NGS-based phenotype determination in sepsis patients with machine learning methods

Master thesis - Exposé

Valentin Noske

26th of October 2023

First supervisor: Prof. Dr. Ulf Leser
Second supervisor: tba.

Institute of Computer Science

Humboldt University of Berlin

In cooperation with the Fraunhofer IGB, Stuttgart

Introduction

According to the third international consensus definition of sepsis, published in 2016, sepsis is a state of a "life-threatening organ dysfunction caused by a dysregulated host response to infection" [1]. Although sepsis is a common condition - between 1979 and 2015, the number of cases was estimated at 50 million per year [2] - the mortality through sepsis is still high. The mortality rate was 17% for sepsis and 26% for severe sepsis during these years [2]. Since sepsis is considered a medical emergency, treatment, and resuscitation, restoring life or consciousness, should start as soon as the diagnosis is made [3]. As a pathogen induces sepsis, often bacteria, early antibiotic treatment is common as part of the initial intervention. However, this also bears the risk of inappropriate treatment linked to higher mortality [4].

In this regard, there are two main challenges: on the one hand, the effective and rapid diagnosis of sepsis that also examines the trigger of sepsis, and on the other hand, the assessment of the current physical state of the sepsis patient. The Fraunhofer IGB's In-vitro Diagnostics research group, where the master thesis research project will be conducted, has worked intensively on the first topic in the past and has developed a pipeline for the diagnosis of sepsis pathogens based on Next-generation-sequencing (NGS) data [5, 6, 7]. The focus of this master thesis will be on the second topic. In the context of determining a patient's current state, we speak of their phenotype. "The word 'phenotype' refers to some deviation from normal morphology, physiology, or behavior" [8]. Phenotypes to be studied include, for example, those of patients with a lower chance of survival or ineffective treatment.

To characterize phenotypes, mainly NGS data of human circulating cell-free DNA (cfDNA) will be used. cfDNA is defined as "DNA liberated from the confinement of cells into any type of extracellular space" [9]. cfDNA has become a useful biomarker in the last two decades [9]. The extraction of cfDNA is minimally invasive as only blood samples are taken [9]. Furthermore, it can serve as a "source of diverse biological and pathological information" [9]. Since human cfDNA is the product of necrosis or apoptosis [10], we can derive information about which cells are going into cell death from features such as fragment length, amount, or methylation patterns [11]. Thereby, we also derive information about the current condition of the patient. cfDNA is particularly interesting in sepsis, as the heterogeneity of patients makes characterization based on classical biomarkers, for example, blood levels or disease-specific assessment scores, difficult [10].

To effectively use the information contained in the cfDNA features, an algorithm is needed to identify highly informative features and predict the phenotype of each patient with high accuracy. Because it is not yet fully understood which attributes of the cfDNA might be the most promising features [10], an algorithm that makes its own selection is preferable. Machine learning (ML) can provide this functionality. "ML models are capable of producing knowledge about domain relationships contained in data, often referred to as interpretations" [12]. Fleuren et al. have shown in their systematic review and meta-analysis of the performance of ML in sepsis diagnostics that the accuracy of the algorithms is often better than that of traditional methods, even if ML models lack interpretability [13]. A lack of interpretability means that it is often not understood how an ML model arrives at a particular solution. As a result, it is not clear whether this judgment can be trusted. The latter is seen as a problem, especially in clinical practice [13]. Although this study does not fully address the research topic, as it deals with sepsis diagnostics, the results can be transferred to ML applications in sepsis phenotyping.

Formulation of objectives

In my master thesis, I will examine cfDNA data of sepsis patients to find features that yield information on the patient's phenotype. The patient's risk of death will be the initial phenotype to be studied, as many predictors for this case can be used as a performance comparison. Different ML algorithms will be used to predict phenotypes since no optimal solution is yet known. A value of 0.9 for the area under the receiver operating characteristic (ROC) curve (AUC) is set as the target for the diagnostic ability of any predictor being developed, bearing in mind that the priority of sensitivity or specificity may vary depending on the context. This is to ensure comparability with state-of-the-art algorithms. Higher targets cannot be set due to the limited size of the data set. Ultimately, the interpretability of each predictor is made a criterion for its evaluation. The contribution of the features to decision-making is analyzed, examining whether these effects can be explained by the literature. The key research question will be how accurately and comprehensibly an ML model can be created for phenotyping sepsis patients based on NGS data from cfDNA.

State of the art

There are several simple predictors of mortality in sepsis patients in the literature, with variations in the number of days mortality was measured. "Simple" means in this context that no ML algorithm is used. In general, two groups of predictors can be differentiated. The first group are scores calculated based on the presence of symptoms, the exceeding of specific levels of biomarkers, or certain biometrical features like old age. If the score crosses a threshold, mortality is predicted within the given time frame. Examples for the first group are the SIRS-criteria (AUC 0.76) [14], the qSOFA score (AUC 0.81) [14] and SOFA score (AUC 0.75) [15], and the SMRS score (AUC 0.789) [16]. It is important to emphasize that these scores were not all invented to characterize sepsis disease. The other group contains important biomarkers that are part of the immune response. If the value of the specific biomarker is outside the normal range, this will be associated with mortality. Examples for this group of predictors are initial blood lactate (AUC 0.664) [17], procalcitonin (PCT) (AUC 0.76) [18], and interleukin 6 (IL-6) (AUC 0.785) [19] levels. Measuring blood lactate levels is one of the simplest approaches. Therefore, an AUC value of 0.664 should be the baseline for any predictor.
ML has already been used to predict mortality in sepsis patients. However, the algorithms here are also mainly based on biometric data, symptom and disease data, and biomarker levels. The time frames in which mortality is predicted are not always identical. Hu et al. trained seven ML algorithms on the sepsis data of the Medical Information Mart for Intensive Care (MIMIC-III) [20], which is a large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital. Their best-performing algorithm is an eXtreme Gradient Boosting (XGBoost) algorithm, which achieves an AUC value of 0.884 [20]. Karlsson et al. used a Balanced Random Forest Classifier and 91 variables reflecting emergency department (ED) presentation to predict 7- and 30-day mortality and reached an AUC value of 0.83, respectively 0.8. The research team led by Park utilized data from the US Nationwide Inpatient Sample (NIS), the largest all-payer inpatient care database in the United States, to train six commonly employed ML algorithms [22]. The neural network achieved the best performance with an AUC value of 0.893 [22]. Wernly et al. took a different approach by using only arterial blood gases (ABG) to predict mortality, as these are routinely determined even in heavily loaded ICUs [23]. This design ensures applicability in all ICUs, not only in the particularly well-equipped ones [23]. Their long-short term memory (LSTM) network, a neural network, could reach an AUC value of 0.93 [23]. Of course, we cannot discuss all existing algorithms here. However, referring to the systematic review by Wu et al., which generally deals with the use of artificial intelligence in sepsis, it can be said that deep learning, particularly, has shown success in mortality prediction in sepsis [24].
The use of human cfDNA analysis is already more widespread in cancer diagnostics. There, it is used as an input to cancer diagnostic tests based on ML. A primary goal is to develop a noninvasive tool that can diagnose cancer and its type, a so-called multi-cancer early detection test (MCED) [25]. Early and simple diagnoses of the kind of cancer would allow more precise treatment of patients and are, therefore, a high priority. The analysis of human cfDNA is promising because cancer cells shed special cfDNA, cell-free circulating tumor DNA (ctDNA) [26]. To identify and interpret the ctDNA, ML algorithms are used. Christiano et al. focused on fragmentation profiles of cfDNA to find significant differences between cancer and non-cancer patients for different cancer types using a gradient tree boosting machine learning model [27]. The fragmentation profile describes, among other things, the length of the fragments and their genomic position. In another study, Jamshidi et al. compared a variety of ML classifiers for MCED, each using a distinct feature of cfDNA [25]. Thereby, cfDNA methylation offered the best results for MCED [25]. Although research in this field mainly focuses on diagnostics, many methodological developments exist for using cfDNA sequencing data in ML models. These may guide the phenotype prediction of sepsis patients with ML models. The work of Jamshidi et al. is most relevant for building an ML model with cfDNA data, as it provides many clues as to what features can be extracted from it. Regarding the use of cfDNA in sepsis, the In-vitro diagnostics research group of the Fraunhofer IGB, Stuttgart has developed an NGS-based "Sepsis Identifying Quantifier" (SIQ) that uses sequences of pathogenic cfDNA to diagnose sepsis [5] respectively septic shock [6]. Besides, the performance of the first algorithm has been shown in a multicentre study involving internal medicine and surgical intensive care units (ICU) in hospitals with the highest possible level of care [7]. In addition, levels of cfDNA have already been used to predict sepsis mortality in some studies (AUC 0.79 [28], AUC 0.97 [10]).

Way of proceeding

An ML model will be created to predict the mortality of sepsis patients. Then, if the applied methods of trait extraction and engineering work well in combination with the ML model, an extension of the model is worked on, or a similar model is trained for other phenotypes and diagnostic questions.

The data used to build an ML model originate from the multicenter, non-interventional, prospective clinical study to evaluate the utility of the SIQ score in sepsis pathogen diagnosis [7]. Fraunhofer IGB participated in this study as the central diagnostic collaborator [7]. This study includes 500 patients with suspected or proven sepsis based on sepsis-3 criteria [30]. A plethora of clinical parameters and metadata were collected for each patient in the study. In addition, there is NGS data of cfDNA based on plasma samples and two sets of blood cultures (2x aerobic / 2x anaerobic) per patient, all taken at baseline (admission to ICU) and 72 hours afterward [30]. All patients in this study are of legal age and have consented to participate in the study [30]. The data was collected in seventeen clinics across Germany [30]. The final outcome assessment of the patients was conducted after 28 days [30].

The most time-consuming step in building an ML model to predict phenotypes of sepsis patients is expected to be feature extraction and selection. The paper of Jamshidi et al. [25] will be used as a methodological blueprint. Features that are mentioned in this paper are whole genome (WG) methylation, single nucleotide variant (SNV) with and without white blood cell (WBC) background removal, somatic copy number alteration (SCNA), fragment endpoints, fragment lengths and allelic imbalance [25]. Since neither methylation data nor white blood cell reference data are available in the dataset for this research, the focus will be on SNVs, SCNAs, fragment endpoints, and fragment lengths. In addition, fragment end motifs and possibly other features will be considered depending on the availability of other literature. Unlike in their paper, the objective will not be to focus on one type of feature only but to combine several into one ML model. Nevertheless, it is not certain that all of these traits are important enough to be included, as their importance has only been demonstrated in cancer. Furthermore, the model design will consider clinical and NGS-based pathogen diagnostic data.

An SNV is a “single nucleotide (nt) substitution” [29]. A subset of these mutations can influence multifactorial diseases or traits and is a potentially interesting feature for the ML model. To use SNV information, polymerase chain reaction (PCR) duplicates of the sequences of DNA fragments are first grouped to be summarized using mean collapsed coverage [25]. SNVs in the sequences are then identified by a Bruijn graph assembler [25]. A noise model gives the SNVs a score representing their distinctiveness from the reference cohort [25]. SNVs with too low a score or similarity to a DNA damage artifact are filtered out [25]. Based on these SNVs, a fixed-length vector is created for each patient by assigning every gene the maximum allele fraction of any SNV [25].

Copy number variations (CNV) are genome sections repeated several times. The number of repeats depends on the individual [31]. SCNAs are “somatic changes to chromosome structure that result in gain or loss in copies of sections of DNA” [32]. These changes are not hereditary but happen during cell division and can also affect regions crucial in other cellular processes [32]. To use SCNAs for ML, the DNA reads are divided into bins of 100 kb (kilobases) [25]. This number of reads per bin is then subtracted from a baseline of the reference cohort [25]. These numbers are finally corrected for GC bias [25], the human genome has a higher GC content than would be possible by random chance [33], and systemic effects [25].

The nuclease activity in the cells mainly determines the fragment endpoints and lengths because the nucleases are responsible for the generation and clearance of cfDNA [34]. As the nuclease activity depends on the current state of the cells, it could be a good indicator of the phenotype of a patient. To extract fragment endpoint information, two count vectors are used for each test and reference cohort, each corresponding to the genome length [25]. Significant endpoint positions are then identified per chromosome using hierarchical Bayesian modeling [25]. Normalized numbers of fragments ending in the selected positions are then used as input vectors [25]. For fragment length analysis, the fragment length distribution is calculated for each patient in the test and reference cohort and summarized using the mean density per fragment length across participants per cohort [25]. The values obtained are then normalized, and a fragment length probability is calculated based on the test and control distribution and controlled for GC bias [25]. To obtain an input vector, the genome is divided into 100kb regions, which are then given a score based on the average probability of the fragments being part of the test cohort [25].

Since there are many potential features to choose from, an embedded supervised feature selection strategy such as Least Absolute Shrinkage and Selection Operator (LASSO) can be used to find the most appropriate feature selection regarding accuracy and interpretability. A tool like the “glmnet” package in R is a candidate for use. As already described, it will not be possible to precisely define the properties of the ML model to be developed. Moreover, this may vary depending on the input the ML model is to process and the output it is to produce. For example, the number of phenotypes to be predicted is not yet fixed. The literature discussed can offer a selection proven effective in exemplary cases. In the work of Jamshidi et al., kernel and elastic binomial logistic regressions, a convolutional neural network, and an XGBoost classifier performed best [25]. Other algorithms that are used are Lasso regression, random forest [22], or support vector machine (SVM) [24]. Supervised learning and reinforcement learning are prevalent. Unsupervised learning is referred to only once with a link to the K-means algorithm [24]. Many papers conclude that deep neural networks or XGBoost classifiers provide the best results. [20, 23, 24, 25]. Neural networks, however, are out of scope with the cohort size of the study to be used.

To assess ML models, the AUC value will be most important as it is “widely recognized as the measure of a diagnostic test’s discriminatory power” [35]. The AUC of the Receiver ROC (Receiver Operating Characteristic) curve is a simple way to compare classifiers capturing both sensitivity (true positive rate) and specificity (true negative rate) [35]. Furthermore, the AUC is calculated in each study, so we achieve a comparison to the current state of the art [17, 20, 21, 22, 23]. But also, the AUC value has “drawbacks, including the decoupling from the class skew” [36], meaning the asymmetry observed in a probability distribution. Therefore, one can also include the AUC of the precision-recall curve (PR) in the evaluation of the model. The AUC-PR combines the precision (positive predictive value) and the recall (true positive rate) of the classifier [36].

Additionally, the ML model shall be judged based on its interpretability and comprehensibility. The value of an ML algorithm in medicine is only given if its functioning can be sufficiently explained. A way to apprehend ML models is SHAP (SHapley Additive exPlanations) [37]. SHAP assigns an importance value to each feature a model uses for a particular prediction [37]. One can also look at the average importance or examine which values of the selected feature have the most influence [20]. By this means, it is easier to understand why the algorithms come to a particular conclusion. SHAP is not the only instrument that was developed for this purpose. Arguments in favor of its usefulness are that SHAP has already been used successfully in several papers on sepsis mortality prediction [20, 22]. Moreover, in their paper, Lundberg and Lee describe SHAP as more intuitive than other comparable instruments [37]. Furthermore, since the interpretability of an ML model does not necessarily lead to reliance, the observed effects of the features on the model need to be justified by a literature review.

References

[1] Singer, Mervyn, et al. ”The third international consensus definitions for sepsis and septic shock (Sepsis-3).” Jama 315.8 (2016): 801-810.

[2] Fleischmann, Carolin, et al. ”Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations.” American journal of respiratory and critical care medicine 193.3 (2016): 259-272.

[3] Evans, Laura, et al. ”Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021.” Critical care medicine 49.11 (2021): e1063-e1143.

[4] Paul, Mical, et al. ”Systematic review and meta-analysis of the efficacy of appropriate empiric antibiotic therapy for sepsis.” Antimicrobial agents and chemotherapy 54.11 (2010): 4851-4863.

[5] Grumaz, Silke, et al. ”Next-generation sequencing diagnostics of bacteremia in septic patients.” Genome medicine 8.1 (2016): 1-13.

[6] Grumaz, Silke, et al. ”Enhanced performance of next-generation sequencing diagnostics compared with standard of care microbiological diagnostics in patients suffering from septic shock.” Critical Care Medicine 47.5 (2019): e394.

[7] Brenner, T., and TIFOnet Critical Care Trials Group. ”Next Generation Sequencing” zur Diagnostik der Bakteriämie bei Sepsis–Next GeneSiS-Trial.” Der Anaesthesist 69 (2020): 593-595.

[8] Robinson, Peter N. ”Deep phenotyping for precision medicine.” Human mutation 33.5 (2012): 777-780.

[9] Bronkhorst, Abel J., et al. ”Towards systematic nomenclature for cell-free DNA.” Human genetics 140 (2021): 565-578.

[10] Dwivedi, Dhruva J., et al. ”Prognostic utility and characterization of cell-free DNA in patients with severe sepsis.” Critical care 16 (2012): 1-11.

[11] Volik, Stanislav, et al. ”Cell-free DNA (cfDNA): clinical significance and utility in cancer shaped by emerging technologies.” Molecular Cancer Research 14.10 (2016): 898-908.

[12] Murdoch, W. James, et al. ”Interpretable machine learning: definitions, methods, and applications.” arXiv preprint arXiv:1901.04592 (2019).

[13] Fleuren, Lucas M., et al. ”Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy.” Intensive care medicine 46 (2020): 383-400.

[14] Serafim, Rodrigo, et al. ”A comparison of the quick-SOFA and systemic inflammatory response syndrome criteria for the diagnosis of sepsis and prediction of mortality: a systematic review and meta-analysis.” Chest 153.3 (2018): 646-655.

[15] Jones, Alan E., Stephen Trzeciak, and Jeffrey A. Kline. ”The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation.” Critical care medicine 37.5 (2009): 1649.

[16] Zhang, Kai, et al. ”Development and validation of a sepsis mortality risk score for sepsis-3 patients in intensive care unit.” Frontiers in Medicine 7 (2021): 609769.

[17] Rabello Filho, Roberto, et al. ”Blood lactate levels cutoff and mortality prediction in sepsis—time for a reappraisal? A retrospective cohort study.” Shock (Augusta, Ga.) 46.5 (2016): 480.

[18] Lee, Chien-Chang, et al. ”Prognostic value of mortality in emergency department sepsis score, procalcitonin, and C-reactive protein in patients with sepsis at the emergency department.” Shock 29.3 (2008): 322-327.

[19] Liu, Junkun, et al. ”Mortality prediction using a novel combination of biomarkers in the first day of sepsis in intensive care units.” Scientific Reports 11.1 (2021): 1275.

[20] Hu, Chang, et al. ”Interpretable machine learning for early prediction of prognosis in sepsis: a discovery and validation study.” Infectious Diseases and Therapy 11.3 (2022): 1117-1132.

[21] Karlsson, Adam, et al. ”Predicting mortality among septic patients presenting to the emergency department–a cross sectional analysis using machine learning.” BMC Emergency Medicine 21.1 (2021): 1-8.

[22] Park, James Yeongjun, et al. ”Predicting sepsis mortality in a population-based national database: Machine learning approach.” Journal of Medical Internet Research 24.4 (2022): e29982.

[23] Wernly, Bernhard, et al. ”Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation.” International journal of medical informatics 145 (2021): 104312.

[24] Wu, Miao, et al. ”Artificial intelligence for clinical decision support in sepsis.” Frontiers in Medicine 8 (2021): 665464.

[25] Jamshidi, Arash, et al. ”Evaluation of cell-free DNA approaches for multi-cancer early detection.” Cancer Cell 40.12 (2022): 1537-1549.

[26] Heitzer, Ellen, Peter Ulz, and Jochen B. Geigl. ”Circulating tumor DNA as a liquid biopsy for cancer.” Clinical chemistry 61.1 (2015): 112-123.

[27] Cristiano, Stephen, et al. ”Genome-wide cell-free DNA fragmentation in patients with cancer.” Nature 570.7761 (2019): 385-389.

[28] Avriel, Avital, et al. ”Admission cell free DNA levels predict 28-day mortality in patients with severe sepsis in intensive care.” *PloS one* 9.6 (2014): e100514.

[29] Wright, Alan F. ”Genetic variation: polymorphisms and mutations.” e LS (2001).

[30] Brenner, Thorsten. “Next-Generation Sequencing Diagnostics of Bacteremia in Sepsis (NextGeneSiS).” CTG Labs - NCBI, University Hospital Heidelberg, 18 May 2022, clinicaltrials.gov/study/NCT03356249.

[31] McCarroll, Steven A., and David M. Altshuler. ”Copy-number variation and association studies of human disease.” Nature genetics 39.Suppl 7 (2007): S37-S42.

[32] Qiagen. (n.d.). Somatic mutations and copy number changes in cancer: finding the right targets - QIAGEN. https://www.qiagen.com/gb/spotlight-pages/newsletters-and-magazines/articles/reviews-online-copy-number-alteration/

[33] US DOE Joint Genome Institute: Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4, et al. ”Initial sequencing and analysis of the human genome.” nature 409.6822 (2001): 860-921.

[34] Han, Diana SC, and YM Dennis Lo. ”The nexus of cfDNA and nuclease biology.” Trends in Genetics 37.8 (2021): 758-770.

[35] Fan, Jerome, Suneel Upadhye, and Andrew Worster. ”Understanding receiver operating characteristic (ROC) curves.” Canadian Journal of Emergency Medicine 8.1 (2006): 19-20.

[36] Keilwagen, Jens, Ivo Grosse, and Jan Grau. ”Area under precision-recall curves for weighted and unweighted data.” PloS one 9.3 (2014): e92209.

[37] Lundberg, Scott M., and Su-In Lee. ”A unified approach to interpreting model predictions.” Advances in neural information processing systems 30 (2017).