VARIATIONAL MODELLING OF TEMPORAL EHRS FOR PHENOTYPIC CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

The recent availability of Electronic Health Records (EHR) has allowed for the development of algorithms predicting inpatient risk of deterioration and trajectory evolution. However, prediction of disease progression with EHR is challenging since these data are sparse, heterogeneous, multi-dimensional, and multimodal time-series. Clustering is an alternative approach to identify similar groups within the patient cohort that can be leveraged to predict patient status evolution. In particular, the identification of phenotypically separable clusters has proven very useful in improving healthcare delivery. Some clustering models have been proposed to identify phenotypically separable clusters; however, these have struggled in clinical settings characterised by several, highly imbalanced event classes. To that end, we propose a generative model to cluster EHR data based on identifying clinically meaningful phenotypes with regard to patient outcome prediction and physiological trajectory. We introduce a novel probabilistic method that is capable of simultaneously a) generating observation data, b) modelling temporal cluster assignments, and c) predicting admission outcomes. Our results show performance similar to state-of-the-art methods, with increased clustering separability, and capability to generate observation data.

1 INTRODUCTION

The healthcare domain is comprised of a variety of settings where cohorts can be separated into multiple patient subgroups, largely distinguished by differences in interventions, response to treatment, etc. Identification of such subgroups can be invaluable to enhance understanding of underlying patient physiological status (including potential clinical conditions or predictive trends), which consequently allows for improvement in care quality and treatment, (e.g. Turner et al. (2015); Vogelmeier et al. (2018)). We tackle the task of identifying clusters with different phenotypes in medical cohorts. Phenotypes typically contain key clinical properties that relate to a patient's future health status - as such, they might not be fully available for new patients. In this paper, we define a cluster phenotype as a combination of a) the physiological trajectory evolution profile of patients within the cluster, and b) the distribution of clinical outcomes (unseen variables of clinical interest) within the cluster.

Data from clinical environments is generally recorded in EHRs. Modelling EHRs, however, is extremely challenging due to severe heterogeneity, Shickel et al. (2017). Clinical variables comprise a mixture of temporal and static (i.e. time-independent such as age, and gender) features, with different signal characteristics. Moreover, EHR temporal variables are multi-modal (i.e. recorded with different collection methods). Time-series features may also not be aligned with respect to observation collection time and may show disparate sampling rates, missing data rates, noise distribution, and evolution trends. Traditional clustering models such as K-means (Lloyd, 1982) and Time-Series K-Means (TSKM, Tavenard et al. (2020), have proven to be lacklustre in dealing with temporal data due to their failure in capturing existing time-dependent feature relationships, Raykov et al. (2016). Recently, Deep Learning (DL) architectures such as Recurrent Neural Networks (RNN, Rumelhart et al. (1985)) or Variational Auto-Encoders (VAE, Fortuin et al. (2019)) have shown promise when applied to time-series data on a variety of domains due to increased modelling capacity. AC-TPC (Lee & Van Der Schaar, 2020) and CAMELOT (Aguiar et al., 2022), in particular, successfully leveraged (future) event information to aid in cluster formation.

There has been an increased focus on variational and generative approaches to aid in learning representations as they are more expressive. For multi-dimensional temporal data, Chung et al. (2015) propose a Gaussian Mixture Model-Variational Recurrent Neural Network (GMM-VRNN) model for clustering highly structured and noisy data. On the other hand, the authors in Jun et al. (2021) introduce a variational recurrent model for estimating missing feature distribution in EHR data. Our contributions include an end-to-end DL-supervised method to identify phenotypically separable clusters in EHR data and a probabilistic model that is capable of a) generating clinical observations data, b) clustering clinical observations over time, and c) modelling admission outcomes.

2 DATA AND PRE-PROCESSING

We validate our proposed model on 2 very distinct medical settings representing disparate environments within the healthcare domain. Our first dataset is a proprietary, secondary care dataset, denoted as **HAVEN**, from the United Kingdom (UK), which contains observations from hospital ward admissions of patients suffering from chronic respiratory disease. It is also the data that mainly drove model development where it pertains to clinical application and feedback. Separately, we also consider a freely-available, public dataset of patients admitted to an Emergency Department (ED) within a hospital in the United States. The latter provides supportive evidence of our model's generalisability and increases the reproducibility of our work. We note this dataset was derived from the MIMIC-IV-ED database (Johnson et al., 2021), and we refer to it solely as **MIMIC** henceforth.

We note that both datasets share common key characteristics. For a single patient, the resulting trajectories (i.e. collection of observations for the same clinical variable over time) display a) heterogeneity, b) multi-modality, and differences in c) noise distributions, d) sampling rates, and e) missing value rates. Such properties are common across EHR settings, and ensure that prediction or learning of useful representations is challenging. Both datasets also contain information about the occurrence of adverse events at the end of an admission, which allows the formation of outcome classes. The outcome classes are indicative of a patient's underlying health status, but are not completely representative, in the sense that the sub-cohorts characterised by the same outcome are not separable (and hence motivating the concept of phenotyping). Finally, the two settings considered here lead to severe imbalances between outcome classes. Pre-processing was conducted similarly with respect to both datasets, and a further description is provided in the Appendix, Section A1.

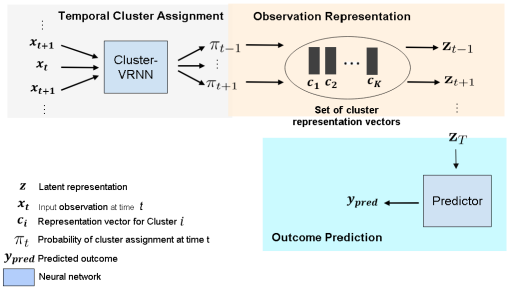
3 METHODS

We propose a novel generative model that is capable of identifying separable clusters. A sketch of our proposed methodology is displayed in Figure 1(a), and a sketch of the Cluster-VRNN network is shown in Figure 1(b)[1]. Consider an input dataset of the form $\mathbb{X} = \{\{\mathbf{x}_{n,t}\}_{t=1}^{T_n}\}_{n=1}^{N}$, where N is the number of patients, and T_n is the number of temporal observations for the n-th patient. Each $\mathbf{x}_{n,t} = [x_{n,t}^1, ..., x_{n,t}^f] \in \mathbb{R}_{N_f}$ is a *set of observations* with at most N_f values, where N_f is the number of features (for instance, representing heart rate, age, etc.). We denote $[x_{n,1}^i, ..., x_{n,T_n}^i]$ as a trajectory. For each patient, we write $\mathbf{x}_n$ to represent the complete sequence of corresponding observations, and $\mathbf{y}_n \in \mathbb{R}_{N_o}$ is a one-hot vector corresponding to the patient outcome.

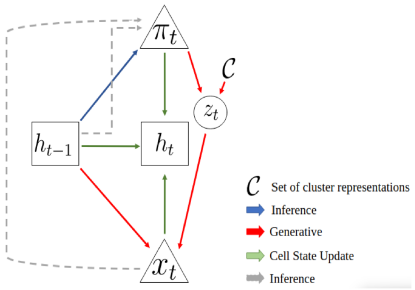
Our representation framework builds on the work of Chung et al. (2015) to model the joint probability $p(\mathbf{x}_{\leq T}, \boldsymbol{\pi}_{\leq T}, \mathbf{y})$, where $\boldsymbol{\pi}_t \in \Delta_{K-1}$ are latent variables representing cluster probability assignments, and K is the total number of clusters. The RNN represents $\mathbf{x}_{\leq t}, \boldsymbol{\pi}_{\leq t}$ via a single cell state vector $\mathbf{h}_t \in \mathbb{R}^s$. Each cluster is represented according to a normal distribution with mean $\boldsymbol{\mu}_k \in \mathbb{R}_l$ and variance $\boldsymbol{\sigma}_k^2 \boldsymbol{I}$, where $\mathbf{I}_l$ is the identity matrix, and l is the dimension of the latent space. Finally, we also associate a (categorical) outcome vector $\mathbf{y}$ for each patient. We define the VRNN component models as follows:

- **Prior** $f_p(t) = p(\boldsymbol{\pi}_t \mid \mathbf{x}_{<t}, \boldsymbol{\pi}_{<t}) \sim \mathcal{D}(\boldsymbol{\alpha}_{p,t}), \quad \boldsymbol{\alpha}_{p,t} = \text{ReLU}(g_1(\mathbf{h}_{t-1}))$;
- **Generation** $f_g(t) = p(\mathbf{x}_t \mid \mathbf{x}_{<t}, \boldsymbol{\pi}_{<t}) \sim \mathcal{N}(\mu_g, \sigma_g^2 \mathbf{I}_\mathbf{f})$, where $\mu_g, \sigma_g = g_2(\mathbf{h}_{t-1}, g_3(\mathbf{z}_t))$ and $\mathbf{z}_t = \sum_k \boldsymbol{\mu}_k \boldsymbol{\pi}_t^k$;

[1] A repository for our code will be shared after review



(a) General overview of the proposed model. Time-series input data passes through a Cluster-Variational Recurrent Neural Network (VRNN), which models a generative approach for both data and latent variables. At each time step, the VRNN estimates a cluster assignment probability, $\boldsymbol{\pi}_t$, which is used to obtain a representation of the input data, $\mathbf{z}_t$. The representation at the last time-step is used to predict the admission outcome, $\hat{\mathbf{y}}_{\mathbf{pred}}$, through another neural network.



(b) Sketch diagram of the (unrolled) Cluster-VRNN network. Probabilistic variables are represented with a triangle, deterministic with a square, and inner computations with a circle. Given a set of cluster representations, $\mathcal{C}$, our RNN-based network implements, at each time step, a generative, prior, inference probabilistic model, and a cell state update criteria.

- **Inference** $q(t) = q(\boldsymbol{\pi}_t \mid \mathbf{x}_{\leq t}) \sim \mathcal{D}(\boldsymbol{\alpha}_{i,t})$, where $\boldsymbol{\alpha}_{i,t} = \text{ReLU}(g_4(\mathbf{h}_{t-1}, g_5(\mathbf{x}_t)))$;
- **Cell State Update** $\mathbf{h}_t = g_6(\mathbf{h}_{t-1}, g_5(\mathbf{x}_t), g_3(\mathbf{z}_t))$;
- **Outcome Prediction** $\hat{\mathbf{y}} = \sigma(g_7(\mathbf{z}_T))$, given $\mathbf{z}_T = \sum_k \boldsymbol{\mu}_k \boldsymbol{\pi}_T^k$,

where $\mathcal{D}$ denotes the Dirichlet distribution, Dirichlet (1969), g_i are 'gate' functions of the RNN (in practice, shallow feed-forward neural networks), ReLU and σ denote, respectively, the rectified linear unit and softmax activation functions.

Optimisation and Training Our model is trained variationally, according to an Evidence Lower BOund (ELBO) approach, due to the intractability of the data log-likelihood integral. For our model, it can be shown[2] that the ELBO can be written as:

$$L = \mathbb{E}_{\boldsymbol{\pi}_{\leq T} \sim q(\boldsymbol{\pi}_{\leq T} | \mathbf{x}_{\leq T})} \big[\log p(\mathbf{y}_T \mid \boldsymbol{\pi}_T) + \sum_{t=1}^{T} \big(\log f_g(t) - \text{KL}(q(t) \| f_p(t)) \big) \big] \tag{1}$$

L is estimated by sampling once from a Dirichlet distribution, and computing the corresponding expression. In the Appendix Section A2, we provide further detail concerning practical implementation with regard to estimating L, namely, our sampling mechanisms, and simplification of the inner term.

3.1 Model Training

Our model is trained to maximise the ELBO lower bound (Equation 1). All model gates were implemented with 2 hidden layers of 30 nodes each, and a hyperbolic tangent activation function. Glorot initialization Glorot & Bengio (2010) was used for all neural network weights, with the exception of bias weights, which are initialized according to a standard normal distribution. Our model was implemented in Python using the Tensorflow, scikit-learn and NumPy packages, and all experiments were run with 1 Tesla v100GPU, and 8 CPUs Intel(R) Xeon(R) Gold 6246 @3.30 GHz.

4 Results and Discussion

For Benchmarking, we regarded TSKM and GMM as classical clustering benchmarks and CAMELOT and VRNN-GMM as state-of-the-art DL methods for clustering multi-dimensional time-series data. We evaluated clustering performance through standard clustering metrics, including Silhouette score

[2] full derivation in the Appendix Lemma A3

(SIL, Rousseeuw (1987)), Davies-Bouldin Index (DBI, Davies (1979)), Variance Ratio Criterion (VRI, Caliński & Harabasz (1974)). To evaluate phenotype identification, we considered the related task of outcome prediction. The prediction task was benchmarked against traditional classifiers for outcome prediction in EHR data, namely Support Vector Machines (SVM), XGBoost (XGB) and NEWS2 (i.e., the National Early Warning Score used in the UK hospitals). Prediction performance was evaluated using a multi-class macro-averaged one-vs-all Area-under-the-Receiver-Operating-Curve (AUROC), unweighted mean F1-score (F1), and unweighted mean Recall (Recall). Their average and standard deviation over 5 experiments, with different seeds, are reported. For purely unsupervised models, a predictive pipeline was constructed by assigning patients to clusters and returning a prediction corresponding to the empirical outcome distribution in the cluster. All models were trained on the same training (60%) and test (40%) sets. Hyper-parameters were selected through grid search and Occam's razor approach, the latter used for integer-valued hyper-parameters. Where applicable, neural network dimensions were kept consistent across all DL models.

We show the performance results of our model and benchmarks on the HAVEN dataset in Tables 1(clustering) and 2(outcome prediction). In Appendix Section A3, we further show results for the MIMIC dataset, which follow a similar trend.

Metric	SIL	DBI	VRI
TSKM	**0.35 (±0.01)**	**1.19 (±0.08)**	**554.6 (±2.5)**
GMM	0.14 (±0.02)	2.39 (± 0.23)	457.93 (± 11.9)
VRNNGMM	0.16 (± 0.30)	12.8 (± 9.32)	66.5 (± 18.7)
CAMELOT	0.11 (± 0.04)	3.12 (± 0.53)	216.7 (± 6.2)
Proposed	0.19 (± 0.10)	2.4 (± 0.34)	366.5 (± 18.7)

Table 1: Clustering separability performance on HAVEN. Average and standard deviation are reported for each metric.

Metric	AUROC	F1-score	Recall
SVM	0.50 (± 0.02)	0.23 (± 0.00)	0.25 (± 0.00)
XGB	0.65 (± 0.01)	0.23 (± 0.00)	0.22 (± 0.00)
NEWS2	0.61	0.29	0.34
TSKM	0.55 (± 0.01)	0.24 (± 0.03)	0.26 (± 0.02)
GMM	0.62 (± 0.1)	0.30 (± 0.08)	0.26 (± 0.02)
VRNNGMM	0.68 (± 0.01)	0.32 (±0.02)	0.33 (± 0.01)
CAMELOT	**0.73 (±0.02)**	**0.36 (±0.01)**	**0.38 (±0.02)**
Proposed	0.72 (± 0.02)	**0.36(±0.02)**	0.35 (± 0.02)

Table 2: Outcome Prediction performance on HAVEN. Average and standard deviation are reported for each metric.

Compared with all other DL-based benchmarks, our model shows a large improvement in clustering performance (see Table 1). Furthermore, our model outperforms the standard probabilistic model, GMM, also by a significant margin (with the exception of DBI). TSKM obtains better scores with regard to clustering metrics, however, this is largely due to a) metric bias towards convex clusters, and b) clustering directly on the input space, as opposed to DL models which do so in a latent space with extracted features.

With regard to identifying separable phenotypes, our model performs similarly to CAMELOT, and outperforms all other benchmarks (DL and standard classifiers) on both HAVEN and MIMIC - an increase of at least 4% (HAVEN) and 6% (MIMIC) in mean AUROC. We did not find statistical evidence (through Friedman testing, Friedman (1937)) to indicate a difference between CAMELOT and our proposed method. On the other hand, our model, which is dynamic and adaptive over time, is tremendously more expressive than CAMELOT, which is static and not probabilistic. Furthermore, the novel generative framework is extremely useful for clinical application as it can be used to better understand the future evolution for both clusters, and also individual patients.

5 Conclusion and Future Work

In this work, we propose a novel deep-learning model for the task of identifying phenotypically separable clusters in temporal EHR data. As part of our proposed model, we propose a sequential, variational generative approach that is capable of a) modelling cluster probability assignment at each time step, b) generating observational data over time, and c) predicting the outcome of a patient. Our experiments show promising results with respect to cluster separability and outcome prediction in two independent datasets, one of which is freely available. The addition of the DL-parametrised probabilistic model likely aided in learning relevant representations and clusters through the increased modelling capacity. Furthermore, our approach provides insight into identifying underlying cluster probability assignments throughout the duration of a patient stay. In the future, we will further explore the relevance of our generative framework and conduct an investigation into the quality of the generated observations. Finally, we will develop methodologies to analyse the temporal cluster assignments, for instance, what motivates cluster changes over time, and how can we accurately evaluate learned cluster assignments' performance.

REFERENCES

Henrique Aguiar, Mauro Santos, Peter Watkinson, and Tingting Zhu. Learning of cluster-based feature importance for electronic health record time-series. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 161–179. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/aguiar22a.html.

Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Curran Associates, Inc., 2015. URL http://www.github.com/jych/nips2015_vrnn.

DL Davies. et dw bouldin. a cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell*, 1 (2), 1979.

Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pp. 260–265, 1986.

Peter Gustav Lejeune Dirichlet. Sur une nouvelle méthode pour la détermination des intégrales multiples. *Werke, Bd.*, 1969.

Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, and Gunnar Rätsch. Deep self-organization: Interpretable discrete representation learning on time series. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rygjcsR9Y7.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101:215–220, 6 2000.

A. Johnson, L. Bulgarelli, T. Pollard, L. A. Celi, R. Mark, and S. Horng. MIMIC-IV-ED (version 1.0). *PhysioNet*, 2021.

Eunji Jun, Ahmad Wisnu Mulyadi, Jaehun Choi, and Heung-Il Suk. Uncertainty-gated stochastic sequential model for ehr mortality prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9):4052–4062, 2021. doi: 10.1109/TNNLS.2020.3016670.

David A Knowles. Stochastic gradient variational bayes for gamma approximating distributions. *arXiv preprint arXiv:1509.01631*, 2015.

Changhee Lee and Mihaela Van Der Schaar. Temporal phenotyping using deep predictive clustering of disease progression. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5767–5777. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/lee20h.html.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. Dirichlet and related distributions: Theory, methods and applications. 2011.

Marco AF Pimentel, Oliver C Redfern, Stephen Gerry, Gary S Collins, James Malycha, David Prytherch, Paul E Schmidt, Gary B Smith, and Peter J Watkinson. A comparison of the ability of the national early warning score and the national early warning score 2 to identify patients at risk of in-hospital mortality: a multi-centre database study. *Resuscitation*, 134:147–156, 2019.

Yordan P Raykov, Alexis Boukouvalas, Fahd Baig, and Max A Little. What to do when k-means clustering fails: a simple yet principled alternative algorithm. *PloS one*, 11(9):e0162259, 2016.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Royal College of Physicians. National early warning score (news) 2. *Standardising the assessment of acute-illness severity in the NHS*, 2017.

Francisco R Ruiz, Titsias RC AUEB, David Blei, et al. The generalized reparameterization gradient. *Advances in neural information processing systems*, 29, 2016.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, et al. Tslearn, a machine learning toolkit for time series data. *J. Mach. Learn. Res.*, 21(118):1–6, 2020.

Alice M Turner, Lilla Tamasi, Florence Schleich, Mehmet Hoxha, Ildiko Horvath, Renaud Louis, and Neil Barnes. Clinically relevant subgroups in copd and asthma. *European respiratory review*, 24 (136):283–298, 2015.

Claus F Vogelmeier, Kenneth R Chapman, Marc Miravitlles, Nicolas Roche, Jørgen Vestbo, Chau Thach, Donald Banerji, Robert Fogel, Francesco Patalano, Petter Olsson, et al. Exacerbation heterogeneity in copd: subgroup analyses from the flame study. *International journal of chronic obstructive pulmonary disease*, 13:1125, 2018.

APPENDIX

1 DATA AND PRE-PROCESSING

HAVEN HAVEN is a dataset of routinely collected clinical observations from March 2014 to March 2018 (HAVEN project, REC reference: 16/SC/0264;16/CAG/0066 and Confidential Advisory Group reference 08/02/1394). HAVEN contains EHR measurements of patients admitted to Oxford University Hospitals NHS Foundation Trust. The database *does not* contains observation data from Intensive Care Units (ICU) (only information about ICU entry/exit), and we exclude ED observations.

We used the protocol defined in Pimentel et al. (2019) to subset the cohort to those patients at risk of developing Type-II Respiratory Failure (T2RF) in hospital (a diagram of the data selection steps can be found in Figure A.1 in the Appendix). Our processing pipeline identifies 4 separate outcomes based on the occurrence of adverse events: i) no event during the hospital stay, leading to successful discharge from the hospital, or the first instance of one of three possible events, ii) unplanned entry to ICU, iii) cardiac arrest (also named 'Cardiac' hereafter) and iv) 'Death'. We show the features of different outcome groups are not clearly separable (see Tables A.2, A.1 in the Appendix). As such, learnt clusters will naturally contain a mix of outcomes. With regard to HAVEN, we represent the clinically relevant component of a cluster phenotype (denoted as *cluster propensity (score)*) as a categorical distribution with the corresponding empirical outcome proportion within the cluster.

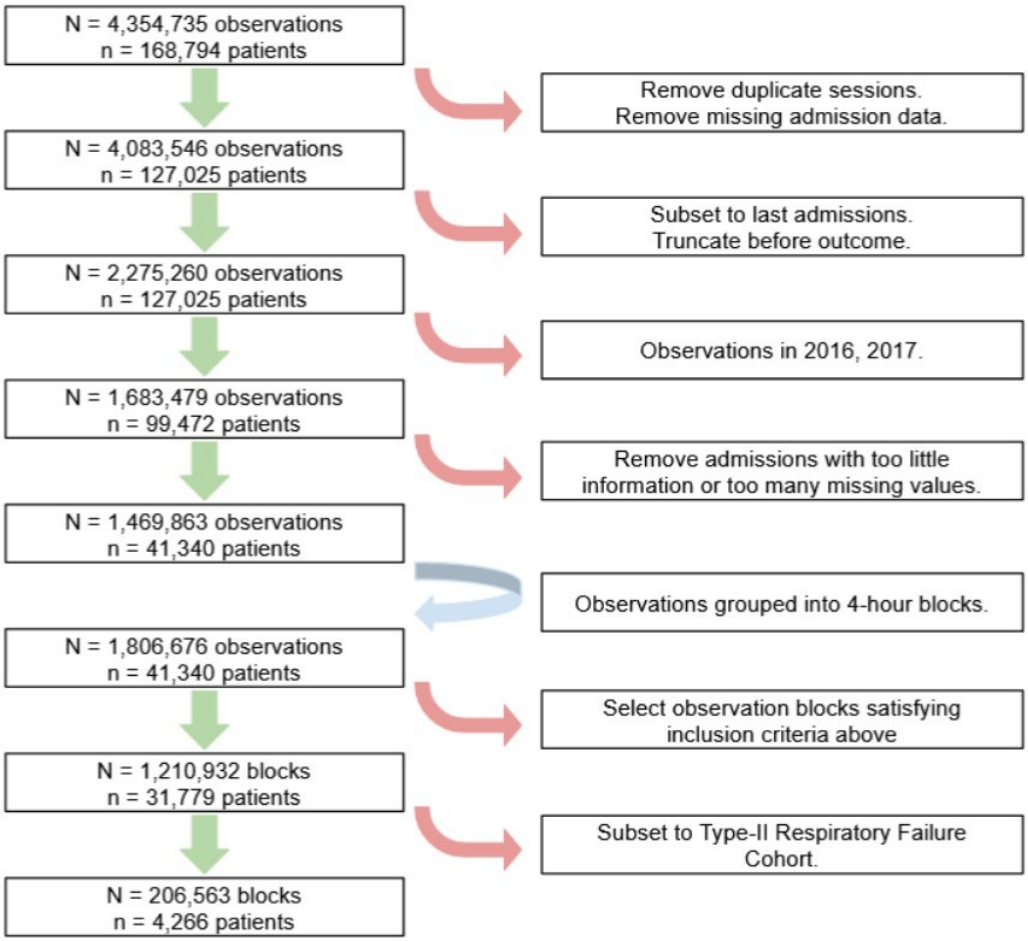
For each admission, observational data were averaged into 4 hour-window blocks, based on the time to outcome (or time to discharge, in the case of no event during stay), to allow uniformisation of the time indices across different features for the same patient. Following the literature validating Early Warning Score (EWS) systems (baseline models used by UK NHS staff to track inpatient physiology, (Royal College of Physicians, 2017)) and clinical input, we restrict observations to those within 24 and 72 hours prior to the occurrence of an event. Our cluster phenotypes, therefore, showcase the outcome propensity in the subsequent 24 hours.

Collected features were all transformed according to min-max normalisation. This was preferred over gaussian normalisation due to the skewness and heterogeneity of most feature distributions. Consequently, any missing values were imputed based on the first previous observation - all remaining missing observations were imputed according to the feature median within the train and validation cohorts. Finally, static features (such as age, gender, etc.) were similarly normalised and passed to all models as a constant time-varying feature with the corresponding value.

After processing, input data contained over 100,000 patient trajectories corresponding to 4,266 unique hospital admissions (only the last admission of each patient was considered in our analysis). Original trajectories for the patient cohort are shown in the Appendix in Figures A.3, A.4, A.5 for different variables/features. It can be observed that the different outcome groups are not clearly separable, both with respect to temporal and static variables. Furthermore, we note the high degree of imbalance in the data - admissions with no event account for over 86.8% of the total number of admissions, while event classes correspond to 10.3% Death, 1.8% ICU and 1.1% Cardiac. A description of the complete pipeline of data pre-processing is shown in Figure A.1.

A total of 26 input features were considered. Firstly, 4-hourly vital-sign sets which included 8 features: Heart Rate (HR), Respiratory Rate (RR), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), peripheral Oxygen Saturation (SpO_2), Temperature (TEMP), level of consciousness via the AVPU scale - Alert, Verbal, Pain, Unresponsive - and estimated Fraction of Inspired Oxygen (FiO_2, available when an oxygen mask is applied to the patient). Each set consisted of a timestamp and the vital-sign numerical values. Secondly, 4 demographic variables were selected (modelled as static variables): age, sex, and admission type (elective or surgical). Thirdly, we included 6 features resulting from biochemistry blood tests, denoted as 'Serum': Serum levels of urea, albumin, creatinine, sodium, potassium and C-reactive protein concentrations. Finally, 8 haematological blood test features were also included: white and haemaglobin cell counts, concentration of eosinophils, basophils, neutrophils, and lymphocytes, as well as eosinophil-to-basophil and neutrophil-to-lymphocyte ratios. These features were selected based on domain knowledge of features related to severity in the prognosis and outcome of inpatients at risk of T2RF.

Descriptive statistics for all input variables is described in Table A.2. Median and inter-quartile range (IQR) is displayed for continuous and categorical variables, while binary variables are shown



A.1: HAVEN data pre-processing steps.

according to number of counts in the dataset and corresponding cohort proportion. Statistics are displayed for the complete data ("All"), but also for each sub-cohort defined by the overall outcome. We can observe that these sub-cohorts are not clearly separable and are hard to identify solely from this information.

		Description	Units	Type	All	Cardiac	ICU	Death	Healthy
n		Patient Count		Integer	4266	48	76	441	3701
N		Trajectory Count		Integer	110,916	1,248	1,976	11,466	96,226
O		Observation Count		Integer	1,286,119	14,095	20,227	184, 504	1,067,293
Vital signs									
	HR	Heart-rate	beats/minute (bpm)		82.50 (72 - 94)	88 (77 - 100)	88.50 (77 - 99)	89 (77 - 100)	81.67 (72 - 92)
	RR	Respiratory-Rate	breaths/minute (Bpm)		18 (16 - 19)	18.29 (17.50 - 20)	18 (16 - 19)	19 (18 - 21)	18 (16 - 18)
	SBP	Systolic Blood Pressure	mmHg		126 (112 - 141)	122 (108 - 138)	119 (104 - 137)	123 (108 - 140)	127 (113 - 142)
	DBP	Diastolic Blood Pressure	mmHg	Continuous	67 (60 - 76)	65 (57 - 73)	64 (57 - 72.33)	66 (58 - 75)	67 (60 - 76)
	SPO2	Estimated Oxygen Saturation	%		95 (94 - 97)	95 (93 - 97)	95 (94 - 97)	94 (91 - 96)	95 (94 - 97)
	FIO2	Fraction of Inspired Oxygen concentration	%		21 (21.00 - 28.67)	21 (21 - 31)	21 (21 - 41)	28 (21 - 45)	21 (21 - 21)
	TEMP	Temperature	°C		36.40 (36.05 - 36.80)	36.25 (36.00 - 36.60)	36.40 (36.00 - 36.83)	36.20 (36. - 36.65)	36.40 (36.10 - 36.80)
	AVPU	Alert, Verbal, Pain, Unresponsive Scale		Categorical (1-4)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)
Static									
	age	Patient age	year	Integer	72 (62 - 81)	76 (69 - 82)	69 (61 - 80)	81 (74 - 88)	71 (61 - 74)
	gender	Male patients			2123 (49.77 %)	28 (58.33 %)	38 (50.00 %)	247 (56.01 %)	1810 (48.01 %)
	Elective	Elective Admissions		Binary	1139 (26.70 %)	2 (4.17 %)	8 (10.53 %)	3 (0.68 %)	1126 (30.42 %)
	Surgical	Surgical admissions			1130 (26.49 %)	6 (12.50 %)	22 (28.95 %)	48 (10.88 %)	1054 (28.48 %)
Serum									
	HGB	Haemaglobin	g/L		11.20 (9.70 - 12.80)	11.00 (9.35 - 12.40)	10.35 (9.00 - 12.03)	10.80 (9.30 - 12.60)	11.30 (9.80 - 12.90)
	WBC	White Blood Cell count (blood)	x10^9/L		10.03 (7.63 - 13.23)	10.90 (8.41 - 15.36)	10.41 (6.79 - 14.04)	11.38 (8.34 - 15.82)	9.80 (7.54 - 12.82)
	EOS	EOSinophil count (blood)	x10^9/L		0.10 (0.02 - 0.22)	0.06 (0.01 - 0.20)	0.04 (0.00 - 0.15)	0.03 (0.00 - 0.11)	0.11 (0.03 - 0.24)
	BAS	BASophil count (blood)	x10^9/L		0.04 (0.02 - 0.05)	0.03 (0.02 - 0.06)	0.03 (0.02 - 0.05)	0.03 (0.02 - 0.05)	0.04 (0.02 - 0.06)
	EBR	Eosinophil-Basophil Ratio		Continuous	2.60 (0.75 - 5.50)	1.50 (0.50 - 4.88)	1.43 (0.17 - 4.50)	1.00 (0.00 - 3.33)	3.00 (1.00 - 5.86)
	NEU	NEUtrophil count (blood)	x10^9/L		7.60 (5.34 - 10.70)	8.88 (6.25 - 12.60)	8.14 (4.75 - 11.25)	9.31 (6.46 - 13.52)	7.31 (5.19 - 10.19)
	LYM	LYMphocyte count (blood)	x10^9/L		1.16 (0.77 - 1.68)	1.00 (0.63 - 1.58)	1.00 (0.55 - 1.65)	0.84 (0.55 - 1.24)	1.24 (0.85 - 1.75)
	NLR	Neutrophil-Lymphocyte Ratio			6.36 (3.76 - 11.67)	9.40 (4.88 - 15.98)	8.41 (4.33 - 15.55)	10.84 (6.20 - 19.59)	5.74 (3.51 - 10.11)
Haematological									
	ALB	ALBumin level (plasma)	g/L		26.00 (22.00 - 30.00)	23.00 (20.00 - 28.00)	23.00 (19.00 - 27.50)	23.00 (19.00 - 28.00)	27.00 (23.00 - 31.00)
	CR	Creatinine level (plasma)	umol/L		77.00 (58.00 - 109.00)	107.00 (78.75 - 154.25)	78.00 (52.00 - 125.75)	93.00 (63.00 - 145.00)	74.00 (58.00 - 102.00)
	CRP	C-Reactive Protein level (plasma)	mg/L	Continuous	63.43 (21.30 - 137.58)	51.50 (25.80 - 112.85)	112.20 (32.45 - 226.05)	85.20 (36.20 - 156.35)	58.00 (18.68 - 131.35)
	POT	POTassium level (plasma)	mmol/L		4.00 (3.60 - 4.40)	4.20 (3.80 - 4.80)	4.10 (3.70 - 4.50)	4.10 (3.70 - 4.80)	4.00 (3.60 - 4.30)
	SOD	SODium level (plasma)	mmol/L		137.00 (134.00 - 140.00)	136.00 (132.00 - 139.00)	136.00 (133.00 - 139.00)	138.00 (134.00 - 142.00)	137.00 (135.00 - 140.00)
	UR	URea concentration levels	ml		6.40 (4.40 - 10.40)	10.50 (6.45 - 18.15)	7.05 (4.40 - 11.67)	10.40 (6.40 - 17.50)	5.90 (4.20 - 9.00)

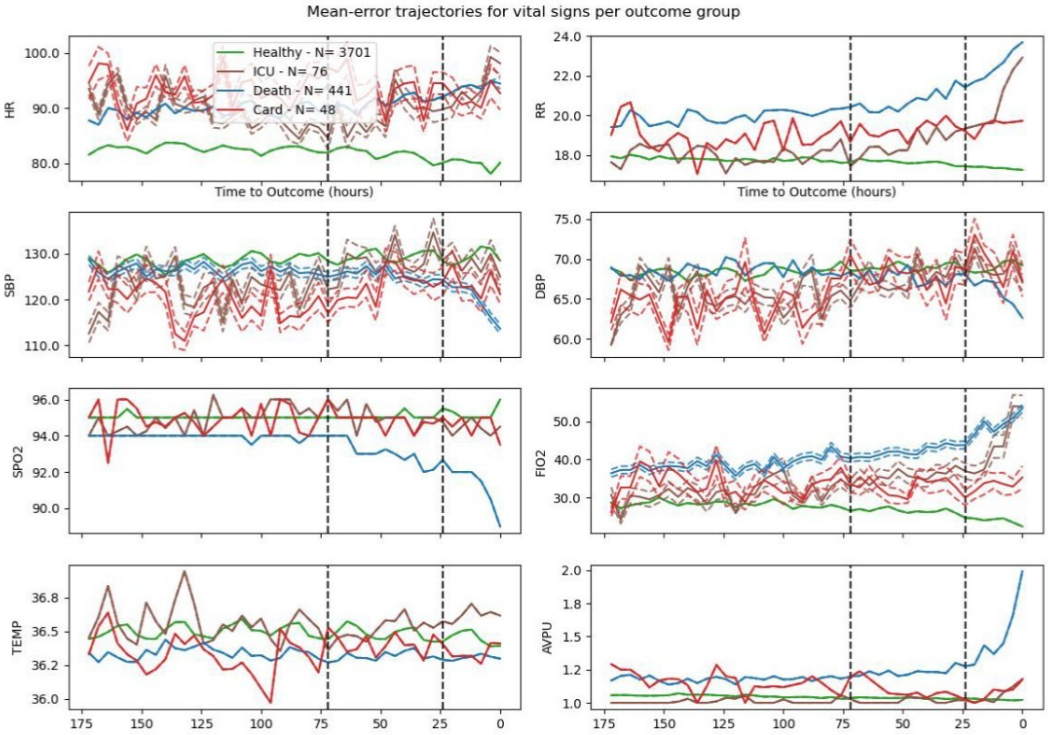
A.2: Descriptive statistics and information of all input data features. Variables are displayed with type, description, units and average statistics. We separate all features according to medical literature, including vital-sign, static, serum and haematological variables, and we also display statistics per outcome sub-cohort, defined as a cohort with those patients assigned to a given outcome.

A summary of the patient cohort in relation to outcomes and target phenotypes can be seen in Table A.1. Challenges with regards to obtaining phenotypically separable clusters can similarly be observed - there is no clear significant difference between the target outcome sub-cohorts with regards to demographic input variables. With regards to outcome distribution, we also note the high degree of imbalance in the dataset - the large majority of the patients in our dataset suffered from no adverse events (over 86%), while only 48 had a Cardiac event, and 76 were re-directed to the ICU.

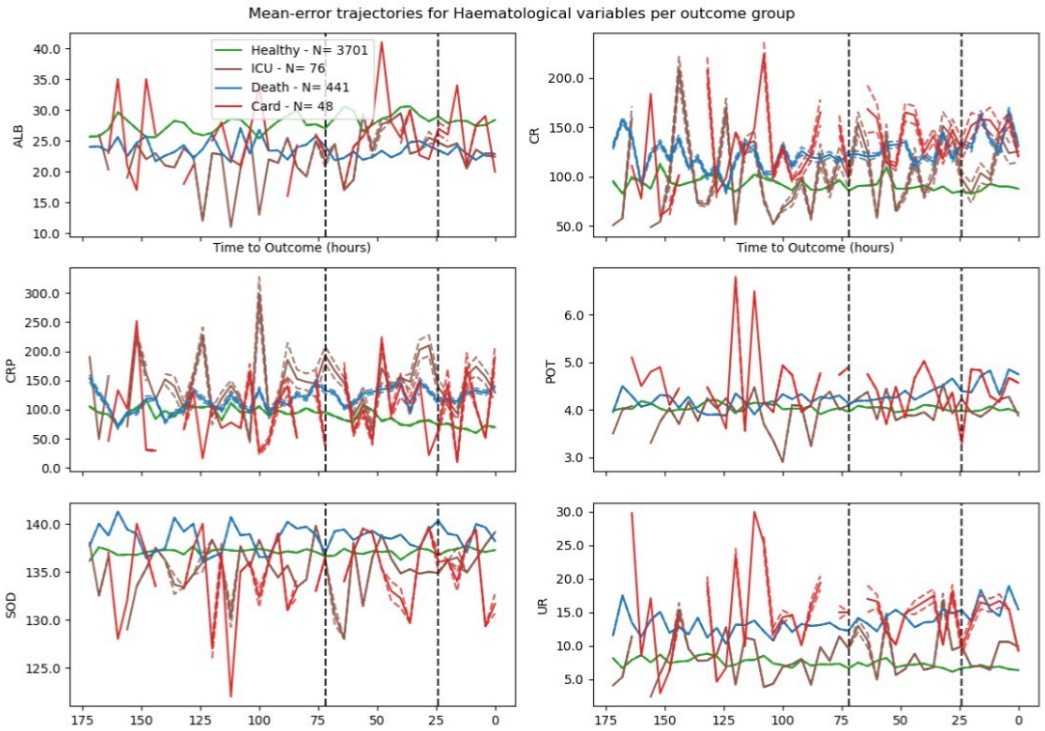
A.1: Descriptive demographic variable information for each outcome sub-cohort.

	No Event	Death	ICU	Cardiac
N	3701	441	76	48
Age (IQR)	71 (61 - 80)	81 (74 - 88)	69 (61 - 74)	76 (69 - 82)
Gender, M	1810 (48.9%)	247 (56.0%)	38 (50.0%)	28 (58.33%)
CCI (IQR)	4 (3 - 13)	14 (4 - 21)	7 (4 - 17)	15 (4 - 23)
Elective	1126 (30.4%)	3 (0.7%)	8 (10.5%)	2 (4.2%)
Surgical	1054 (28.5%)	48 (10.1%)	22 (29.0%)	6 (12.5%)

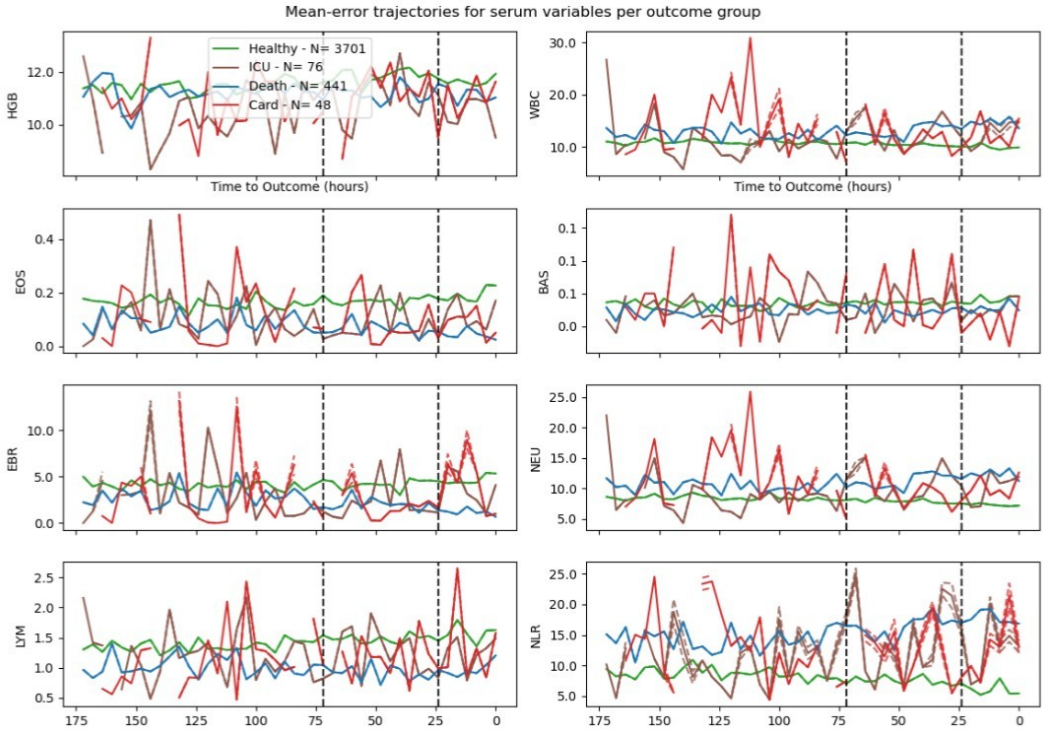
The lack of outcome sub-cohort separability can be further observed in a temporal domain. Figures A.3, A.4, A.5 plot the mean trajectories of different temporal variables sets for each outcome sub-cohort, respectively, according to vital signs, haematological and serum features. Mean is calculated based on the time to outcome, and missing observations are disregarded and ignored.



A.3: Plot of mean vital-sign trajectories (median with respect to SpO_2) in solid line as given by the 4 outcome groups: admissions with a) Cardiac-, b) Death- , or c) ICU-, and d) No-events. The corresponding feature units are: HR (bpm), RR (breaths-per-minute), SBP and DBP (mmHg), SPO_2 and FIO_2 (%), TEMP (C) and AVPU is unitless. The respective standard errors are represented by the dashed lines. We visualised trajectories from up to 7 days prior to an outcome event or discharge - the black lines represent the time window (72 - 24 hours prior to an event or discharge) considered for input to all models.



A.4: Plot of mean haematological trajectories in solid line as given by the 4 outcome groups: admissions with a) Cardiac-, b) Death- , or c) ICU-, and d) No-events. The $y-$ axis representations concentration in g/L (ALB), μ mol/L (CR), mmol/L (POT, SOD), mg/L (CRP) and mL (UR). The respective standard errors are represented by the dashed lines. We visualised trajectories from up to 7 days prior to an outcome event or discharge - the black lines represent the time window (72 - 24 hours prior to an event or discharge) considered for input to all models.



A.5: Plot of mean serum trajectories in solid line as given by the 4 outcome groups: admissions with a) Cardiac-, b) Death- , or c) ICU-, and d) No-events. The $y-$ axis denotes concentration in g/L (HGB) or $10^9/L$ (all other features). The respective standard errors are represented by the dashed lines. We visualised trajectories from up to 7 days prior to an outcome event or discharge - the black lines represent the time window (72 - 24 hours prior to an event or discharge) considered for input to all models.

MIMIC-IV-ED MIMIC-IV-ED (abbreviated henceforth as MIMIC) is a large, freely available database of ED admissions at the Beth Israel Deaconess Medical Center, introduced in (Johnson et al., 2021; Goldberger et al., 2000). MIMIC contains 448,972 total admissions between 2011 and 2019, and for each admission, the dataset has records of vital-sign variables, triage measurements, medications and consequent hospital journey.

Our pre-processing method was identical to that used for HAVEN. Admissions whose cause of entry was psychiatric and/or childbirth were excluded as they comprise significantly distinct cohorts. We considered 1 hour time block uniformisation, and only observations at most 6 hours before ED discharge time were kept. These values were decided based on clinical input and empirically-observed distribution. For each patient, we define their outcome according to their potential admissions within the following 12 hours: a) whether the patient died ('Death'), b) whether the patient was admitted to any ICU ('ICU'), c) whether the patient remained in hospital, at any ward ('Ward'), and d) whether the patient was discharged ('Discharge'). The post-processing cohort had a similarly imbalanced outcome distribution, with a cohort distribution of 0.30% Death, 16.53% ICU, 2.11% Discharge and 81.06% Ward.

2 Model Properties

The objective of our probabilistic model is the maximisation of the data likelihood:

$$\begin{aligned} p(\mathbf{x}_{\leq T}, \mathbf{y}_T) &= \int p(\mathbf{x}_{\leq T}, \mathbf{y}_T \mid \boldsymbol{\pi}_{\leq T})\, d\boldsymbol{\pi}_{\leq T} \\ &= \int p(\mathbf{y}_T \mid \boldsymbol{\pi}_T) f_j(T) d\boldsymbol{\pi}_{\leq T} \\ &= \int p(\mathbf{y}_T \mid \boldsymbol{\pi}_T) \prod_{t=1}^{T} f_g(t) f_p(t)\, d\boldsymbol{\pi}_{\leq T} \end{aligned} \tag{2}$$

Unfortunately, Equation 2 is generally impossible to optimise directly due to the intractability of the integral expression. On the other hand, direct approximations can be computationally expensive. As such, we use an Evidence Lower BOund (ELBO) approach with our VRNN; in particular, we use q, defined above, as the family of variational approximations, and we focus on maximising the ELBO lower-bound.

Lemma 3. *The ELBO lower-bound corresponding to Equation (2) is given by:*

$$\begin{aligned} L = \mathbb{E}_{\boldsymbol{\pi}_{\leq T} \sim q(\boldsymbol{\pi}_{\leq T} \mid \mathbf{x}_{\leq T})} \Bigg[& \log p(\mathbf{y}_T \mid \boldsymbol{\pi}_T) + \\ & \sum_{t=1}^{T} \Big(\log f_g(t) - KL(q(t) \| f_p(t)) \Big) \Bigg] \end{aligned} \tag{3}$$

Proof. We start from Equation 2. We have:

$$p(\mathbf{x}_{\leq T}, \mathbf{y}_T) = \int p(\mathbf{x}_{\leq T}, \mathbf{y}_T \mid \boldsymbol{\pi}_{\leq T})\, d\boldsymbol{\pi}_{\leq T}$$

Define new set of variables $\mathbf{x}'$ to be such that $\mathbf{x}'_t = \mathbf{x}_t$ for $t < T$, and $\mathbf{x}'_T = [\mathbf{x}_T, \mathbf{y}_T]$. Then our goal can be written as $p(\mathbf{x}'_{\leq T}) = \int p(\mathbf{x}'_{\leq T} \mid \boldsymbol{\pi}_{\leq T})\, d\boldsymbol{\pi}_{\leq T}$. Similarly, we also extend our variational approximation family, $q(\boldsymbol{\pi}_{\leq T} \mid \mathbf{x}_{\leq T})$, by naturally incorporating $\mathbf{y}_T$. Denote the derived variational family by $q' = q'(\boldsymbol{\pi}_{\leq T} \mid \mathbf{x}'_{\leq T})$. In this case, we can use the expression derived in Chung et al. (2015):

$$L = \mathbb{E}_{\boldsymbol{\pi}_{\leq T} \sim q(\boldsymbol{\pi}_{\leq T} | \mathbf{x}'_{\leq T})} \left[\sum_{t=1}^{T} \Big(- \mathrm{KL}(q'(t) \| f'_p(t)) + \log f'_g(t) \Big) \right]$$

where f'_p, f'_g are analogously defined. Note that L represents a sum over time of two opposite goals - a) minimisation of the KL divergence between variational distribution and prior on latent variables, and b) log-likelihood.

For $t < T$, we note that all expressions are unaltered, as $x'_t = x_t$ for $t < T$, y_T depends solely on $\mathbf{z}_T$, and does not interfere with computation of other variables.

For $t = T$, we consider each term separately. Note that $f'_p(T) = p(\boldsymbol{\pi}_T \mid \boldsymbol{\pi}_{<T}, \mathbf{x}'_{<T}) = p(\boldsymbol{\pi}_T \mid \boldsymbol{\pi}_{<T}, \mathbf{x}_{<T}) = f_p(T)$. Similarly, re-arranging the terms and expanding the definition on conditional distributions, we get the same for the other KL term. On the other hand:

$$f'_g(t) = p(\mathbf{x}_T, \mathbf{y}_T \mid \mathbf{h}_{T-1}, \boldsymbol{\pi}_T) = p(\mathbf{y}_T \mid \mathbf{h}_{T-1}, \boldsymbol{\pi}_T, \mathbf{x}_T) p(\mathbf{x}_T \mid \mathbf{h}_{T-1}, \boldsymbol{\pi}_T) = p(\mathbf{y}_T \mid \boldsymbol{\pi}_T) f_g(t)$$

Thus $\log f'_g(T) = \log f_g(T) + \log p(\mathbf{y}_T \mid \boldsymbol{\pi}_T)$. Our Lemma then follows by re-arranging the terms. □

Note that the positive log terms represent the quality of the data generation component, while the second term measures the difference between the variational approximation, q, and the latent prior, f_p, on cluster assignment. Maximising this expression, therefore, can be achieved by accurately capturing both goals.

To compute the lower bound, it is necessary to estimate the expectation in Equation (3). Not only that, we need to achieve this so we can back-propagate through our model weights. We estimate the expectation by sampling from the Dirichlet distribution, and computing the corresponding expression. This approach has been shown to be a sufficiently good approximation to the expectation, Ruiz et al. (2016). In order to back-propagate the lower-bound through the model weights (in particular, the inference weights), we re-parametrise the Dirichlet distribution. We use the following results:

Lemma 4. *Suppose we have M independent gamma distributed random variables, $X_i \sim \Gamma(\alpha_i, 1)$, where $\alpha_i > 0$. Let X_0 denote the sum of all random variables. Then, the variable $P = (X_1/X_0, ..., X_M/X_0)$ is Dirichlet distributed with parameter $(\alpha_1, ..., \alpha_M)$.*

Proof. The proof relies on the Jacobian formula for transforming random variables. Further details in Devroye (1986). □

Lemma 4 is not yet sufficient - the gamma variables $\Gamma(\alpha_i, 1)$ are dependent on $\boldsymbol{\alpha}_i$, and therefore it is not possible to back-propagate through the sampling. We further re-parametrise the gamma distributions to counter this effect. There currently is no provably correct method for sampling from gamma random variables, however, useful approximations exist, Knowles (2015). We use inverse transform sampling (Devroye, 1986) - a) sample u is generated from a $[0, 1]$ uniform distribution, b) approximate the inverse cumulative distribution as $F^{-1}_{\alpha,\beta}(z) \approx \beta^{-1}(z\alpha\Gamma(\alpha))^{1/\alpha}$, c) transform u according to inverse approximation. Note that we can differentiate through the generated sample with respect to the gamma distribution parameters. As such, we combine this approach with Lemma 4, to obtain a Dirichlet-generating algorithm that is also differentiable with respect to its parameters.

On the other hand, we can also simplify the KL-divergence term in Equation 3 through the following lemma:

Lemma 5. *Suppose that random variables W_1, W_2 are respectively given by Dirichlet distributions with parameters $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}_m$. Then, we have:*

$$KL\Big(W_1 \| W_2\Big) = \log\Gamma(\alpha_0) - \log\Gamma(\beta_0) - \sum_{i=1}^{d} \Big(\log\Gamma(\alpha_i) - \log\Gamma(\beta_i) \Big) + \sum_{i=1}^{d} (\alpha_i - \beta i)(\psi(\alpha_i) - \psi(\alpha_0))$$

where Γ denotes the Gamma function, and ψ is the digamma function.

Proof. We use some known properties of the Dirichlet distribution (Ng et al., 2011), specifically that:

- Marginal distributions $W_1^j \sim Beta(\alpha_j, \alpha_0 - \alpha_j)$;
- $\mathbb{E}[-lnW_1^j] = \psi(\alpha_0) - \psi(\alpha_j)$;

Then:

$$\begin{aligned}
\mathrm{KL}(W_1 \| W_2) = -\int_{\Delta_{d-1}} \log p_{W_1}(w) \frac{p_{W_2}(w)}{p_{W_1}(w)} dw = \mathbb{E}_{w \sim p_{W_1}} \left[\log \frac{p_{W_1}}{p_{W_2}} \right] = \\
\mathbb{E}_{x \sim p_{W_1}} \left[\log p_{W_1} - \log p_{W_2} \right] \\
= \mathbb{E}_{w \sim p_{W_1}} \Bigg[\log \Gamma(\alpha_0) - \sum_i \log \Gamma(\alpha_i) + \sum_i (\alpha_i - 1) \log(x_i) - \\
\log \Gamma(\beta_0) + \sum_i \log \Gamma(\beta_i) - \sum_i (\beta_i - 1) \log(W_1^i) \Bigg] \\
= \log \Gamma(\alpha_0) - \log \Gamma(\beta_0) - \sum_i \big[\log \Gamma(\alpha_i) - \log \Gamma(\beta_i) \big] + \\
\sum_i (\alpha_i - \beta_i) \mathbb{E}_{w \sim W_1^i} \left[\log W_1^i \right]
\end{aligned}$$

and the result follows from known Dirichlet properties. □