January 2019

Searching For Phenotypes Of Sepsis: An Application Of Machine Learning To Electronic Health Records

Michael Jarvis Boyle

Searching for Phenotypes of Sepsis:

An Application of Machine Learning to Electronic Health Records

A Thesis Submitted to the

Yale University School of Medicine

In Partial Fulfillment of the Requirements for the

Degree of Doctor of Medicine

by

Michael Jarvis Boyle

2019

SEARCHING FOR PHENOTYPES OF SEPSIS: AN APPLICATION OF MACHINE LEARNING TO ELECTRONIC HEALTH RECORDS. Michael J. Boyle (Sponsored by R. Andrew Taylor). Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT.

Sepsis has historically been categorized into discrete subsets based on expert consensus-driven definitions, but there is evidence to suggest it would be better described as a continuum. The goal of this study was to perform an exhaustive search for distinct phenotypes of sepsis using various unsupervised machine learning techniques applied to the electronic health record (EHR) data of 41,843 Yale New Haven Health System emergency department patients with infection between 2013 and 2016. Specifically, the aims were to develop an autoencoder to reduce the high-dimensional EHR data to a latent representation amenable to clustering, and then to search for and assess the quality of clusters within that representation using various clustering methods (partitional, hierarchical, and density-based) and standard evaluation metrics. Autoencoder training was performed by minimizing the mean squared error of the reconstruction. With this exhaustive search, no convincing consistent clusters were found. Various clustering patterns were produced by the different methods but all had poor quality metrics, while evaluation metrics meant to find the ideal number of clusters did not agree on a consistent number but seemed to suggest fewer than two clusters. Inspection of one promising arrangement with eight clusters did not reveal a statistically significant difference in admission rate. While it is impossible to prove a negative, these results suggest there are not distinct phenotypic clusters of sepsis.

Acknowledgements

I am indebted to my thesis advisor, Dr. R. Andrew Taylor, for his constant support and insight, and to my friends and colleagues for their willingness to discuss these ideas and serve as valuable sounding boards. This work was made possible through the generous support of the Yale Summer Research Grant.

None of this would be possible, however, without the love and support of my wife, Shirin Jamshidian. This work is dedicated to her.

Introduction

Sepsis, defined as "life-threatening organ dysfunction caused by a dysregulated host response to infection" (1), affects an estimated 30 million people worldwide every year, potentially resulting in 5.3 million deaths annually (2). In one 2017 study of 409 hospitals encompassing 10% (2,901,019) of all hospital admissions in the United States, the incidence of sepsis was 6.0% with a mortality rate of 15% (3). Another study of two large cohorts including nearly 7 million adult hospitalizations in the United States between 2010 and 2012 found that sepsis contributed to between 34.7% and 55.9% of all inpatient deaths (4). According the Agency for Healthcare Research and Quality, in 2013 sepsis was the most costly condition in the United States, responsible for 23.6 billion dollars of healthcare expenditure that year alone. That expense amounts to 6.2% of national hospital costs resulting from nearly 1.3 million hospital stays (5). These staggering statistics are why in 2017 the WHA, the decision-making body of the WHO, adopted a resolution declaring the importance of improving diagnosis and management of sepsis (6), and why in 2018 there were more than 2,300 publications mentioning sepsis in the title when searched via PubMed.

Sepsis Definitions

Despite the interest in and impact of sepsis, it remains poorly understood. Its etiology is likely multifactorial, dependent upon both host and pathogenic factors, pro- and anti-inflammatory mediators, and the coagulation and neuroendocrine systems (7). But lacking a precise understanding of its pathophysiological mechanism, the task of

defining the syndrome has been left to expert-led consensus groups which have reviewed and revised their recommendations three times since 1991 with no shortage of controversy (1, 8-11).

While terms like “sepsis syndrome” were proposed earlier by researchers like Bone et al. in a 1989 trial of methylprednisolone for sepsis (12), the first consensus-based sepsis definitions were proposed at the 1991 American College of Chest Physicians/Society of Critical Care Medicine Sepsis Definitions Conference and published in 1992 (13, 14). Those definitions differentiated between infection, the invasion of host tissue by microorganisms, from sepsis, defined as the systemic host response to that infection as identified by having greater than one of the Systemic Inflammatory Response (SIRS) criteria (8). The SIRS criteria, which had been previously defined and which even then were acknowledged as not specific to sepsis, were composed of: 1) a temperature greater than 38°C or less than 36°C; 2) tachycardia greater than 90 beats per minute; 3) tachypnea greater than 20 breaths per minute or a $PaCO_2$ of less than 32 mm Hg; and 4) a white blood cell count greater than 12,000/mm^3 or less than 4,000/mm^3, or the presence of more than 10 percent immature neutrophils. The experts proposed the term “severe sepsis” to define the pathological condition where the adaptive response known as sepsis became maladaptive by causing organ dysfunction, hypoperfusion (lactic acidosis, oliguria, or acutely altered mental status), or sepsis-induced hypotension. They further defined “septic shock” as a more extreme subset of “severe sepsis” where the maladaptive response produced fluid-unresponsive hypotension or tissue hypoperfusion. Although the consensus group explicitly acknowledged that

"sepsis and its sequelae represent a continuum of clinical and pathophysiologic severity", they also defined transition points between these states which were subsequently used for nearly two decades to guide patient care and recruitment into clinical trials. Infection was differentiated from sepsis by two or more SIRS criteria; the adaptive host response (sepsis) became maladaptive (severe sepsis) with the presence of organ dysfunction, hypoperfusion, or hypotension; and fluid unresponsive hypotension marked the transition point between severe sepsis and septic shock.

The 1992 definitions were criticized almost immediately. The use of the SIRS criteria was criticized for its rigid cutoffs that narrowly excluded potentially septic patients from clinical trials, its lack of specificity for sepsis and the consequent heterogeneity of the patients it captured (68% of one study group including ICU and general wards patients met SIRS criteria), its uselessness for guiding clinical care, and its superficial relationship with underlying pathophysiology (10, 15).

In response to these criticisms, in 2001 a second sepsis definitions conference was held. However, citing a lack of new evidence, the expert consensus group merely reaffirmed the 1991 definitions with the additional acknowledgement that more clinical and laboratory variables could be used to identify systemic illness than just the four SIRS criteria. They did not provide specific guidance about how to use these additional variables to make the diagnosis (9).

Over the subsequent decade, the same criticisms of the definitions persisted and new studies clarified existing shortcomings. More researchers pointed out the need for

objective principles and biomarkers (16), while others suggested that organ dysfunction become part of the criteria for sepsis to prevent confusion between the terms sepsis and severe sepsis (17). Significantly, in 2015 Kaukonen et al. showed that among more than 100,000 ICU patients with infection and organ failure, one in eight did not meet SIRS criteria and mortality increased in a linear stepwise fashion with each additional SIRS criterion. There was no transitional increase in mortality at the threshold of two SIRS criteria, challenging "the sensitivity, face validity, and construct validity of the rule regarding two or more SIRS criteria in diagnosing or defining severe sepsis in patients in the ICU" (18).

Finally, in 2016 a group of critical care specialists met once more to develop the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). The task force determined that limitations of previous definitions included "excessive focus on inflammation, the misleading model that sepsis follows a continuum through severe sepsis to shock, and inadequate specificity and sensitivity of the systemic inflammatory response syndrome (SIRS) criteria" (1). They created the current definition for sepsis, "life-threatening organ dysfunction caused by a dysregulated host response to infection," and operationalized this definition as the increase of two or more points in the ICU-centric Sequential Organ Failure Assessment (SOFA) score. Severe sepsis was discarded as a redundant term, and septic shock was defined as a higher-mortality subset of sepsis in requiring vasopressors to maintain a mean arterial pressure of 65 mm Hg or greater and a serum lactate level greater than 2 mmol/L (>18 mg/dL) in the absence of hypovolemia. The consensus article and two accompanying analyses

determined the in-hospital mortality rates of these new definitions to be greater than 10% for sepsis and greater than 40% for septic shock (19, 20). The group also published a new scoring system, the quick Sequential Organ Failure Assessment (qSOFA) score, meant to be used to identify patients with a mortality equivalent to that of sepsis outside the ICU setting.

While the most recent criteria were analyzed with data in the papers that accompanied their release, they were still expert consensus-based and not derived a priori from an understanding of the pathophysiology (21). The group did not delineate distinct phenotypes of patients within the heterogeneous group captured by the non-specific organ dysfunction criteria. Moreover, they retained a categorical distinction between normal physiology, sepsis, and septic shock with discrete laboratory and clinical cutoffs. This categorical approach has been criticized as far back as the early literature prior to the release of the first sepsis definitions. In their 1992 critique of Bone et al.'s proposed "sepsis syndrome" definition, Knaus and colleagues wrote of their own analysis: "these findings led us to our major conclusion that while categoric definitions of sepsis may be useful in selecting patients for entry into clinical trials, they may not be useful in characterizing individual, or perhaps even group, risks. What our results suggest rather is that the current clinical condition of sepsis, at least as it is applied to a subset of critically ill patients admitted to ICUs, is a continuous state with the prognosis determined, in large part, by the degree of physiologic imbalance at the time of admission" (22).

This debate over definitions has significant real-world implications for patients because definitions can drive management. One of the major turning points in the management of sepsis was the 2001 trial of early goal-directed therapy (EGDT) for severe sepsis and septic shock, frequently referred to as the Rivers trial after its first author (23). The trial showed that when severe sepsis or septic shock were managed with specific goals for central venous oxygen saturation and pressure, hematocrit, and mean arterial pressure, mortality dropped from 46% to 30% compared to standard of care. The intervention was validated in a population of patients meeting severe sepsis and septic shock criteria as determined by the 1992 consensus definitions (two or more SIRS criteria with hypotension or elevated lactate). More contemporary trials of EGDT for septic shock have also used as entry criteria two SIRS criteria with refractory hypotension or elevated lactate (24). Since interventions validated in clinical trials are often applied only to the validated patient population, and in light of recent findings describing the stepwise linear increase in mortality with each additional SIRS criterion and the lack of a major transitional increase in mortality with two SIRS criteria, there may have been many patients that could have benefited from trial-validated interventions but did not receive them.

Based on all this prior work and debate, it stands to reason that if smaller groups of distinct pathophysiological processes or phenotypes could be identified amongst the heterogeneous group captured by expert consensus-defined diagnostic criteria, we might better be able to discover and deliver effective interventions. That is the motivation of this thesis.

Machine Learning and Electronic Health Records

The advent of widespread use of electronic medical records has created significant opportunities for large-scale data mining in healthcare (25). The sheer quantity of data available makes it amenable to analysis with a set of statistical inference algorithms known as machine learning.

Machine learning techniques applied to electronic health record data provide a potential solution to the problem of sepsis categorization by enabling phenotype discovery without the manual selection of features. The realm of machine learning is generally divided into two types of learning algorithms: supervised and unsupervised. Supervised learning aims to make predictions from data with a model trained on examples where the predicted value is known. Data where the target variable is known is called *labeled data*. A well-known example of a supervised task is the identification of objects within an image. To make accurate predictions, these models are trained on images where the object within the image has already been labeled.

On the other hand, unsupervised machine learning aims to discover patterns in data that has no labels (26). There are several types of unsupervised learning tasks, but one of the most common is called clustering, which is the attempt to separate unlabeled data into distinct clusters so that similar instances are grouped closely in space. Clustering techniques can be broadly be divided into hierarchical and partitional methods. Hierarchical methods function by creating a nested series of partitions, forming a dendrogram, whereas partitional methods only have one high-level partition

(27). Whatever the method, clustering applied to electronic medical record data provides an opportunity to discover distinctly different subsets of patients and disease states that are more similar to each other than they are to those in other clusters. This categorization can enable prediction and risk-stratification, can inform development of future therapies, and has even been used to discern subtypes of sepsis (28-32).

One of the challenges of applying clustering techniques to EHRs is that the data is very high-dimensional, has frequently missing values, and is highly heterogeneous combining both continuous and categorical variables (33-35). Traditional clustering techniques, like the *k-means* algorithm, do not perform well on very high-dimensional data. Thus, prior to clustering, high-dimensional data is often reduced to fewer dimensions using techniques that try to preserve the high dimensional relationships in a lower-dimensional *latent space*. Principle component analysis is an oft used method that attempts to find a transformation of the variable space that accounts for the variance within the distribution of data with the fewest possible orthogonal dimensions, known as principal components. More recently however, the development of a type of deep learning called the autoencoder has provided a more robust method for dimensionality reduction that is ideally suited for EHR data due to its ability to "learn" highly abstract features which can be represented in fewer dimensions (36).

Deep learning is a relatively new field that loosely emulates the structure neurons in the human brain – an "artificial neural network" -- to create computational models that learn abstract representations of data (37). They offer multiple advantages over more traditional learning algorithms, one of which is their ability to model complex non-linear

functions. Deep learning is responsible for numerous breakthroughs in computer vision, speech-to-text transcription, and even self-driving cars.

Invented by one of the fathers of artificial neural networks, Geoff Hinton, autoencoders are a type of deep learning where the input data is sequentially forced to be represented in fewer and fewer dimensions with each layer of the network before being allowed to expand again to the original number of dimensions with an architecture mirroring the reducing side. The network is then optimized so that the error between the input data and output data, known as the reconstruction error, is minimized. Once training is complete, new data can be fed through the first half of the network, the encoder, which outputs a latent representation that can subsequently be used for clustering. Essentially, the data is forced through a bottleneck that acts to compress the representation of the high dimensional data into fewer dimensions with minimal loss (38). Already, this technique has been applied to gain new insights from EHR data, including diagnosis prediction and the imputation of missing data (39, 40). These recent advances, from EHRs to machine learning and deep learning, provide researchers with powerful new tools to gain novel insights that could help patients.

In this thesis, I perform an exhaustive search for distinct phenotypes of infection by applying various clustering techniques to the latent (i.e. low-dimensional) representation of EHR data. If clusters can be identified within the data and these clusters have distinct features and mortalities, they could enable more precise clinical management and inform future investigations into targeted therapeutic approaches. If, however, an exhaustive search fails to reveal clusters, it would support the notion that

sepsis exists as a continuum and thus ought to be treated as such in clinical management. For example, a computer model that could project likelihood of in-hospital mortality might enable more precise clinical management than the current categorical classification of simply sepsis or septic shock. This effort is motivated by the aforementioned shortcomings of the expert-defined sepsis definitions, namely their use of cutoffs within continuous variables such as respiratory rate; their limitation to a small number of variables amenable to bedside rules; their muddied purpose of both clinical trial inclusion criteria and framework for clinical management; and ultimately their categorical classification of mortality despite the evidence for a continuum of disease severity (18, 22).

Aims

The purpose of this thesis is to perform an exhaustive search for clusters corresponding to distinct phenotypes of infection within the EHR data of patients in the emergency department with infection. I hypothesize that no clusters will be found. Because machine learning has a degree of art to it in addition to science, there is no way I can definitively prove that clusters do *not* exist; what I aim to do is to try multiple approaches to reasonably demonstrate that such clusters are unlikely.

Thus, my specific aims are the following:

1. Develop an autoencoder to reduce the high-dimensional EHR data to a latent space amenable to clustering while minimizing reconstruction error.
2. Use multiple partitional and hierarchical clustering methods to cluster the data.

3. Evaluate the proposed clusters with a variety of cluster validity metrics.

Methods

Study Design

This was a retrospective study of ED visits to three Yale-New Haven Health System (YNHHS) emergency departments between March 1, 2013 and May 1, 2016. The study was approved by the institutional review board.

Study Setting and Population

This study was performed across three sites: 1) the YNHHS York Street ED, 2) the YNHHS Saint Raphael ED, and 3) the YNHHS Shoreline ED. All hospitals used the Epic ASAP (Verona, WI) EHR.

This study included all emergency department encounters with patients at least 18 years old having a primary encounter diagnosis considered to be of infectious etiology, determined by ICD-10 code membership in a list of predetermined "infectious" ICD-10 codes. In order to include all patient encounters that were potentially septic, I reviewed all ICD-10-CM codes and generated a list of codes corresponding to diagnoses that could elicit a host response to infection. The decision to include or exclude a certain diagnosis was made based on my thesis advisor's and my clinical knowledge of the potential for that diagnosis to lead to sepsis. So, for example, "appendicitis" was included while "acute tubulo-interstitial nephritis" was not. Each included diagnosis was further categorized as one of the following types: "bacterial", "viral", "fungal/protozoal/parasitic", or "unspecified". The "unspecified" category was applied

when the diagnosis description was insufficient to determine the type of infectious process, e.g. “Pharyngitis”, or when the infection was specifically labeled as of unspecified origin, e.g. “Pneumonia, unspecified organism”. It was additionally found that because the study timeframe included the transition from the ICD-9 to the ICD-10 standard, certain diagnoses within the Yale-New Haven Health System’s Epic deployment lacked an ICD-10 code but possessed an ICD-9 code. In order to capture patient encounters associated with these diagnoses, I broadened the inclusion list to include any diagnoses where there was both no ICD-10 code and one of the following conditions were met: 1) the ICD-9 code was explicitly for an infectious or parasitic disease (ICD-9 001-139) or 2) the diagnosis name (as listed in the Epic deployment’s table) contained one of several keywords I defined, e.g. “infectious” or “cellulitis”. These additional diagnoses were also further categorized as with the ICD-10 codes.

I was motivated to cast a wide net with any potentially “infectious” ICD-10 codes rather than using physician-diagnosed sepsis in order to avoid biasing the included population towards those that met consensus-defined criteria. The objective was to capture all potential phenotypes of sepsis, including those that may have yet been unknown.

Study Protocol

An overview of the study protocol can be seen below in Fig. 1. Briefly, data was extracted from the EHR and reduced to one measurement per variable per encounter within a four-hour window starting with the first recorded measurement of any type for that patient. The data was then limited to only include variables not more than 50%

missing with the exception of a few that are part of the SOFA or septic shock criteria which I was motivated to retain due to prior work showing their importance in sepsis mortality prediction. Values were then imputed for all missing values. For each variable, an additional binary variable was added designating whether the value had been imputed or not. The now-complete dataset with 41,843 encounters and 290 variables/dimensions was used to train an autoencoder that compressed the dataset to a latent space of 16 dimensions. This compressed dataset was then used as the input for various clustering techniques which were subsequently evaluated. With the exception of the initial SQL query, all data analysis and autoencoder training was performed with the Python programming language with Jupyter notebooks. The Python packages Pandas, Sci-kit Learn, Keras with Tensorflow were used extensively for the data processing, clustering, and deep learning respectively. A detailed explanation follows below.

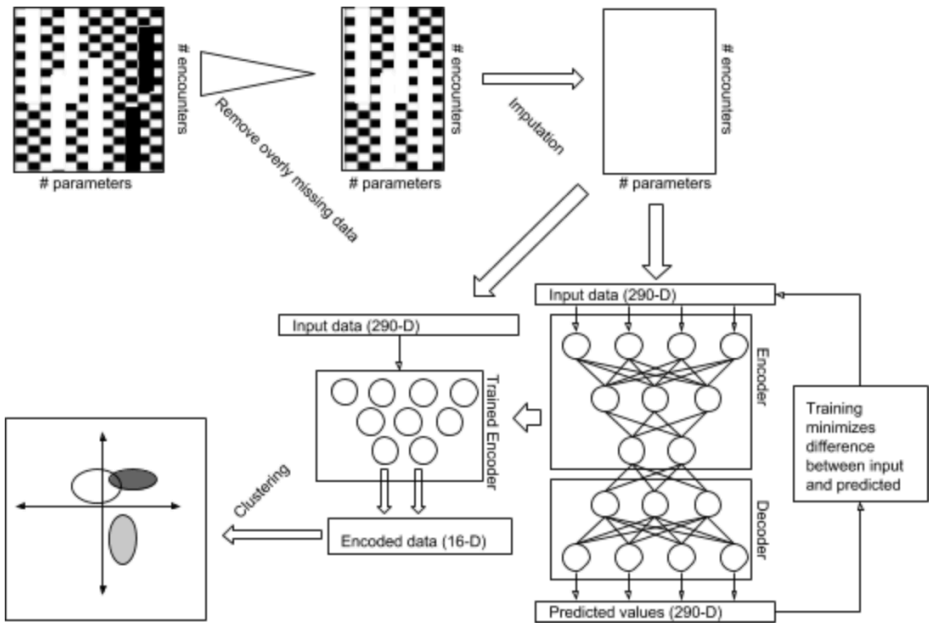

Fig. 1: Overview of the study protocol. Starting in the top left: rows and columns of data with some missing values (black) are restricted to only include columns without overly-missing data. The remaining missing data is imputed (all white), and then is used to train the autoencoder. When the autoencoder is trained, the encoding layers are extracted and used to generate a compressed representation of the data that is amenable to clustering.

Data Set Creation

All data was extracted from the Clarity enterprise data warehouse (Epic) with Structured Query Language (SQL) queries. For each patient encounter, these queries extracted demographic information (age, sex, ethnicity), social history (smoking status, alcohol use status, illicit drug use status), vital signs and oxygen requirement while in the ED, labs obtained in the ED, home medications, and past medical history.

Encounters missing disposition (1,146) were removed leaving a total of 41,843 encounters. Ages above 115 were converted to missing (NA) because 116 is the age used in Epic for unidentified patients.

For social history, if more than one response was recorded for a patient (e.g., smoking list as *never smoker* and *every day smoker*), the more severe value was chosen because it is less likely that was entered in error.

Past medical history for each patient was extracted in the form of ICD-10 code. In order to group the numerous possible diagnoses into meaningful and relevant abstract categories, each ICD-10 diagnosis was mapped to categories defined by the AHRQ Clinical Classification Software (CCS). For each encounter, this list of retained CCS codes was limited to those determined by my thesis advisor and me to affect the immune response. This determination was made by consulting various clinical scoring systems (SOFA, APACHE II/III, Charlson comorbidity score) and individual parameters used for sepsis criteria or sepsis mortality prediction (1, 19, 41-47). Finally, the list of CCS codes was condensed to form a more abstracted list of 17 classes of relevant past medical history (**Error! Reference source not found.**). Ultimately, each encounter was associated with 17 binary values, each indicating the presence of one of the types of relevant past medical history.

Table 1: Past medical history categories

HIV infection	Cancer
Immunity disorders	Maintenance chemotherapy or radiotherapy
Asthma	Chronic obstructive pulmonary disease and bronchiectasis
Other Respiratory	Liver disease (alcohol-related)
Thyroid disorders	Kidney disease
Diabetes	Other nutritional, endocrine, and metabolic disorders
Arrhythmias	FEN (electrolyte and nutritional disorders)
CHF	Hypertension with complications and secondary hypertension
Heart Disease	

Similarly, patient home medications were grouped into categories based on the YNHHS medication type schema. There were a total of 48 types of medication classes**Error! Reference source not found.**, and as with past medical history, each patient encounter was associated with 48 binary values, each indicating whether the patient was using one or more medications of that class. An additional variable was added to each encounter which corresponded to the total number of home medications in order to add additional information to the otherwise binary encoding.

In developing the "number of medications" variable, it became apparent that this section of the EHR may be particularly prone to user error or infrequent updating since many patients were using an inordinately large number of medications (Fig. 2). It is also possible that our SQL query failed to distinguish between active medications and ones that the patient was no longer using. Rather than decide upon an arbitrary cutoff for what a reasonable number of medications is, I decided to leave it as is with the understanding that if it is particularly noisy or meaningless, it will be deemphasized in the latent space representation after passing through the autoencoder.

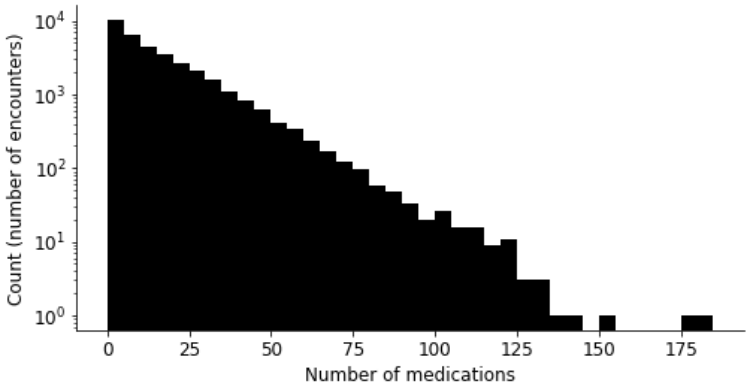

Fig. 2: Distribution of number of home medications. Note the logarithmic scale.

Laboratory values and vital sign measurements required a different approach. Whereas the other data, like demographics or medications, only had one allowable value per encounter, vital signs and laboratory values could be measured multiple times. With the motivation to try to capture phenotypes as they initially presented without the influence of therapeutic intervention, we chose to limit labs and vitals to those recorded within a few hours of arrival to the emergency department. On the one hand, if the time window was too short we risked losing valuable data that was reported later (e.g., a lab that was drawn early in the visit but had not been reported by the laboratory until several hours later). On the other hand, too long a window risked retrieving labs and vitals that had been influenced by therapeutic interventions. To determine an ideal time window, I examined the fraction of common labs and vitals missing as a function of time since arrival. The point at which the curve begins to flatten is the point at which extending the window does not provide substantially more data to warrant inclusion of biased values (Fig. 3). Ultimately, I decided that four hours produced a reasonable tradeoff since extending beyond that did not appreciably decrease the amount of missing data.

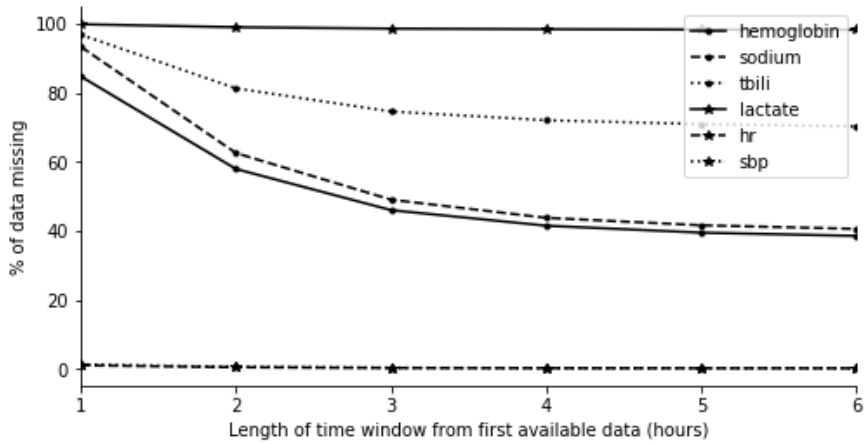

Fig. 3: Percentage of data missing as a function of time since first data point. This plot illustrates the effect of different time window cutoffs on the percentage of data available. Too short a cutoff results in a lot of missing data.

Since vital sign observations are manually entered by nursing staff, one can expect aberrant values and nonsensical outliers. It becomes more difficult to discern real values from mistakes when the data entered is theoretically possible, but improbable (e.g. a systolic blood pressure of 300). To try to limit the effect of outliers on vital signs data, I tried a number of techniques commonly used for dealing with outliers. Limiting vitals to three standard deviations of the mean proved too restrictive; the distribution of healthy vital signs is so narrowly distributed that even aberrant values seen commonly in the emergency department (e.g., a heart rate of 144 beats per minute) would have been excluded. I then attempted to limit vitals to 1.5 and 3.0 times the interquartile range (IQR) above the third quartile and below the first quartile, which are common definitions of outliers and extreme outliers. This method also proved too limiting as it discarded values like a respiratory of 28 as an extreme outlier. Distributions of vital signs are shown as boxplots in Fig. 4.

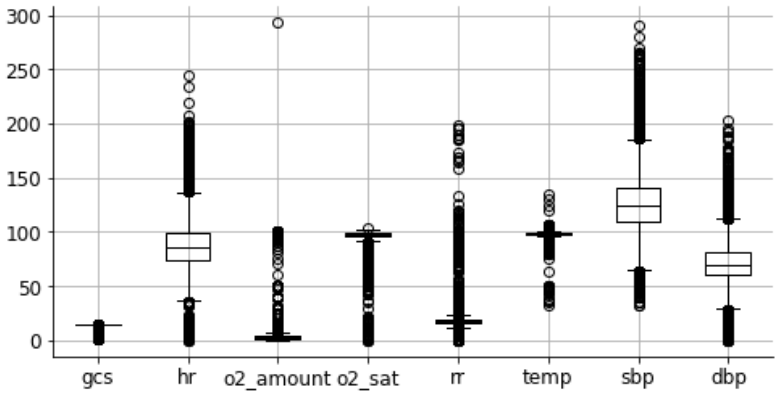

Fig. 4: A boxplot of the distribution of vital signs. gcs = Glasgow Coma Score, hr = heart rate, o2_amount = oxygen requirement (L/min), o2_sat = SpO2, rr = respiratory rate, temp = temperature (f), sbp = systolic BP (mm Hg), dbp = diastolic BP (mm Hg).

Ultimately, the best solution was to limit the vital signs to estimated physiological limits based on the experience of my thesis advisor and an examination of the values listed (e.g., a respiratory rate above 70 is more likely to be a heart rate entered in the wrong field than a respiratory rate. Table 2 below shows the cutoffs that were used for each vital sign. Values greater than the maximum or less than the minimum were set as missing values (NA).

After clipping vitals, most encounters had multiple values for each vital sign recorded during the four-hour time window. In order to reduce these observations to a single observation per encounter, vital summary statistics were creating. For each vital sign, a new variable was generated corresponding to the first, last, minimum, mean, and maximum values during the

Table 2: Cutoffs for vital signs

Vital sign	Min	Max
Glasgow coma scale	0	15
Heart rate	30	300
Respiratory rate	8	70
Temperature (F)	80	110
Systolic blood pressure (mm Hg)	30	300
Diastolic blood pressure (mm Hg)	20	250
Oxygen amount (L/min)	0	60
Oxygen saturation (SpO2)	40	100

time window. Another vital sign not shown in Fig. 4 or Table 2 is the oxygen dependency status. This was a categorical variable based upon a free-text field that required coercing into a limited number of possible options. These final categories, in order of increasing demand, were *room air, other, nasal, mask, positive pressure,* and *mechanical ventilation*. Since this variable was categorical instead of continuous, the mean summary statistic was replaced with the mode statistic.

Laboratory values were extracted only if the result was posted within the four-hour time window. If more than one measurement was posted for a given lab within that timeframe, only the first value was extracted in accordance with the goal of having a snapshot of the patient before therapeutic interventions influenced measurements. Any laboratory tests that had not posted a result in the four-hour time window were marked as missing (NA).

After windowing was complete, the degree of missing data was assessed. To avoid creating a dataset that was overall greater than 50% missing, I chose to retain only variables less than 50% missing with the exception of variables that feature prominently in the SOFA score or sepsis definitions (e.g., bilirubin and lactate).

The full list of labs that were retained and the percentage missing in the full dataset is listed in Table 6 in the appendix.

Imputation

After all the data was merged together and there was only one value per variable per encounter, missing data was addressed by imputing the column mode for each variable. Both mean and mode imputation were considered, but many of the variables, especially vitals and labs, were distributed in Poisson distribution with long tails towards abnormal values. Choosing mean imputation in these cases would have unreasonably skewed the imputation towards abnormal values. For example, lactate would have been imputed with a value greater than 2 mmol/L, which is greater than the threshold for inclusion in the septic shock criteria with the Sepsis-3 definitions.

In addition to imputing the mode, for each variable an additional column was added to mark with it was missing or not. The intent was for the autoencoder to learn to associate the missing marker with the missing variable itself and thus learn to ignore or discount that imputed variable.

Autoencoder Training

To make the dataset amenable to consumption by a neural network, all variables had to become numeric. Any Boolean variables (e.g., “uses alcohol”) and categorical variables (e.g., “O2 dependency” which could be *room, nasal,* etc.) were one-hot encoded. One-hot encoding transforms a single column of categorical values into a binary matrix

where each column corresponds to a single category and the binary value marks whether this category is present or not.

The data was then randomly split into a training (90%) and validation set (10%). One of the risks of training a machine learning model is overfitting the training data so that the model "memorizes" the training data but generalizes to new data poorly. To evaluate the model's generalizability, which is also a proxy for the degree to which it is learning a meaningful latent representation of the input data, the model is trained on one set of data but evaluated on another (48).

After splitting, each variable was zero-centered and scaled to unit variance by subtracting the mean and dividing by the standard deviation. This is common practice because many machine learning estimators behave badly if individual features do not resemble normally distributed data. One can imagine that if one feature had significantly more variance than another, it would dominate training because it would have more proportional explanatory power of variance compared to other variables (48).

With the data prepared ready for training, the next task was to find a combination of autoencoder parameters which, after training, would produce the lowest reconstruction error on the validation set. For this purpose, reconstruction error was measured as the mean squared error between the autoencoder input and output. A total of 16 encoding dimensions was chosen from the set of [2 8, 16, 32] because initial experiments training on a small subset of the data showed that 16 dimensions produced an acceptable

tradeoff between reconstruction error and a small enough number of dimensions to be easily amenable to clustering. A useful comparison is the dimensionality reduction from PCA. PCA applied to the dataset showed that 119 dimensions were required to explain 95% of the variance, so the autoencoder should at least be able to reduce the number of dimensions to 119 without much loss. For further comparison, I took the first 2, 4, 8, 16, and 32 principal components and projected the dataset into each and then reversed the transformation to create a lossy reconstruction from the compressed data. The reconstruction error from each of these compressed representations served as a useful benchmark for comparing to the autoencoder. If the autoencoder is "learning" an abstract representation of the data, it should outperform PCA when encoded with the same number of dimensions. I examined the difference in reconstruction error between PCA and a prototype of the autoencoder for the same number of compressed dimensions and observed where the difference between reconstruction errors began to stabilize (Fig. 5 and Fig. 6). This occurred around 16 dimensions, validating this choice.

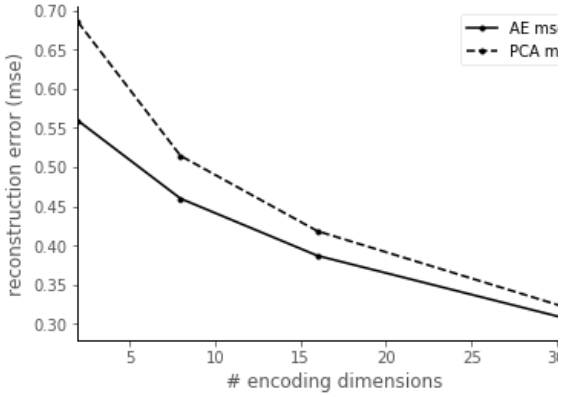

Fig. 5: Reconstruction mean-squared error (mse) as a function of # compressed dimensions with a prototype autoencoder (AE) and PCA.

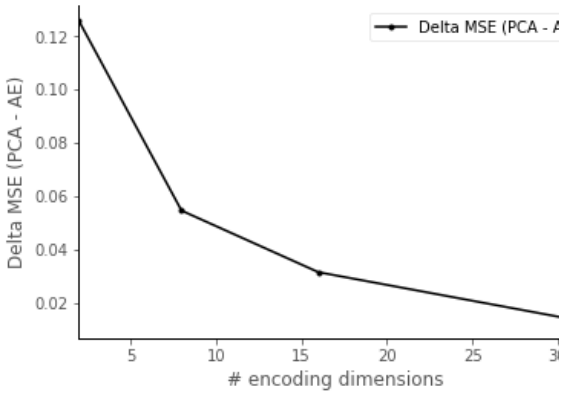

Fig. 6: Difference in reconstruction mean-squared error between PCA and autoencoder (AE). The curve flattens between 8 and 16 dimensions, suggesting diminishing benefit of AE over PCA beyond 16 dimensions.

After deciding on the number of encoding dimensions, I varied the network architecture (number of layers, number of neurons per layer), the optimizer and its learning rate (Adam and Adadelta), and various regularization parameters which serve to prevent overfitting (L2 normalization, batch normalization, and dropout)(48, 49). Ultimately, the combination that had the lowest reconstruction error on the validation set was an architecture with 2048-1024-512-16-512-1024-2048 neurons in each layer, with the 16-neuron layer serving as the encoding layer. The ideal optimizer was the Adam optimizer with a learning rate of 0.0001, and the network was trained with a batch size of 4096 training examples per batch. To prevent overfitting, the ideal input dropout rate was 5% with a dropout rate of 20% between hidden layers except between the encoding layer and the subsequent 512-neuron layer, along with batch normalization. After training for 1700 epochs the validation loss began to climb signifying overfitting, so training was halted and the best scoring model was saved and used for all future work. When the full dataset was put through the autoencoder and compared to its reconstruction (this time without the effect of dropout), the network produced a reconstruction error of 0.118,

which compares very favorably to the PCA reconstruction mean-squared error for the same number of dimensions (a total of 16), which was 0.419.

Clustering

Once the autoencoder was trained, the entire original dataset was transformed via the encoder to its latent 16-dimensional representation. From this point, an exhaustive search was performed with various clustering algorithms. For each algorithm attempted, several cluster validation metrics were applied to analytically determine cluster fit, and the clustering results were visualized in 2-dimensional PCA and t-SNE transformations of the latent space to permit visual assessment of cluster fit. The t-SNE algorithm is a non-parametric mapping algorithm used to project higher dimensional data into lower spaces while preserving higher-level relationships between points (50). The clustering methods attempted were *k-means, agglomerative hierarchical clustering* with multiple distance metrics (*linkages*), and the density-based *DBSCAN* algorithm. Prior to any clustering, the overall propensity for cluster-ability was assessed by projecting the original and latent data into 2 dimensions with PCA and t-SNE, as well as by calculating the Hopkins statistic (51) which returns values greater than 0.5 when the data is clumped. It does not differentiate between data that is in one big cluster versus several smaller clusters.

After clustering, the fit of each cluster arrangement and number of clusters was assessed with the elbow method (52), silhouette score (53), gap statistic (54), Calinski and Harabasz score (55) where possible.

Results and Discussion

Quality of dimensionality reduction and latent representation

The PCA two-dimensional projection of the latent representation can be seen in Fig. 7. While the PCA projection of the latent features does not contain any visually discernable clusters, because it is such a compressed representation it may certainly have a structure that is just not captured in two dimensions. Another way to visually assess the presence of distributions amenable to clustering within data of greater than a few dimensions is to plot a histogram of the distributions along the primary principle components of the latent space (Fig. 9). With the exception of some secondary peaks in the third and fifth principal components, there are no obvious separate distributions outside the primary Poisson and Gaussian distributions. On the other hand, the t-SNE projection shows some very distinct clustering. It is important to note t-SNE does not have any linear relationship to the dimensions that it represents. Rather, it is a mapping that is learned from the higher dimensional space while optimizing for representing differences between groups of points in that space.

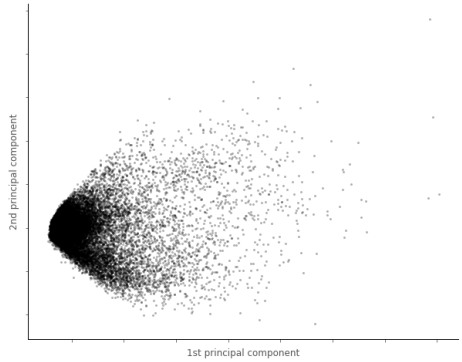

Fig. 7: PCA projection of the first 2 principal components of the 16-dimensional latent space produced by the autoencoder.

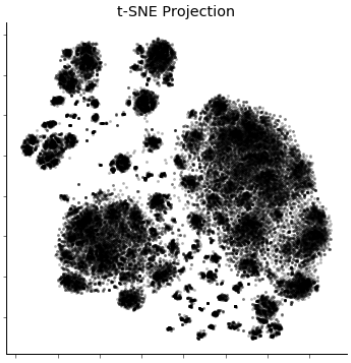

Fig. 8: t-SNE projection of the latent space.

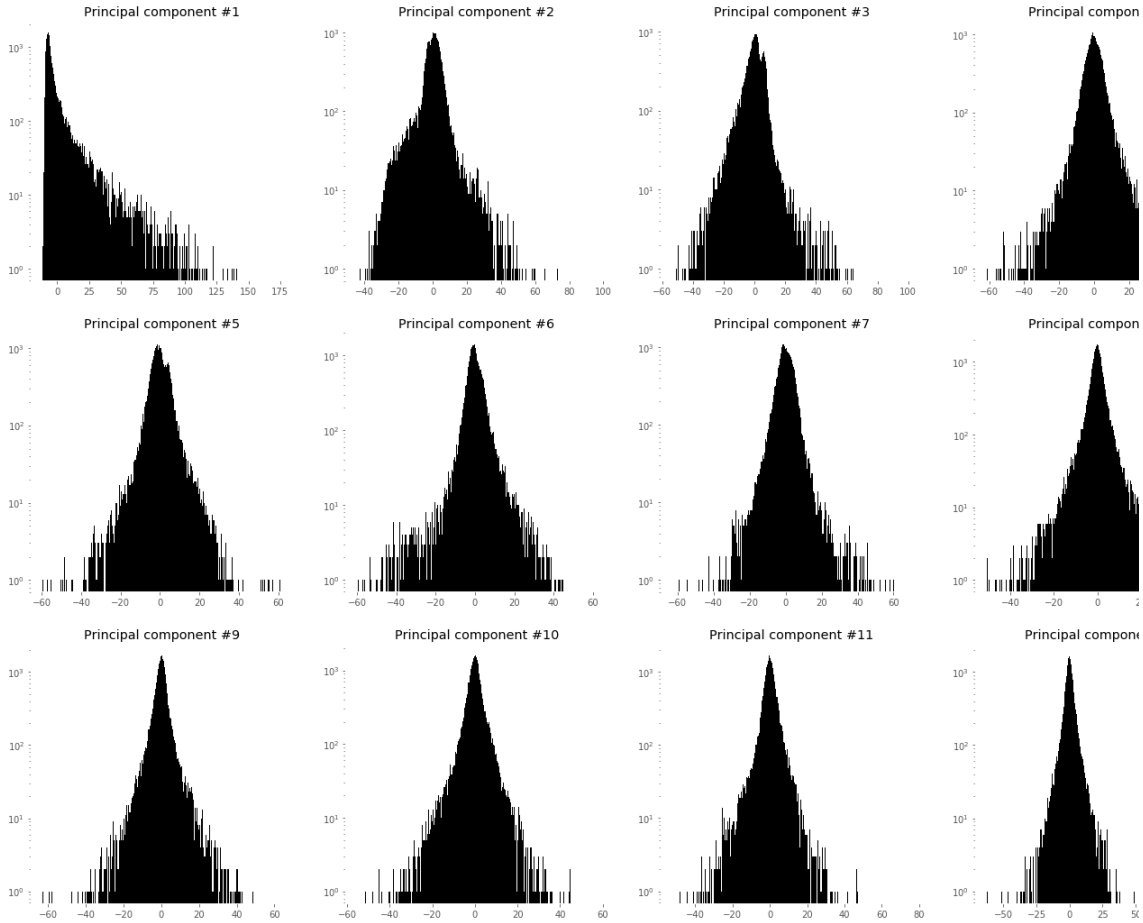

Fig. 9: PCA projection of the first 12 principal components of the latent representation, shown with a logarithmic scale to better visualize smaller groups. Together these components explain 91% of the variance in this space.

Clustering

Assessing clustering propensity

Prior to clustering, the propensity for the data to form clusters was analyzed. This can be predicted with a Hopkins statistic (51, 52), which compares the distribution of the data to what one would expect from a uniformly distributed dataset within the feature space. Values closer to one indicate that the data is aggregated whereas a value of 0.5 indicates the data is uniformly distributed. The Hopkins statistic for the latent feature space was 0.94 suggesting it is highly aggregated. The disadvantage of the Hopkins

statistic is that aggregation does not imply useful clustering because one single large clump would also have a very high Hopkins statistic.

Assessing ideal number of clusters

To cluster with the *k-means* algorithm and other partitional clustering algorithms, one must know the desired number of clusters. A common technique used to provide a best-guess is the so-called *elbow method* (52). The elbow method plots the sum of within-cluster squared distances from each point in each cluster to its cluster centroid. When there are insufficient clusters, each additional cluster helps lower the sum of within-cluster squared distances. But eventually with the addition of too many cluster centers, they begin to break up preexisting clusters into smaller clusters without a significant drop in the sum of within-cluster squared distances. The "elbow" in the graph marks this point of diminishing returns. The elbow method plot for *k-means* applied to the latent feature space can be seen in Fig. 10. No elbow is visible, indicating that with *k-means* there is not an obviously ideal number of clusters.

The gap statistic is another widely used method for ascertaining the ideal number of clusters within a distribution of data (54). The ideal number of clusters is indicated by the point on the curve where the gap drops for the first time, which is located at 8 clusters in the case of this data (Fig. 11).

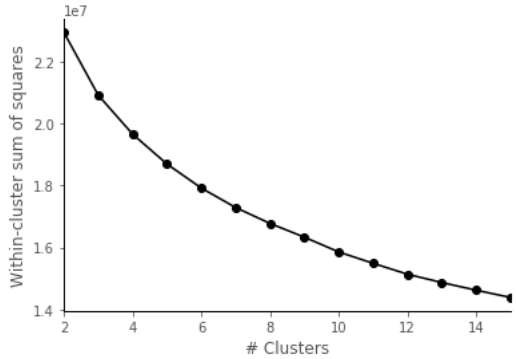

Fig. 10: The elbow method plots the within-cluster sum of squares against number of clusters. A marked bend in the curve would indicate a point of diminishing explanatory ability of additional clusters. No elbow is visible.

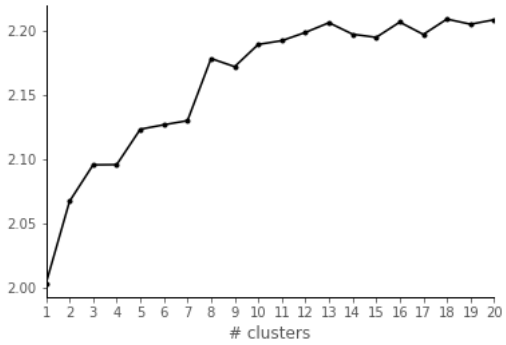

Fig. 11: The gap statistic decreases with the addition of one more cluster to the ideal arrangement, signifying that the explanatory power of the model is decreasing. Here, the gap indicates that 8 clusters is ideal.

A third method for evaluating the proper number of clusters is provided by the Calinski and Harabasz score (55). This score measures the ratio of the between-cluster dispersion mean to the within-cluster dispersion, so a higher score identifies a model with better-defined clusters. The Calinski and Harabasz score is shown in Fig. **12** with a peak score at two clusters and diminishing from there.

Table 3: Ideal number of clusters, by method

Method (clustering)	Ideal # of clusters
Elbow method (k-means)	N/A
Gap statistic (k-means)	8
Calinski and Harabasz (k-means)	2
Silhouette score (k-means)	2

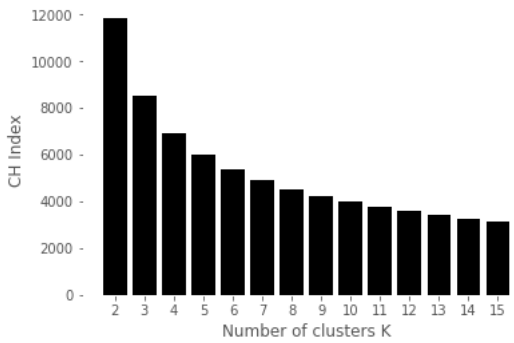

Fig. 12: Calinski and Harabasz score as a function of number of k-means clusters. A higher score indicates a better-formed cluster.

In summary, the elbow method did not provide any guidance; the gap statistic, considered a more standardized version of the elbow method, indicated that eight clusters would be ideal. On the other hand, the Calinski and Harabasz score indicated that two (or possibly one, though this is not calculable) is the ideal number of clusters. Both of these options are examined in the next section. A summary of ideal cluster number analysis is shown in Table 3. It is important to note that these methods of clustering propensity are based on clustering with k-means, which is ideally suited for convex (i.e. spherical) clusters. Thus, the value they provide in identifying ideal number of clusters is limited to these types of convex clusters.

Partitional Methods

K-means

K-means clustering with anywhere from 2 to 16 clusters was performed and each arrangement was assessed for quality of fit with the silhouette score (52) which

measures intra-cluster cohesion against inter-cluster dispersion with a value near one indicating maximal clustering. The k-means algorithm was repeated for every number of clusters between two and 15 and the results of the silhouette score were plotted to assess quality (Fig. 13). Projections of two, five, and eight clusters (as suggested by the ideal cluster analysis, as well as an intermediate) into two dimensions via PCA and t-SNE are plotted in Fig. 14. The PCA projection does not provide any new insights; indeed, the pattern of separation of clusters looks similar to one would expect were one to attempt to cluster a spherical distribution of points. However, t-SNE projection does seem to mirror what one might anticipate, especially the 8-cluster arrangement, which nicely separates the three main groups (colored in red, green, and blue). It is difficult to reconcile the apparently nice clustering in the t-SNE projection with the lack of other quantitative evidence in the form of the silhouette score.

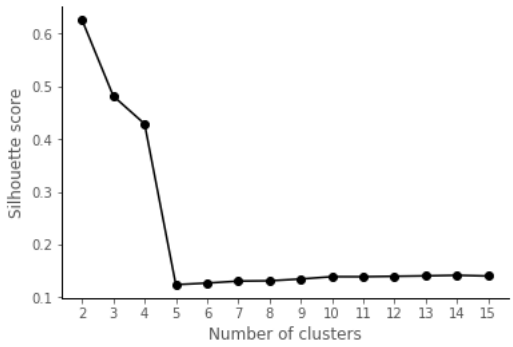

Fig. 13: Silhouette score for k-means clustering. A higher score indicates better cluster separation. Here, the maximum score is at k=2.

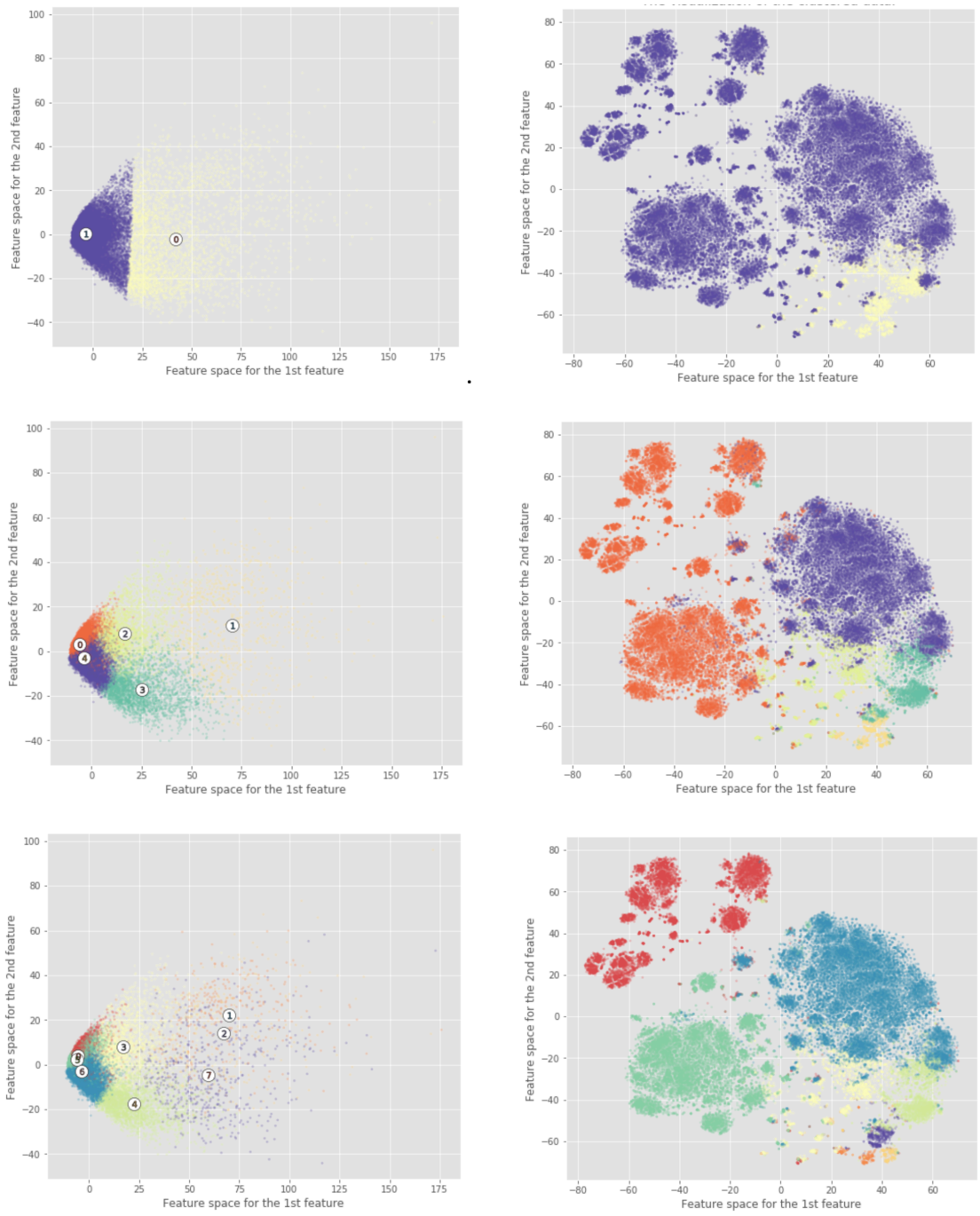

Fig. 14: K-means clustering results in PCA (left) and t-SNE (right) representations with 2 (top), 5 (middle), and 8 (bottom) clusters. Cluster centers are displayed as small white circles in the PCA projections.

However, that does not explain why k-means clustering did manage to align with splitting of clusters evident in the t-SNE projection. Another interesting observation is that when only two clusters are utilized, the second cluster is skewed significantly by the

outliers along the first principal component. This is even more evident in the t-SNE projection where one can see the second cluster constitutes a very small portion of the total data points. This could very well be a result of k-means sensitivity to outliers (52). To address this concern, I also conducted k-medoids clustering in the following section.

K-medoids

K-medoids (also known as Partitioning Around Medoids) is much like the k-means algorithm except that instead of allowing arbitrary points in space to be the cluster centers, only actual data points can serve as cluster centers. This mitigates the risk of an outlier dragging the mean of a cluster far out in one direction and placing a cluster center far away from most of its points (52). Like k-means, one must specify the number of clusters *k*, and it produces convex clusters amenable to analysis with the silhouette score. With the insight from initial k-means clustering that eight clusters produces nice separation of the groups in the t-SNE projection, I chose to try k-medoids with eight clusters. The projections and a silhouette plot are shown in Fig. 15. The mean silhouette score was 0.008, hardly an indication of good clustering where a score of one is idea. However, again on t-SNE one can see nice separation of cluster 0, 3, and 6 (red, yellow, blue) while on PCA it is impossible to discern. It is important to remember that a limitation of PCA in two dimensions is that much information is lost in the projection to two dimensions so one cannot say there are not viable clusters just because they are not appreciable in the projection; in fact, in this case two dimensions only accounts for 41% of the variance. But it is also important to recognize the mapping of points in space in t-SNE can change based on the hyperparameters chosen for training (perplexity and

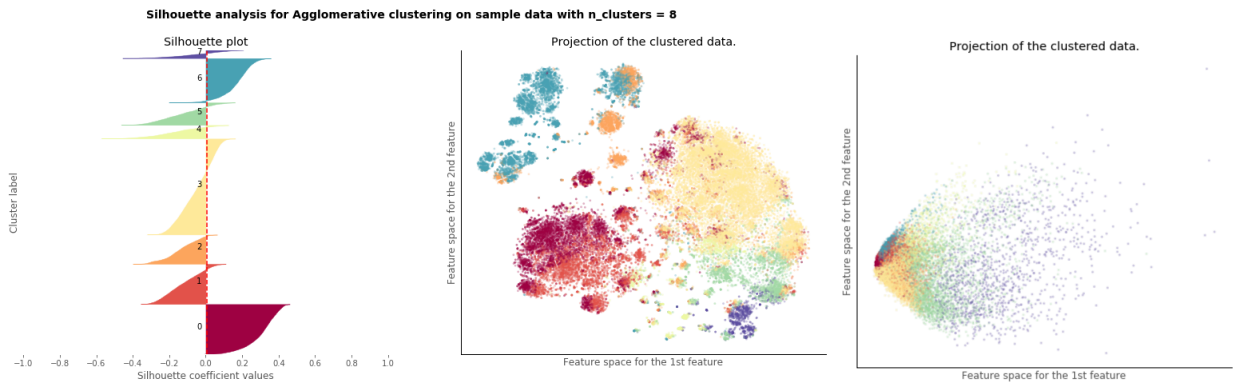

Fig. 15: K-medoids clustering silhouette plot, t-SNE projection, and PCA projection with 8 clusters. The silhouette plot shows the individual silhouette score of each point in each cluster. Scores closer to 1 indicate good clustering, whereas scores less than 0 indicate poor clustering (i.e. the point does not have more affinity for neighbors in its own cluster than for neighbors in other clusters). The red vertical dotted line indicates the mean silhouette score (0.008).

number of iterations). While I did do a search of the hyperparameter space to find the ideally separated groups, it is entirely possible that this mapping is truly representative of the data in 16 dimensions. I did not appreciate significant changes in the appearance of the plot as I changed the hyperparameters, but it is possible that another set of hyperparameters would have produced a mapping less convincing of clusters. In sum, the t-SNE projection must be taken with a grain of salt. The quantitative methods like the silhouette score should be trusted in the case where there are convex clusters, and it is clear that by that metric the ideal number of clusters is two or less. To search for non-convex clusters, I also tried hierarchical and density-based methods.

Hierarchical Methods

Agglomerative clustering with ward linkage

Agglomerative hierarchical clustering is another approach to clustering altogether, where the process happens from the bottom up rather than top down. Agglomerative hierarchical clustering works by successively grouping groups points into a hierarchy of

trees. In this manner, it may be able to find non-convex shapes of clusters that would not be found by partitional methods (52). For this implementation, I chose *ward* linkage as the minimization objective, which equates to minimizing the variance of two clusters being merged. Like the other methods, the number of clusters must be specified beforehand. Results can be seen in Fig. 16. Again, eight clusters separate the t-SNE data nicely. The silhouette score is inapplicable in this case as it is not guaranteed to form convex clusters.

Agglomerative clustering with single and complete linkage

There are other linkage metrics that can be used with hierarchical clustering, like single and complete. Single linkage allows merging of clusters based on the distance between their two closest points and tends to optimize clusters defined by local proximity, whereas complete linkage merges clusters based on the distance between their farthest points and tends to optimize clusters for global proximity (52). Outliers are harshly penalized in complete linkage. Both of these linkages were applied to the latent representation of the data, and both produced essentially one giant cluster encompassing all the data, with seven additional imperceptible clusters so small they were not visible in the projections. For brevity, these figures are not shown here.

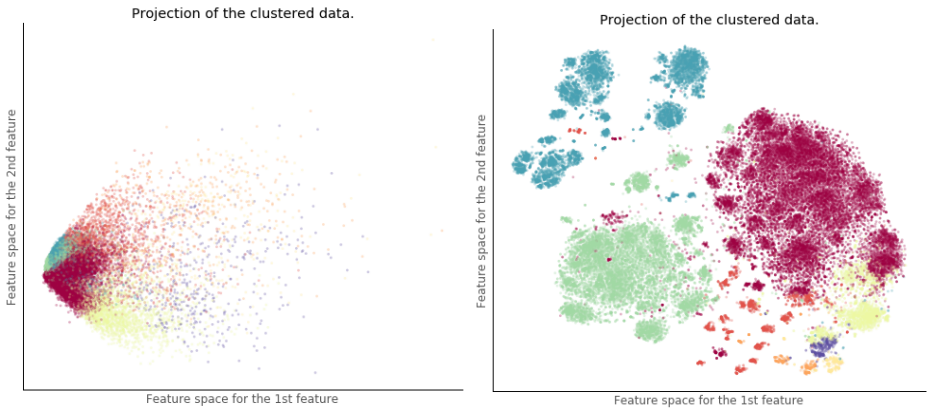

Fig. 16: PCA and t-SNE projection of agglomerative hierarchical clustering with ward linkage.

Density-Based Methods

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm which, as its name suggests, clusters not by distance as k-means does, but by density. Briefly, it finds data points that meet a minimum threshold of having *n*

neighbors within distance ϵ, called core objects. It sequentially builds up clusters by joining core objects if one is within distance ϵ

of the other. The advantage of a density-based approach to clustering is that the clusters need not be convex (52). One disadvantage is that both n and ϵ are user-defined, thus the potential search space is much greater than when searching for k clusters with k-means. An additional noteworthy feature of the algorithm is that any points outside a dense region will not become part of any cluster and will instead be marked as outliers. Several iterations of DBSCAN with multiple hyperparameter settings are shown in Fig. 17. The DBSCAN results interestingly bridge points across what seem to be different clusters from visual inspection, suggesting that these points are actually close by in the latent space. Not shown are the PCA projections, which

demonstrate that in all of these arrangements, half if not most of the points are considered outliers. It demonstrates that the majority of the points in the latent space are clustered tightly together towards on side of the first principal component as seen in Fig. 9.

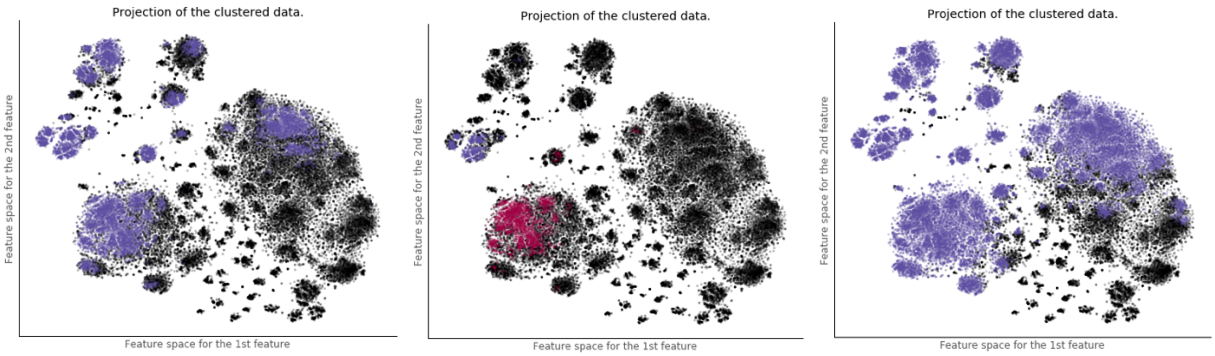

Fig. 17: DBSCAN clustering with three hyperparameter settings. From left to right: E=5, min_samples=100; E=5, min_samples=1000; E=10, min_samples=1000. They arrived at 1 cluster, 2 clusters, and 1 cluster respectively. Black points are outliers.

Making Sense of the Clustering

Overall, the general pattern was that quantitative methods and projection by PCA did not convincingly demonstrate distinct clusters. Interestingly, an eight-cluster arrangement brought about the same separation in the t-SNE projection with both k-means, k-medoids, and agglomerative hierarchical clustering with ward linkage. But one must take the overall picture; if k-means and k-medoids clusters were high quality clusters as they appear to be in the t-SNE projection, one would expect that the quantitative methods, especially the silhouette score, would have shown more promising results. While the gap statistic did recommend eight clusters, there was not a significant drop in gap between eight and nine clusters, while the silhouette score showed very strikingly the drop in average score from two clusters onwards. The hypothesis that there are no clusters is also supported by the PCA projection along the 12 principal components in Fig. 9. Likewise, the density-based clustering supports the notion that the t-SNE is misleading in that dense regions in a single cluster bridge the apparent "clusters" shown in that projection. The DBSCAN results very much mirror what one might expect looking at the PCA projection along the first 12 principal components. Moreover, both complete- and single-linkage metrics used for hierarchical agglomerative clustering created essentially one large cluster with seven imperceptible small clusters.

To better illuminate possible differences in clusters, I examined the centroid example for each cluster in the k-medoids clustering. I examined the 20 columns with the greatest index of dispersion (σ^2/μ), along with their final disposition, and compared

them in Table 4. Clusters 0, 3, and 6 (corresponding approximately to clusters 5, 6, and 0 in the k-means clustering with k=8; green, magenta, and blue in the agglomerative clustering with k=8) correspond roughly to the major groupings seen in the t-SNE projection. I was unable to discern salient differences except that all three are middle aged or older, have higher creatinine, take more medications, and cluster 0 is centered on an elderly person with a high white count with a neutrophilic predominance. The centroids of clusters 0 and 3 were admitted while the rest were discharged. In Table 5, I show the admission rates of these clusters. A *chi-squared* test did not find any statistically significant difference in admission rate between them, with a test-statistic of 9.0 and a p=0.25. In summary, there is reasonable doubt as to whether these are, indeed, distinct clusters with distinct differences. They do not differ significantly by admission rate, although this may be because there are differences imperceptible to the physicians making those decisions. This explanation is less likely however.

Table 4: K-medoids centroids and variables with greatest dispersion.

A=admit, D=discharge

Variable	Cluster centroid							
	0	1	2	3	4	5	6	7
platelets	1118	207	278	229	207	348	226	207
age	92	38	24	43	54	21	58	32
num_meds	19	3	10	21	1	1	31	9
creatinine	2.4	0.7	0.7	1.5	0.7	0.6	13.1	0.7
bun	27	11	14	20	11	11	34	11
vitals_dbp__min	57	93	109	77	95	55	78	81
vitals_dbp__last	58	93	109	80	95	55	78	81
lymphocytes	6	10	25	24	10	26	17	10
anc	19.5	4.8	5.6	3.7	4.8	6.7	4.6	4.8
vitals_dbp__mean	65.5	93	112.3	78.5	95	60.5	89	83.5
vitals_hr__first	106	98	100	72	81	64	72	75
vitals_dbp__first	76	93	114	77	95	66	106	86
vitals_dbp__max	76	93	114	80	95	66	106	86
vitals_sbp__last	123	135	159	154	155	105	128	132
wbc	21.4	8.4	8.4	6.5	8.4	10.6	6.6	8.4
vitals_hr__min	93	98	81	70	81	64	62	75
vitals_sbp__min	123	135	154	131	155	105	128	129
vitals_o2_amount__max	2	0	0	0	0	0	0	0
vitals_o2_amount__last	2	0	0	0	0	0	0	0
monocytes	3	7	6	16	7	7	11	7
Disposition	A	D	D	A	D	D	D	D

Table 5: Admission rate by cluster

Cluster	Admit rate (%)
0	42.43
1	40.31
2	39.92
3	39.95
4	39.61
5	39.79
6	39.89
7	39.90

In summary, based upon these data, it does not appear that there are salient clusters. Though this thesis has attempted to perform a thorough search with multiple techniques, many more remain to be tested. So, while I cannot conclusively determine that no clusters exist (with enough data and the right representation, they probably do), these results reasonably demonstrate that no obvious clusters exist.

Limitations and Advantages

There several key limitations to this study. First, the dataset is a highly heterogeneous clinical dataset with a significant amount of missing data (see Table 6 in Appendix A). While it is commonplace in real-world clinical datasets, missing data provides a serious challenge to machine learning algorithms that learn relationships between different variables because new relationships (i.e. bias) can be introduced through the process of imputation. In clinical data, missing data is usually not missing not at random. In other words, there is information in the fact that the data is missing; a physician might not

have ordered a laboratory test because she did not anticipate that the value would be abnormal. In this manner, physician insight leaks into the dataset. Then, one must decide how to impute the missing values. As discussed in the methods section, mean imputation introduces problems when the data do lie in a normal distribution. In this thesis, I tried to mitigate these influences by imputing the column mode for each value, and by introducing an "is missing" variable for each variable. The intention is that the autoencoder would come to learn the relationship between the mode of a variable and the presence of the missing flag, thus discounting its reliance on this value for prediction. There is evidence that the autoencoder did learn well considering the reconstruction error compared to PCA. State of the art imputation methods use other machine learning techniques, like a Random Forest classifier or regressor to impute missing values by learning from data where that value is not missing. Though this approach is vulnerable to data missing not at random, it may provide better performance for this model in the future. In this thesis, it could not be employed due to technical issues.

Another limitation of this thesis is the interpretability of the autoencoder latent representation. Because an autoencoder learns a non-linear mapping of the original data to the latent space, it is very difficult to discern the significance of the original variables in the latent representation as one could with PCA. Inspection of cluster differences based upon the medoids shows some differences, but despite this the overall admission rate was unchanged between clusters. Further analysis will be needed to understand any differences between these putative clusters.

A third limitation is the representation of the data for training by the autoencoder. Because binary and continuous variables were treated equivalently, with the training minimizing the mean squared error between the original data and its reconstruction, it is possible that the binary variables overwhelmingly dominated the loss function and the encoder was not forced to learn a good representation of the continuous variables. This could potentially be mitigated in future work by building an autoencoder with two output layers, one for continuous variables and one for binary variables, which are trained together but with different loss functions (mean squared error and cross-entropy, respectively) which are then combined in a weighted sum to produce an overall loss function.

Overall, there are several advantages of the approach taken in this thesis. By not including physician notes as other EHR deep learning has (39), this approach reduces the potential for physician bias to leak into the data. Moreover, the use of an autoencoder enables the discovery of highly abstract features and non-linear relationships that would not be apparent with the traditional regression techniques used in the seminal sepsis definition papers (19). It also obviates the need for feature selection, thereby enabling the discovery of new important features that may have previously been overlooked.

Conclusions

This thesis sought to characterize phenotypes of infection amongst potentially septic patients in the emergency department through a variety of unsupervised machine learning techniques. I created an autoencoder, a type of deep learning architecture, to

reduce the dimensionality of the electronic health record data. The reconstruction error of this reduction compared very favorably to PCA, suggesting the latent representation had captured salient abstract features of the dataset. When clustering, however, results were not as clear. The sum of evidence did not point to distinct clusters. If the 8 putative clusters identified by several methods are indeed real, there was no difference in admission rate amongst them suggesting any differences may not be salient enough to produce a clinical effect (or that physicians are not noticing the differences). The implication of this lack of clusters is significant for clinical care, and was articulated clearly by Knaus et al. in 1992 (22):

> *"Sepsis is a complex clinical entity and could be viewed as a continuum with substantial variation in initial severity and risk of hospital death. One accurate description of sepsis is the continuous measure of hospital mortality risk estimated primarily from physiologic abnormalities… These findings led us to our major conclusion that while categoric definitions of sepsis may be useful in selecting patients for entry into clinical trials, they may not be useful in characterizing individual, or perhaps even group, risks. What our results suggest rather is that the current clinical condition of sepsis, at least as it is applied to a subset of critically ill patients admitted to ICUs, is a continuous state with the prognosis determined, in large part, by the degree of physiologic imbalance at the time of admission."*

If potentially septic patients were scored directly with a continuous mortality prediction tool, that might better inform their management. Categorization by bedside rules is helpful when a clinical condition can be reduced to such a scoring system, but it is unreasonable to expect that something as complex as pathophysiology can always be summarized with an easily-memorized rule, despite what Vincent et al. have argued (10). With the advent of EHRs and increasing computing power, complex models can potentially be included in the

physician workflow without added effort. One can even imagine these prediction tools running on all patients and only alerting a physician when mortality prediction reaches a certain threshold. This would spare the debate over what category a patient falls into for the time being. In the future, a better pathophysiological understanding of sepsis may make this categorization possible, but for now it may be best for patients to wait until then to use categorical classification with sepsis.

References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA.* 2016;315(8):801-10.
2. Fleischmann C, Scherag A, Adhikari NK, Hartog CS, Tsaganos T, Schlattmann P, et al. Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations. *Am J Respir Crit Care Med.* 2016;193(3):259-72.
3. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, et al. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. *JAMA.* 2017;318(13):1241-9.
4. Liu V, Escobar GJ, Greene JD, Soule J, Whippy A, Angus DC, et al. Hospital Deaths in Patients With Sepsis From 2 Independent Cohorts. *JAMA.* 2014;312(1):90-2.
5. Torio CM, and Moore BJ. Rockville, MD: Agency for Healthcare Research and Quality; 2016.
6. Reinhart K, Daniels R, Kissoon N, Machado FR, Schachter RD, and Finfer S. Recognizing Sepsis as a Global Health Priority - A WHO Resolution. *NEJM.* 2017;377(5):414-7.
7. Yao YM, Luan YY, Zhang QH, and Sheng ZY. Pathophysiological aspects of sepsis: an overview. *Methods Mol Biol.* 2015;1237:5-15.
8. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, et al. American-College of Chest Physicians Society of Critical Care Medicine Consensus Conference - Definitions for Sepsis and Organ Failure and Guidelines for the Use of Innovative Therapies in Sepsis. *Crit Care Med.* 1992;20(6):864-74.
9. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, et al. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. *Crit Care Med.* 2003;31(4):1250-6.
10. Vincent JL. Dear SIRS, I'm sorry to say that I don't like you. *Crit Care Med.* 1997;25(2):372-4.
11. Abraham E, Matthay MA, Dinarello CA, Vincent JL, Cohen J, Opal SM, et al. Consensus conference definitions for sepsis, septic shock, acute lung injury, and acute respiratory distress syndrome: time for a reevaluation. *Crit Care Med.* 2000;28(1):232-5.
12. Bone RC, Fisher CJ, Jr., Clemmer TP, Slotman GJ, Metz CA, and Balk RA. A controlled clinical trial of high-dose methylprednisolone in the treatment of severe sepsis and septic shock. *N Engl J Med.* 1987;317(11):653-8.
13. Bone RC, Fisher CJ, Jr., Clemmer TP, Slotman GJ, Metz CA, and Balk RA. Sepsis syndrome: a valid clinical entity. Methylprednisolone Severe Sepsis Study Group. *Crit Care Med.* 1989;17(5):389-93.
14. Marshall JC. Sepsis Definitions: A Work in Progress. *Crit Care Clin.* 2018;34(1):1-14.

15. Rangel-Frausto MS, Pittet D, Costigan M, Hwang T, Davis CS, and Wenzel RP. The natural history of the systemic inflammatory response syndrome (SIRS). A prospective study. *JAMA.* 1995;273(2):117-23.
16. Gaieski DF, and Goyal M. What is sepsis? What is severe sepsis? What is septic shock? Searching for objective definitions among the winds of doctrines and wild theories. *Expert Review of Antiinfective Therapy.* 2013;11(9):867-71.
17. Vincent J-L, Opal SM, Marshall JC, and Tracey KJ. Sepsis definitions: time for change. *Lancet.* 2013;381(9868):774-5.
18. Kaukonen KM, Bailey M, Pilcher D, Cooper DJ, and Bellomo R. Systemic inflammatory response syndrome criteria in defining severe sepsis. *N Engl J Med.* 2015;372(17):1629-38.
19. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA.* 2016;315(8):762-74.
20. Shankar-Hari M, Phillips GS, Levy ML, Seymour CW, Liu VX, Deutschman CS, et al. Developing a New Definition and Assessing New Clinical Criteria for Septic Shock: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA.* 2016;315(8):775-87.
21. Simpson SQ. New Sepsis Criteria: A Change We Should Not Make. *Chest.* 2016;149(5):1117-8.
22. Knaus WA, Sun X, Nystrom O, and Wagner DP. Evaluation of definitions for sepsis. *Chest.* 1992;101(6):1656-62.
23. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N Engl J Med.* 2001;345(19):1368-77.
24. Mouncey PR, Osborn TM, Power GS, Harrison DA, Sadique MZ, Grieve RD, et al. Trial of Early, Goal-Directed Resuscitation for Septic Shock. *NEJM.* 2015;372(14):1301-11.
25. Murdoch TB, and Detsky AS. The Inevitable Application of Big Data to Health Care. *Jama-Journal of the American Medical Association.* 2013;309(13):1351-2.
26. Mohammed M, Khan MB, and Bashier EBM. *Machine Learning: Algorithms and Applications.* Boca Raton, FL: CRC Press; 2017.
27. Jain AK, Murty MN, and Flynn PJ. Data clustering: A review. *Acm Computing Surveys.* 1999;31(3):264-323.
28. Marlin BM, Kale DC, Khemani RG, and Wetzel RC. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. Miami, Florida, USA: ACM; 2012:389-98.
29. Cerna AEU, Wehner G, Hartzel DN, Haggerty C, and Fornwalt B. Data Driven Phenotyping of Patients With Heart Failure using a Deep-learning Cluster Representation of Echocardiographic and Electronic Health Record Data. *Circulation.* 2017;136.
30. Knox DB, Lanspa MJ, Kuttler KG, Brewer SC, and Brown SM. Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome. *Intensive Care Med.* 2015;41(5):814-22.

31. Nowak RM, Reed BP, Nanayakkara P, DiSomma S, Moyer ML, Millis S, et al. Presenting hemodynamic phenotypes in ED patients with confirmed sepsis. *Am J Emerg Med.* 2016;34(12):2291-7.
32. Mayhew MB, Petersen BK, Sales AP, Greene JD, Liu VX, and Wasson TS. Flexible, cluster-based analysis of the electronic medical record of sepsis with composite mixture models. *J Biomed Inform.* 2018;78:33-42.
33. Hripcsak G, and Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association.* 2013;20(1):117-21.
34. Jensen PB, Jensen LJ, and Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics.* 2012;13(6):395-405.
35. Luo J, Wu M, Gopukumar D, and Zhao YQ. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights.* 2016;8:1-10.
36. Miotto R, Wang F, Wang S, Jiang XQ, and Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics.* 2018;19(6):1236-46.
37. LeCun Y, Bengio Y, and Hinton G. Deep learning. *Nature.* 2015;521(7553):436-44.
38. Hinton GE, and Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science.* 2006;313(5786):504-7.
39. Miotto R, Li L, Kidd BA, and Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep.* 2016;6:26094.
40. Beaulieu-Jones BK, and Moore JH. Missing Data Imputation in the Electronic Health Record Using Deeply Learned Autoencoders. *Pacific Symposium on Biocomputing 2017.* 2017:207-18.
41. Mazzone A, Dentali F, La Regina M, Foglia E, Gambacorta M, Garagiola E, et al. Clinical Features, Short-Term Mortality, and Prognostic Risk Factors of Septic Patients Admitted to Internal Medicine Units: Results of an Italian Multicenter Prospective Study. *Medicine (Baltimore).* 2016;95(4):e2124.
42. Ford DW, Goodwin AJ, Simpson AN, Johnson E, Nadig N, and Simpson KN. A Severe Sepsis Mortality Prediction Model and Score for Use With Administrative Data. *Crit Care Med.* 2016;44(2):319-27.
43. Drumheller BC, Agarwal A, Mikkelsen ME, Sante SC, Weber AL, Goyal M, et al. Risk factors for mortality despite early protocolized resuscitation for severe sepsis and septic shock in the emergency department. *J Crit Care.* 2016;31(1):13-20.
44. Zhang Z, Chen K, and Chen L. APACHE III Outcome Prediction in Patients Admitted to the Intensive Care Unit with Sepsis Associated Acute Lung Injury. *PLoS One.* 2015;10(9):e0139374.
45. Whittaker SA, Fuchs BD, Gaieski DF, Christie JD, Goyal M, Meyer NJ, et al. Epidemiology and outcomes in patients with severe sepsis admitted to the hospital wards. *J Crit Care.* 2015;30(1):78-84.

46. Rathour S, Kumar S, Hadda V, Bhalla A, Sharma N, and Varma S. PIRO concept: staging of sepsis. *J Postgrad Med.* 2015;61(4):235-42.
47. Roest AA, Tegtmeier J, Heyligen JJ, Duijst J, Peeters A, Borggreve HF, et al. Risk stratification by abbMEDS and CURB-65 in relation to treatment and clinical disposition of the septic patient at the emergency department: a cohort study. *BMC Emerg Med.* 2015;15:29.
48. Chollet F. *Deep Learning with Python.* Shelter Island, NY: Manning Publications; 2017.
49. Ioffe S, and Szegedy C. *arXiv e-prints*. 2015.
50. Maaten Lvd, and Hinton G. Visualizing data using t-SNE. *Journal of machine learning research.* 2008;9(Nov):2579-605.
51. Hopkins B, and Skellam JG. A New Method for determining the Type of Distribution of Plant Individuals. *Annals of Botany.* 1954;18(2):213-27.
52. Han J, Kamber M, and Pei J. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers Inc.; 2011.
53. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics.* 1987;20:53-65.
54. Tibshirani R, Walther G, and Hastie T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2001;63(2):411-23.
55. Caliński T, and Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods.* 1974;3(1):1-27.

Appendix

Table 6: Retained variables and % missing

Variable	% missing	Variable	% missing
ethnicity	0.0	medtype_BIOLOGICALS	19.2
gender	0.0	medtype_PSYCHOTHERAPEUTIC DRUGS	19.2
age	0.0	medtype_PRE-NATAL VITAMINS	19.2
vitals_hr__max	0.2	medtype_MUSCLE RELAXANTS	19.2
vitals_hr__min	0.2	medtype_ANTIDOTES	19.2
vitals_hr__mean	0.2	medtype_MISCELLANEOUS MEDICAL SUPPLIES, DEVICES, NON-DRUG	19.2
vitals_hr__last	0.2	medtype_INVESTIGATIONAL	19.2
vitals_hr__first	0.2	medtype_IMMUNOSUPPRESANT	19.2
vitals_sbp__first	0.3	medtype_HORMONES	19.2
vitals_sbp__last	0.3	medtype_HERBALS	19.2
vitals_sbp__mean	0.3	medtype_CARDIAC DRUGS	19.2
vitals_sbp__min	0.3	medtype_CARDIOVASCULAR	19.2
vitals_sbp__max	0.3	medtype_GASTROINTESTINAL	19.2
vitals_dbp__last	0.3	medtype_ELECT/CALORIC/H2O	19.2
vitals_dbp__mean	0.3	medtype_CNS DRUGS	19.2
vitals_dbp__min	0.3	medtype_COLONY STIMULATING FACTORS	19.2
vitals_dbp__first	0.3	medtype_EENT PREPS	19.2
vitals_dbp__max	0.3	medtype_DIURETICS	19.2
vitals_o2_sat__first	0.4	medtype_DIAGNOSTIC	19.2
vitals_o2_sat__max	0.4	medtype_BLOOD	19.2
vitals_o2_sat__last	0.4	medtype_ANALGESICS	19.2
vitals_o2_sat__mean	0.4	medtype_COUGH/COLD PREPARATIONS	19.2
vitals_o2_sat__min	0.4	medtype_ANTIHISTAMINE AND DECONGESTANT COMBINATION	19.2
vitals_rr__max	0.6	medtype_ANTIARTHRITICS	19.2
vitals_rr__first	0.6	medtype_ANTIASTHMATICS	19.2
vitals_rr__last	0.6	medtype_ANESTHETICS	19.2
vitals_rr__min	0.6	medtype_ANTIBIOTICS	19.2

vitals_rr__mean	0.6	medtype_ANTIHYPERGLYCEMICS	19.2
vitals_temp__max	1.5	medtype_ANTIINFECTIVES	19.2
vitals_temp__first	1.5	medtype_ANTIHISTAMINES	19.2
vitals_temp__last	1.5	medtype_ANTIINFECTIVES/MISCELLANEOUS	19.2
vitals_temp__min	1.5	medtype_CONTRACEPTIVES	19.2
vitals_temp__mean	1.5	medtype_ANTIPARKINSON DRUGS	19.2
altered	3.0	medtype_ANTIFUNGALS	19.2
vitals_o2_dependency__mean	4.3	medtype_ANTIPLATELET DRUGS	19.2
vitals_o2_dependency__max	4.3	medtype_ANTI-OBESITY DRUGS	19.2
vitals_o2_dependency__first	4.3	medtype_ANTICOAGULANTS	19.2
vitals_o2_dependency__last	4.3	medtype_ANTINEOPLASTICS	19.2
vitals_o2_dependency__min	4.3	rdw	41.5
vitals_o2_amount__max	5.0	wbc	41.5
vitals_o2_amount__first	5.0	hematocrit	41.5
vitals_o2_amount__last	5.0	mcv	41.5
vitals_o2_amount__min	5.0	mpv	41.5
vitals_o2_amount__mean	5.0	hemoglobin	41.5
use_etoh	5.1	rbc	41.5
use_illicit	5.1	platelets	41.5
smoking	5.3	mchc	41.5
pmh_arrhythmias	10.4	mch	41.5
pmh_cancer	10.4	anc	41.8
pmh_other_respiratory	10.4	lymphocytes	41.9
pmh_diabetes	10.4	absolute lymphocyte count	41.9
pmh_other_nutritional_endocrine_and_metabolic_disorders	10.4	neutrophils	41.9
pmh_maintenance_chemotherapy_radiotherapy	10.4	monocytes	42.0
pmh_chf	10.4	eosinophils	42.0
pmh_liver_disease_alcohol_related	10.4	basophils	42.0
pmh_chronic_obstructive_pulmonary_disease_and_bronchiectasis	10.4	calcium	43.8
pmh_immunity_disorders	10.4	chloride	43.8
pmh_hypertension_with_complications_and_secondary_hypertension	10.4	sodium	43.8
pmh_hiv_infection	10.4	co2	43.8

pmh_heart_disease	10.4	anion gap	43.8
pmh_fen	10.4	bun	43.8
pmh_thyroid_disorders	10.4	creatinine	43.8
pmh_kidney_disease	10.4	glucose	43.8
pmh_asthma	10.4	potassium	44.8
medtype_ANALGESIC AND ANTIHISTAMINE COMBINATION	19.2	vitals_gcs__max	59.1
num_meds	19.2	vitals_gcs__mean	59.1
medtype_ANTIVIRALS	19.2	vitals_gcs__last	59.1
medtype_VITAMINS	19.2	vitals_gcs__first	59.1
medtype_UNCLASSIFIED DRUG PRODUCTS	19.2	vitals_gcs__min	59.1
medtype_THYROID PREPS	19.2	total bilirubin	72.0
medtype_SMOKING DETERRENTS	19.2	lactate	81.7
medtype_AUTONOMIC DRUGS	19.2		
medtype_SKIN PREPS	19.2		
medtype_SEDATIVE/HYPNOTICS	19.2		

Table 7: Medication Type Categories

ANALGESIC AND ANTIHISTAMINE COMBINATION	ANTIPARKINSON DRUGS	GASTROINTESTINAL
ANALGESICS	ANTIPLATELET DRUGS	HERBALS
ANESTHETICS	ANTIVIRALS	HORMONES
ANTI-OBESITY DRUGS	AUTONOMIC DRUGS	IMMUNOSUPPRESANT
ANTIARTHRITICS	BIOLOGICALS	INVESTIGATIONAL
ANTIASTHMATICS	BLOOD	MISCELLANEOUS MEDICAL SUPPLIES, DEVICES, NON-DRUG
ANTIBIOTICS	CARDIAC DRUGS	MUSCLE RELAXANTS
ANTICOAGULANTS	CARDIOVASCULAR	PRE-NATAL VITAMINS
ANTIDOTES	CNS DRUGS	PSYCHOTHERAPEUTIC DRUGS
ANTIFUNGALS	COLONY STIMULATING FACTORS	SEDATIVE/HYPNOTICS
ANTIHISTAMINE AND DECONGESTANT COMBINATION	CONTRACEPTIVES	SKIN PREPS
ANTIHISTAMINES	COUGH/COLD PREPARATIONS	SMOKING DETERRENTS
ANTIHYPERGLYCEMICS	DIAGNOSTIC	THYROID PREPS
ANTIINFECTIVES	DIURETICS	UNCLASSIFIED DRUG PRODUCTS
ANTIINFECTIVES/MISCELLANEOUS	EENT PREPS	VITAMINS
ANTINEOPLASTICS	ELECT/CALORIC/H2O	