scientific reports

Check for updates

OPEN

Empirical phenotyping of joint patient-care data supports hypothesis-driven investigation of mechanical ventilation consequences

J. N. Stroh[1,2]✉, Peter D. Sottile[3], Yanran Wang[4], Bradford J. Smith[2,5], Tellen D. Bennett[1,6,8], Marc Moss[3] & David J. Albers[1,7]

Analyzing patient data under current mechanical ventilation (MV) management processes is essential to understand MV consequences over time and to hypothesize improvements to care. 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. Ventilator waveform data originate from patient-ventilator interactions within the LVS while care processes manage both patients and ventilator settings. This study develops a computational pipeline to segment joint waveform and care settings timeseries data into phenotypes of the data generating process. The modular framework supports many methodological choices for representing waveform data and unsupervised clustering. The pipeline is generalizable although empirical output is data- and algorithm-dependent. Applied individually to 35 ARDS patients including 8 with COVID-19, a median of 8 phenotypes capture 97% of data using naive similarity assumptions on waveform and MV settings data. Individual's phenotypes organize around ventilator mode, PEEP, and tidal volume with additional delineation of waveform behaviors. However, dynamics are not solely driven by setting changes. Fewer than 10% of phenotype changes link to ventilator settings directly. Evaluation of phenotype heterogeneity reveals LVS dynamics that cannot be discretized into sub-phenotypes without additional data or alternate assumptions. Individual phenotypes may also be aggregated for use in scalable analysis, as behaviors in the 35 patient cohort comprise 16 cohort-scale LVS types. Further, output phenotypes compactly discretize the data for longitudinal analysis and may be optimized to resolve features of interest for specific applications.

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 effort. Both PVD and VILI may contribute to acute respiratory distress syndrome (ARDS) and ARDS-related mortality[7–9]. Protective strategies depend on mechanistic understanding to guide ventilator settings, including positive end-expiratory pressure (PEEP), tidal volume, and driving pressure[10–12]. Despite the effectiveness of protective advances, association

[1]Department of Biomedical Informatics, University of Colorado Anschutz, Aurora, CO 80045, USA. [2]Department of Biomedical Engineering, University of Colorado Denver | Anschutz Medical Campus, Aurora, CO 80045, USA. [3]Division of Pulmonary Sciences and Critical Care Medicine, University of Colorado Anschutz, Aurora, CO 80045, USA. [4]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO 80045, USA. [5]Pediatric Pulmonology and Sleep Medicine, University of Colorado Anschutz, Aurora, CO 80045, USA. [6]Pediatric Intensive Care Unit, Children's Hospital Colorado, Aurora, CO 80045, USA. [7]Department of Biomedical Informatics, Columbia University, New York, NY 10023, USA. [8]Department of Pediatrics (Critical Care Medicine), University of Colorado Anschutz, Aurora, CO 80045, USA. ✉email: jn.stroh@cuanschutz.edu

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

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[16–18] 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 and MV consequences over longer therapeutic timescales are limited and hindered by patient- and care-specific heterogeneity. A method for 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[19–22], identifying PVD waveform characteristics[23], and estimating event severity[24]. However, these labels may be ill-suited for MV research involving temporal analysis: they are stationary, are not mutually exclusive[21], depend on MV mode characteristics[8], and vary in organization[25,26]. Another research avenue uses interpretable model-based parametrizations to analyze waveform data[27–31], potentially allowing for a wider and more flexbile exploration of breath behavior.

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[29]. Moreover, MV management changes ventilator settings; these care factors contextualize the waveform data within LVS trajectories. The assemblages of coupled LVSs and applied management processes are the data generating process and the objects of interest for improving MV.

Quantifying clinical consequences of MV on patient health are necessary to evaluate and improve MV care. This requires linking outcomes to MV descriptors that include both ventilator setting as well as patient-ventilator interaction, but these LVS categories do not exist. Analysis based on ventilator settings and care processes alone will not incorporate patient-heterogeneous responses observed in pressure-volume, while those based on PVD labels that omit ventilator settings. This work addresses the methodological gap in quantifying MV consequences by developing an unsupervised categorization process to define joint LVS state categories as suitable targets for consequence association. Namely, it digitizes joint LVS data into interpretable phenotypes based on data similarities[32]. This approach reduces the dimensionality of the problem, enabling scalability to larger datasets, while incorporating the essential data components needed for consequence attribution.

Clinical validation of phenotypes requires linking breath behaviors to outcomes or other MV consequences. This work develops a generalized process to produce validatable phenotypes with clinical validation an intended downstream application ("Discussion").

In this work, phenotype trajectories of individuals are scrutinized in relation to MV management changes and timeseries of PVD labels to evaluate their consistency and ability to differentiate important characteristics. The phenotyping examples assign equal weight to LVS feature components to be agnostic about data element importance. The resulting data segmentation follow ventilator settings changes and persistent variations in waveform behavior within individuals individuals timeseries, while aggregate cohort-scale analysis shows they are general enough to mix patients. The main result is a generalized phenotyping pipeline whose empirical results are data-specific phenotypes. These outputs are not anticipated not generalize beyond the 2-day snippets of 35 ARDS patients on one ventilator model, because the data do not represent the broad diversity of MV breaths. The data-specific classification approach is tied to context of the data, providing benefits of informativeness and accuracy generally lost in a universal scheme[33].

A robust and systematic process for phenotyping the diverse LVS behaviors is a necessary step toward quantifying MV consequences and optimization of respiratory management to mitigate VILI. Phenotypes output by the developed pipeline may be used as a basis for explaining impacts on respiratory health. For example, one could investigate the distribution of phenotype occurrences, combining both ventilator settings and patient-ventilator response to them, with temporal changes in driving pressure[34] or gas ratios[35,36]. Importantly, phenotypes in this context data mask low-level heterogeneity to reduce trajectory complexity ("Patient-level phenotyping") and provide a standard basis for comparison across patients ("Cohort-scale phenotyping").

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 enables adjustments to the source data, waveform representation, feature definition, and segmentation strategy. This permits phenotype generation to accommodate different hypotheses about which aspects of the patient-ventilator-care system matter when evaluating MV consequence or other targets.

Data

Data including airway pressure, volume, and ventilator settings were captured for a cohort of intubated at University of Colorado ICUs with ARDS diagnoses, were mechanical ventilated using Hamilton G5 ventilators (https://www.hamilton-medical.com), and who had substantial risk of VILI. The collection effort and data use were approved by Colorado Multiple Institutional Review Board protocol (COMIRB, protocol #18-1433) and follow ethical standards set by COMIRB and the Helsinki Declaration of 1975. Children, pregnant women, and age-censored elders (> 89 years), and the imprisoned were excluded. Enrollment targeted collection of esophageal pressures, which are not analyzed in this work, and imposed additional exclusion criteria (viz. esophageal fistula, variceal bleeding or banding, facial fracture, and recent gastric/esophageal surgery). Eligibility for recording was contingent on active MV therapy, so patients were necessarily unconscious at the time of enrollment. Each

patient's identified proxy decision-maker provided informed consent as unconscious patients could not consent directly.

Following esophageal balloon placement, continuous recordings up to 48 hours were made directly from ventilators for 35 MV encounters satisfying enrollment criteria[22]. The cohort includes 14 women and 21 men with median age 58 years and interquartile range (IQR) 24.9 years; 71.4% are white, 34% of which identify as Hispanic or Latino. Pre-processing comprised 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

Breath-wise PVD identified by supervised ML in previous work[19,22] are used to enrich LVS evolution context and provide comparison for extracted categories. PVD types were assigned breath-wise 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 as defined and applied in previous analysis[21,37]. One-minute moving averages of these breath labels communicate PVD occurrence over time.

Pipeline

The computational pipeline described is a process for developing phenotypes from joint waveform and MV care-related data. The method is depicted in Fig. 1 and follows Wang et al.[32] 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. Code developed in MATLAB® for the implemented pipeline is available by reasonable request to the authors.

(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[38], clinical parameters[19], model parametrization[31], or signal processing methods such as polynomials and wavelets.

Detail	Count	%	Median	IQR
Monitored (h)			47.1	38.2
Recorded (h)			43.7	39
Age (years)			58.0	24.8
Gender				
Female	14	40	54.5	25.0
Male	21	60	58.0	26.5
Race/ethicity				
White	25	71.4		
Unknown/NA	5	14.3		
Black/AA	3	8.6		
AI or AK native	1	2.9		
More than one race	1	2.9		
ARDS risk				
Pneumonia	11	31.4		
COVID	11	31.4		
Sepsis	6	17.1		
Other	3	8.6		
Pancreatitis	2	5.7		
Aspiration	2	5.7		
P:F ratio			136.0	77.0
Mortality	8	22.9		
NMB use	9	25.7		

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.

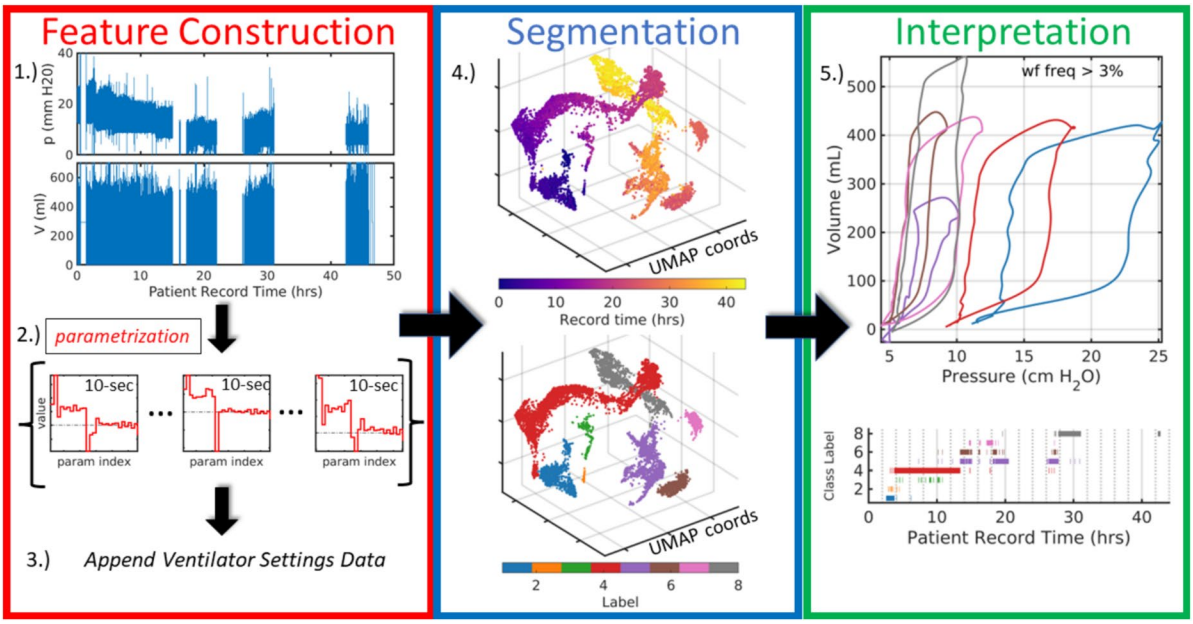

Fig. 1. Broad pipeline organization. Raw data (1) are digitally parametrized (2) over short windows, typically satisfying stationarity assumptions. Distributional parameter estimates are summarized and augmented with the contextual data of ventilator settings (3) which include information such as ventilator 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 (5).

One empirical parametrization developed for LVS data analysis[29] uses an asynchronous ensemble Kalman smoother (wEnKS)[39] 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

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

Here, g is a fixed smoothing parameter, θ is constant breath rate over the interval so that $t \,(\mathrm{mod}\,\theta)$ gauges time within each breath cycle, and y_0 is the baseline value of the signal (PEEP when y is pressure). The function φ in Eq.(1) is a piecewise constant function over the breath cycle with heights determined by parameter vector $\mathbf{a}$, whose optimal values digitize the observational waveform data (details found in SI A.1 and past work[29]).

The model construction ensures parameter identification throughout a bayesian ensemble-based inversion. Consequently, data-informed parameters are unique[40] and quantify uncertainties[41] 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 a reference value and parameters triplet $(y_0, \theta, \{\mathbf{a}\})$.

(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-s 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-s windows reduce the temporal sampling rate from 31.25 Hz 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 maximum volume 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. When available, other care-originating factors such as patient sedation level, paralytic use, and patient posture are easily included and may be used to target analysis of specific care regimes. The current implementation considers only MV settings but future effort might incorporate other care-originating factors like patient sedation level, paralytic use, and patient posture.

(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[42] include analyses of factors[43,44] as well manifold methods such as Uniform Manifold Approximation and Projection (UMAP[45,46]) and *t*-distributed Stochastic Neighborhood Embedding (tSNE[47]). Group labels are assigned by a clustering process[48,49] applied to LVS descriptors in the reduced coordinate system that describe similarity in a non-dimensional way. Options include space partition methods (*k*-means, *k*-medoids, etc.), kernel-based methods like Support Vector Clustering[50], and density methods like Density-based Spatial Clustering of Applications with Noise (DBSCAN[51,52]).

This study employs UMAP for reduction because it preserves local and global similarity structure of its input and has a numerically efficient MATLAB® implementation[53]. In this instance, the projection assesses similarity using the uniform Gower distance[54,55] because feature vectors contain mixed-type variables. Non-uniform feature weights may be added to modulate the influence of specific data elements 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 the dimensionless UMAP coordinates are often irregular and non-convex (Fig. 1, 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). Every time point in an MV record carries a phenotype label, which applies to all LVS observables of the patient at that time regardless of whether they were included among features. For example, downstream analysis of e.g., FiO_2 or SpO_2 could use phenotype identity to stratify data.

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.

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 ("Pipeline", 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[43,56]. This tool reveals the degree of LVS variance occurring under during stationarity to investigate non-ventilator temporal changes not identified by segmentation.

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-s intervals of continuous multi-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.

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 ("Patient-level phenotyping") and the assembly of cohort-scale phenotypes ("Cohort-scale phenotyping"). 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.

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] h), over-segmentation is needed to address the more diverse behavior of longer patient records compared with shorter ones. Over-segmentation addresses the need for record length variation (median[IQR] 47[37] h) 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 B). Changes in settings are typically $(\mathrm{mean}(s2l) > 60\%)$ reflected in label changes, with ∼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 Table 1, respectively). Their cases are typical of experiments in record length (∼24 h),

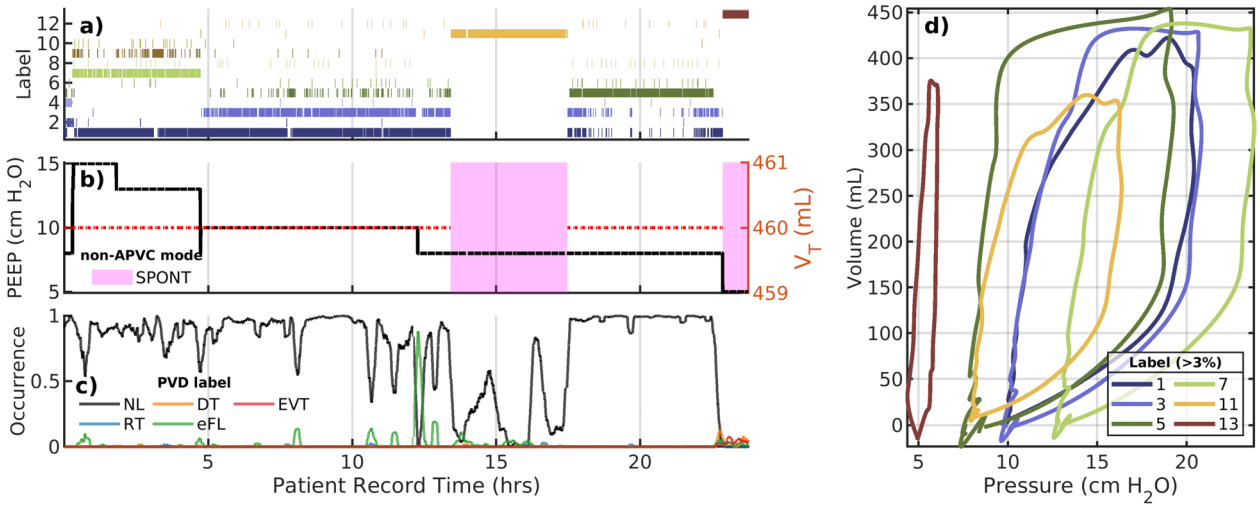

Fig. 2. LVS evolution of patient 34. (**a–c**) The trajectory of phenotype labels, ventilator settings, and externally identified PVD, respectively with a common horizontal axis of patient record hours. In (**b**), non-APVCMV ventilator modes are indicated by shaded regions. In (**c**), PVD identification over time is depicted by label occurrence percentage within 1-minute moving windows. The (**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.

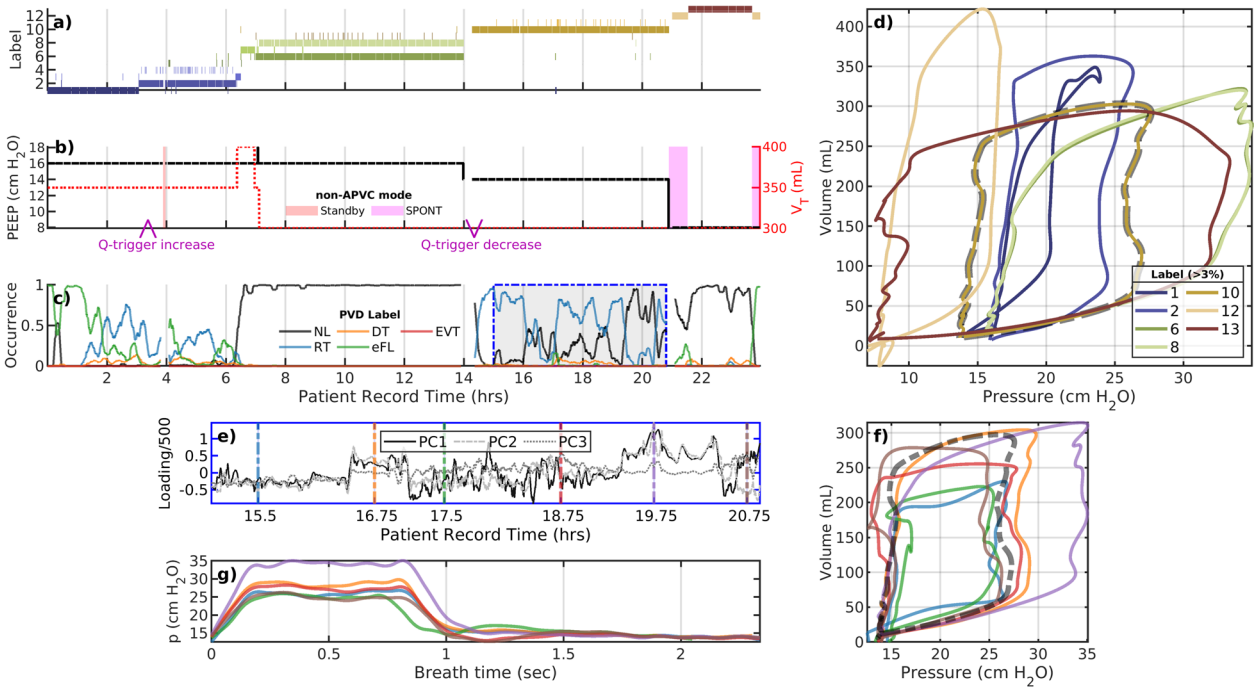

Fig. 3. A representative example: patient 11. The plot panels (**a–d**) are the same as the previous figure. A flow trigger ((**b**), at purple arrows) near 3 and 13 hours are the only MV settings changes besides mode, PEEP, or tidal volume. The lower panels (**e–g**) examine the variability during the record interval 15–21 h 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 (**a**) and (**d**), and among (**e–g**). Labeling and coloring in this figure do not relate to other figures.

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. Most breaths are identified as non-dyssynchronous in externally labeled PVD (c) apart from flow-limited behavior around 12.5 hours following reduction of PEEP from 10 to 7 cm H_2O. 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.

Example 2: Figure 3 illustrates an analysis of patient 11 whose LVS undergoes multiple changes over a 24-h data period. A flow trigger increase near 3 h (**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 h 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 ('Q-trigger') is returned to its initial value, near 14.25 h. Breaths during brief changes to spontaneous breathing around 20 and 23 h have markedly different *pV* characterizations (discontinuous label #12, tan). The interim period (20.5–22.5 h) 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 h. 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. Figure SI 3 demonstrates a case where intra-label variability may be discretely resolved.

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 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. Figure 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.

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 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 Fig. 2a, during hours 5–12, 17–18, 22–23) and Figure 3a (hours 3–14). Groupings can also mask important breath heterogeneity if phenotypes are too coarse (e.g., Figure SI 1e–g and Figure SI 3. Resolution of behavior categories 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., Fig. 1, step 4, label #4 in red). Weighting feature components in the similarity metric can improve sub-type resolution of such behaviors.

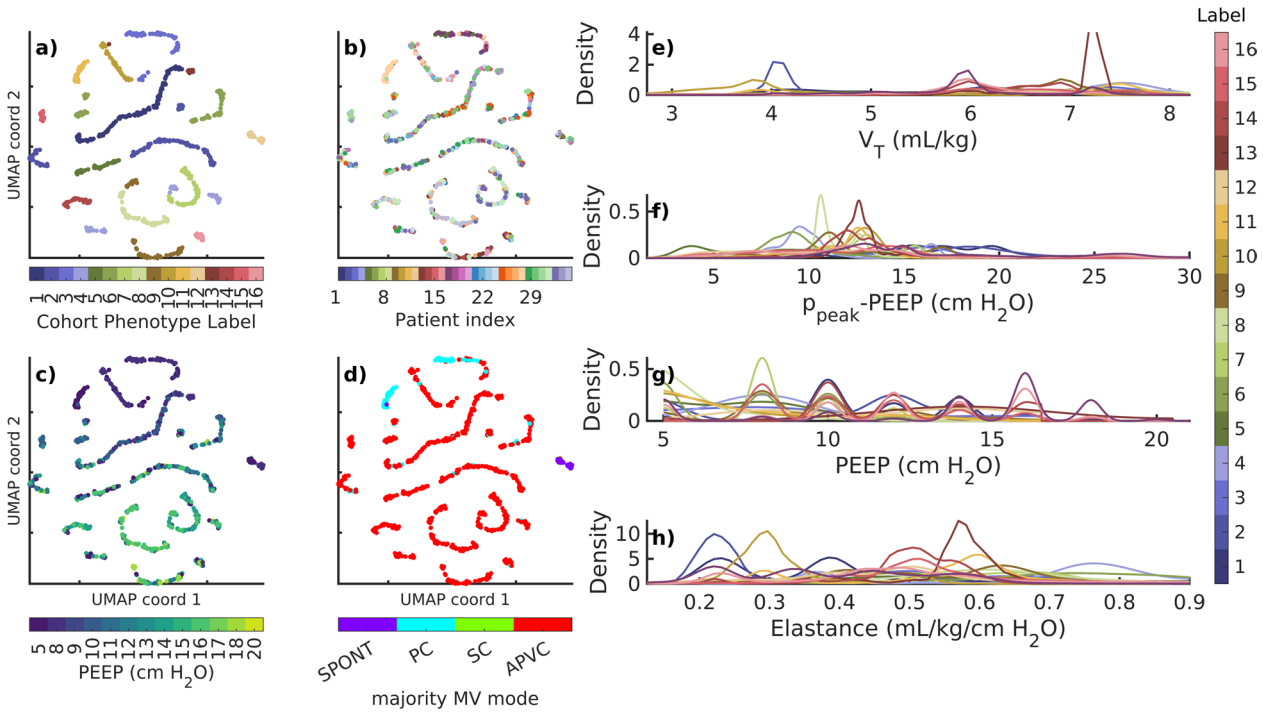

Fig. 4. Membership and data properties associated with cohort phenotypes. Points in (**a–d**) correspond to 721 individual phenotypes shown in unitless 3D UMAP coordinates that describe similarity (only two axes depicted, for simplicity). Labels (**a**) mix patients (**b**) while defining empirical partitions 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 (**a**). Labels and colors of (**a**) define the those of (**e–h**) and Fig. 5. Modes: spontaneous (SPONT), pressure controlled (PC), synchronized controlled (SC), and adaptive pressure volume controlled (APVC).

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

Table 2. Cohort label properties. Columns identify: cohort-level label, percentage of 10-s 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 ($\Delta p := p_{\mathrm{peak}} - p_{\mathrm{base}}$) 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.

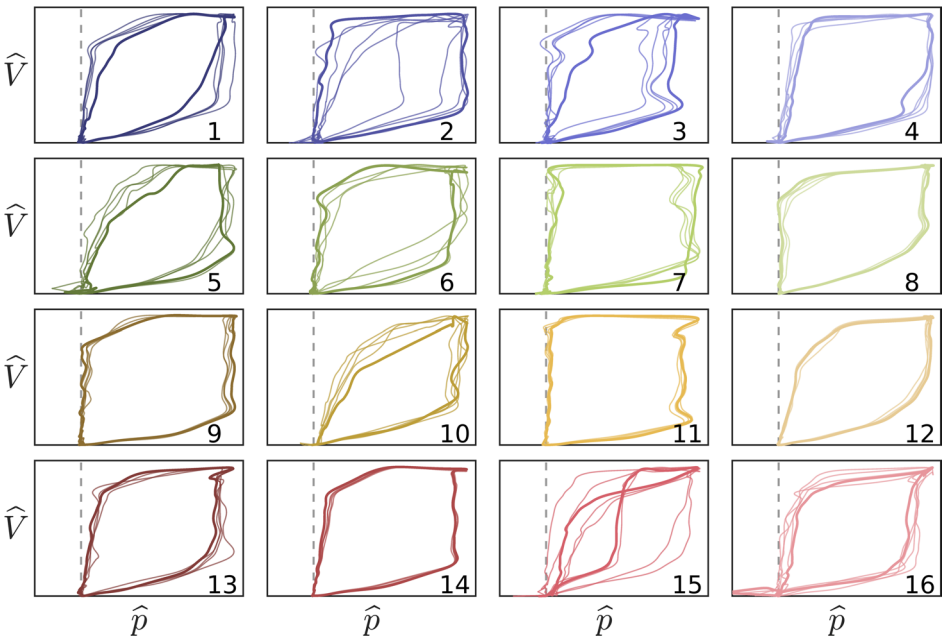

Fig. 5. Non-dimensional waveform characterizations. Pressure-volume traces correspond to median (bold) and nearby (thin) window characterizations of each cohort phenotype. Labels and colors correspond to Fig. 4a. Vertical and horizontal scales axes correspond to $\widehat{V} := V/V_T$ and $\widehat{p} := (p(t) - p_{\mathrm{base}})/(p_{\mathrm{peak}} - p_{\mathrm{base}})$, respectively, per Fig. 4e–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.

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.

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 and hypotheses about which LVS factors are important in a given application. This general methodology contrasts with the construction of a generalized output, 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. Research toward improving and personalizing MV benefits from phenotypes that include both ventilator settings and patient-ventilator interaction. Both LVS factors are necessary to accurately analyze MV consequences in light of patient and time heterogeneity. 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, performed 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. LVS categories reflected null hypotheses that prioritized no particular features of the included data in phenotype definition. 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 (Fig. 3), transient behavior (Fig. SI 1), and alternation between both similar and non-similar breath patterns (Fig. 3). Phenotype resolution could be adjusted to further delineate certain discrete variations, 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.

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[57], 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 behaviors observed during 1–2-day segments 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 (Table SI 1) 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. "Patient-level phenotyping" and SI 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 ∼1.5 M 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.

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 pipeline outputs 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. What the phenotyping process provides, however, are categories to which clinical consequences may be attributed in further study.

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. However, these data were used to manually identify PVD in breaths used in the supervised PVD labeling[22] featured in validation. 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 narrow 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 incorporate background hypotheses to target feature weights that emphasize key LVS data features of interest. Further improvements can easily involve outer-loop targeting of application-specific objectives beyond the generalized scope of this work.

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. These ongoing efforts are inextricably linked with clinical validation, which requires that quantified clinical consequences of MV be tied to phenotypes through hypotheses involving both phenotype definitions (viz., the data and pipeline choices defining them) and the nature of behavior-to-consequence association. Converting LVS dynamics to sequential progression through finite states allows the application of symbolic dynamics[58–60], game theory[61], discrete-time Markov chains, and large language models. Such tools can extend this work's investigation to patient trajectory patterns and their consequences to improve understanding of dynamical effects of current MV protocols on patient-ventilator systems.

Data availability

The clinical datasets used and/or analyzed during the current study are not publicly available due to ongoing collection, lack of patient consent for broad dissemination of their data, and data size (25 GB). Data are available by reasonable request to the author (J.N. Stroh, jn.stroh@cuanschutz.edu) and will require a data use agreement with the data owners.

Received: 4 February 2025; Accepted: 14 October 2025
Published online: 18 November 2025

References

1. National Heart, Lung, and Blood Institute ARDS Clinical Trials Network. Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. *N. Engl. J. Med.* **351**, 327–336 (2004).
2. Gattinoni, L., Citerio, G. & Slutsky, A. S. Back to the future: ARDS guidelines, evidence, and opinions. *Intensive Care Med.* **49**, 1226–1228 (2023).
3. Fan, E., Villar, J. & Slutsky, A. S. Novel approaches to minimize ventilator-induced lung injury. *BMC Med.* **11**, 1–9 (2013).
4. Karbing, D. S. et al. An open-loop, physiologic model-based decision support system can provide appropriate ventilator settings. *Crit. Care Med.* **46**, e642–e648 (2018).
5. Curley, G. F., Laffey, J. G., Zhang, H. & Slutsky, A. S. Biotrauma and ventilator-induced lung injury: clinical implications. *Chest* **150**, 1109–1117 (2016).
6. Slutsky, A. S. & Ranieri, V. M. Ventilator-induced lung injury. *N. Engl. J. Med.* **369**, 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. *N. Engl. J. Med.* **342**, 1301–1308 (2000).
8. Enrico, B., Cristian, F., Stefano, B. & Luigi, P. Patient-ventilator asynchronies: Types, outcomes and nursing detection skills. *Acta Bio Med. Atenei Parmensis* **89**, 6 (2018).
9. Blanch, L. et al. Asynchronies during mechanical ventilation are associated with mortality. *Intensive Care Med.* **41**, 633–641 (2015).
10. Brower, R. G. & Rubenfeld, G. D. Lung-protective ventilation strategies in acute lung injury. *Crit. Care Med.* **31**, S312–S316 (2003).

11. Petrucci, N. & Iacovelli, W. Lung protective ventilation strategy for the acute respiratory distress syndrome. *Cochrane Database Syst. Rev.* (2007).
12. Sutherasan, Y., Vargas, M. & Pelosi, P. Protective mechanical ventilation in the non-injured lung: Review and meta-analysis. *Annu. Update Intensive Care Emerg. Med.* **2014**, 173–192 (2014).
13. Moss, M. & Mannino, D. M. Race and gender differences in acute respiratory distress syndrome deaths in the United States: An analysis of multiple-cause mortality data (1979–1996). *Crit. Care Med.* **30**, 1679–1685 (2002).
14. Dreyfuss, D. & Hubmayr, R. What the concept of VILI has taught us about ARDS management. *Intensive Care Med.* **42**, 811–813 (2016).
15. Cochi, S. E., Kempker, J. A., Annangi, S., Kramer, M. R. & Martin, G. S. Mortality trends of acute respiratory distress syndrome in the United States from 1999 to 2013. *Ann. Am. Thoracic Soc.* **13**, 1742–1751 (2016).
16. Seneff, M. G., Zimmerman, J. E., Knaus, W. A., Wagner, D. P. & Draper, E. A. Predicting the duration of mechanical ventilation: The importance of disease and patient characteristics. *Chest* **110**, 469–479 (1996).
17. Zilberberg, M. D., Nathanson, B. H., Ways, J. & Shorr, A. F. Characteristics, hospital course, and outcomes of patients requiring prolonged acute versus short-term mechanical ventilation in the United States, 2014–2018. *Crit. Care Med.* **48**, 1587–1594 (2020).
18. Rose, L. & Messer, B. Prolonged mechanical ventilation, weaning, and the role of tracheostomy. *Crit. Care Clin.* **40**, 409–427 (2024).
19. Sottile, P. D., Albers, D., Higgins, C., Mckeehan, J. & Moss, M. M. The association between ventilator dyssynchrony, delivered tidal volume, and sedation using a novel automated ventilator dyssynchrony detection algorithm. *Crit. Care Med.* **46**, e151 (2018).
20. Gholami, B. et al. Replicating human expertise of mechanical ventilation waveform analysis in detecting patient-ventilator cycling asynchrony using machine learning. *Comput. Biol. Med.* **97**, 137–144 (2018).
21. Sottile, P. D. et al. Ventilator dyssynchrony-detection, pathophysiology, and clinical relevance: A narrative review. *Ann. Thorac. Med.* **15**, 190 (2020).
22. Sottile, P. D., Smith, B., Stroh, J. N., Albers, D. J. & Moss, M. Flow-limited and reverse-triggered ventilator dyssynchrony are associated with increased tidal and dynamic transpulmonary pressure. *Crit. Care Med.* **52**, 743–751 (2024).
23. Bakkes, T. et al. Automated detection and classification of patient-ventilator asynchrony by means of machine learning and simulated data. *Comput. Methods Prog. Biomed.* **230**, 107333 (2023).
24. Loo, N., Chiew, Y. S., Tan, C. P., Mat-Nor, M. & Ralib, A. M. A machine learning approach to assess magnitude of asynchrony breathing. *Biomed. Signal Process. Control* **66**, 102505 (2021).
25. Mellott, K. G. et al. Patient ventilator asynchrony in critically ill adults: Frequency and types. *Heart Lung* **43**, 231–243 (2014).
26. De Haro, C. et al. Patient-ventilator asynchronies during mechanical ventilation: Current knowledge and research priorities. *Intensive Care Med. Exp.* **7**, 1–14 (2019).
27. Agrawal, D.K., Smith, B.J., Sottile, P.D. & Albers, D.J. A damaged-informed lung ventilator model for ventilator waveforms. *Front. Physiol.* **12** (2021).
28. Zhou, C. et al. Reconstructing asynchrony for mechanical ventilation using a hysteresis loop virtual patient model. *BioMed. Eng. Online* **21**, 1–20 (2022).
29. Stroh, J. N., Smith, B.J., Sottile, P.D., Hripcsak, G. & Albers, D.J. Hypothesis-driven modeling of the human lung-ventilator system: A characterization tool for acute respiratory distress syndrome research. *J. Biomed. Inform.* 104275 (2022).
30. Chen, Y., Zhang, K., Zhou, C., Chase, J. G. & Hu, Z. Automated evaluation of typical patient-ventilator asynchronies based on lung hysteretic responses. *BioMed. Eng. Online* **22**, 102 (2023).
31. Agrawal, D.K., Smith, B.J., Sottile, P.D., Hripcsak, G. & Albers, D.J. Quantifiable identification of flow-limited ventilator dyssynchrony with the deformed lung ventilator model. *Comput. Biol. Med.* 108349 (2024).
32. Wang, Y. et al. A methodology of phenotyping ICU patients from EHR data: High-fidelity, personalized, and interpretable phenotypes estimation. *J. Biomed. Inform.* **148**, 104547 (2023).
33. Yona, G., Aharoni, R. & Geva, M. Narrowing the knowledge evaluation gap: Open-domain question answering with multi-granularity answers. arXiv preprint arXiv:2401.04695 (2024).
34. Guérin, C. et al. Effect of driving pressure on mortality in ards patients during lung protective mechanical ventilation in two randomized controlled trials. *Crit. Care* **20**, 1–9 (2016).
35. Rice, T. W. et al. Comparison of the SpO2/FiO2 ratio and the PaO2/FiO2 ratio in patients with acute lung injury or ards. *Chest* **132**, 410–417 (2007).
36. Chen, W. et al. Clinical characteristics and outcomes are similar in ards diagnosed by oxygen saturation/fio2 ratio compared with PaO2/FiO2 ratio. *Chest* **148**, 1477–1483 (2015).
37. Sottile, P.D., Smith, B., Moss, M. & Albers, D.J. The development, optimization, and validation of four different machine learning algorithms to identify ventilator dyssynchrony. *medRxiv* https://doi.org/10.1101/2023.11.28.23299134 (Preprint, 2023). https://www.medrxiv.org/content/early/2023/11/29/2023.11.28.23299134.full.pdf.
38. Park, C. & Lee, D. Classification of respiratory states using spectrogram with convolutional neural network. *Appl. Sci.* **12**, 1895 (2022).
39. Sakov, P., Evensen, G. & Bertino, L. Asynchronous data assimilation with the EnKF. *Tellus Ser. A Dyn. Meteorol. Oceanogr.* **62**, 24–29. https://doi.org/10.1111/j.1600-0870.2009.00417.x (2010).
40. Latz, J. Bayesian inverse problems are usually well-posed. *SIAM Rev.* **65**, 831–865 (2023).
41. Stuart, A. M. Inverse problems: A Bayesian perspective. *Acta Numer.* **19**, 451–559 (2010).
42. Anowar, F., Sadaoui, S. & Selim, B. Conceptual and empirical comparison of dimensionality reduction algorithms. *Comput. Sci. Rev.* **40**, 100378 (2021).
43. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417 (1933).
44. Lawley, D. N. & Maxwell, A. E. Factor analysis as a statistical method. *J. R. Stat. Soc. Ser. D (Stat.)* **12**, 209–229 (1962).
45. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018).
46. Healy, J. & McInnes, L. Uniform manifold approximation and projection. *Nat. Rev. Methods Primers* **4**, 82 (2024).
47. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** (2008).
48. Omran, M. G., Engelbrecht, A. P. & Salman, A. An overview of clustering methods. *Intell. Data Anal.* **11**, 583–605 (2007).
49. Hastie, T., Tibshirani, R. & Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2 (Springer, 2009).
50. Ben-Hur, A., Horn, D., Siegelmann, H. T. & Vapnik, V. Support vector clustering. *J. Mach. Learn. Res.* **2**, 125–137 (2001).
51. Ester, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD* **96**, 226–231 (1996).
52. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst. (TODS)* **42**, 1–21 (2017).
53. Meehan, C., Meehan, S. & Moore, W. Uniform manifold approximation and projection (UMAP v4.2). In *MATLAB Central File Exchange* (2022).
54. Gower, J. C. A general coefficient of similarity and some of its properties. *Biometrics* 857–871 (1971).
55. Tuerhong, G. & Kim, S. B. Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert Syst. Appl.* **41**, 1701–1707 (2014).
56. Rao, C.R. The use and interpretation of principal component analysis in applied research. *Sankhyā Indian J. Stat. Ser. A* 329–358 (1964).

57. Goldsack, J. C. et al. Verification, analytical validation, and clinical validation (v3): the foundation of determining fit-for-purpose for biometric monitoring technologies (BIOMETS). *npj Digit. Med.* **3**, 55 (2020).
58. Amigó, J. M., Keller, K. & Unakafova, V. A. Ordinal symbolic analysis and its application to biomedical recordings. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **373**, 20140091 (2015).
59. Lind, D. & Marcus, B. *An Introduction to Symbolic Dynamics and Coding*. 2 Ed. (2021).
60. Hirata, Y. & Amigó, J. M. A review of symbolic dynamics and symbolic reconstruction of dynamical systems. *Chaos Interdiscip. J. Nonlinear Sci.* **33** (2023).
61. Bauso, D. *Game Theory with Engineering Applications* (SIAM, 2016).

Acknowledgements

This work is supported by National Heart Lung and Blood Institute awards 5R01HL151630 “Predicting and Preventing Ventilator-Induced Lung Injury” (DJA, BJS), K23HL145011 “The Detection, Quantification, and Management of Ventilator Dyssynchrony” (PDS), and K24HL168225 “Mentoring and Patient-Oriented Research in Clinical Informatics and Data Science” (TDB). Additional support provided by National Library of Medicine award R01LM006910 "Discovering and Applying Knowledge in Clinical Databases" (DJA). Thanks as always to Meg Rebull for local administrative support.

Author contributions

DJA, JNS, YW - methodology and conceptualization; PDS, JNS - formal analysis; JNS - development, writing, visualization; DJA, MM, BJS, PDS - project management; MM, PDS - data acquisition and preparation; DJA, TDB, BJS - resources and funding. All authors reviewed the manuscript.

Funding

The authors acknowledge the indirect funding they received from the following grant: NLM R01 LM006910 "Discovering and Applying Knowledge in Clinical Databases"

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-24489-4.

Correspondence and requests for materials should be addressed to J.N.S.