Empirical phenotyping in coupled patient+care systems: Generating low-dimensional categories for hypothesis-driven investigation of mechanically-ventilated patients

J.N. Stroh, Peter D. Sottile, Yanran Wang, Bradford J. Smith, Tellen D. Bennett, Marc Moss, David J. Albers

January 9, 2025

Abstract

Background. Analyzing patient data under current mechanical ventilation (MV) management processes is essential to develop hypotheses about improvements and to understand MV consequences over time. However, progress is complicated by the complexity of lung-ventilator system (LVS) interactions, patient-care and patient-ventilator heterogeneity, and a lack of classification schemes for observable behavior.

Method. Ventilator waveform data arise from patient-ventilator interactions within the LVS while care processes manage both patient and ventilator settings. This study develops a computational pipeline that segments these joint waveform data and care settings timeseries to phenotype the data generating process. The modular method supports many methodological choices for representing waveform data and unsupervised clustering.

Results. Applied to 35 ARDS patients including 8 with COVID-19, typcially 8[6.8] (median[IQR]) phenotypes capture 97[3.1]% of data using naive similarity assumptions on waveform and MV settings data. Individual phenotypes organized around ventilator mode, PEEP, and tidal volume with additional segmentation reflecting waveform behaviors. Few ($< 10\%$ of) phenotype changes tie to ventilator settings, indicating considerable dynamics in LVS behaviors. Evaluation of phenotype heterogeneity reveals LVS dynamics that cannot be discretized into sub-phenotypes without additional data or alternate assumptions. Suitably normalized individual phenotypes may be aggregated into coherent groupings suitable for analysis of cohort data.

Conclusions. The pipeline is generalizable although empirical output is data- and algorithm-dependent. Further, output phenotypes compactly discretize the data for longitudinal analysis and may be optimized to resolve features of interest for specific applications.

1. Introduction

Mechanical ventilation (MV) of critical care patients provides life-saving support over periods typically lasting days to weeks. Over these timescales, ventilator management strategies significantly impact patient outcome [1, 2]. Modern care protocols and technologies [3, 4] emphasize lung-protective strategies [5] to minimize potentially deleterious consequences of MV. These include ventilator-induced lung injury (VILI, [6]) and patient-ventilator dyssynchrony (PVD), a disagreement between ventilator action and patient demand. Both PVD and VILI may contribute to acute respiratory distress syndrome (ARDS) and ARDS-related mortality [7, 8, 9]. Protective strategies depend on mechanistic understanding to guide ventilator settings, including positive end-expiratory pressure (PEEP), tidal volume, and driving pressure [10, 11, 12]. Despite

the effectiveness of these advances, association between MV and ARDS-related mortality remains unacceptably high. Reducing these highly negative outcomes motivates continued improvement and personalization of VILI-minimizing ventilator strategies [13, 14, 15].

Hypotheses about current care are essential for improving MV, but such scientific inquiry suffers from a lack of general breath categories and understanding of patient-ventilator variability. MV applied in critical care typically lasts 3–7 days [16, 17, 18] amidst the context of non-stationary patient conditions and other care procedures. While the short-duration physiology under MV is understood, analysis of those relationships over longer therapeutic timescales is limited and hindered by patient- and care-specific heterogeneity. A method of labeling and classifying MV breaths based on characteristics is desirable to reduce data complexity and facilitate temporal analysis. Currently, the most accessible classification scheme identifies PVD types from waveform characteristics. PVD research is rich with ML applications primarily focused on extending manual labels to larger datasets through supervised methods [19, 20, 21], identifying PVD waveform characteristics [22], and estimating event severity [23]. However, these labels may be ill-suited for general MV research involving temporal analysis: they are stationary, are not mutually exclusive [21], depend on MV mode characteristics [8], and vary in organization [24, 25]. Another research avenue uses interpretable model-based parametrizations to analyze waveform data [26, 27, 28, 29, 30], potentially allowing for a wider exploration of breath behavior over time.

The clinical observables from MV include airway pressure (p), volume (V), and flow timeseries that record the dynamic interaction between patient lungs and care-managed machine. The human lung-ventilator system (LVS), rather than an isolated human lung, underlies the data generating process when investigating MV from waveform-sourced data[28]. Moreover, MV management changes ventilator settings; these care factors must be included to appropriately frame LVS trajectories. The assemblages of coupled LVSs and applied management processes are the data generating process and the objects of interest for improving MV.

This work addresses methodological gaps by developing an unsupervised categorization process with time-dependent analysis of joint LVS data in mind. Specifically, it adapts an interpretable segmentation pipeline [31] to construct state categories, or phenotypes, derived from distributional similarity of data characteristics. This empirically generates low-dimensional categories of LVS behavior that are amenable to analysis, facilitating the study of changes over time in the state of MV patients. The workflow reveals the structure and complexity of breath evolution in individual patients from heterogeneous LVS factors and extends these classes to the cohort scale. Such categories simplify hypothesis-driven inquiry and analysis rather than serving immediate clinical information purposes.

Canonical validation of computed phenotype results is impossible because 'gold standard' LVS state classes are not presently known. Further, a classification scheme tied to context of the data and the question of interest gains informativeness and accuracy [*e.g.,* 32], benefits potentially lost in a universal scheme. In this work, empirical phenotypes are evaluated by whether they identify cohesive groupings that differentiate important characteristics. This means segmentation should follow ventilator settings changes and persistent variations in waveform behavior in individuals, while mixing similar patient behaviors at the cohort level. The following investigation scrutinizes individual phenotype trajectories in relation to MV management changes. The phenotyping process assigns equal weight to LVS feature components to be agnostic about data element importance. Constructed phenotype results are data-specific and may not generalize. The methodological results offer a flexible and generalizable path to optimize phenotype resolution, highlighting

the approach's analytical potential and its use for large dataset investigation.

Developing a successful method for phenotyping the diverse behaviors in LVS data enables future hypothesis-driven investigation into optimizing and personalizing MV patient management to reduce VILI. The results of this work could, for example, be used to statistically quantify the consequences of different MV breath behavior by analyzing the relationship between LVS phenotypes and localized costs such as driving pressure [33] or gas ratios [34, 35]. Importantly, use of phenotypes in this context data masks low-level heterogeneity to reduce trajectory complexity (§3.1) and provide a standard basis for comparison across patients (§3.2).

2. Method

Phenotype identification analyzes LVS data, including waveforms and ventilator settings, using an unsupervised computational pipeline. This section develops a specific implementation while framing it in a general way. The process is generalized but its output may not be: *empirical phenotypes reflect the data and methods defining them.* The modular workflow allows for changing the source data, waveform representation, feature definition, and segmentation to make the impact of these choices accessible to analysis.

2.1. Data

Data including airway pressure, volume, and ventilator settings were captured for a cohort of patients admitted to University of Colorado ICUs with ARDS diagnoses, required mechanical ventilation, and who had substantial risk of VILI (Colorado Multiple Institutional Review Board protocol #18-1433). Children, pregnant women, and age-censored elders ($>$ 89 years), and the imprisoned were excluded. Collection of esophageal pressures, which are not analyzed in this work, imposed additional exclusion criteria (*viz.* esophageal fistula, variceal bleeding or banding, facial fracture, and recent gastric/esophageal surgery). Partial recordings from Hamilton G5 ventilators (`https://www.hamilton-medical.com`) were made of 36 MV encounters satisfying these criteria following esophageal balloon placement. Source patients include 14 women and 22 men with median[IQR] age 59[25] years; 72% are white, 35% of which identify as Hispanic or Latino. One patient record with a 20-hour gap was dropped during pre-processing, comprising removal of breaths with ventilator calibration artifacts and carrying forward last-known settings values within ventilator modes. Table 1 summarizes clinical and demographic characteristics of included patients. Data total 1.74 million breaths over 71.14 recording-days (median 1.97[1.56] days per patient) recorded at 31.25 Hz. Adaptive pressure volume-controlled and pressure-controlled mandatory ventilation modes (APVCMV and P-CMV, respectively) account for 84% and 10% of breaths, respectively, with the remainder in spontaneous/supported (SPONT), synchronized controlled (SCMV), and standby modes. Care and ventilator management follow the ARDSnet protocols [7].

Dyssynchrony labels. An existing supervised ML technique [19] identifies breath-wise PVD to enrich LVS evolution context and provide comparison for newly calculated labels. PVD types were assigned breathwise to the data using a gradient boosted decision tree (XGBoost) based on a manually labeled subset. PVD categories include normal (NL), reverse triggered (RT), early flow limited (eFL), double trigger (DT), and early vent termination (EVT) types, per definitions of [21] and application in [36]. This work uses moving averages of breath labels to communicate PVD occurrence during model analysis windows.

Table 1: Tabular summary of the patient cohort and associated data. 'Monitored' and 'Recorded' denote the duration spanned by data and length continuous data contents, respectively, in hours. P:F ratio is the PaO_2/FiO_2 ratio at admission used to qualify ARDS and need for MV, AA = African-American, AI = American-Indian, AK = Alaska, NMB = Neuromuscular blockade

Detail	*Count*	%	*Median*	*IQR*
Monitored (hrs)			47.0	37.2
Recorded (hrs)			43.1	40
Age (years)	36		58.5	24.5
Gender				
Female	14	38.9	54.5	25.0
Male	22	61.1	58.5	26.0
Race/Ethicity				
White	26	72.2		
Unknown/NA	5	13.9		
Black/AA	3	8.3		
AI or AK Native	1	2.8		
More than one race	1	2.8		
ARDS risk				
Pneumonia	12	33.3		
COVID	11	30.6		
Sepsis	6	16.7		
Other	3	8.3		
Pancreatitis	2	5.6		
Aspiration	2	5.6		
P:F ratio			135.9	81.0
Mortality	9	25.0		
NMB use	9	25.0		

2.2. Pipeline

The computational pipeline described is a process for developing phenotypes from joint waveform and MV care-related data. The method is depicted in Figure 1 and follows [31] by using data-informed parameter distributions to uncover latent similarities in observed ICU patients. The three main phases (feature construction, segmentation, and interpretation) involve methodological decisions, which are discussed below both in general terms and in terms of specific implementation applied to individual patient data.

1.) Waveform parametrization. Waveform observations can optionally be digitized to improve comparison of breath-level data through approximation and regularization of the continuous time signals. Digitization examples in other work include spectrograms [37], clinical parameters [19], model parametrization [30], or signal processing methods such as polynomials and wavelets.

One empirical parametrization developed for LVS data analysis [28] uses an asynchronous ensemble Kalman smoother (wEnKS) [38] to transform waveform data segments into parameter distributions (SI A). Specifically, observations y^o (representing pressure or volume data) are mapped to M-dimensional parameter vector samples $\{\mathbf{a}\}$ of the bayesian posterior distribution $p(\mathbf{a}|y^o)$. The likelihood function $p(y^o|\mathbf{a})$ describes the RMS error between data y^o and simulated counterparts defined by the ordinary differential equation

$$\frac{dy}{dt} + g \cdot (y - y_0) = \varphi\left(t \left(\operatorname{mod} \theta\right); \mathbf{a}\right). \tag{1}$$

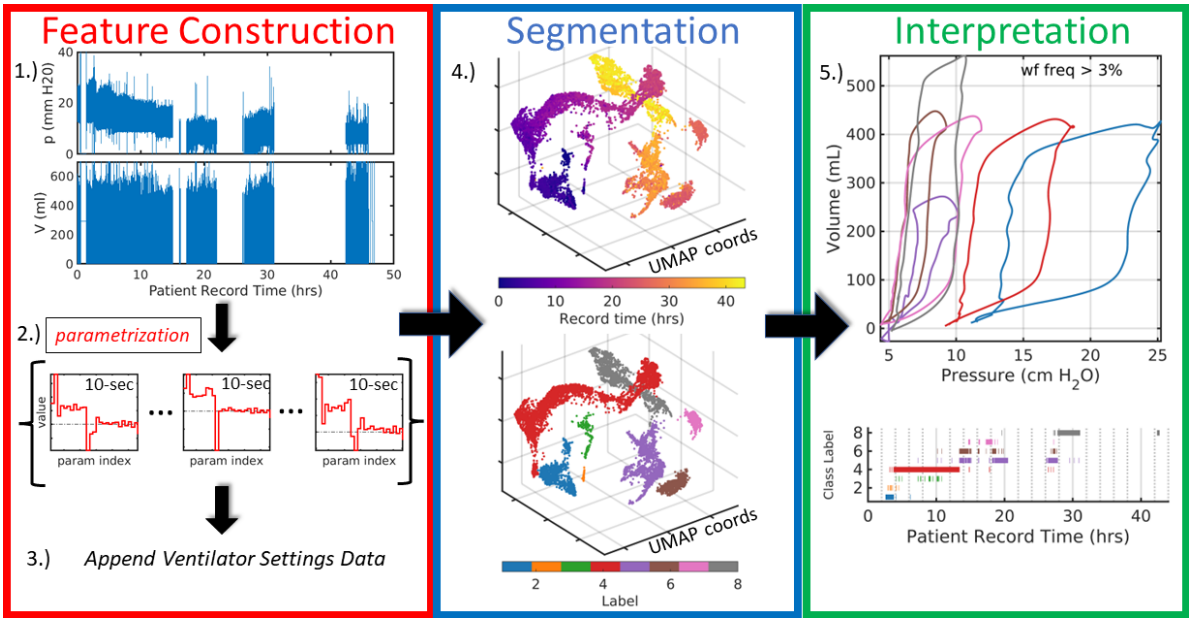

Figure 1: Broad pipeline organization. Raw data (1.) are digitally parametrized (2.) over short windows, typically satisfying stationarity assumptions. Distributional parameter estimates are summarized and augmented with the contextual data of ventilator settings (3.) which include information such as ventilator operation mode, positive-end expiratory pressure (PEEP) or other baseline pressure, flow and pressure triggers, and minimum mandatory breath rate. Feature vectors, defined by the augmented LVS descriptors, are reduced to three dimensions (4.) where they can be analyzed based on time ordering (top) and structural similarity via segmentation (bottom). Finally, in (5.), temporal evolution of the system is compactly encoded in the time-ordered LVS descriptor labels and their associated waveform characterizations in an interpretable and explainable way. The process transforms raw data (1.) into a more easily comprehensible form such as (5.).

Here, g is a fixed smoothing parameter, θ is constant breath rate over the interval, and y_0 is the baseline value of the signal (PEEP when y is pressure). The model construction (Eq.1) ensures parameter identification throughout its bayesian ensemble-based inversion. Consequently, data-informed parameters are unique [39] and quantify uncertainties [40] associated with observation noise, waveform variability, and model resolution. Breath rate θ and signal baseline y_0 are assumed to be stationary during parameter inference, thereby constraining the window length. Informed by these considerations, this work uses a moderate resolution model ($M = 28$) to encode and discriminate essential waveform features over 10-second windows to satisfy stationarity requirements of the inference. Each 10-second window is associated with reference values and parameter samples $(y_0, \theta, \{\mathbf{a}\})$.

2.) Parameter Distribution Summaries. Statistically summarizing waveform parameters estimates over longer stationary timescales, rather than detailing each breath, is an optional step in the pipeline. This is computationally advantageous for large dataset phenotyping because it invokes fewer pairwise comparisons of more detailed objects. The chosen parametrization process (above) samples the data-informed posterior distribution of parameters on each 10-second window. The estimators used to summarize these samples include mean, quartiles, variance, and mode, plus non-gaussian measures (skewness, kurtosis, and Kolmogorov-Smirnov distance). The latter items capture bimodal or asymmetric properties characterizing non-stationary LVS behavior. The stationary parameters, such as mean period and baseline pressure, are also included in these summary vectors. Statistical parameter summaries of 10-second windows reduce the

temporal sampling rate from 31.25Hz in raw data to $\frac{1}{10}$Hz while 2D raw data become ∼400-dimensional vectors of parameter estimators.

3.) Including care and context information. Appending ventilator settings data to each statistical waveform parameter summary contextualizes them in the health-care process. Ventilator settings detail the mode of operation (APVCMV, PCMV, SCMV, SPONT, standby), targeted quantities (set inspiratory pressure or set tidal volume) as well as various machine settings (trigger thresholds, ramp time, mandatory minimum breath rate). Some ventilator settings such as PEEP and I:E ratio are represented implicitly in waveform descriptors and need not be explicitly included. Other available factors such as ventilator delivery power are not considered here but may be included in other applications. Ventilator mode is a nominal variable that is one-hot encoded into a set of binary variables. However, not all settings are properties of each mode. Waveform properties proxy for ventilator settings (*i.e.*, observed V_T for V_T in PCMV), while missing data with no observable analogs (*e.g.*, trigger settings) are filled with zero values. MV settings are included in the LVS window summaries as static properties because settings change infrequently compared with the number of windows or breaths. The current implementation only considers MV settings. Other care-originating factors could be easily included here, such as patient sedation level, paralytic use, and patient posture, and used to subset analysis on specific care regimes or other contexts of interest.

4.) Phenotype Labeling. Phenotypes are defined through labels assigned to joint waveform-MV settings based on content similarity, which can be performed at individual patient or aggregated cohort levels. LVS descriptor vectors are reduced to lower dimensions so that labeling and assessment occurs in an easily visualized geometry. Dimensional reduction methods [41] include analyses of factors [42, 43] as well modern embedding/projection methods such as Uniform Manifold Approximation and Projection (UMAP, [44, 45]) and t-distributed Stochastic Neighborhood Embedding (tSNE,[46]). Group labels are assigned by a clustering process [47, 48] applied to LVS descriptors in the reduced coordinates. Options include space partition methods (k-means, k-medoids, etc.), kernel-based methods like Support Vector Clustering [49], and density methods like Density-based Spatial Clustering of Applications with Noise (DBSCAN, [50, 51]).

This study employs UMAP for reduction because it preserves local and global similarity structure of its input and has a numerically efficient MatLab implementation [52]. In this instance, the projection assesses similarity using the uniform Gower distance [53, 54] because feature vectors contain mixed-type variables. The metric supports feature weights to specify the influence of feature components in future applications. For example, the impact of volume waveform components can be given less weight when investigating MV in volume control modes. DBSCAN was chosen to label groups for its ability to identify clusters of arbitrary shape, as LVS feature clouds in UMAP coordinates are often irregular and non-convex (Fig1, step 4).

UMAP hyperparameters are fixed (neighborhood size 5 points, minimum distance 0.01) to maximize the equivalence of similarity-based projections across patients. During DBSCAN labeling, a brief grid-search over hyperparameters (core point requirement 4–12; neighborhood radius 1.5–5 by 0.5) finds the grouping that minimize the total distance between centroids to balance group consistency with the number identified of sets. Segmentation quality depends on data variability which increases over time; this search improves phenotype resolution uniformity across varying record lengths.

5.) Phenotype interpretation. The terms ‘label’ and ‘’phenotype’ below are synonyms because cluster labels identify consistent groups of 10-second data windows that characterize phenotypes. For example, the model

images of group median parameters characterize the central behavior of pressure and volume waveforms (Fig.1, step 5 top). Each time point in a MV patient record is associated with a phenotype presence so that labeled extend to all LVS observables of an individual patient. For example, downstream analysis of *e.g.,* FiO_2 or SpO_2 could use phenotype identity to stratify data.

2.3. Phenotypes and Characterizations of LVS data

The pipeline organizes data into discrete phenotypes based on similarity of windowed LVS states. An objective is to reveal LVS changes without corresponding ventilator settings changes. Such changes suggest the presence of factors that influence LVS trajectory including changes in patient expectation and breathing pattern (*e.g.*, patient effort, respiratory drive), lung mechanical function (*e.g.*, VILI progression or recovery from ARDS), or another aspect of physiology. Other factors like resistance and compliance of ventilator tubes, accumulated moisture, and changes in sedation and posture influence observed waveforms; these data are not available and remain potential confounders.

2.4. Experimental Phenotyping of LVS data

The pipeline described above is applied individually to 35 LVS records defining a context of 2-day periods of a few ARDS patients. This narrow context provides an opportunity to demonstrate the complexity of characterized behaviors as phenotype diversity is expected to be limited. There is no a priori reason to expect phenotypes to organize around particular data elements because the similarity metric (§2.2, item 4) weights components equally. *These experimental applications of the pipeline investigate what data elements impact phenotype structure and what variability remains in phenotypes.* The LVS trajectories are presented as a timeseries of phenotype labels contextualized by ventilator settings and shown in relation to classified PVD. Pressure-volume (pV) loop characterizations of each phenotype, computed from the median of relevant parameter estimates, provide visually summarize waveform data. Such visualizations intend to summarize key features and notable changes defining the LVS trajectory. Subsequent analysis and discussions employ principal component analysis (PCA), an empirical signal factorization based on variance minimization [42, 55]. This tool reveals the degree of LVS variance occurring under during stationarity to investigate non-ventilator temporal changes not identified by segmentation.

2.4.1. Cohort-scale phenotyping

Direct application of the individual pipeline to cohort data is a computationally expensive problem due to the data volume ($O(10^6)$ 10-second intervals of continuous multivariate variables). A simple alternative is to develop cohort-scale meta-labels for the population of individual phenotypes. To achieve commensurable features across the cohort, volume data (in mL) are standardized by separating the scaled magnitude (in mL/kg) from volume waveform parameters. Pressure waveforms are likewise standardized by zeroing on PEEP or other support pressure and scaling by driving (peak-minus-baseline) pressure within each window. Feature vectors for cohort clustering are individual phenotype statistics of baseline pressure, driving pressure, scaled tidal volume, estimated parameters of normalized waveform data, and associated ventilator settings. Segmentation uses UMAP-DBSCAN as in the individual case but with different hyperparameter values.

3. Results

The clinical data associated with ARDS patients (Table 1) are an important and practical phenotyping application because attentive MV management and patient instability instigate diverse LVS behaviors. This section reports experiment results from phenotyping individual ARDS patient data records (§3.1) and the assembly of cohort-scale phenotypes (§3.2). Within individual experiments, the temporal structure of LVS data labels is examined for consistency and resolution. Phenotypes aggregated across the cohort produce generalized LVS descriptor characterizations. PVD labels provide additional context for phenotype labels derived from ventilator settings and waveform characteristics.

3.1. Patient-level Phenotyping

LVS patient data are identified with 20[14] (median[IRQ]) individual phenotypes, totaling 721 groups across the cohort. Many of these patient-specific phenotypes capture less than 1% of a given patient LVS record. Over 97[3.1]% of data are captured by 8[6.5] core phenotypes that represent more than 3% of an individual record. Low-occurrence phenotypes often identify outliers and brief events such as suctioning that may be eliminated by reducing label specificity via UMAP-DBSCAN hyperparameters. Because record length varies (median[IQR] 47[37] hours), over-segmentation is needed to address the more diverse behavior of longer patient records compared with shorter ones.

Phenotypes primarily organize around changes in MV settings such as tidal volume, PEEP, and ventilator mode in all experiments, but they also reflect changes in the patient-ventilator (or LVS) behavior. There is high correspondence between changes in ventilator settings and persistent changes (lasting longer than 30 seconds) in individual patient phenotype labels (SI Table B.3). Changes in settings are typically (mean($s2l$) $>$ 60%) reflected in label changes, with $\sim$92% of changes in PEEP, MV mode. and V_T inducing label changes. The former assessment is biased by few settings changes in some patients and by counting changes unlikely to discretely affect waveform behavior (*e.g.,* trigger sensitivity or mandatory breath rates). However, label-to-settings change coherence ($l2s$) is much lower; less than 10% of label changes are associated with ventilator settings changes. Phenotypes capture LVS variability resulting from care-related changes to MV settings in this data-limited context.

Figures 2 and 3a–d visualize particular aspects of the low-dimensional trajectories in phenotypes for two patients (#34 and #11 of Table1, respectively). Their cases are typical of experiments in record length ($\sim$24 hours), number of ventilator settings changes, and number of identified phenotypic breaths. SI C provides additional examples that show the complexity and heterogeneity of the joint LVS-care processes in time. Each joint LVS-care record shows significant variation over time with complex and non-stationary patterns.

Example 1:. Figure 2 illustrates the low-dimensional trajectory of patient 35. Here, the system is driven by a progression of PEEP reductions and mode changes from APVCMV to supported spontaneous breaths for several hours. Externally labeled PVD types show little identified dyssynchrony as most breaths are identified as normal. However, there is also heterogeneous behavior indicated by labels (a) during the period from 5 to 12 hours under stationary ventilator settings (b). The LVS state vacillates between labels #1 and #3 with notably distinct pV characterization (d) during this period. Irregularity in delivered tidal volume under the APVCMV and in breath length are likely explanations for variability here and in other cases.

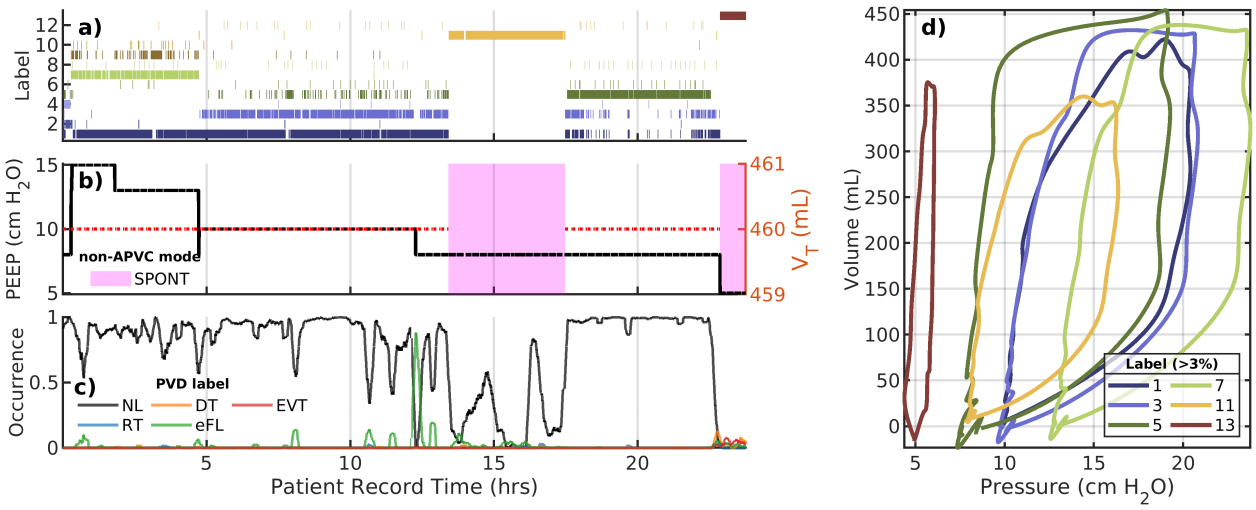

Figure 2: LVS evolution of patient #34. Panels a–c show the trajectory of phenotype labels, ventilator settings, and externally identified PVD, respectively with a common horizontal axis of patient record hours. In b), non-APVCMV ventilator modes are indicated by shaded regions. The panel (d) shows the model image of waveform parameters nearest to the group median, which characterizes breath pV loops of that phenotype (shown with the same color as (a)). The occurrence of label #1 is discontinuous in time and occurs under different PEEP values suggesting waveform shapes vary only in baseline pressure. The PVD-less evolution of the LVS shows much waveform variation separate from ventilator settings changes. Labeling and coloring in this figure do not relate to other figures.

Example 2:. Figure 3 illustrates an analysis of patient #11 whose LVS undergoes multiple changes over a 24-hour data period. A flow trigger increase near 3 hours (b) prompts a phenotype change, and the association with eFL and RT PVD types is reflected in inspiratory coving in pV loops (d, label #1, #2, and #10). The behavior over 7–14.5 hours is identified as normal breaths characterized by quite similar phenotypes (#6 and #8); these could be merged by modifying label specificity (via hyperparameters) or via post-processing. Dyssynchronies return when the flow trigger is returned to its initial value, near 15 hours. Breaths during brief changes to spontaneous breathing around 20 and 23 hours have markedly different pV characterizations (discontinuous label #12, tan). The interim period (20.5–22.5 hours) consists of primarily normal breaths (label #13, brown) under the default adaptive pressure volume control mode.

Intra-label variability: A closer look at label #10 of patient #11. The record of patient #11 (Fig.3) indicates no MV settings changes during 15–21 hours. One phenotypic breath dominates this period (c, blue dashed outline) while various PVD labels intimate variability worth scrutiny. Principal components during this interval (e) reveal structural waveform changes (f,g) that are not clearly identified as sub-phenotypes. While pressure characterizations (f) suggest the differences are attributable to pressure plateau pressure, full characterization indicates ∼35% variability in tidal volume (g) as well. This *continuous* variation lacks a natural discretization without altering the similarity metric, such as including other data. SIC.8 demonstrates a case where intra-label variability may be discretely resolved.

3.2. Cohort-scale phenotyping

The collection of 721 phenotypes generated from pipeline application to 35 individual records can be used to identify systemic LVS features of the cohort. Figure 4 presents key properties and labeled data distributions of 16 similarity-based clusters identified in pool of individual phenotypes. This approach

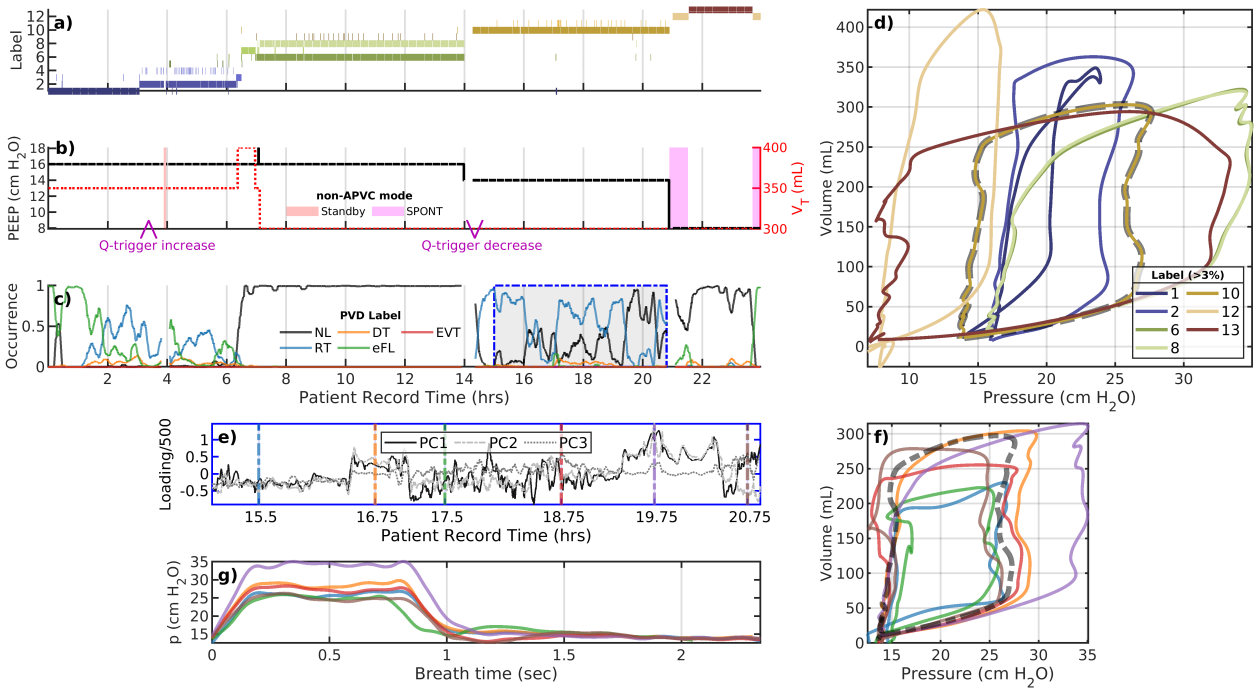

Figure 3: A representative example: patient #11. The plot panels (a–d) are the same as the previous figure. A flow trigger (b, at purple arrows) near 3 and 13 hours are the only MV settings changes besides mode, PEEP, or tidal volume. The lower panels (e–g) examine the variability during the record interval 15–21 hours under stationary ventilator settings. The mean (dashed black line) coincides with the golden pV loop (label #10) in the upper plot. The many distinct breath sub-types identified are more similar than to other main types in the upper plot; as a result, they are grouped together at this choice of hyper-parameters. Internal phenotype variability suggests continuous LVS changes that may not admit a natural discretization. Colors coordinate between panels a) and d), and among e), f), and g). Labeling and coloring in this figure do not relate to other figures.

to segmenting the full dataset, relying on hyperparameters and not generalizing from these data, further minimizes variability while displaying consistent waveform properties and ventilator settings. Labels mix patients (b) while separating PEEP (c), with exceptions for uncommon ventilation modes (d) applied to few patients. Table 2 and Fig 4 quantitatively validate labeling of original data in the general settings. Specifically, labels consistently align with structured properties of the LVS data. Fig 5 shows the associated non-dimensional waveform characterizations; PEEP, tidal volume, and peak pressure features are used to normalize these waveform data elements across patients.

Granularity of cohort meta-characterization depends on UMAP-DBSCAN hyper-parameters (UMAP: neighborhood size 12, minimum distance 1; DBSCAN: epsilon 2.7, min points 5). The small sample size ($N = 721$) lead to robust UMAP representation but high sensitivity to neighborhood size in DBSCAN (SI D). Chosen parameters aimed to maximize the number of phenotypes while easily communicating waveform characterizations in an array of figures; the results are qualitatively similar for nearby parameters. Table 2 summarizes the occurrence and properties of the 16 cohort phenotypes.

3.3. Synthesis

Experiment results provide insight into the structure and heterogeneity of empirically phenotypes constructed for 2-day ARDS patient LVS+care data. Based on equally weighted data elements, these groupings

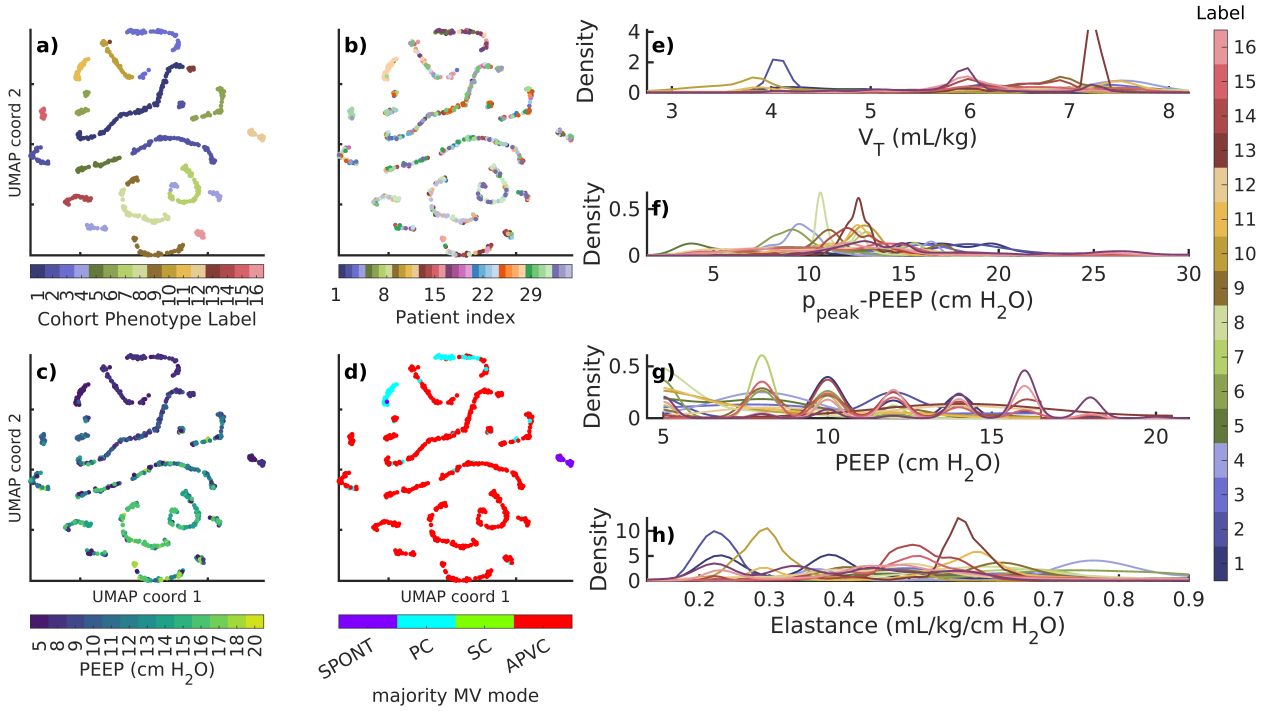

Figure 4: Membership and data properties associated with cohort phenotypes. Points in panels (a–d) correspond to 721 individual phenotypes shown in UMAP coordinates. Labels (a) mix patients (b) while defining empirical partitions of other factors of patient data (c–h). Groupings separate PEEP (c,g) and ventilator modes (d), which are arguably among the most important ventilator feature elements. Structured distributional separation occurs for continuous breath variables such as tidal volume (e), driving pressure (f), and elastance ($V_T/(p_{\max} - p_{\text{base}})$, g). PEEP (c) and ventilator mode (d) of UMAP labels identify the median value of each individual phenotype; probability densities (e–h) are computed from original data and colored according to panel (a). Labels and colors of panel a) define the those of panels e–h) and Fig5. *Modes: spontaneous (SPONT), Pressure controlled (PC), Synchronized controlled (SC), and Adaptive Pressure Volume Controlled (APVC)*

Table 2: Cohort label properties. Columns identify: cohort-level label, percentage of 10-second windows, number of contained patients (N_{pat}), number of contained individual phenotypes (N_{pheno}), median[IQR] of baseline pressures (p_{base}, typically PEEP) and pressure change ($\Delta p := p_{\text{peak}} - p_{\text{base}}$) in cm H_2O, median[IQR] of tidal volumes (V_T) in mL/kg, and dominant associated ventilator mode. Values are determined from breath-level source data whose individual phenotypes share a given cohort phenotype. *=5–10% SPONT, **=10–20% SPONT, ***= 40% SPONT

Label #	Total%	N_{pat}	N_{pheno}	p_{base}	Δp	V_T	$\Delta p/V_T$	MV mode
1	15.5	23	101	10	12.1[3.7]	6.3[1.0]	1.9[0.6]	APVCMV
2	13.8	22	101	12	14.2[3.3]	6.0[1.1]	2.3[1.0]	APVCMV
3	11.4	11	52	8	12.7[4.1]	7.9[1.3]	1.5[0.4]	PCMV*
4	8.4	12	37	14	12.2[6.9]	5.9[0.2]	2.0[1.3]	APVCMV*
5	7.3	11	32	12	15.1[13.6]	5.9[0.1]	2.7[2.4]	APVCMV
6	6.9	17	58	11	13.1[2.9]	6.2[1.3]	2.1[0.5]	APVCMV
7	6.3	8	49	14	12.6[2.9]	6.2[1.3]	1.9[0.7]	APVCMV
8	6.2	11	49	16	13.4[2.3]	6.0[0.6]	2.2[0.6]	APVCMV
9	6.2	9	51	16	15.9[6.0]	5.9[2.8]	2.6[2.4]	APVCMV
10	4.1	11	34	8	9.7[2.7]	6.8[1.5]	1.6[0.4]	APVCMV
11	3.7	5	25	5	10.7[0.2]	6.5[0.7]	1.7[0.2]	PCMV**
12	3.4	14	22	5	8.9[4.1]	7.0[2.4]	1.2[0.8]	APVCMV***
13	2.5	6	10	8	11.1[1.4]	6.6[1.0]	1.7[0.2]	APVCMV
14	1.7	11	27	14	13.3[2.9]	6.0[1.3]	2.0[0.8]	APVCMV
15	1.5	5	14	10	13.5[1.9]	6.5[0.3]	2.1[0.4]	APVCMV
16	1.1	10	16	14	21.3[9.5]	5.6[1.8]	3.7[3.1]	APVCMV

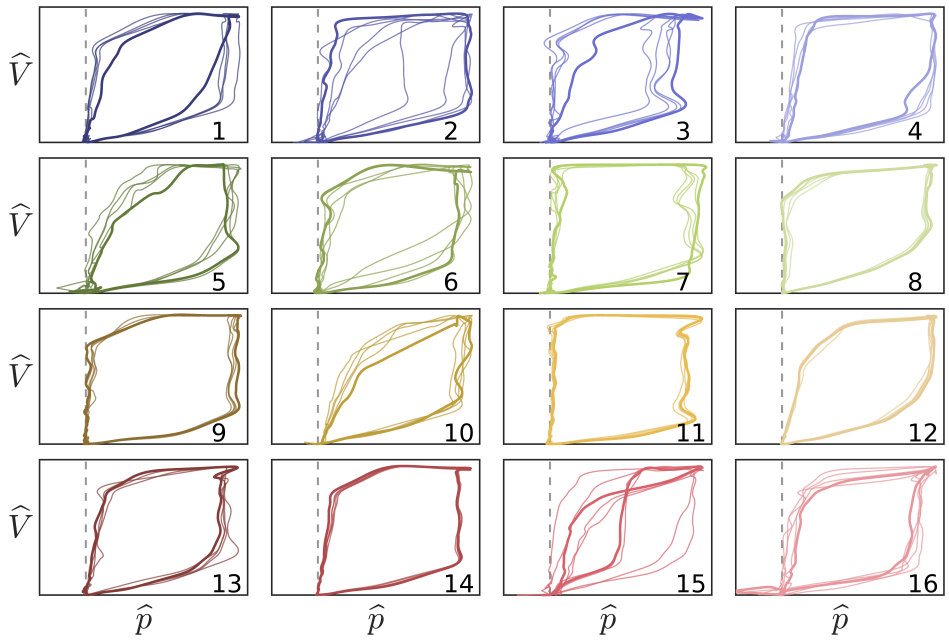

Figure 5: Non-dimensional waveform characterizations. Pressure-volume traces correspond to median (bold) and nearby (thin) window characterizations of each cohort phenotype. Labels and colors correspond to Fig4a. Vertical and horizontal scales axes correspond to $\widehat{V} := V/V_T$ and $\widehat{p} := (p(t) - p_{\text{base}})/(p_{\text{peak}} - p_{\text{base}})$, respectively, per Fig4e–g. The dashed line indicates baseline pressure. Cohort phenotypes differentiate waveform shape characteristics and pressure-volume coordination in conjunction with associated scaling factors. Intra-group variation is naturally high given the low specificity of each type.

primarily organized around ventilation mode and PEEP followed by tidal volume and internal LVS variability, giving a hierarchical organized by key MV settings and waveform instability. LVS behavior is shown to vary in several ways under MV settings stationarity that may be of particular interest for VILI detection and tracking of ARDS progression. This local variability can be resolved by phenotypes as in Fig2a, during hours 5–12, 17–18, 22–23) and Fig3a (hours 3–14). Groupings can also mask important breath heterogeneity if phenotypes are too coarse (*e.g.*, SI Figs C.6e–g and C.8). Resolution of behaviors is improvable through UMAP-DBSCAN parameter optimization or local analysis, such as application of PCA. Heterogeneous phenotypes (*e.g.*, Fig 3e–g) also arise in diverse groups of breaths with continuous local similarity (*e.g.*, Fig1, step 4, label #4 in red). Weighting feature components in the similarity metric can improve sub-type resolution of such behaviors. Additionally, cohort scale analysis is easily computed from a cohort of individual phenotypes. The resulting phenotypes are specific to the data and algorithm parameters and do not generalize, but the modular pipeline process may be adapted and manipulated for more specific purposes. Such categorization provides a coarse but scalable and unified basis for analyzing the evolution of LVSs in terms of their consistent statistical properties.

4. Discussion

This study presents a framework for extracting meaningful, low-dimensional characterizations of lung-ventilator system (LVS) states from observable data of managed patient-ventilator systems. Consequently, the observable LVS data, whose contents define a limited context of application, are reduced to a discrete set

of patient-level phenotypes. The phenotyping pipeline was developed as a generalized process to accommodate many methodological choices. This contrasts with a construction of generalized results, as source data insufficiently sample the breadth of MV processes including other ventilators, sites, patient characteristics, and care protocols. The approach instead provides a practical means for researchers to explore LVS evolution from data with significant heterogeneity caused by ventilator adjustments and changes in patient-ventilator dynamics. In particular, the data-specific phenotypes naturally disentangle effects of MV settings from the human-machine interaction of the LVS which typology. This is an important development for research aimed at personalizing MV to support patients with ARDS and to mitigate VILI because it facilitates hypothesis development about the effects of current care over time. The phenotype identification process is intended for research rather than for MV decision support. The need for a framework to develop hypotheses about temporal effects of MV and local outcome validation targets from retrospective data motivated the developments presented in this work.

Experiments on 35 ICU ARDS patients, including 8 COVID-19 cases from 2020, used naive segmentation with uniform LVS feature weights. Fixed similarity assumptions and hyperparameter optimization ranges were used to define individual-scale phenotypes, which were later compared in cohort-scale phenotyping. Individual results showed that phenotypes primarily organize by ventilator mode, PEEP, and tidal volume. This effectively separated care processes from the patient-ventilator component of the data generating system, while analysis of periods under persistent VM settings showed LVS changes to be more complicated. Whereas MV settings changes are abrupt, LVSs exhibit a variety of behaviors including continuous but non-monotonic progression (Fig3), transient behavior (FigC.6), and alternation between both similar and non-similar breath patterns (Fig.C.8). Some variation can be refined into separate phenotypes by adjusting phenotype resolution, while continuous changes resist discretization without feature weighting or other source data. Continuous changes may result from apparatus properties (such as changes in tube compliance, resistance from accumulated moisture and bends, and leakage) as well as effects of patient sedation. However, they may also suggest progressive effects of lung physiology under MV. Investigating such behaviors first requires identifying categories, like those developed in this work, for which heterogeneity can be calculated.

4.1. Validation and Interpretation

The created typologies are based on similarities among observable data, making them phenotypes of patient-ventilator-care data representative of LVS behaviors. Clinical validation of output requires quantitative comparison with patient state or conditions [56], but such biomarkers of breath behavior do not currently exist. Further, global outcomes (discharge disposition, 30-day mortality, *etc.*) are unlikely to relate to local behavior during 1–2-day partial records of MV encounters. Investigation demonstrated label consistency in relation to changes in PEEP, ventilator mode, and tidal volume for phenotypes based on naive hyperparameters and an uninformative similarity metric. The analysis qualitatively validates practical application by identifying what LVS behaviors computed phenotypes did and did not differentiate as well as what variability can be isolated via hyperparameter tuning.

Phenotypes are more granular than a ventilator settings-based classification (SI Table B.3) and more generalized than PVD labels, which target specific behaviors. Label heterogeneity suggests potential label sub-types so that hierarchical or multi-stage clustering are important refinements in future applications. Although 10-second window scalar phenotypes are directly incomparable to breath-wise vector types of PVD,

changes in label-described behavior strongly coordinate with changes in PVD type. Notably, phenotype variability analysis and PVD labels identified qualitatively similar temporal patterns (*e.g.* §3.1 and SISI C) without dyssynchrony labels informing LVS descriptors. Additionally, esophageal pressures were not encoded into phenotypes but are required to confirm certain PVD types [21].

Cohort labels demonstrably partition data into groups albeit with an expected high degree of variability given the reduction of ∼1.5M breaths to 16 categories. Their identification required waveform component normalization to ensure patient comparability that would benefit from stratified analysis based on mode, PEEP, and primary control variables. Despite the coarseness of categories, signs of dyssynchrony are apparent in these median pV shapes such as ineffective triggering (sub-baseline pressures in #5, #15, and #16) and flow limitation (inspiratory coving in #3, #11, and #15). This indicates that some of the cohort scale phenotypes, while broader and less specific than PVD types, center on elements of dyssynchronous behavior. Including PVD labels or other physiological information in feature descriptors may better align phenotypes with PVD labels in applications targeting LVS specific behaviors.

4.2. Innovations, Limitations, and Improvements

This work discretized joint patient-ventilator-care system data as holistic units to overcome limitations on analysis imposed by data complexity and heterogeneity. This phenotyping approach is generalizable, suitable for other datasets, and can accommodate different feature and clustering options. The process and results are geared toward data-driven research use rather than clinical informatics or clinical decision support. In scientific application, practitioner guidance is required to inform data features and their importance relative to target observable investigated via LVS phenotypes. Nevertheless, the developed method and specific implementation assumes certain conditions and has limitations.

The presented phenotyping results are neither generalizable nor clinically validated. This is a consequence of insufficient data to sample all LVS behavior, lack of a "gold standard" MV breath typology, and absence of LVS state biomarkers. Target biomarkers would aid in feature design of the pipeline through clinical knowledge and physiology regarding how such observables relate to LVS data used here.

The pipeline ignored uncommon esophageal pressure data, which are essential to confirm certain dyssynchronies, because they require high model resolution to resolve and have inconsistencies (gaps, drift) that limit continuous time characterization. The waveform parametrization also relies on ventilator-identified breath cycles, so the pipeline lacks the flexibility needed to identify double-triggered PVD events that occur over multiple ventilator cycles. Analysis omitted important potential influences such as neuromuscular blockade use, position/posture, and airway secretions whose data were not available. LVS descriptors easily can incorporate these factors to better resolve care-stationary periods and more precisely resolve LVS variability. Additionally, data reflect only one ventilator model; additional harmonization is needed to compare breaths generated by different ventilators because mode settings and pV observation points may differ.

Finally, the group identities of empirical phenotypes depend directly on hyperparameters that govern similarity and specificity. Fixed UMAP parameters reflected a constant local similarity assumption, while DBSCAN parameters were optimized over a small domain to account for differences in record length. The dimensional reduction process employed a similarity metric with uniform feature weights to limit external assumptions. Practical applications should use optimizing weights to emphasize key features of interest. Further improvements can easily involve outer-loop targeting of application-specific objectives beyond the generalized scope of this work.

4.3. Concluding Remarks

This work developed a flexible categorization process for context-constraining data timeseries from patient-ventilator system under managed care. The research outlined a process of empirically discretizing relevant observational data capable of isolating patient-ventilator dynamics from care processes and labeling data subsets based on similarity. Assessing phenotype local heterogeneity is an essential first step in temporal analysis of MV patient data within the context of applied care. Ongoing work toward formulating hypotheses about system trajectories related to applied care, local variability, and outcome motivated providing a shared low-dimensional basis for LVS comparison, which also motivated the construction of cohort-scale phenotypes Converting LVS dynamics to sequential progression through finite states allows the application of symbolic dynamics [57, 58, 59], game theory [60], discrete-time Markov chains, and large language models. Such tools can extend this works' investigation to patient trajectory patterns –*temporal* phenotypes– to improve understanding of dynamical effects of current MV protocols on patient LVSs.

Acknowledgments

This work is supported by National Heart Lung and Blood Institute awards 5R01HL151630 "Predicting and Preventing Ventilator-Induced Lung Injury" (DJA), K23HL145011 "The Detection, Quantification, and Management of Ventilator Dyssynchrony" (PDS), and K24HL168225 "Mentoring and Patient-Oriented Research in Clinical Informatics and Data Science" (TDB). Thanks as always to Meg Rebull for local administrative support.

Declarations of Interest

None. The authors have no conflicts of interest to disclose.

Declaration of Generative AI and AI-assisted Technology Use

The author used ChatGPT 3.5 to suggest rephrasings of complex sentences in early drafts. No generative AI output was used directly. No other AI-assisted technologies were employed.

References

[1] Lung National Heart and Blood Institute ARDS Clinical Trials Network. Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. *New England Journal of Medicine*, 351(4):327–336, 2004.

[2] Luciano Gattinoni, Giuseppe Citerio, and Arthur S Slutsky. Back to the future: Ards guidelines, evidence, and opinions. *Intensive Care Medicine*, 49(10):1226–1228, 2023.

[3] Eddy Fan, Jesus Villar, and Arthur S Slutsky. Novel approaches to minimize ventilator-induced lung injury. *BMC medicine*, 11(1):1–9, 2013.

[4] Dan Stieper Karbing, Savino Spadaro, Nilanjan Dey, Riccardo Ragazzi, Elisabetta Marangoni, Francesca Dalla Corte, Federico Moro, David Lodahl, Niklas Schurmann Hansen, Robert Winding, et al. An open-loop, physiologic model–based decision support system can provide appropriate ventilator settings. *Critical Care Medicine*, 46(7):e642–e648, 2018.

[5] Gerard F Curley, John G Laffey, Haibo Zhang, and Arthur S Slutsky. Biotrauma and ventilator-induced lung injury: clinical implications. *Chest*, 150(5):1109–1117, 2016.

[6] Arthur S Slutsky and V Marco Ranieri. Ventilator-induced lung injury. *New England Journal of Medicine*, 369(22):2126–2136, 2013.

[7] Acute Respiratory Distress Syndrome Network. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *New England Journal of Medicine*, 342(18):1301–1308, 2000.

[8] Bulleri Enrico, Fusi Cristian, Bambi Stefano, and Pisani Luigi. Patient-ventilator asynchronies: types, outcomes and nursing detection skills. *Acta Bio Medica: Atenei Parmensis*, 89(Suppl 7):6, 2018.

[9] Lluís Blanch, Ana Villagra, Bernat Sales, Jaume Montanya, Umberto Lucangelo, Manel Luján, Oscar García-Esquirol, Encarna Chacón, Anna Estruga, Joan C Oliva, et al. Asynchronies during mechanical ventilation are associated with mortality. *Intensive care medicine*, 41:633–641, 2015.

[10] Roy G Brower and Gordon D Rubenfeld. Lung-protective ventilation strategies in acute lung injury. *Critical care medicine*, 31(4):S312–S316, 2003.

[11] Nicola Petrucci and Walter Iacovelli. Lung protective ventilation strategy for the acute respiratory distress syndrome. *Cochrane Database of Systematic Reviews*, (3), 2007.

[12] Yuda Sutherasan, Maria Vargas, and Paolo Pelosi. Protective mechanical ventilation in the non-injured lung: review and meta-analysis. *Annual Update in Intensive Care and Emergency Medicine 2014*, pages 173–192, 2014.

[13] Marc Moss and David M Mannino. Race and gender differences in acute respiratory distress syndrome deaths in the United States: an analysis of multiple-cause mortality data (1979–1996). *Critical care medicine*, 30(8):1679–1685, 2002.

[14] Didier Dreyfuss and Rolf Hubmayr. What the concept of vili has taught us about ards management. *Intensive care medicine*, 42:811–813, 2016.

[15] Shea E Cochi, Jordan A Kempker, Srinadh Annangi, Michael R Kramer, and Greg S Martin. Mortality trends of acute respiratory distress syndrome in the United States from 1999 to 2013. *Annals of the American Thoracic Society*, 13(10):1742–1751, 2016.

[16] Michael G Seneff, Jack E Zimmerman, William A Knaus, Douglas P Wagner, and Elizabeth A Draper. Predicting the duration of mechanical ventilation: the importance of disease and patient characteristics. *Chest*, 110(2):469–479, 1996.

[17] Marya D Zilberberg, Brian H Nathanson, Judy Ways, and Andrew F Shorr. Characteristics, hospital course, and outcomes of patients requiring prolonged acute versus short-term mechanical ventilation in the united states, 2014–2018. *Critical care medicine*, 48(11):1587–1594, 2020.

[18] Louise Rose and Ben Messer. Prolonged mechanical ventilation, weaning, and the role of tracheostomy. *Critical Care Clinics*, 40(2):409–427, 2024.

[19] Peter D Sottile, David Albers, Carrie Higgins, Jeffery Mckeehan, and Marc M Moss. The association between ventilator dyssynchrony, delivered tidal volume, and sedation using a novel automated ventilator dyssynchrony detection algorithm. *Critical Care Medicine*, 46(2):e151, 2018.

[20] Behnood Gholami, Timothy S Phan, Wassim M Haddad, Andrew Cason, Jerry Mullis, Levi Price, and James M Bailey. Replicating human expertise of mechanical ventilation waveform analysis in detecting patient-ventilator cycling asynchrony using machine learning. *Computers in biology and medicine*, 97:137–144, 2018.

[21] Peter D Sottile, David Albers, Bradford J Smith, Marc M Moss, et al. Ventilator dyssynchrony–detection, pathophysiology, and clinical relevance: A narrative review. *Annals of Thoracic Medicine*, 15(4):190, 2020.

[22] Tom Bakkes, Anouk van Diepen, Ashley De Bie, Leon Montenij, Francesco Mojoli, Arthur Bouwman, Massimo Mischi, Pierre Woerlee, and Simona Turco. Automated detection and classification of patient–ventilator asynchrony by means of machine learning and simulated data. *Computer Methods and Programs in Biomedicine*, 230:107333, 2023.

[23] NL Loo, Yeong Shiong Chiew, Chee Pin Tan, MB Mat-Nor, and Azrina Md Ralib. A machine learning approach to assess magnitude of asynchrony breathing. *Biomedical signal processing and control*, 66:102505, 2021.

[24] Karen G Mellott, Mary Jo Grap, Cindy L Munro, Curtis N Sessler, Paul A Wetzel, Jon O Nilsestuen, and Jessica M Ketchum. Patient ventilator asynchrony in critically ill adults: frequency and types. *Heart & Lung*, 43(3):231–243, 2014.

[25] Candelaria De Haro, Ana Ochagavia, Josefina López-Aguilar, Sol Fernandez-Gonzalo, Guillem Navarra-Ventura, Rudys Magrans, Jaume Montanyà, and Lluís Blanch. Patient-ventilator asynchronies during mechanical ventilation: current knowledge and research priorities. *Intensive care medicine experimental*, 7:1–14, 2019.

[26] Deepak K Agrawal, Bradford J Smith, Peter D Sottile, and David J Albers. A damaged-informed lung ventilator model for ventilator waveforms. *Frontiers in physiology*, 12, 2021.

[27] Cong Zhou, J Geoffrey Chase, Qianhui Sun, Jennifer Knopp, Merryn H Tawhai, Thomas Desaive, Knut Möller, Geoffrey M Shaw, Yeong Shiong Chiew, and Balazs Benyo. Reconstructing asynchrony for mechanical ventilation using a hysteresis loop virtual patient model. *BioMedical Engineering OnLine*, 21(1):1–20, 2022.

[28] JN Stroh, Bradford J Smith, Peter D Sottile, George Hripcsak, and David J Albers. Hypothesis-driven modeling of the human lung-ventilator system: A characterization tool for acute respiratory distress syndrome research. *Journal of Biomedical Informatics*, page 104275, 2022.

[29] Yuhong Chen, Kun Zhang, Cong Zhou, J Geoffrey Chase, and Zhenjie Hu. Automated evaluation of typical patient–ventilator asynchronies based on lung hysteretic responses. *BioMedical Engineering OnLine*, 22(1):102, 2023.

[30] Deepak K Agrawal, Bradford J Smith, Peter D Sottile, George Hripcsak, and David J Albers. Quantifiable identification of flow-limited ventilator dyssynchrony with the deformed lung ventilator model. *Computers in Biology and Medicine*, page 108349, 2024.

[31] Yanran Wang, JN Stroh, George Hripcsak, Cecilia C Low Wang, Tellen D Bennett, Julia Wrobel, Caroline Der Nigoghossian, Scott W Mueller, Jan Claassen, and DJ Albers. A methodology of phenotyping ICU patients from EHR data: high-fidelity, personalized, and interpretable phenotypes estimation. *Journal of biomedical informatics*, 148:104547, 2023.

[32] Gal Yona, Roee Aharoni, and Mor Geva. Narrowing the knowledge evaluation gap: Open-domain question answering with multi-granularity answers. *arXiv preprint arXiv:2401.04695*, 2024.

[33] Claude Guérin, Laurent Papazian, Jean Reignier, Louis Ayzac, Anderson Loundou, Jean-Marie Forel, Investigators of the Acurasys, and Proseva Trials. Effect of driving pressure on mortality in ards patients during lung protective mechanical ventilation in two randomized controlled trials. *Critical Care*, 20:1–9, 2016.

[34] Todd W Rice, Arthur P Wheeler, Gordon R Bernard, Douglas L Hayden, David A Schoenfeld, Lorraine B Ware, ARDS Network, National Institutes of Health, et al. Comparison of the spo2/fio2 ratio and the pao2/fio2 ratio in patients with acute lung injury or ards. *Chest*, 132(2):410–417, 2007.

[35] Wei Chen, David R Janz, Ciara M Shaver, Gordon R Bernard, Julie A Bastarache, and Lorraine B Ware. Clinical characteristics and outcomes are similar in ards diagnosed by oxygen saturation/fio2 ratio compared with pao2/fio2 ratio. *Chest*, 148(6):1477–1483, 2015.

[36] Peter D Sottile, Bradford Smith, Marc Moss, and David J Albers. The development, optimization, and validation of four different machine learning algorithms to identify ventilator dyssynchrony. *medRxiv*, 2023.

[37] Cheolhyeong Park and Deokwoo Lee. Classification of respiratory states using spectrogram with convolutional neural network. *Applied Sciences*, 12(4):1895, 2022.

[38] Pavel Sakov, Geir Evensen, and Laurent Bertino. Asynchronous data assimilation with the EnKF. *Tellus, Series A: Dynamic Meteorology and Oceanography*, 62(1):24–29, 2010.

[39] Jonas Latz. Bayesian inverse problems are usually well-posed. *SIAM Review*, 65(3):831–865, 2023.

[40] Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

[41] Farzana Anowar, Samira Sadaoui, and Bassant Selim. Conceptual and empirical comparison of dimensionality reduction algorithms. *Computer Science Review*, 40:100378, 2021.

[42] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[43] David N Lawley and Adam E Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229, 1962.

[44] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[45] John Healy and Leland McInnes. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1):82, 2024.

[46] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

[47] Mahamed GH Omran, Andries P Engelbrecht, and Ayed Salman. An overview of clustering methods. *Intelligent Data Analysis*, 11(6):583–605, 2007.

[48] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[49] Asa Ben-Hur, David Horn, Hava T Siegelmann, and Vladimir Vapnik. Support vector clustering. *Journal of machine learning research*, 2(Dec):125–137, 2001.

[50] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[51] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[52] Connor Meehan, Stephen Meehan, and Wayne Moore. Uniform manifold approximation and projection (UMAP v4.2). *MATLAB Central File Exchange*, 2022. `https://www.mathworks.com/matlabcentral/fileexchange/71902`.

[53] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

[54] Gulanbaier Tuerhong and Seoung Bum Kim. Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert systems with applications*, 41(4):1701–1707, 2014.

[55] C Radhakrishna Rao. The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 329–358, 1964.

[56] Jennifer C Goldsack, Andrea Coravos, Jessie P Bakker, Brinnae Bent, Ariel V Dowling, Cheryl Fitzer-Attas, Alan Godfrey, Job G Godino, Ninad Gujar, Elena Izmailova, et al. Verification, analytical validation, and clinical validation (v3): the foundation of determining fit-for-purpose for biometric monitoring technologies (biomets). *npj digital Medicine*, 3(1):55, 2020.

[57] José M Amigó, Karsten Keller, and Valentina A Unakafova. Ordinal symbolic analysis and its application to biomedical recordings. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2034):20140091, 2015.

[58] Douglas Lind and Brian Marcus. *An introduction to symbolic dynamics and coding.* 2nd edition, 2021.

[59] Yoshito Hirata and José M Amigó. A review of symbolic dynamics and symbolic reconstruction of dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(5), 2023.

[60] Dario Bauso. *Game theory with engineering applications.* SIAM, 2016.

[61] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation.* SIAM, 2005.

[62] François-Xavier Le Dimet and Olivier Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A: Dynamic Meteorology and Oceanography*, 38(2):97–110, 1986.

[63] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011.

SI A. Inference of model-based parameters

The waveform parametrization adopted in this work is briefly outlined below; further details, error analysis, and validation of the parametrization are found in [28]. Briefly, the model transforms waveform data y^{obs} on a continuous time window I into a parameter vector $\mathbf{a}$ under the assumption that PEEP and breath period θ are locally constant. A differential equation models the state variable y (pressure, volume, or another waveform variable of interest):

$$\frac{dy}{dt} + g \cdot (y - y_0) = \varphi(t; \mathbf{a}, \theta) \tag{A.1}$$

where t is time, g is a smoothing parameter, y_0 is baseline or reference state (such as PEEP when y is pressure, and zero when y is volume), and $\varphi(t)$ is a time-dependent function of local amplitude parameters $a \in \mathbb{R}^M$ and breath period θ. The model, Eq(A.1), defines a map $\mathbf{a} \mapsto y(t)$ that simulates a state trajectory y from a periodic step function φ modulated by parameters $\mathbf{a}$:

$$\varphi(t; \mathbf{a}, \theta) = \frac{1 - e^{-g\Delta t}}{g} \sum_{i=1}^{M} a_i \left[\!\left[(i-1) \leq \frac{\widehat{t}}{\Delta t} < i \right]\!\right]. \tag{A.2}$$

where $\widehat{t} := t(\bmod \theta)$ is time within the breath, which is divided into M equal epochs of width $\Delta t = \theta/M$. Note that φ is define as piecewise constant EqA.2; on interval $I_i := \widehat{t} + \Delta t \cdot [i-1, i]$, Eqn(A.1) is simply $y' + gy = C_i$ for some constant C_i (regardless of whether epochs are equal).

Structurally, the model can be written as $y' + gy = f(t;\omega)$ where $\omega = (a, y_0, \theta) \in \mathbf{R}^{M+2}$ is the full parameter vector. The function $f(a, y_0) = \varphi(a) - gy_0$ is a simple linear function that is constant on subintervals $\{I_i : i = 1..M\}$ that partition the local breath domain I. Thus on each subinterval, the equation is $Ly_i = C_i;\ y(0) = y_{i,0}$ where $L = \frac{d}{dx} + g\mathcal{I}$ is the linear differential operator, C_i is the value of φ on I_i, $y_{i,0}$ is the initial data for I_i, and $\mathcal{I}$ is the identity function. The initial value problem has a unique solution

$$y_i = \frac{C_i}{g}(1 - \exp[-gt]) + y_{i,0}\exp[-gt] \text{ for } t \in I_i. \tag{A.3}$$

The solution on I is found by solving over I_i in sequence using terminal values as initial data for the next subinterval. That is, $y_{i+1,0} = y_i(T_i)$ where T_i is the left endpoint of I_i for $i = 1..M-1$, and $y_1(0) := y^{obs}(0)$ This solution $L^{-1}f$ is absolutely continuous and uniquely specified by parameters.

SI A.1. Parameter identification via model inversion

The model defines a map from parameters to trajectories, but what is desired to parameterize data is a map from observable data to parameter distributions $y^{obs} \mapsto \{\mathbf{a}\}$ which requires model inversion [61]. Optimal parameter distributions for pressure and volume data are generated by applying an ensemble Kalman-like smoother using a moderate resolution model ($M = 28$) to sequentially estimate parameters for each 10-second block of data using 1.6 second windows with 0.8 second overlaps.

Specifically, a windowed application of the Ensemble Kalman Smoother (EnKS) [38] is used for parameter estimation. This technique approximates a Bayesian update by minimizing the discrepancy between a model forecast trajectory and corresponding external observations $y^{obs}_{\Omega(k)}$ $(k = 1..L)$ for the kth window $\Omega(k)$ using an ensemble forecast over the local times $\{t_k\},\ k = 1..L$. The target parameters a are those that minimize the standard 4Dvar cost function[62]

$$\mathcal{J}(y(\mathbf{a})) = \frac{1}{2\gamma}\left\|\mathbf{P}_0^{-1/2}(y - \overline{y}_0)\right\|^2 + \frac{1}{2}\sum_{k\in\Omega}\left\|\mathbf{R}^{-1/2}(y_k^{obs} - y_k)\right\|^2 \tag{A.4}$$

Here, k indexes time points of data within an inversion subwindow and γ is a covariance inflation factor that adjusts the weight of the data in the analysis. The initial data and parameters for the ensemble are characterized by the mean $\overline{y}_0$ and matrices $\mathbf{P}_0^{1/2}\mathbf{P}_0^{T/2}$, and $\mathbf{R} = \mathbf{R}^{1/2}\mathbf{R}^{T/2}$ represents model and observation error covariances, respectively . The symbols y_k^f denote simulated observations generated by the model using the initial state y and parameters a, and Ω is a subset of time indices for data assimilation. The ensemble of parameter variations around the optimal mean is retained to reinitialize the model forecast at time t_S $(0 < S \leq L)$ for the next forecast-correction iteration.

Computationally, the ensemble state anomaly matrix $\mathbf{P}^{1/2}$ is constructed column-wise by deviations from their mean and scaled by $\sqrt{N}$. A feature of ensemble-Kalman type approximations to Eq(A.4) is that each potential solution lies within the subspace $\mathcal{W}_N = \text{span}(\overline{y}^f + \mathbf{P}^{1/2}w)$. Therefore, each possibility is determined by a control vector $w \in \mathbb{R}^N$ where N is the number of elements in the ensemble. Exploiting this detail, an optimal subspace solution w^* is defined as the variable that minimizes the cost function over $\mathcal{W}_N$. This is

given by

$$\widehat{\mathcal{J}}(w) = \frac{1}{2\gamma}\|w\|^2 + \frac{1}{2}\sum_{k\in\Omega}\left\|c_k^f - \mathbf{Z}_k w\right\|^2 \tag{A.5}$$

where $c_k^f := \mathbf{R}^{-1/2}\left(y_k^{obs} - \bar{y}_k^f\right)$ are the mean forecast error and $\mathbf{Z}_k := \mathbf{R}^{-1/2}\mathbf{P}^{1/2}$ are ensemble forecast anomalies at the time of observation y_k. This has a simple quadratic form

$$\widehat{\mathcal{J}}(w) = \frac{1}{2}w^T\mathbf{Q}_\gamma w + b^T w + \frac{1}{2}\epsilon^2 \tag{A.6}$$

where $\mathbf{Q}_\gamma = \gamma^{-1}\mathcal{I}_N + \sum_{k\in\Omega}\mathbf{Z}_k^T\mathbf{Z}_k$ is the Hessian matrix, $b := -\sum_{k\in\Omega}\mathbf{Z}_k c^f$ is the linear coefficient vector, and $\epsilon^2 = \sum_{k\in\Omega}\left\|c_k^f\right\|^2$ represents the initial trajectory error. The minimum of the cost function is characterized the vanishing gradient of $\widehat{\mathcal{J}}(w)$ with respect to w. The corresponding solution of $\nabla\widehat{\mathcal{J}}(w) = \mathbf{Q}_\gamma w + b = 0$ is the optimal coefficient

$$w^* = -\mathbf{Q}_\gamma^{-1}b = \left(\gamma^{-1}\mathcal{I}_N + \sum_{k\in\Omega}\mathbf{Z}_k^T\mathbf{Z}_k\right)^{-1}\left(\sum_{k\in\Omega}\mathbf{Z}_k^T c^f\right). \tag{A.7}$$

Once the optimal trajectory $y^* = y^f + \mathbf{P}^{1/2}w^*$ is known, so are the model parameters used to generate it. To update the prior distribution for the next forecast window, the ensemble variations are recentered around this optimal solution point at some future time point t_S by reweighting the forecast covariance factors by those of the inverse Hessian matrix: $\mathbf{P}_S^{1/2} \leftarrow \mathbf{P}_S^{1/2}\mathbf{Q}_\gamma^{-1/2}$.

SI B. Label changes in relation to vent settings changes

Table B.3 shows that most vent settings changes are accompanied by changes in labels. However, very few phenotype label changes correspond to changes in vent settings. *Over 64% of ventilator settings changes are identified in label changes.* A larger proportion (>87%) are identified when limited to PEEP, tidal volume, and model changes, which induce significant waveforms changes compared with other settings such as mandatory breath rate (set_rate). *Few (8%) identified changes in label, however, are directly associated in time with ventilator settings changes.*

SI C. Individual Experiments, continued

This supplement continues illustrated examples of §3.1.

Figure C.6 panels a–d illustrate the analysis of Patient #2 whose data consists of 7 record hours with one simple ventilator setting change. Only ventilator PEEP (a) is changed while there are three primary behaviors identified (b,d). The reduction of PEEP occurs about 2 hours following a rise in early flow limited breaths (eFL, panel c). This PEEP change (from 8 to 5 cm H_2O) shifts peak pressure from 16 to 12 cm H_2O for about an hour, at which time higher esophageal pressures returns. These breaths are identified as normal (NL) [19]. Increased specificity may be pursued by local segmentation or other dimensional reduction methods.

Table B.3: N_s indicate the number of ventilator settings changes in set_PEEP, set_ptrigger, set_qtrigger, set_rate, set_fio2, set_ie, set_flowpat, set_mode, set_vt, and vt_set. N_l indicates the number of persistent label changes, counting those lasting longer than 30 seconds, to omit isolated transient changes and variability occurring as mixed-breath types (*e.g.*, FigC.7 during 11–14 hours, characterized by both alternation between labels #10 and #13 labels and changes in ML-identified PVD type). Column 's2l' indicates the percentage of vent settings changes that occur with a label change within 100 seconds. Column 'l2s' indicate the number of label changes that occur within 100 seconds of vent changes.

Patient #	N_s	s2l (%)	N_l	l2s (%)	ID	N_s	s2l (%)	N_l	l2s (%)
1	45	73.33	66	16.67	18	11	45.45	138	3.62
-	1	-	1	-	19	14	85.71	296	6.42
2	2	100.00	35	8.57	20	10	70.00	25	20.00
3	25	36.00	356	1.97	21	37	78.38	154	18.83
4	1	100.00	46	4.35	22	76	5.26	175	0.57
5	99	56.57	288	9.72	23	104	86.54	695	6.47
6	20	75.00	342	2.92	24	222	59.01	590	6.78
7	3	33.33	73	1.37	25	137	54.01	166	14.46
8	23	78.26	328	7.01	26	10	60.00	38	21.05
9	59	45.76	177	6.21	27	24	70.83	451	3.99
10	14	78.57	260	6.15	28	47	100.00	629	1.75
11	48	29.17	50	24.00	29	73	45.21	431	7.42
12	0	-	4	0	30	37	62.16	713	2.66
13	83	80.72	370	12.70	31	83	77.11	421	9.03
14	13	76.92	1000	1.40	32	50	62.00	380	8.42
15	51	98.04	265	8.30	33	220	42.27	650	9.38
16	57	40.35	246	8.94	34	39	25.64	296	4.73
17	18	88.89	460	3.91	35	12	75.00	424	2.12
					mean	52	64.57	325	8.00

A closer look at label 1 of patient #2:. The first principal component loadings (panel e, black) for LVS descriptors over the first 5-hour period track the sequence of normal and eFL PVD labels (f, shown as 5-minute statistics for clarity). Within the same breath phenotype (label 1), the *sign* of the component loading statistically the eFL PVD labels (AUROC=0.8718); high positive values are associated with eFL breaths (f,g; green) where pressure maxima proceed volume maxima. These LVS variations result from changes in the patient component, as there is no change of ventilator settings. Note that direct correlation between continuous loading values on 10 second windows and statistical breath-wise binary PVD label is not well-defined while binary-to-binary comparison is.

The patient #10 (Figure C.7) dataset is nearly twice as long with again only one PEEP change occurring after 10.5 hours of the 15.6 hour record. Breaths are stably identified as normal-type until about 8 hours, occupying two cluster-identified similar breath shapes. This is followed briefly by eFL breaths and a transition to a new characterization (label #8, light green) for about 30 minutes. In the following period (9–14 hours), breaths are characterized by lower pressure maxima (label #10, gold); these are associated/identified with reverse-trigger breaths (primarily RTm) and waveforms featuring pronounced inspiratory pressure drop. The reduction in PEEP slightly increases the incidence of normal breaths during 11–14 hours although this results in the more frequent appearance of shallow breaths (label #13, red).

SI C.1. Intracluster variability in relation to PVD labels

This final example shows a case where principal component loadings strongly tie to difference between normal and eFL breaths in patient #8, label #2.

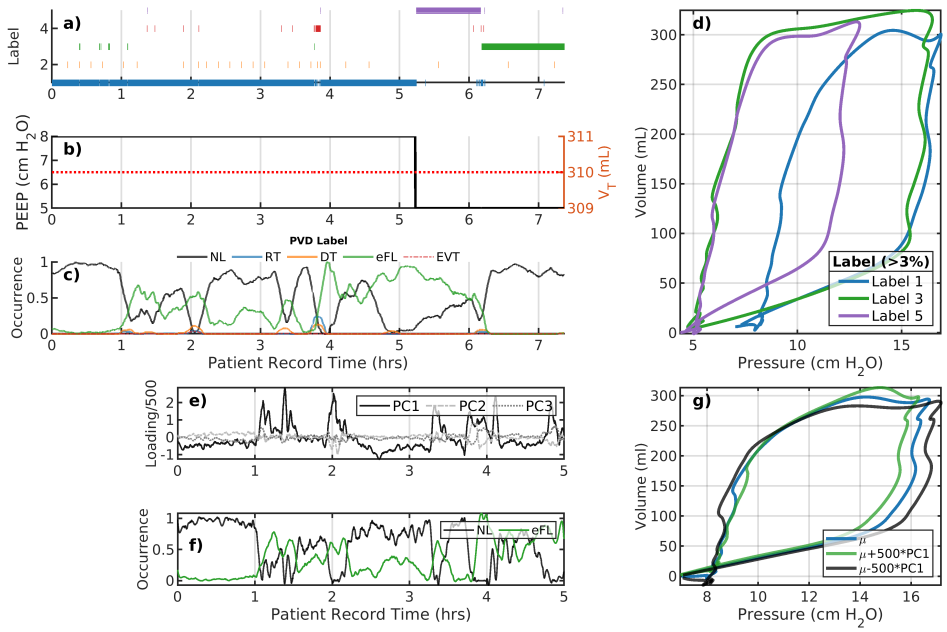

Figure C.6: Analysis of patient #2 LVS data (a–d) and the initial a 5-hour interval (e–g). Panels a–c correspond to changes in ventilator settings, segmentation labels, and identified PVD type, respectively. The horizontal axis for these panels is the patient record time in hours. The panel (d) shows the model image of segmented data median parameters, which characterize the pV loops of breaths with that label (shown with the same color). Evolution of the LVS can be parsed pictorially from these figures. Large positive variations in the first principal component loading (e, black) for the initial 5-hour period align with PVD labels indicating eFL type breaths (f) for this period. Specifically, this suggests discrimination of breaths shapes (g) can be differentiated using qualitatively criterion on local loadings or other segmentation. Labeling and coloring in this figure do not relate to other figures.

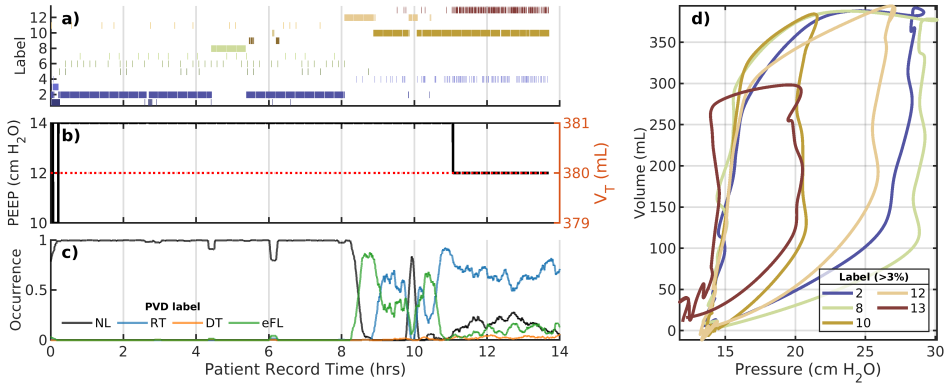

Figure C.7: The patient #10 evolution includes only PEEP change. The layout is the same as panels a–d of the previous figure. Under constant ventilator settings, breaths undergo transition several times including intervals of PVD prior to PEEP change around 10.5 hours. A 1-hour long shift from label #2 to #8 occurs around 8 hours during which breaths decrease peak pressure and includes an increase in eFL and RT PVD occurrence. After the PEEP change, LVS behavior is dyssynchronous and primarily centered around the characterization with label #10. Labeling and coloring in this figure do not relate to other figures.

SI D. Influence of Hyperparameter choices on cohort phenotypes

For each of the 721 individual phenotypes, feature vectors defined by the 5-number summaries of period, PEEP, maxima of volume and pressure, ventilator settings, and estimated parameters of range-normalized

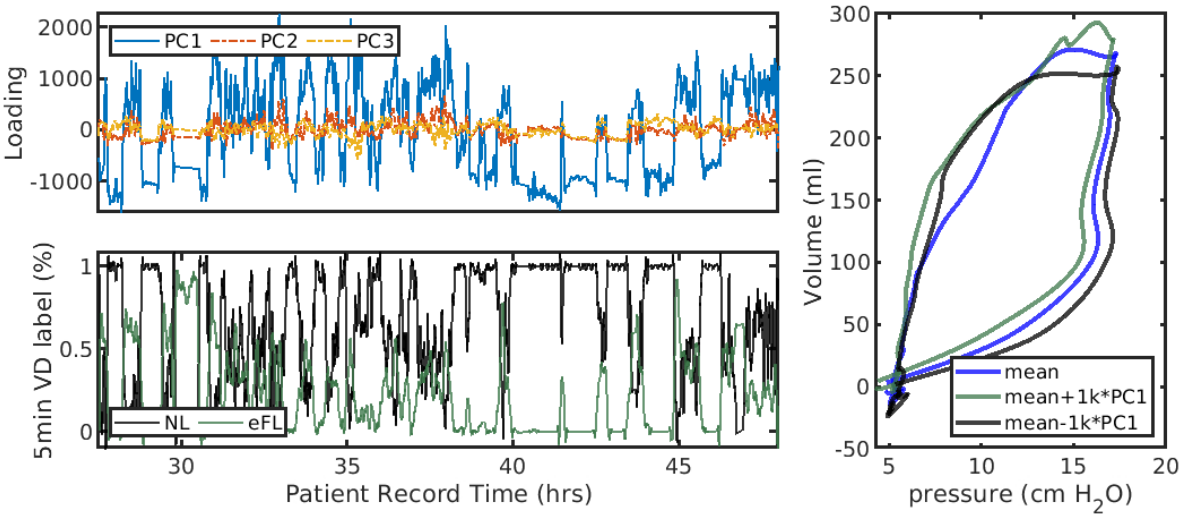

Figure C.8: The sign of PC1 loading roughly divides the PVD classes of patient #8, label #2. A threshold for the PC1 loading at zero roughly separates NL and eFL labels by 34%/65% and 85%/14%, respectively, with NL labels strongly associated with negative loadings. The optimal threshold (∼0.05) offers only subtle improvement. The right panel illustrates low fidelity changes in the cluster median pV loop (blue) when modified by these negative (black, more associated with NL) and positive (green, eFL) loadings. Note that this involves comparing 10-second properties (representing typically ∼3–4 breaths) to breath-wise labels, and some representation errors thus arise from summarizing binary PVD labels over all breaths intersecting a 10-second analysis window. Labeling and coloring in this figure do not relate to other figures.

waveform were assembled from the population of LVS windows with a given label. Ventilator mode was represented as a vector of percentages of each mode rather than a vector of binary categories. UMAP applied to these cohort feature vectors with the scaled-euclidean metric produced a relatively stable point configuration across various hyper-parameter choices; 12 point neighborhoods (2% of data) with a minimum distance of 1 unit were adopted as values. Identified groups were more sensitive to DBSCAN labeling hyper-parameters. Figure D.9a shows the possibilities of different groupings based on the search neighborhood size (ε). Subsequent results in the main section employ a hyper-parameter choice at the 'knee-point' [63] to balance generalizability and specificity. Panel b shows a more specific labeling ($\varepsilon = 2.5$) is qualitatively similar to that in the main text.

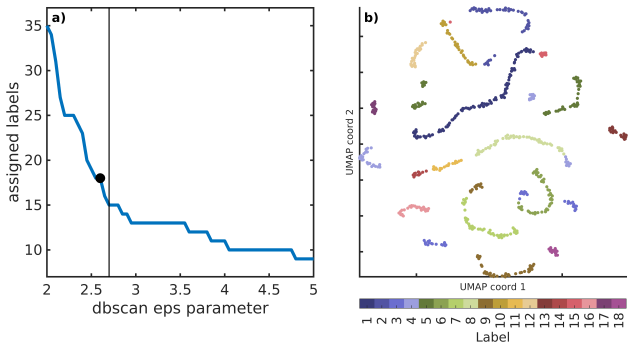

Figure D.9: DBSCAN search radius (ε) $v.$ the number of identified groups. The black line indicates $\varepsilon = 2.7$ selected for cohort clustering. Choices of $\varepsilon \in [2.67, 2.82]$ yield equivalent results following increased granularity of groupings at ε lower values.

Table D.4: The equivalent of Table2 for the alternate choice of hyper-parameter $\varepsilon = 2.5$

Label	Total%	N_{pat}	N_{pheno}	$p_{\min}$	p_{drive}	$V_{\max}$	dp/dV	MV mode
1	15.5	11	101	8	12.7[4.1]	7.9[1.3]	1.5[0.4]	APVCMV
2	11.4	16	52	12	13.3[3.6]	6.2[2.1]	2.3[1.2]	PCMV*
3	8.4	5	37	10	13.5[1.9]	6.5[0.3]	2.1[0.4]	APVCMV*
4	7.7	8	45	14	12.6[2.9]	6.2[1.3]	1.9[0.7]	APVCMV
5	6.9	11	58	16	13.4[2.3]	6.0[0.6]	2.2[0.6]	APVCMV
6	6.3	7	49	11	16.6[12.5]	5.9[0.7]	3.4[2.4]	APVCMV
7	6.2	6	49	8	11.1[1.4]	6.6[1.0]	1.7[0.2]	APVCMV
8	6.2	23	56	10	12.1[3.7]	6.3[1.0]	1.9[0.6]	APVCMV
9	6.2	10	51	14	21.3[9.5]	5.6[1.8]	3.7[3.1]	APVCMV
10	4.1	12	34	14	12.2[6.9]	5.9[0.2]	2.0[1.3]	APVCMV*
11	4.0	14	20	5	8.9[4.1]	7.0[2.4]	1.2[0.8]	APVCMV
12	3.7	16	25	12	14.3[3.2]	6.0[0.4]	2.3[0.9]	PCMV**
13	3.4	17	22	11	13.1[2.9]	6.2[1.3]	2.1[0.5]	APVCMV***
14	3.3	11	12	8	9.7[2.7]	6.8[1.5]	1.6[0.4]	APVCMV
15	2.5	8	10	12	15.1[12.5]	5.9[0.1]	2.4[2.2]	APVCMV
16	1.7	11	27	14	13.3[2.9]	6.0[1.3]	2.0[0.8]	APVCMV
17	1.5	9	14	16	15.9[6.0]	5.9[2.8]	2.6[2.4]	APVCMV
18	1.1	5	16	5	10.7[0.2]	6.5[0.7]	1.7[0.2]	APVCMV