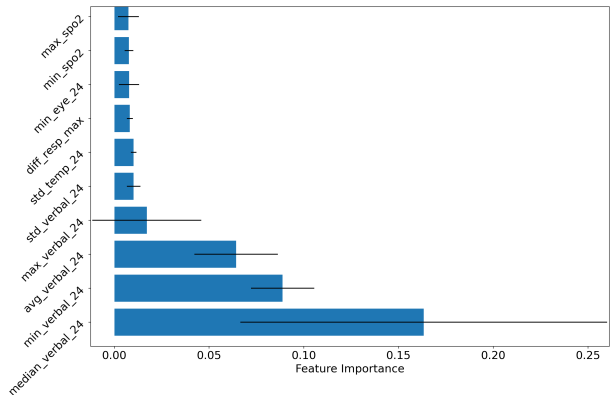

Fig. 4. All Feature Model - XGBoost Feature Importance

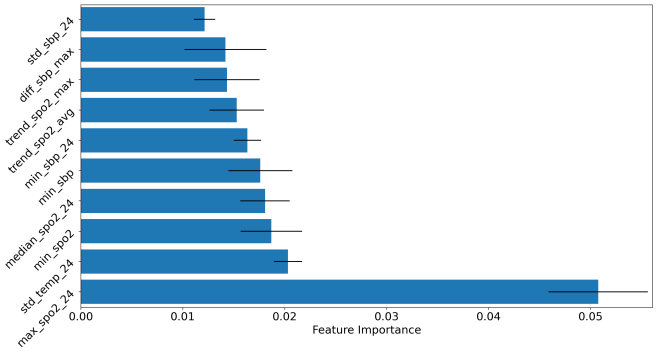

Fig. 5. Vitals-Only Model - XGBoost Feature Importance

severity on a more dynamic daily basis with promising accuracy and precision.

CRS is a noninfectious SIRS condition that is common in oncology, and, if it progresses to severe grades, it is dangerous and costly for the affected patient [3], [5], [6]. The study extracted a cohort of oncology patients who could be at risk for CRS from the MIMIC3 dataset. These patients did not have ICD9 codes indicating treatment for infection or sepsis, but many of them developed fever along with signs of hypotension and/or hypoxia. These patients' days were labelled with CRS grades following the latest guidelines [9]. The extracted data was analyzed to understand the shortcomings of existing data for understanding and predicting this condition. Models were constructed to predict the patients' CRS grade in the following 24 hours. The different models incorporated data from different types of environments to allow for predictions to be done most accurately depending on where a patient might be monitored.

As shown in Table 3, the data extracted from MIMIC3 had patients, in general, staying at the same CRS grade day-to-day or getting better. Over 86% of no CRS patient days stayed the same one day to the next day (159/185). Patients with mild CRS often improved (806/2144) or stayed the same (1215/2144). Severe CRS grade days stayed the same into the next day 75.1% (2537/3378) of the time. This makes sense because the MIMIC3 dataset contains patient records from ICUs. Closer examination of these

TABLE III
TRANSITION MATRIX FOR CRS GRADES - SHOWING PATIENT GRADE ONE DAY TO THE NEXT - COHORT 2

		Previous Day's CRS Grade			
		No CRS	Mild CRS	Severe CRS	Total
Current Day's CRS Grade	No CRS	159	806	169	1134 (19.9 %)
	Mild CRS	19	1215	672	1906 (33.4 %)
	Severe CRS	7	123	2537	2667 (46.7 %)
	Total	185	2144	3378	5707

patients EHRs showed that many of these patients were receiving surgical interventions for their cancers. Patients were receiving treatment the entirety of the data and closely monitored by clinical professionals.

There is a need for continued model development and validation in novel datasets as well as continuous and remote patient monitoring. It will be important in subsequent iterations of model development to have a more continuous view of the patients data. This will require continuous and remote monitoring of oncology patients who are at risk for CRS. Critical data will include information about: the patient's type of cancer, the tumor burden, precise information about the treatments being employed and a continuous picture of the patient's vital signs. Disease burden and treatment type make a substantial difference in the rate and severity of CRS [14]. Many of the patients in the MIMIC3 dataset were receiving surgical interventions and exhibiting CRS like exaggerated immune responses. CRS is incredibly common and needs to be better understood particularly for the emerging novel immunotherapy treatments. Subsequent iterations of model development will focus on patients receiving these treatments.

There are easily obtainable vital sign-based features that allow for remote patient monitoring of CRS. The models developed in this work showed high predictive value for CRS grades for the 24 hours following the time of prediction. We showed an AUROC statistic of of 0.76 for identifying when a patient was going to have severe grade CRS when only incorporating HR, RR, SpO2, and body temperature-related features. This is a promising first step that supports patients being continuously and remotely monitored for severe CRS.

Clinical parameters beyond basic vital signs, that can be incorporated episodically, will enhance model accuracy. GCS and blood pressure enhanced the predictive performance. In all six of our model types, there was a significant difference ($p < .05$) in our models' discriminatory accuracy depending on the features incorporated. The discrimination was the highest using all features, vitals and GCS scores, and higher using the full complement of vitals than HR, RR, SpO2 and body temperature alone. Further, when looking at the top 10 features with respect to feature importance for our most practical models, we saw verbal GCS and SBP appear repeatedly. Models can benefit from periodically attending caregivers by incorporating this clinical information.

There are additional applications to target with our models and ways in which the models could be extended. The patients incorporated were oncology patients who were not treated for infection or sepsis. There are other patients within the dataset who could have been included as no CRS examples to increase the sample of no CRS days.

SIRS conditions can be inseparable with respect to clinical parameters. Machine learning-based solutions could assess SIRS severity more generally. Solutions should also target separating SIRS conditions without using clinical tests. These tasks will be undertaken as we further develop remote and continuous monitoring for CRS and SIRS patients.

REFERENCES

[1] R. A. Balk, "Systemic inflammatory response syndrome (sirs) where did it come from and is it still relevant today?" *Virulence*, vol. 5, no. 1, pp. 20–26, 2014.

[2] I. T. Schrijver, H. Kemperman, M. Roest, J. Kesecioglu, and D. W. de Lange, "Myeloperoxidase can differentiate between sepsis and non-infectious sirs and predicts mortality in intensive care patients with sirs," *Intensive care medicine experimental*, vol. 5, no. 1, pp. 1–9, 2017.

[3] A. Shimabukuro-Vornhagen, P. Gödel, M. Subklewe, H. J. Stemmler, H. A. Schlößer, M. Schlaak, M. Kochanek, B. Böll, and M. S. von Bergwelt-Baildon, "Cytokine release syndrome," *Journal for immunotherapy of cancer*, vol. 6, no. 1, pp. 1–14, 2018.

[4] P. Comstedt, M. Storgaard, and A. T. Lassen, "The systemic inflammatory response syndrome (sirs) in acutely hospitalised medical patients: a cohort study," *Scandinavian journal of trauma, resuscitation and emergency medicine*, vol. 17, no. 1, pp. 1–6, 2009.

[5] "Nci dictionary of cancer terms," National Cancer Institute. [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cytokine-release-syndrome

[6] J. S. Abramson, T. Siddiqi, J. Garcia, C. Dehner, Y. Kim, A. Nguyen, S. Snyder, N. McGarvey, M. Gitlin, C. Pelletier *et al.*, "Cytokine release syndrome and neurological event costs in lisocabtagene maraleucel–treated patients in the transcend nhl 001 trial," *Blood Advances*, vol. 5, no. 6, pp. 1695–1705, 2021.

[7] B. Santomasso, C. Bachier, J. Westin, K. Rezvani, and E. J. Shpall, "The other side of car t-cell therapy: cytokine release syndrome, neurologic toxicity, and financial burden," *American Society of Clinical Oncology Educational Book*, vol. 39, pp. 433–444, 2019.

[8] M. J. Pettinati, G. Chen, K. S. Rajput, and N. Selvaraj, "Practical machine learning-based sepsis prediction," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 4986–4991.

[9] D. W. Lee, B. D. Santomasso, F. L. Locke, A. Ghobadi, C. J. Turtle, J. N. Brudno, M. V. Maus, J. H. Park, E. Mead, S. Pavletic *et al.*, "Astct consensus grading for cytokine release syndrome and neurologic toxicity associated with immune effector cells," *Biology of Blood and Marrow Transplantation*, vol. 25, no. 4, pp. 625–638, 2019.

[10] M. S. Thakar, T. J. Kearl, and S. Malarkannan, "Controlling cytokine release syndrome to harness the full potential of car-based cellular therapy," *Frontiers in oncology*, vol. 9, p. 1529, 2020.

[11] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[12] S. Rassekh, M. Lorenzi, L. Lee, S. Devji, M. McBride, and K. Goddard, "Reclassification of icd-9 codes into meaningful categories for oncology survivorship research," *Journal of cancer epidemiology*, vol. 2010, 2010.

[13] "Common terminology criteria for adverse events (ctcae) version 5.0. 2017," US Department of Health and Human Services and others, 2017. [Online]. Available: ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/CTCAE_v5_Quick_Reference_5x7.pdf

[14] R. Hong, H. Zhao, Y. Wang, Y. Chen, H. Cai, Y. Hu, G. Wei, and H. Huang, "Clinical characterization and risk factors associated with cytokine release syndrome induced by covid-19 and chimeric antigen receptor t-cell therapy," *Bone marrow transplantation*, vol. 56, no. 3, pp. 570–580, 2021.