Identifying Structure in the MIMIC ICU Dataset

Zad Chin
Harvard University
Cambridge, MA, 02138
zadchin@college.harvard.edu

Shivam Raval
Harvard University
Cambrdige, MA, 02138
sraval@g.harvard.edu

Leo Anthony Celi
Massachusetts Institute of Technology
Cambridge, MA, 02138
lceli@mit.edu

Martin Wattenberg
Harvard University
Cambridge, MA, 02138
wattenberg@seas.harvard.edu

Finale Doshi-Velez
Harvard University
Cambridge, MA, 02138
finale@seas.harvard.edu

Abstract

The MIMIC-III dataset, containing trajectories of 40,000 ICU patients, is one of the most popular datasets in machine learning for health space. However, there has been very little systematic exploration to understand what is the *natural* structure of these data—most analyses enforce some type of top-down clustering or embedding. We take a bottom-up approach, identifying consistent structures that are robust across a range of embedding choices. We identified two dominant structures sorted by either fraction-inspired oxygen or creatinine — both of which were validated as the key features by our clinician co-author. Our bottom-up approach in studying the macro-structure of a dataset can also be adapted for other datasets.

1 Introduction

Understanding the structure of clinical data can help us identify important insights into disease subtypes and process variation. Within medicine, clustering techniques have been commonly applied to the diagnosis of breast cancer (Chen [2014]), Parkinson's disease (Polat [2012];Nilashi et al. [2016]), mental health and psychiatric disorders (Trevithick et al. [2015]), heart and diabetes diseases (Yilmaz et al. [2014]), among many others.

In this paper, we focus on understanding the natural structure of the MIMIC-III Intensive Care Units (ICU) electronic health record data set (Johnson et al. [2016]) via projection-based clustering methods. MIMIC-III contains timeseries of measures of 40,000 patients in the ICU. With 4,564 citations as of October 2022, it is one of the most commonly used data sets in machine learning and health. Thus, its structure is of inherent interest. However, to date, most works have imposed a top-down approach to finding structure—e.g. imposing a particular clustering method (Shea [2020]; Sharafoddini et al. [2021];Zhang et al. [2021]; Fang et al. [2020]) or creating a task-specific embedding (Beaulieu-Jones et al. [2018]; Galozy [2018]; Chaudhary et al. [2020]).

Most of the literature identified above flattens the temporal structure of MIMIC-III into summary statistics (e.g. mean, median, or extreme values) before clustering, and none take measurement decisions into account. MIMIC-III data consists of irregularly measured time series and the care

Workshop on Learning from Time Series for Health, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

of critically ill patients relies heavily on laboratory data, and by extension, the laboratory reference ranges associated with them (Tyler et al. [2018]). However, these laboratory reference ranges are typically created by surveying healthy outpatients (Rifai [2017]). Hence, although the strategy of flattening the time series and using standard scaling in machine learning is computationally convenient, it leads to a significant loss of information (potentially introducing bias and making certain clinical insights inaccessible) and also significant inconsistency in the results.

Moreover, for clinical questions which are already frequently asked and well-defined (e.g. predicting the course of a specific disease or patient outcome), need-driven clustering methods, such as those explored in previous studies, have shown promise for generating insights. However, there are also potentially important clinical questions “hidden" in the data which have never been asked (e.g. the existence of coherent, discrete subtypes of disease with independent progression mechanisms), and in that sense may be more novel. Furthermore, when we first cluster the data and then distinguish the characteristic of the clusters with the clinical labels of interest such as mortality, we are ignoring other possible informative factors that are distinguishing the clusters, which are important for the interpreting the resulting clustering.

In our work, we take the opposite approach: What structure in the MIMIC-III data is persistent across a range of processing choices? Specifically, we identify patterns that persist across a range of data standardizations and parameters for projection methods (in the context of tSNE, UMAP). Our approach provides 3 major benefits: (1) It summarizes the dataset in a tractable way, without imposing any labelling, and gives a "representative sample" of patients for physicians. (2) A more informative and comprehensive interpretation of the clusterings rather than coercing a supervised label (such as mortality) to the clustering (3) A pioneering study in how different clinical range affects the structure of the data relating to critically ill patient.

Within the range of our exploration, we find that the dominant structures found either by fraction-inspired oxygen (Fi02) or by creatinine levels—both of which were validated as key patient features by our clinical co-author and are known to be important in clinical decisions. Specifically, creatinine is an important indicator for physicians during the earlier stage of admission to the ICU while Fi02 is an important indicator for the later stage of diagnosis and treatments. Moreover, we found a much greater sensitivity to standardization approaches than projection parameters. Our work provides important insights about the MIMIC-III data set and suggests ways to identify robust structures in other clinical data sets.

2 Methods

Our MIMIC-III cohort is the first 48 hours of seven commonly measured features of each of 10,184 patients: The features are heart rate, respiratory rate, creatinine, hematocrit, fraction inspired oxygen, sodium and mean blood pressure. When seeking structure, we explored the effect of (a) different ways to standardize the data and (b) different projection parameters. We identified commonalities between the induced embeddings for further inspection.

2.1 Description of Different Standardization Methods

Different features have different natural ranges and variations. For example, the heart rate normally varies between 60 and 100 beats per minute (BPM); the respiratory rate varies between 12 and 18 breaths per minute (BPM). Thus, decisions have to be made regarding how to standardize the data. For each feature d, we consider standardizations of the form:

$$z_d = f((x_d - x_{d,\text{min}})/(x_{d,\text{max}} - x_{d,\text{min}})), \tag{1}$$

For the values of $x_{d,\text{min}}$ and $x_{d,\text{max}}$, we randomly sample values between *valid* values—a range that feasibly exists in the human physiology—and *normal* values—a range that that is normal for healthy outpatients, but not necessarily normal or safe for ICU patients. These values are in Table 1. We consider two forms for f: *continuous*, that is, $f(x) = x$, or *discrete*, that is $f(x) = -1$ if $x < 0$; $f(x) = 0$ if $x \in [0, 1]$, and $f(x) = 1$ if $x > 0$. Finally, we include the "standard" standardization $z_d = (x_d - \bar{x}_d)/\sigma_{x,d}$ based on the feature's mean and variance.

2.2 Description of Different Projection Parameters

We consider two types of projections, t-Stochastic Neighbor Embedding, t-SNE, (Van der Maaten and Hinton [2008]) and Uniform manifold approximation and projection, UMAP, (McInnes et al. [2018]). For t-SNE, we vary the perplexity, that parameter that impacts the resulting projections the most. We selected a range of values, with `perplexity` = $\{5, 30, 50, 80\}$. For UMAP, the different parameters we explored were `min_dist` and `n_neighbors`, where we tested `n_neighbors` = $\{5, 15, 50, 100\}$ and `min_dist` = $\{0.1, 0.25, 0.5, 0.99\}$. (Note: we also explored using PCA, but the linear projection did not reveal substructure in the data—see Appendix Figure 2.)

2.3 Embedding Similarity Metric

Each standardization and choice of parameters in projection resulted in an embedding of the data set in a two-dimensional space. To compare the similarities between the embeddings, we tested a total of 42 different settings: 5 discrete randomized ranges and 5 continuous randomized range for both t-SNE and UMAP, where the randomized clinical range is scaled with $e \sim$Unif(0,1) from valid range and normal range, with the detailed formula in Appendix A, UMAP `n_neighbors = 5, 15, 50, 100`, UMAP `min_dist = 0.1, 0.25, 0.5, 0.99` and t-SNE `perplexity = 5, 30, 50, 80` for both continuous and discrete range, resulted in a total of $10 + 10 + 3 \times (16) = 42$ different methods tested. For each of the methods, we compare the labelling that we get for each method of standardization and metrics, and compute a similarity matrix between them (put simply: How many points stay at the same clusters comparing two different methods). We visualize the similarity matrix with a heat map for clarity.

3 Results

Across the different standardization and projection parameters, we find two major structures.

Figure 7 and 8 shows the similarities between the embeddings across the different standardization and projection approaches for 10 and 100 randomized ranges respectively. We see that embeddings cluster into two main groups, distinguished by

- Fraction Inspired Oxygen, for the data that are continuously standardized with valid range or randomized continuous range
- Creatinine, for the data that is only discretely standardized against the normal range

Examples of projections from each group are given in Figures 3b and 3c, as well as the associated timeseries. For embedding that shows interesting structures and visually identifiable clusters, we aim to uncover the important features that result in the particular structure. Statistically, we observed that for the Continuous-Normal scheme, we observe that cluster 1 (the small blue cluster identified in Figure 3b and 4b) has a mean and median Fi02 $= 0.21$ across 24 hours while the other clusters have FiO2 values maintained at a mean of 0.8 and a median of 0.6 for the first 5 hours at maintained at a mean and median of 0.5 after the 5th hour, as shown in the Figure 5a. For the Discrete-Normal scheme, we observe that cluster 1 (the small blue cluster identified in Figures 3c and 4c) a mean and median creatinine≈ 0.8mg/dL across 48 hours while cluster 2 has creatinine > 0.8mg/dL across 48 hours, as shown in the Figure 5b.

Separation Based on Features We further validated our result using the Simple Decision Tree Classifier to investigate whether there are simple human-understandable rules that can be inferred from this structure that may reliably predict whether a data point may lie in a particular cluster. We observe that a single-layer decision tree is sufficient to find a distinguishing feature between the cluster that cleanly separates more than 97% of the points inside a cluster.

The decision tree result can be found in Figure 1. For both cases where we observe identifiable clusters, we find a decisive feature that separates the clusters and is consistent with the statistical analysis. For the Continuous-Valid scheme, the most important feature for cluster separation is Fraction-Inspired Oxygen at hour 10, the decisive value being 0.24 and for the Discrete-Normal scheme, it is Creatinine at hour 20, with the decisive value being 1.35 mg/dL. For the Continuous-Valid scheme, it is clear that patients with lower values of FiO2 are in one small cluster, with the other cluster having larger values.

```
|--- fraction_inspired_oxygen10 <= 0.24
|   |--- class: 1
|--- fraction_inspired_oxygen10 >  0.24
|   |--- fraction_inspired_oxygen16 <= 0.27
|   |   |--- heart_rate8 <= 91.50
|   |   |   |--- class: 2
|   |   |--- heart_rate8 >  91.50
|   |   |   |--- class: 1
|   |--- fraction_inspired_oxygen16 >  0.27
|   |   |--- class: 2
```

```
|--- creatinine21 <= 0.50
|   |--- creatinine17 <= 0.50
|   |   |--- class: 2
|   |--- creatinine17 >  0.50
|   |   |--- class: 1
|--- creatinine21 >  0.50
|   |--- creatinine30 <= 0.50
|   |   |--- class: 1
|   |--- creatinine30 >  0.50
|   |   |--- class: 1
```

Figure 1: Decision Tree Output: The first level features important for separating clusters are clinically relevant in early stages of treatment

Effect of Standardization Methods on Distribution We further investigated why different standardization methods give rise to different clustering and structure by investigating the distribution plot for the identified feature. From Figure 6, we observed that the standardization methods that produce meaningful results are those capture the distribution of the original data closely.

Clinical Relevance of Features In the critical care setting, Fraction Inspire Oxygen (Fi02) describes the population of patients that are on/off respirators, with patients on respirators having higher FiO2. Creatinine, an important feature about kidney function, has been identified by the medical domain expert to be crucial for physicians' clinical decisions in the early stages of patients' treatment outcomes.

The "Typical" Ways To Standardize Lead to Very Different Projections. Critically ill patients have numerous laboratory abnormalities. The laboratory reference intervals that define normal values established by sampling healthy outpatients might not provide insights. While we checked for robust patterns across a different range of choices as well as across different models, we found out that the typical ways of standardizing, which is based on the feature's mean-variance, as well as standardizing against normal ranges, resulted in the loss of structure in the projection, as shown in figure 3e, 3a, 4e, 4a. Hence, future research should be more careful in selecting a clinical range to standardize against.

4 Conclusion

In this work, we identified that across seven different combinations of ways to standardize and project the MIMIC-III data, two key structures emerged, one dependent on fraction-inspired oxygen and one dependent on creatinine. While we explored the effect of standardizations in depth, there are additional directions one could explore to identify if other key structures exist. In particular, we always assumed that timesteps aligned in the 48-hour timeseries. Future work could more closely relate outcomes to the topologies we discovered, further explore the substructures in our projections, and explore ways beyond standardization to compare different patient timeseries.

5 Acknowledgement

We would like to thank our colleague, Andrew Ross, for his personal notes on literature reviews on different clustering algorithms on the MIMIC-III data set. Funding for FDV was made possible in part by R01MH123804. The views expressed do not necessarily reflect the official policies of the NIH; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

References

B. K. Beaulieu-Jones, P. Orzechowski, and J. H. Moore. Mapping patient trajectories using longitudinal extraction and deep learning in the mimic-iii critical care database. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pages 123–132. World Scientific, 2018.

K. Chaudhary, A. Vaid, Á. Duffy, I. Paranjpe, S. Jaladanki, M. Paranjpe, K. Johnson, A. Gokhale, P. Pattharanitima, K. Chauhan, et al. Utilization of deep learning for subphenotype identification in sepsis-associated acute kidney injury. *Clinical Journal of the American Society of Nephrology*, 15(11):1557–1565, 2020.

C.-H. Chen. A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. *Applied Soft Computing*, 20:4–14, 2014.

S. Curto, J. P. Carvalho, C. Salgado, S. M. Vieira, and J. M. Sousa. Predicting icu readmissions based on bedside medical text notes. In *2016 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, pages 2144–a. IEEE, 2016.

T. Desautels, R. Das, J. Calvert, M. Trivedi, C. Summers, D. J. Wales, and A. Ercole. Prediction of early unplanned intensive care unit readmission in a uk tertiary care hospital: a cross-sectional machine learning approach. *BMJ open*, 7(9):e017199, 2017.

C.-H. Fang, V. Ravindra, S. Akhter, M. Adibuzzaman, P. Griffin, S. Subramaniam, and A. Grama. Identifying and analyzing sepsis states: A retrospective study on patients with sepsis in icus. *arXiv preprint arXiv:2009.10820*, 2020.

A. S. Fialho, F. Cismondi, S. M. Vieira, S. R. Reti, J. M. Sousa, and S. N. Finkelstein. Data mining using clinical physiology at discharge to predict icu readmissions. *Expert Systems with Applications*, 39(18):13158–13165, 2012.

A. Galozy. Towards understanding icu procedures using similarities in patient trajectories: An exploratory study on the mimic-iii intensive care database, 2018.

C. Hennig, M. Meila, F. Murtagh, and R. Rocci. *Handbook of cluster analysis*. CRC Press, 2015.

A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

B. Mirkin. *Clustering for data mining: a data recovery approach*. Chapman and Hall/CRC, 2005.

M. Nilashi, O. Ibrahim, and A. Ahani. 'accuracy improvement for predicting parkinson's disease progression,"sci, 2016.

K. Polat. Classification of parkinson's disease using feature weighting method on the basis of fuzzy c-means clustering. *International Journal of Systems Science*, 43(4):597–609, 2012.

N. Rifai. *Tietz textbook of clinical chemistry and molecular diagnostics*. Elsevier Health Sciences, 2017.

G. Ritter. *Robust cluster analysis and variable selection*. CRC Press, 2014.

A. Sharafoddini, J. A. Dubin, and J. Lee. Identifying subpopulations of septic patients: A temporal data-driven approach. *Computers in Biology and Medicine*, 130:104182, 2021.

A. Shea. *Patient clustering using electronic medical records*. PhD thesis, Massachusetts Institute of Technology, 2020.

L. Trevithick, J. Painter, and P. Keown. Mental health clustering and diagnosis in psychiatric in-patients. *BJPsych Bulletin*, 39(3):119–123, 2015.

P. D. Tyler, H. Du, M. Feng, R. Bai, Z. Xu, G. L. Horowitz, D. J. Stone, and L. A. Celi. Assessment of intensive care unit laboratory values that differ from reference ranges and association with patient mortality and length of stay. *JAMA network open*, 1(7):e184521–e184521, 2018.

A. Ultsch. Self organizing neural networks perform different from statistical k-means clustering. *Proc. GfKl, Basel, Swiss*, 1995.

L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

V. van Unen, N. Li, I. Molendijk, M. Temurhan, T. Höllt, A. E. van der Meulen-de, H. W. Verspaget, M. L. Mearin, C. J. Mulder, J. van Bergen, et al. Mass cytometry of the human mucosal immune system identifies tissue-and disease-associated immune subsets. *Immunity*, 44(5):1227–1239, 2016.

J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(2), 2010.

M. Wattenberg, F. Viégas, and I. Johnson. How to use t-sne effectively. *Distill*, 2016. doi: 10.23915/distill.00002. URL `http://distill.pub/2016/misread-tsne`.

N. Yilmaz, O. Inan, and M. S. Uzer. A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases. *J Med Syst*, 38:48–59, 2014.

K. Zhang, Y. Wang, J. Du, B. Chu, L. A. Celi, R. Kindle, and F. Doshi-Velez. Identifying decision points for safe and interpretable reinforcement learning in hypotension treatment. *arXiv preprint arXiv:2101.03309*, 2021.

A Appendix

A.1 Figures & Tables

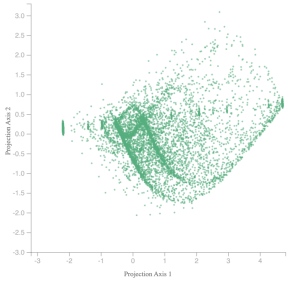
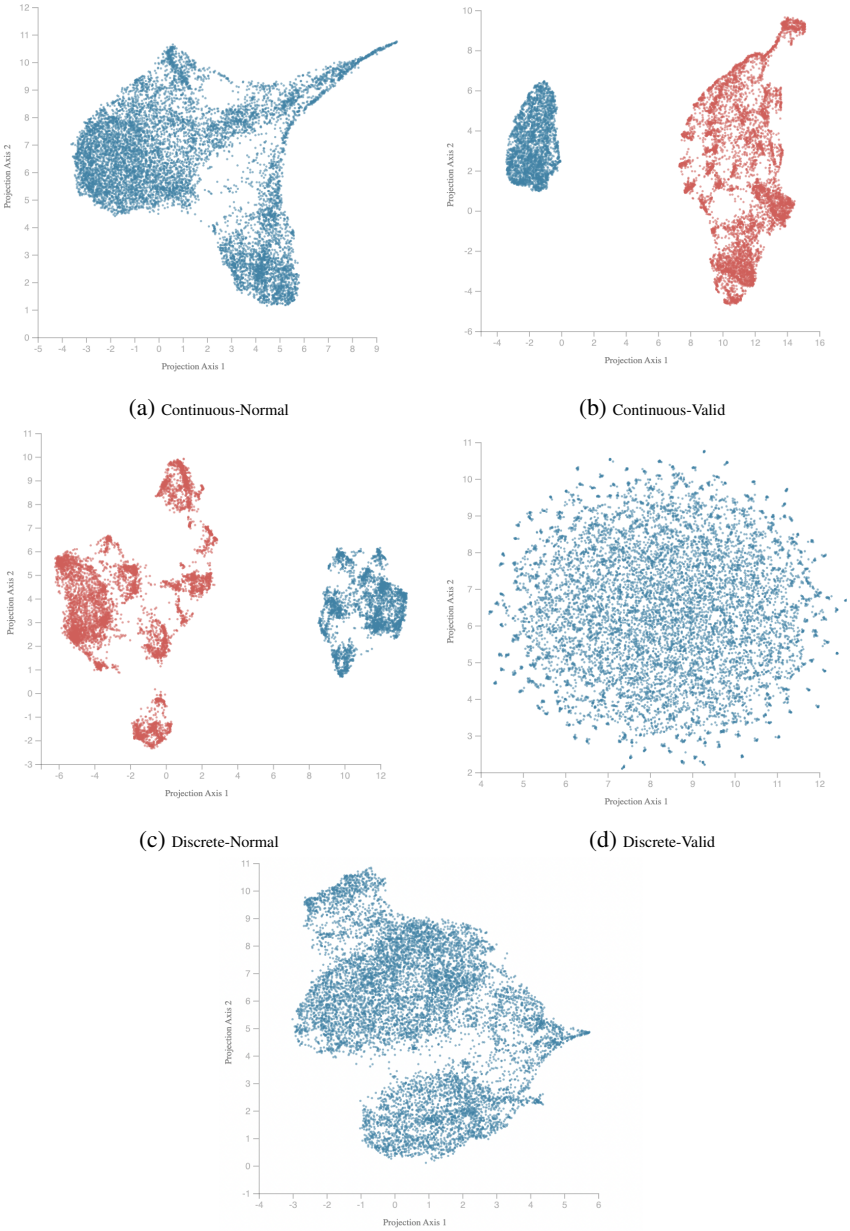


Figure 2: PCA Projections for Continuous Valid Scheme, other schemes result in qualitatively similar plots



(a) Continuous-Normal

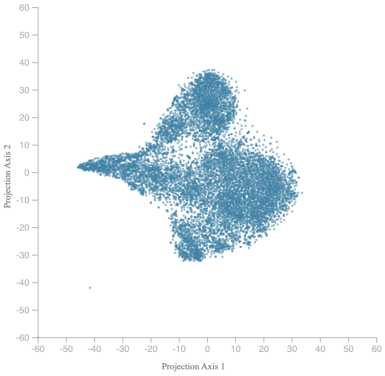
(b) Continuous-Valid

(c) Discrete-Normal

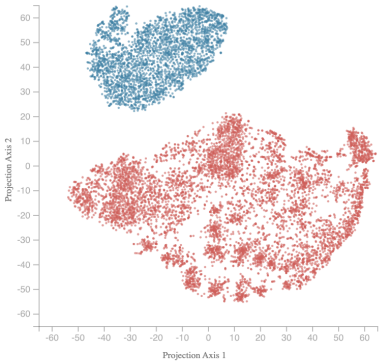
(d) Discrete-Valid

(e) Standard Scaling

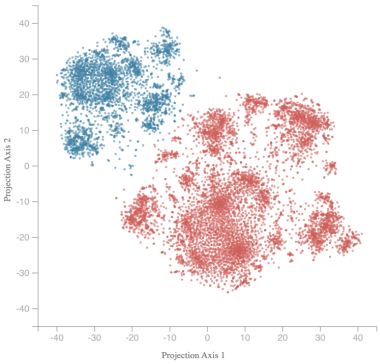
Figure 3: UMAP Projections for different standardization schemes. The Continuous-Valid and Discrete-Normal schemes have separable clusters (colored in red and blue). The Continuous-Normal and Continuous-Standard Scaled schemes have interesting substructures but no clear clustering. The choice of feature range for standardization has important effects on the structure of the resulting projections.



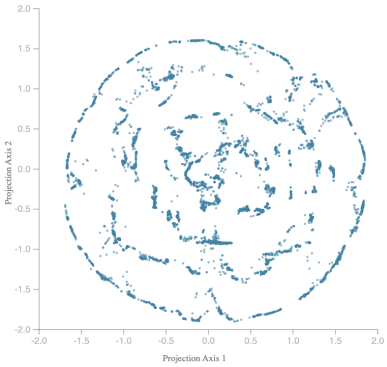
(a) Continuous-Normal



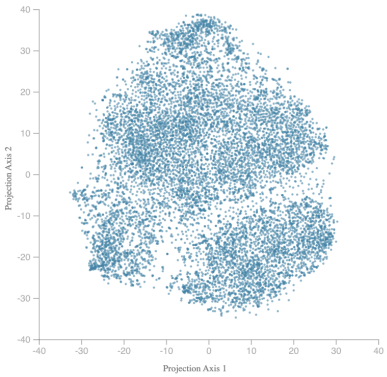
(b) Continuous-Valid



(c) Discrete-Normal



(d) Discrete-Valid



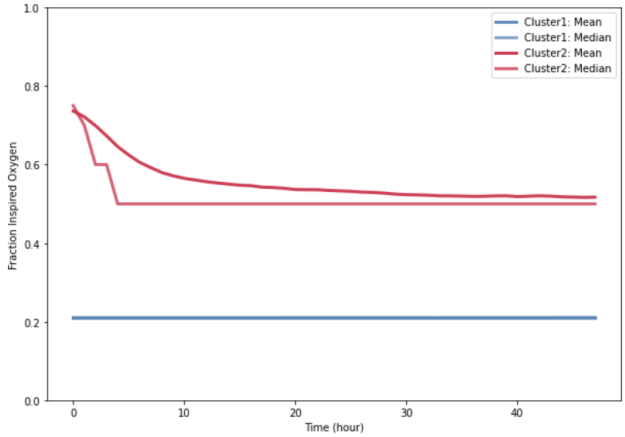
(e) Continuous Standard Scaled

Figure 4: TSNE Projections for different standardization schemes. Similar to UMAP Projections, the Continuous-Valid and Discrete-Normal Schemes have separable clusters (colored in red and blue).

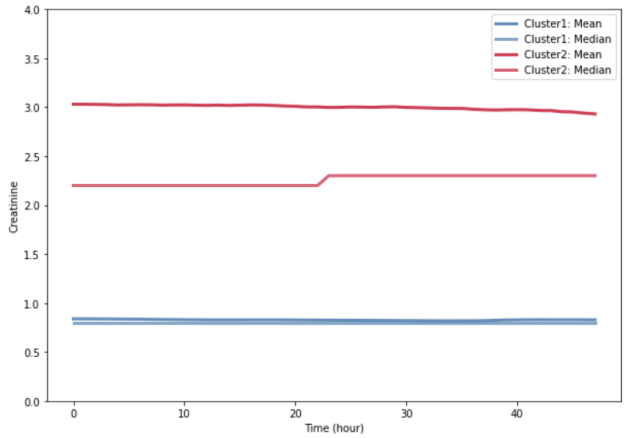
A.2 Data

In this study, we used data from the Medical Information Mart for Intensive Care (MIMIC-III) database, which is a large, freely available database comprising de-identified health-related data associated with over forty thousand patients who stayed in critical care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. (Johnson et al. [2016]). For our study, we sample from the MIMIC-III dataset based on the following criteria:

- Patients with minimum 48 hours of records;
- Patients with age over 18 years old;
- Patients with complete age, gender, ethnicity group, insurance, readmission, and mortality record



(a) FiO2 Timeseries



(b) Creatinine Timeseries

Figure 5: Timeseries for different clusters for UMAP: For Continuous-Valid, the clusters are clearly separated by Fi02, with cluster 1 having mean and median of 0.21 across the 48 hours while cluster 2 has FiO2 mean and median approximately of 0.5 after the first 10 hours. For Discrete-Normal, clusters are separated by Creatinine, with Cluster 1 having a mean and median below 1.0 while cluster 2 having a mean below 2.5 and above 2.0 and a median of 3.0 across 48 hours.

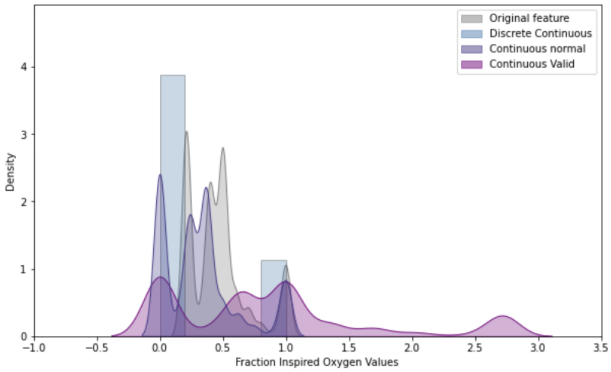
A total of 10,184 patients fit the criteria and are considered. For each of the patients in the selected cohort, we have further selected seven laboratory test measurements for their first 48 hours in the ICU. These 7 clinical features are a key indicators of important organ functions and have been commonly mentioned in previous research in the ICU. These clinical features are respiratory rate, heart rate, mean blood pressure, sodium, creatinine, fraction inspired oxygen, and hematocrit (Fialho et al. [2012]Curto et al. [2016]Desautels et al. [2017]).

For each feature, we record its value at each hour of the patients' first 48 hours of admission to the ICU. If a value is not found, we carry forward the value from the hour before. To minimize model variance in our study, we also remove the patients with outliers values, which are values that are physically impossible to achieve by human standards, as shown in Table 1. All in all, we get a data frame that has a dimension of $10,184 \times 336$.

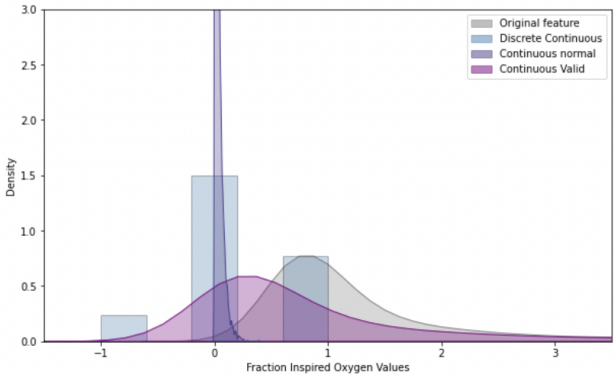
A.3 An Deeper view on Methodology

Projection Based Dimensionality Reduction Algorithm

For high-dimensional datasets, finding global structure directly is difficult as in high-dimensional, the curse of dimensionality applies, where there is a loss of meaningful differentiation between similar



(a) Fraction Inspired Oxygen at hour 10



(b) Creatinine at hour 20

Figure 6: Effect of Standardization on Distribution of important features: For Fraction Inspired Oxygen at Hour 10, we observe that the continuous valid distribution closely captures the peaks observed in the original feature but in a more normalized manner. For Creatinine we observe that the original feature has a slightly skewed normal curve, which is similarly observed in the discrete-normal case. Our hypothesis is standardization methods that closely capture the distribution of the original data produce meaningful results.

Vitals	Valid Low	Valid High	Normal Low	Normal High	Ourlier Low	Outlier High
Heart Rate	0	350	60	100	0.0	390.0
Respiratory Rate	0	300	12	18	0.0	330.0
Creatinine	0.10	60	0.59	1.35	0.0	66.0
Hematocrit	0	75	37	52	0.0	100.0
Fraction Inspired Oxygen	0.21	1.00	0.21	0.5	0.2	1.1
Mean Blood Pressure	14	330	70	100	0.0	375.0
Sodium	50	225	135	145	0.0	250.0

Table 1: Valid and Normal clinical laboratory values for different features.

and dissimilar objects observed. Dimension reduction techniques reduce the dimension of the input space to facilitate finding structures in the data. A novel method for improving structure finding in the high-dimensional dataset is using projection-based clustering. A projection is used as a method for visualizing high-dimensional data in 2D space such that the distance and the density of the data are captured. One approach is to use projections as conventional methods of dimensionality reduction for information visualization in order to transform high-dimensional data into low-dimensional space (Venna et al. [2010]). Thus, in our paper, we experimented with 3 unsupervised dimension reduction techniques, namely Principal Component Analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). We projected our data to two-dimensional, which resulted in a scatter plot that we can analyze. Scatter plots

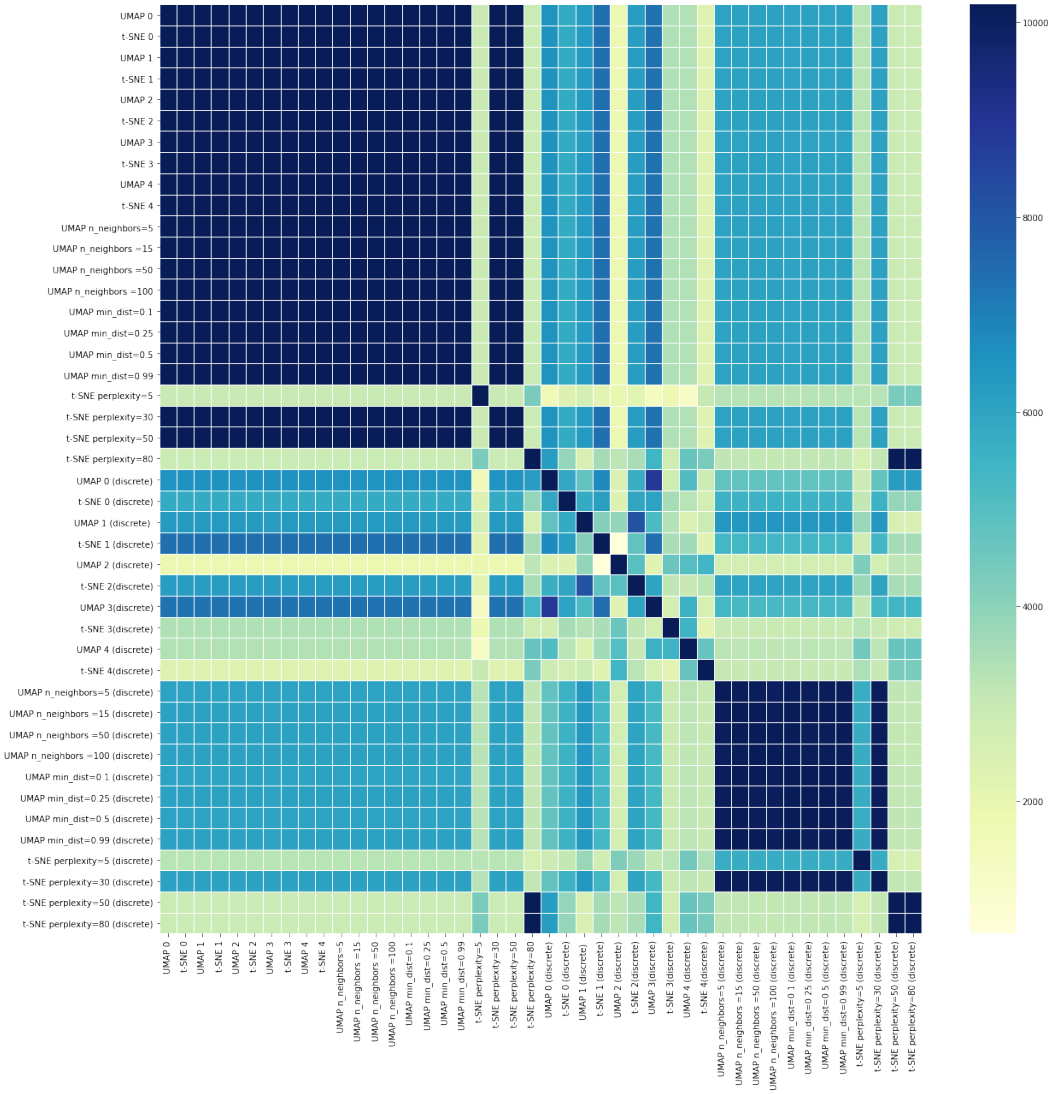


Figure 7: Heatmap for 10 randomized sample ranges for both discrete and continuous. The top left corner dark blue clusters represent that (1) different randomized continuous range shows consistency throughout different models and UMAP and t-SNE parameters (2) consistency of clustering is observed with different UMAP and t-SNE parameters for continuous range, with the exception of t-SNE perplexity = 5 and 80. The bottom right dark blue clusters represent that for discrete ranges have consistent labelling across different UMAP and t-SNE parameters, except for t-SNE perplexity - 5, 30 and 80. The other inconsistency in the heatmap indicates that discrete methods of standardization are sensitive to the scale of the clinical ranges. Similar observation can be found when tested against 100 randomized samples as shown in Figure 8 below.

generated by a projection method remain the state-of-the-art approach in cluster analysis to visualize data structures. (e.g., Mirkin [2005]; Ritter [2014]; Hennig et al. [2015]). We will discuss each of the dimension reduction techniques mentioned above below.

PCA is the oldest and most commonly used linear projection technique that reduces the dimensions of the data. PCA works by dropping the least important variables and preserving the most valuable variables, projecting them onto a two-dimensional space. However, for very high dimensional data, PCA may fail to find interesting structures and patterns. This is because since all linear projections are orthonormal rotations of the data coordinates, clusters that are linear nonseparable entanglements, such as the Chainlink data (Ultsch [1995]) or overlapping convex hulls, cannot be separated. With this

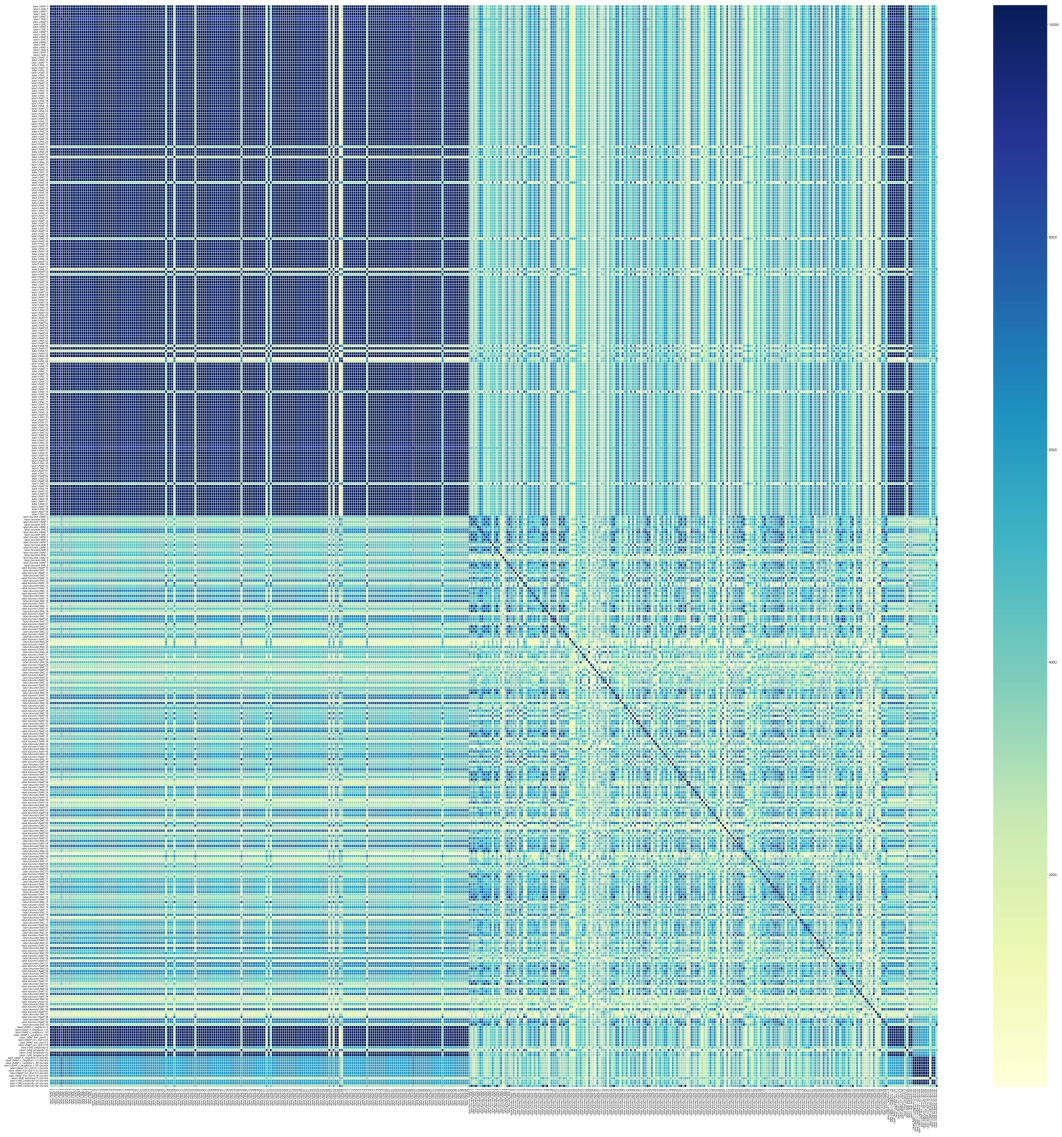


Figure 8: Heatmap for 100 randomized sample ranges for both discrete and continuous. Similar observations from Figure 7 can be observed in the 100 randomized ranges as well, indicating the consistency in the result we found. A clearer analysis on each analysis can be found in the Figure 9 below.

type of projection, it is unavoidable that at some locations remote data are erroneously superimposed in the output space.

In such cases, non-linear stochastic methods like t-SNE and UMAP are particularly beneficial. t-SNE uses points in high-dimensional space to find a faithful representation of those points in a lower-dimensional space. The algorithm is non-linear and adapts to the underlying data, performing different transformations on different regions. t-SNE however suffers from limitations such as loss of large-scale information (the inter-cluster relationships), slow computation time and inability to meaningfully represent very large datasets (van Unen et al. [2016]). UMAP, created by McInnes and Healy (McInnes et al. [2018]), uses the data to construct the initial high-dimensional weighted graph with the nearest neighboring points in a local region. Compared to t-SNE it preserves as much of the local and more of the global data structure, with a shorter runtime.

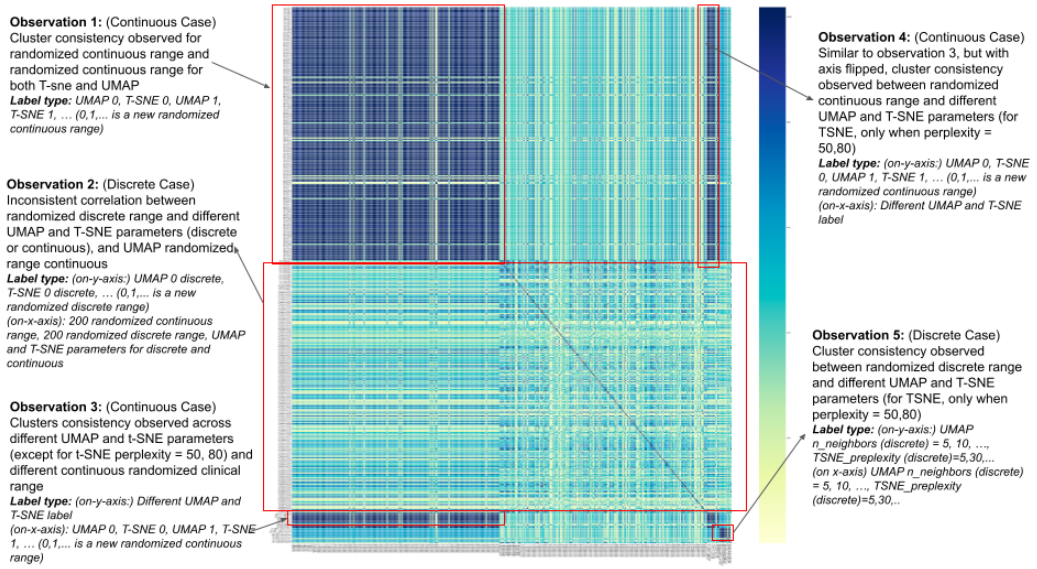


Figure 9: Heatmap Analysis for 100 different randomized ranges. 5 Observations, coherent with the one observed with 10 different randomized ranges above in Figure 7.

A.4 Testing Persistency of Clustering

Once clusters are identified, we want to know how persistent the clusterings identified are across different methods and parameters. Borrowing ideas from persistent homology, we test how sensitive are these clusters toward the clinical range identified in Table 1, as well as towards different model hyperparameters.

Randomized Clinical Range

To test how the role that the predefined clinical range in Table 1, we defined a randomized clinical range, where we scale the clinical range with $e \sim Unif(0, 1)$ as follows:

$$x_{\text{min-random}} = x_{\text{min-valid}} + e(x_{\text{min-normal}} - x_{\text{min-valid}}) \tag{2}$$

$$x_{\text{max-random}} = x_{\text{max-normal}} + e(x_{\text{max-valid}} - x_{\text{max-normal}}), \tag{3}$$

where the values for min-normal, max-normal, min-valid and max-valid can be found in 1 above. With the above range, we redefine continuous and discrete standardization methods with the randomized clinical range as follows:

- **Continuous Randomized:** Continuously standardized against randomized range, defined in equation (1) & (2) in a continuous manner where

$$z = (x - x_{\text{min-random}})/(x_{\text{max-random}} - x_{\text{min-random}}),$$

- **Discrete Randomized:** Discretely standardized against randomized range, defined in equation (1) & (2):

$$f(x) = \begin{cases} -1, & \text{for } x < x_{\text{min-random}} \\ 0, & \text{for } x_{\text{min-random}} \leq x \leq x_{\text{max-random}} \\ 1, & \text{for } x > x_{\text{max-random}} \end{cases}$$

Model Parameters

Previous works on projections using nonlinear and stochastic methods show that they are sensitive to the hyperparameters of the techniques (Wattenberg et al. [2016]). Therefore, we tested against different model hyperparameters across different ranges to check how they affect the projections. Namely,

- UMAP `n_neighbors`: UMAP `n_neighbors` parameters investigate local versus global structure in the data. A small `n_neighbors` will force the algorithm to concentrate on very local structure, and a larger `n_neighbors` will cover broader structure of the data. In our case, we tested `n_neighbors` = $\{5, 15, 50, 100\}$
- UMAP `min_dist`: UMAP `min_dist` affects how tightly points are allow to pack together. A lower `min_dist` will allow us to have clumpier embeddings, which lead us to discover finer topological structure, whereas a higher `min_dist` provides a global view. In our case, we tested `min_dist` = $\{0.1, 0.25, 0.5, 0.99\}$
- TSNE `perplexity`: TSNE `perplexity` balance attention between local and global structure, and different in `perplexity` values have found to produce different clusterings (Wattenberg et al. [2016]). In our paper, we tested `perplexity` = $\{5, 30, 50, 80\}$

Once we identified the standardization methods and model parameters to test against to ensure the consistency of the structure identified, we combine them to test the consistency. Namely, we tested a total of 42 different parameters: 5 discrete randomized ranges and 5 continuous randomized range for both t-SNE and UMAP, UMAP `n_neighbors = 5, 30, 50, 80` UMAP `min_dist = 0.1, 0.25, 0.5, 0.99` and t-SNE `perplexity = 5, 15, 50, 100` for both continuous and discrete range, resulted in a total of $10 + 10 + 3 * (16) = 42$ different methods tested. For each of the methods, we compare the labeling that we get for each method of standardization and metrics, and compute a similarity matrix between them (put simply: How many points stay at the same clusters comparing two different methods). We visualize the similarity matrix with a heat map.

A.5 Additional Results