Discovery of signatures of fatal neonatal illness in vital signs using highly comparative time-series analysis

Justin Niestroy [1,2]
Maxwell A Levinson [1,2]
Sadnan Al Manir [1,2]
Timothy W Clark [1,2,6]
J Randall Moorman [2,3 *]
Karen D Fairchild [2,4]
Douglas E Lake [2,3,5]

1. Department of Public Health Sciences, University of Virginia
2. Center for Advanced Medical Analytics, University of Virginia
3. Department of Medicine, University of Virginia
4. Department of Pediatrics,University of Virginia
5. Department of Statistics, University of Virginia
6. School of Data Science, University of Virginia

* Correspondence: J Randall Moorman, rm3h@virginia.edu

Conflict of interest: JRM and DEL own stock in Medical Predictive Science Corporation, Charlottesville, VA; JRM owns stock and is an officer of Advanced Medical Predictive Devices, Diagnostics and Displays, Charlottesville, VA.

Abstract

Objective
Signatures of illness in vital signs of Neonatal Intensive Care Unit (NICU) patients can inform on future adverse events and outcomes. We implemented highly comparative time-series analysis to discover features and predictive analytics tools for all-cause mortality in the next 7 days, using the ubiquitous HR and SpO_2 vital sign data from bedside monitors.

Design
We populated a Time Series Commons with the complete HR and SpO_2 data from all infants in the University of Virginia NICU from 2009 to 2019. We calculated the results of applying over 80 members of 11 mathematical families to random ten-minute segments of 0.5Hz data each day for each infant, with varying parameter sets, resulting in 4998 algorithmic operations on each infant. We used an unsupervised mutual information-based method to cluster the results, and we selected a single representative operation from each cluster. We used our FAIRSCAPE framework to compute a detailed provenance of all computations, and we constructed a complete software library with links to the analyzed data for reproducibility and reuse. We made multivariable logistic regression models using the lasso to assay the usefulness of the algorithms.

Setting
Neonatal ICU

Patients
5957 NICU infants, of whom 206 died.

Measurements and main results
3555 algorithmic operations returned usable results. Twenty representative operations, selected from each of 20 unsupervised clusters, held more than 81% of the information predicting death. A multivariable model had an AUC of 0.81 for predicting death in the next 7 days. In addition, five algorithms outperformed others: moving threshold, successive increases, surprise, and a random walk model.

Conclusions
Highly comparative time-series analysis revealed new vital sign metrics to identify NICU patients at the highest risk of death in the next week. This approach can facilitate the discovery of signatures of impending, potentially actionable, clinical decompensation in monitored patients.

Introduction

Continuously monitored vital signs of patients in intensive care units hold untapped information about risk for adverse events and outcomes [1]. For example, the display of a score based on analysis of abnormal heart rate characteristics was shown by our group to reduce sepsis-associated mortality by 40% in preterm infants in the Neonatal Intensive Care Unit (NICU) [2–5]. That approach was tailored to detect specific phenomena that we observed in the heart rate data, reduced variability and transient decelerations, in the days prior to sepsis diagnosis[2–5], and we used algorithms optimized for the task, including sample asymmetry [6] and sample entropy [7–9].

Here, we asked a more general question - what if we did not know all the characteristics we wish the algorithms to detect? That is, if we used a very large number of algorithms designed for general use in time-series, would we discover some that were more effective than our tailored design? This approach has been described by Fulcher and coworkers, who called it *highly comparative time-series analysis* [10–12]. The fundamental idea is to extract features from many time series, using many algorithms, each operating with many sets of parameter values. We then apply this ensemble to a data set to determine which algorithms perform best for predicting a specific outcome. Clustering of algorithms can, eventually, simplify this approach for clinical applications. [13,14]

As an example of our approach, the familiar sample entropy algorithm [7,8] requires two parameters, an embedding dimension *m* and a tolerance window *r*. A highly comparative time-series analysis entails many operations of the sample entropy algorithm that vary *m* and *r*. The result of each operation is treated as a potential predictor. Since the results are expected to be highly correlated, we can represent the family of sample entropy results as a cluster, and choose an operation of sample entropy with a single optimal combination of *m* and *r* for use in multivariable statistical models.

Furthermore, rather than simply clustering methods we know to be in the same family ("sample entropy" etc.), but with differing parameters, we can expand clustering to include many families of methods, and their parameters, defining the clusters using an outcome similarity metric. The cluster that contains sample entropy might then also contain related measures detected by clustering, all of which can be represented by a single outcome measure or feature. In all, this is an efficient way to screen many time-series algorithms, to discover features that are predictive of an outcome, without domain knowledge of prior known specific characteristics such as reduced heart rate variability, that might be related to that outcome.

To test these ideas, we selected death in the next 7 days for infants in the NICU as the outcome of interest. This is a topic of clinical interest and usefulness - identification of infants at high risk, especially where risk appears to be rising quickly, can alert clinicians to the possibility of imminent clinical deterioration from illnesses such as sepsis or respiratory failure. The heart rate characteristics score noted above, which is targeted toward a specific time-series phenotype, has modest performance in this area [15]. In this work, the question is whether an examination - and potentially a combination - of many time series feature extraction algorithms may improve on this targeted approach.

Materials and Methods:

Study design

We collected all bedside monitor data from all patients in the Neonatal ICU (NICU) at the UVA Hospital since 2009 using the BedMaster Ex™ system (Excel Medical, Jupiter FL). HR derived from the bedside monitor electrocardiogram signal is sampled at 0.5 Hz. SpO_2 is measured using Masimo SET® pulse oximetry technology (Masimo Corporation, Irvine CA) with a sampling rate of 0.5 Hz and averaging time of 8 seconds. For this analysis, we included all infants admitted from 2009 through 2019 who had HR and SpO_2 data available for analysis. Clinical data were abstracted from a NICU database (NeoData, Isoprime Corporation, Lisle IL). This work was approved by the UVA Institutional Review Board with waiver of consent.

Patient population

From January 2009 through December 2019, 6837 infants were admitted to the UVa NICU, with median gestational age (GA) 35 weeks. Of these, 5957 infants had HR and SpO_2 data available for analysis, and 206 died and had HR and SpO_2 data available within 7 days of death. Median GA of the infants included in the death analysis was 34 weeks and median age at death was 6 days (range 1-204 days). Causes of death included but were not limited to birth asphyxia, severe acute or chronic lung disease of prematurity, sepsis, pulmonary hypertension, congenital (including cardiac) anomalies, and intracranial hemorrhage.

Software and computing environment

We prepared a library of software in Python3 consisting of our implementations of 111 published algorithms [10] described for use in medical and non-medical domains. Table 1 shows the families of algorithms employed, together with a description and examples of each.

We ran these routines and some additional special-purpose MATLAB algorithms in Docker containers designed to run in a horizontally scalable secure cluster environment under the OpenStack cloud operating system, using the FAIRSCAPE data lake environment. [16] We issued persistent identifiers for all software, datasets and analysis results using Archival Resource Keys (ARKs) [17], associated with computational provenance metadata [18] for reproducibility and reusability.

Operations of algorithms

We speak of specific algorithms such as sample entropy with specific values of *m* and *r* as *operations*. We note that operations of a single algorithm might be repeated many times with varying parameters.

The raw vital signs data were stored as vectors of time stamps 2 seconds apart with the corresponding measurement of HR or oxygen saturation. We grouped the vital signs data into more than 18 million 10-minute non-overlapping windows, each with 300 measured values. In each group, we computed 111 time-series algorithms with varying parameters for a total of 2499 operations. The result was a matrix of results with more than 18 million rows and 2499 columns, as illustrated in the workflow diagram in **Figure 1**.

We randomly sampled the *Processed Vitals* dataset taking one 10-minute record per day per patient. This step resulted in 130,000 days of samples, each containing the result of 4998 operations from the HR and SpO_2 data. We removed single-valued results, those with imaginary numbers, and samples with missing values, and were left with 3555 of the 4998 viable candidate algorithmic operations. To adjust for the wide range in scales, we used an outlier-robust sigmoid transform [10,19] to convert operation ranges to the interval [0,1].

We clustered results to reduce dimensionality. We divided the 130,000 results of individual algorithms into ten equiprobable bins and calculated all possible distances using mutual information [20 21]. We organized these results into a distance matrix and determined clusters with k-medoids using the *pam* function of the R *cluster* package [22]. We represented each cluster by a single operation, as shown in **Figure 2**.

Statistical analysis and modeling
A binary outcome of death within the next 7 days was used to evaluate algorithm performance. We selected the twenty medoids of the clustered algorithms for use in multivariate LASSO logistic regression.

Results

AUC for death prediction for each of the 3555 operations

In total, there were 912 daily ten-minute samples within a week of death for 206 infants, for a sample incidence rate of 0.73%. **Figure 3** shows the number of algorithms as a function of their univariate predictive performance for death in the next week, measured as AUC. The top performing algorithm, a symbolic logic count of successive increases in heart rate that is discussed further below, had an AUC of 0.74, substantially higher than that of the traditional algorithms like standard deviation of HR (0.70) and mean SpO_2 (0.65).

Algorithmic results clustered, allowing data reduction
We sought correlations among the results. **Figure 4** shows two heat maps based on the absolute value of the correlation coefficients for 3555 algorithmic operations on the left and 20 identified by cluster medoids on the right.

These results justified an analysis of clusters of results, which we undertook by measuring mutual information among all the operational results. A representative result is shown in **Figure 5**. We sought a number of clusters that was large enough to explain most of the predictive performance of a multivariable statistical model for death but in keeping with the practice of having a reasonable number of predictors for 200 events. [23,24] We found that 20 clusters satisfied these conditions. The findings were robust in repeated experiments with different random sampling of one record per patient per day as well as daily averaged data.

We examined the clusters for interpretability, and found that clusters of algorithmic operations reported on identifiable and interpretable time-series characteristics. For example, one cluster held the results of operations that report on the mean, another cluster held operations that report on the minimum, and so on. As expected, very near neighbors represented the results of operations that are closely related. We concluded that each cluster might be represented by a

single operation and based our further calculations on those 20, reasoning that they comprehensively and efficiently captured the characteristics of neonatal heart rate and oxygen saturation time series.

Multivariable statistical models using the new measures of vital signs to predict death
To understand the relationship of these 20 operations to the outcome of death, we used multivariable logistic regression with the lasso. [25,26] Each patient was represented on each day by 20 results - 10 for each heart rate and each oxygen saturation time-series - calculated on a 10-minute record chosen at random once per day. The purpose of the analysis was to identify time-series analytics that are new to the study of infant heart rates and oxygen saturations that were more effective predictors of death than commonly used measures. Since we picked the operations to be representative of the clusters, we ensured interpretability of the results.

We made models with only HR features, only SpO_2 features, and both HR and SpO_2 features. The multivariate HR-only model had 9 features with non-zero coefficients and had an AUC of 0.77 compared to the best HR univariate model, successive increases, which had AUC of 0.79. The SpO_2 only model had 4 features with non-zero coefficients and had an AUC of 0.75. The combined HR SpO_2 model had 18 features with non-zero coefficients (9 HR, 6 SpO_2, 3 demographic) and had AUC of 0.81.

To better understand how these models performed leading up to the day of death, AUCs for each model were calculated daily from 7 days to 1 day prior to death, as shown in **Table 2**. As expected, model performance is highest one day prior to death. The model that combined HR and SpO_2 measures had the best performance, with AUC near 0.84 the day before death and about 0.76 a week prior to death.

New measures of vital signs associated with NICU death
We found new measures of heart rate and oxygen saturation signals associated with NICU deaths. In the heart rate time-series, the top performing measures were fewer occurrences of *successive increases* (illustrated in **Figure 6**) and larger *surprise.* In the oxygen saturation time-series, the most predictive measure was a small difference from a *random walk with momentum.* A *moving threshold* calculation [27] that showed fewer extreme events was informative in both the heart rate and the oxygen saturation time-series.

The most informative new predictor for death risk was a small number of successive increases in the heart rate, and **Figure 6** shows four records with increasing numbers of successive increases. The value that would be observed in a set of 300 random numbers, 75 ($300/0.5^2$), is approached in the lower right panel. Qualitatively, the finding is that low heart rate variability is associated with higher risk of death. However, a more direct measure of variability, the standard deviation of the heart rate,was less predictive (AUC 0.74 for successive increases compared with 0.70 for HR STD).

Discussion

Much progress has been made in the use of continuous time-series data from the bedside continuous cardiorespiratory monitors in the Neonatal ICU. [28] We tested the idea that we might improve the current art through a systematic study of our very large set of time series using an exhaustive number of analytical measures. We draw from a prismatic work describing the method of highly comparative time-series analysis, applying many time series algorithms to many time series examples of all kinds. [10] We applied the principles of highly comparative time series analysis to our domain, continuous cardiorespiratory monitoring in NICU patients. This work extends the study of highly comparative time-series analysis with its focus on clinical data sets, clinical events that are important to clinicians, and domain-specific knowledge of the physiologic origins of the data and how clinicians use it at the bedside.

In this example of highly comparative time-series analysis of a large clinical data set, we studied vital sign data from Neonatal ICU patients and discovered algorithms not previously reported in this domain that identified infants at higher risk of death. These algorithms report generally on the absence of heart rate variability and low oxygen saturation, features known to inform on poor health. Other major findings were that only 20 clusters of algorithms explained the great majority of the variance, in keeping with another study of highly comparative time-series analysis, [14] that downsampling of the data to single 10-minute records daily did not affect the overall results, and that the newly revealed algorithms outperformed standard measures of vital sign variability.

A small number of clusters explain the variance of the results

This is not entirely unexpected, because many of the operations entail the same algorithm repeated with different arguments and parameters. For example, the sample entropy algorithm requires the choice of an embedding dimension *m* and a tolerance window *r*. [7–9] In all, we performed 12 sample entropy operations with combinations of these variables. Thus, our clusters were to some extent explainable. For example, one held many operations that report on the center of the distributions, another held many reporting on the width of distributions, another held many entropy operations, and so on. These findings are important with regard to the interpretability of statistical models that use the results, avoiding the problem of balck boxes. [29]

Downsampling of the data to single 10-minute records daily did not affect the overall results for algorithmic operations clustering

A downside to the massive exploration of algorithmic operations is the computing time. To begin our investigation, we accordingly massively reduced the data set to a single 10-minute record daily, <1% of the total. To test the fidelity of the results, we repeated the procedure on 3 other single 10-minute records, and on the daily average. The results were not significantly different in the nature of the clusters or their constituents, suggesting that a manageable subsample of the data can be used for exploratory purposes in the highly comparative time-series analysis, and the results verified afterward.

Newly revealed algorithms outperformed canonical measures of vital sign variability

We found 4 algorithms that out-performed the others. They were not medoids of clusters, and we examined them separately. The top-performing algorithms have not previously been applied to vital sign time series analysis. Importantly, they were interpretable in the light of domain knowledge about neonatal clinical cardiovascular pathophysiology. These top-performing 4

algorithms lead to a model with an AUC similar to the model with all 20 features.

Successive increases in heart rate is the result of a symbolic dynamics analysis, [30] and represent small accelerations. Individual vital sign measurements are replaced by symbols that reflect whether they have increased, decreased, or stayed the same compared to the preceding measurement. Our finding was that the number of consecutive increases in every-2-second HR was reduced in the time series of infants at higher risk of death, as illustrated in **Figure 6**. This finding is consonant with reduced heart rate variability, a known marker of abnormal clinical status. It is interesting to speculate why the absence of successive increases should improve upon ordinary measures of variability for prediction of death.

Moving threshold [27] was an approach developed in the field of extreme events in dynamical systems. An example is the human reaction to floods in rivers - when there are no floods, barriers are allowed to lapse. A flood, though, leads to new, higher barriers that could contain the recent event. The moving threshold model examines a time series for points that exceed an initial threshold, increases the threshold after a crossing event, and allows the threshold to decay over time. The parameters that are measured include the rate of events, the intervals between threshold crossings, and statistical measures of the distribution of the thresholds. Our finding was that the vital sign time series of infants at higher risk of death were characterized by lower moving threshold for HR (reflecting low HR variability) and lower moving threshold for SpO_2 (illustrated in **Figure 7** as a gradual decline in SpO_2).

Random walk with momentum measures the fit of the time-series data to a biased random walk. [31] The phenotype of high risk of death for this algorithm is a monotonic downward decline in the SpO_2, which fits well with deteriorating pulmonary status clinically.

Surprise [32] calculates the distribution of points of a subsection of the time series. The surprise of the point following that subsection is then measured by how likely the new point was given the calculated distribution, given as $1/p$. The phenotype associated with mortality here, was a low value of surprise in the HR, consistent with reduced HR variability.

Relationship to prior work

In 2001, [33] we showed that heart rate characteristics of low variability and transient decelerations added information to clinical findings quantified by the SNAP (Score for Acute Neonatal Physiology) [34] and NTISS (Neonatal Therapeutic Intervention Scoring System) [35] in the early detection of sepsis. A heart rate characteristics index predicted sepsis and all-cause mortality in preterm NICU patients [2,15,33,36]. We found that the AUC for the heart rate characteristics index developed at the University of Virginia and tested at Wake Forest University was 0.73. Subsequently we broadened the analyses to include conventional measures of HR andSpO_2 in the first week after birth which we showed predict mortality among preterm NICU patients better than the validated and commonly accepted SNAPPE-II score (Score for Neonatal Acute Physiology-Perinatal Extension). [37–39] and combined HR and SpO_2. [40] In 2010, [41] Saria and coworkers showed that short- and long-term variability of HR and SpO_2 in the first 3 hours of life were useful in classifying premature infants at risk for high-morbidity courses. In the current work we showed that running an enormous number of operations on a single daily random 10 minute window of HR and SpO_2 data uncovered new metrics that predict mortality better than our prior models.

Limitations
To increase the reliability of finding such an analytic, we tested a very large number of algorithmic operations. Since this approach increases the probability of false positives, we tested them in a very large data set, nearly 6000 infants. Even so, there is the limitation of using many predictors per event. External validation will be important.

We found that only 3555 of the 4998 algorithms consistently returned non-null results, and we note that other data sets from other sources might fare differently. This was expected as many of the algorithms were from different domains and may not work on all signals.

We used only logistic regression to test the association of the algorithmic operations with clinical outcome, and other machine learning and deep learning methods might have had different results. We note, though, recent works that point to a similarity of results of logistic regression compared to other methods including recurrent neural networks. [42,43]

Conclusion:

Highly comparative time-series analysis of clinical data reduced thousands of algorithms to an interpretable set of 20 that represent the character of neonatal vital signs dynamics.

This framework will be useful for future work analyzing bedside monitor data for signatures associated with various imminent or future adverse events and outcomes. The terabytes of vital sign data passing over monitors just at a single NICU such as ours, together with electronic medical record data on clinical and laboratory variables, hold valuable insights into actionable outcomes. Developing platforms and systems for sharing data with other investigators so that algorithms can be tested in large and diverse populations is another worthy goal. Harnessing these data could lead to preemptive strategies that improve patient outcomes.

Data and Software Availability

Anonymized data that support the findings of this study, with the evidence graph for the clustering, are openly available in the University of Virginia's LibraData archive at https://doi.org/10.18130/V3/VJXODP. Python code used to process this data is archived in Zenodo at https://doi.org/10.5281/zenodo.4321332. This version and any future versions are also available in Github at https://github.com/fairscape/hctsa-py. Our code is licensed under terms of the MIT license (https://opensource.org/licenses/MIT), and is a reimplementation in Python of most of Ben Fulcher's original MATLAB code, available here: https://github.com/benfulcher/hctsa. Software for clustering analysis and cross-implementation testing, together with the test data, may be found here: https://doi:10.5281/zenodo.4627625 .

Table 1: Families of algorithms implemented in highly comparative time series analysis

Family	**Description**	**Example(s)**
Distribution	Moments and other descriptive statistics	Mean, median, standard deviation
Correlation	Similarity of data points as a function of the time between them	Linear and nonlinear autocorrelation
Stationarity	Statistical properties do not change over time	Standard deviation of moments measured on different window lengths
Symbolic transforms	Convert ranges to letters and analyze their sequence	Frequency of successive increases
Entropy	Order and regularity	Sample entropy
Trend analysis	Fitting lines through data	Slope and intercept
Heart Rate Variability	Canonical analyses	Power spectral density ratios
Time Series Modelling	Fits time series model to data	Surprise
Wavelet	Properties of the time series wavelet spectrum	Wavelet decomposition of time series
Nonlinear Analysis	Nonlinear analysis methods	False nearest neighbors, Information dimension
Other	Extreme values	Moving threshold model

Table 2: Model performances as a function of days until death. AUC for 6 models of a single random daily 10 minute segment of HR, SpO_2, or both are shown at various intervals before death. All models included 3 demographic variables: birth weight, gestational age, and days since birth.

Days until Death	**# of Features**	**ALL <=7 days**	**1 day**	**3 days**	**7 days**
Infants with data		*206*	*182*	*85*	*68*
HR SpO_2 Cluster Centers	23	0.81	0.82	0.81	0.78
HR SpO_2 Top Performers	8	0.82	0.84	0.81	0.77
HR	13	0.77	0.77	0.75	0.76
HR Successive Increases	4	0.79	0.81	0.76	0.73
HR Mean/SD SpO_2 Mean/SD	7	0.76	0.77	0.72	0.72
SpO_2	13	0.75	0.78	0.76	0.71

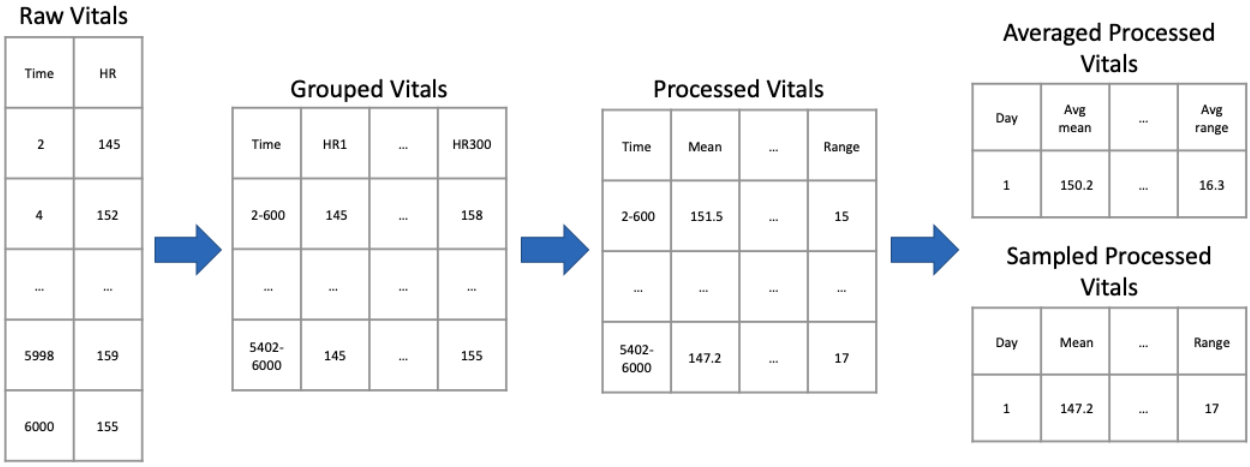

Figure 1. Data processing workflow for HR data. From left to right: Each row of the *Raw Vitals* table contains measured vital signs at 2 second intervals. These values are transposed in the *Group Vitals* table so that each row has a 600 second time range and up to 300 measured values. Algorithms operate on the *Group Vitals* table producing the *Processed Vitals* table of the same size - in the example, the first algorithm is the mean, and the last is the range. The *Averaged Processed Vitals* table holds the average of each result for a day; the *Sampled Processed Vitals* table holds the results for a randomly selected 600 second record.

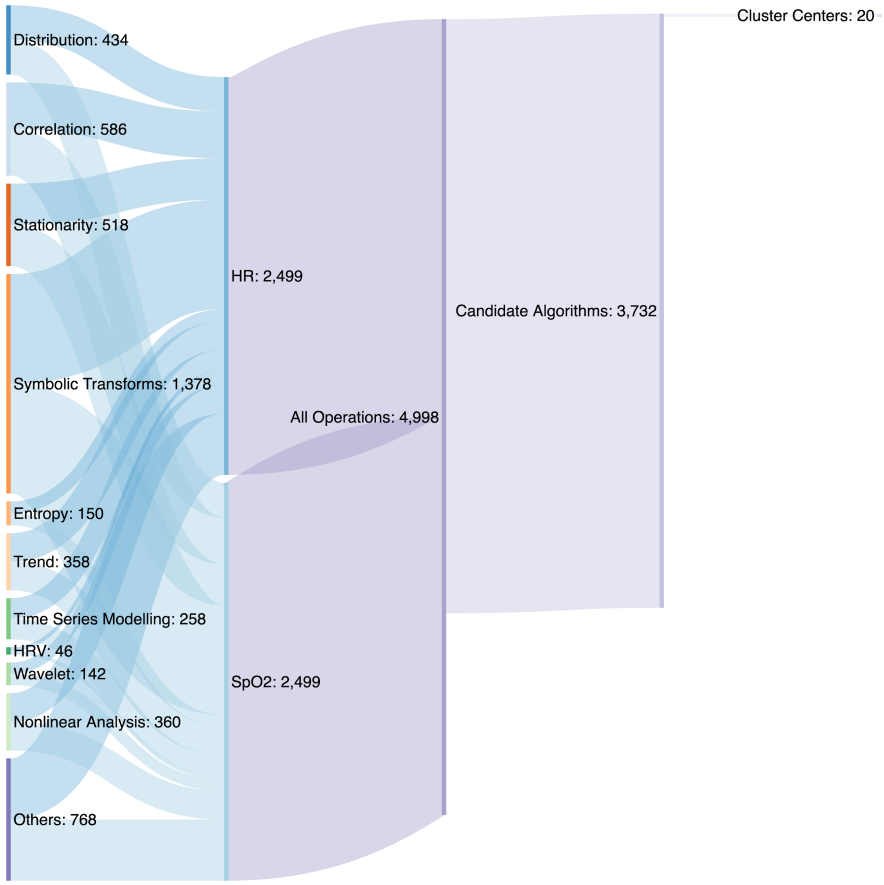

Figure 2. Sankey plot of the computational approach. Beginning with 3555 operations representing 11 families of algorithms, we ended with 20 clusters each represented by a single operation.

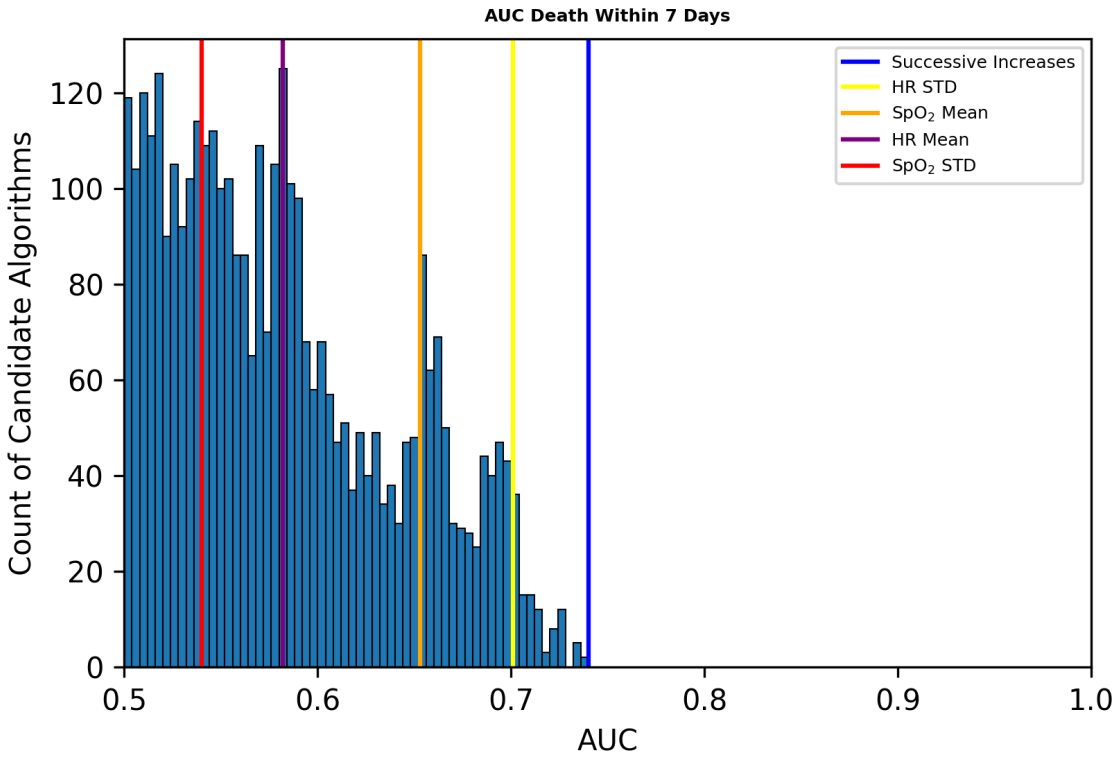

Figure 3. **AUCs of 3555 operations for predicting death in the next 7 days.** Colored vertical bars from left to right indicate the AUCs of the standard deviation of oxygen saturation, mean heart rate, mean oxygen saturation, standard deviation of heart rate, and a novel measure, successive increases of heart rate.

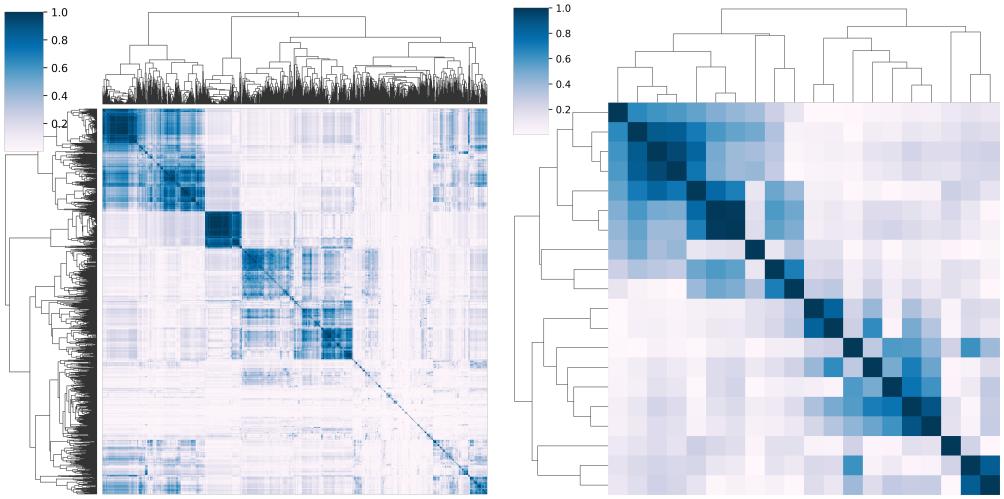

Figure 4. Heat maps of the absolute values of the correlation coefficients between results of operations. Left: correlations between all 3555 candidate algorithmic operations. Right: Correlations between 20 cluster medoids. The reduced feature set explains 81% of the variance in the full set.

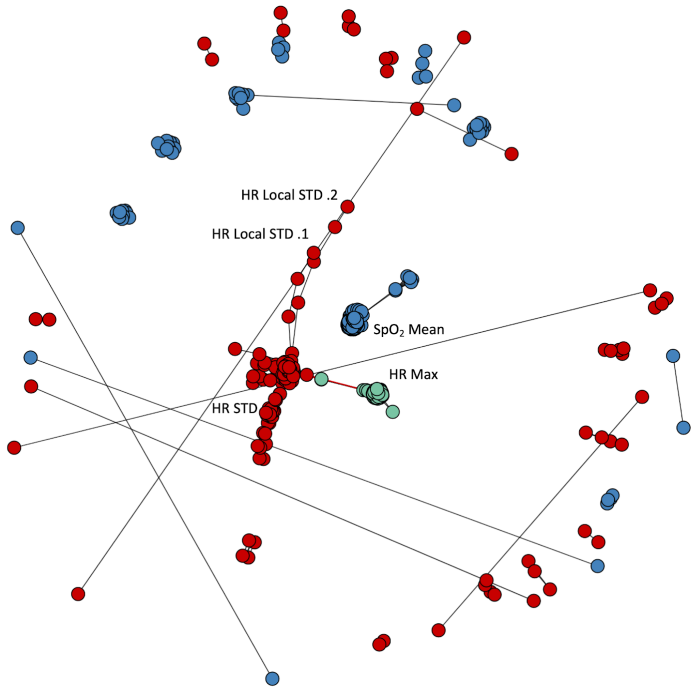

Figure 5. Three representative clusters of operations and their relations to each other. Each dot represents an individual operation. The colors - green, blue, and red - represent the three clusters; several individual operations are labeled. The green group represents measures of the maximum of the heart rate, the blue group reports on the mean oxygen saturation, and the red group reports on the standard deviation of the heart rate. The black lines indicate pairings between operations in the same cluster, whereas the very short red line indicates the small number of pairings between operations from different clusters. HR heart rate; STD standard deviation.

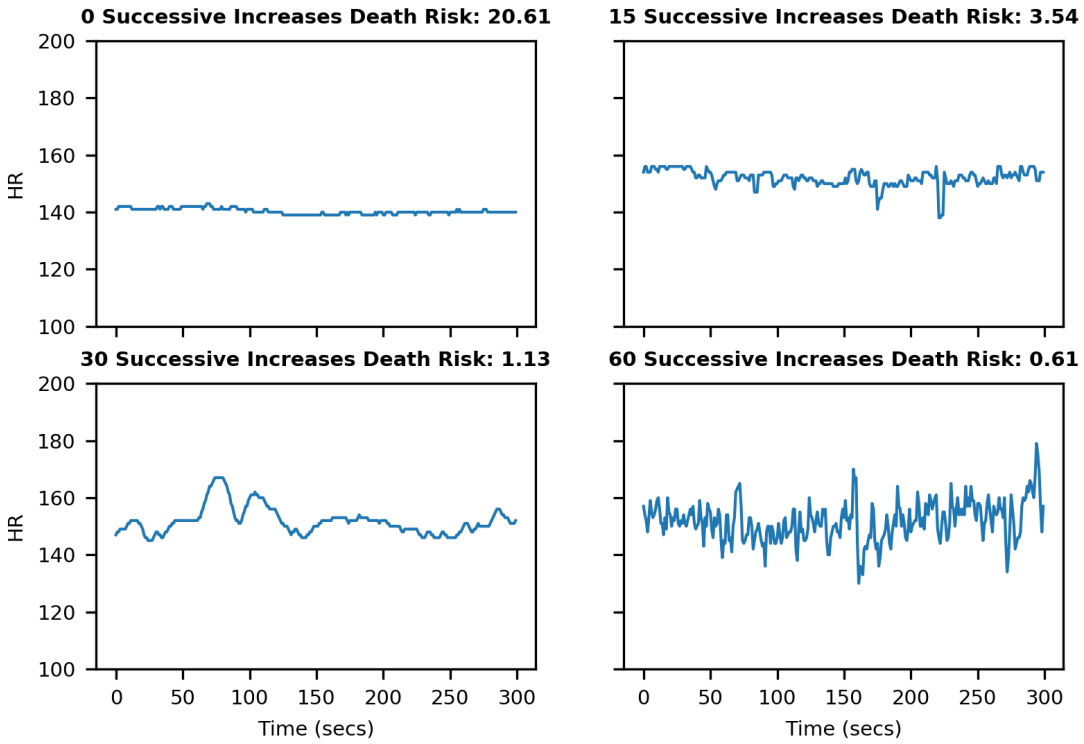

Figure 6. Lack of successive increases in HR predict increased risk of death. Four 10-minute heart rate records that correspond to increasing death risk with a decreasing number of *successive increases* in heart rate.

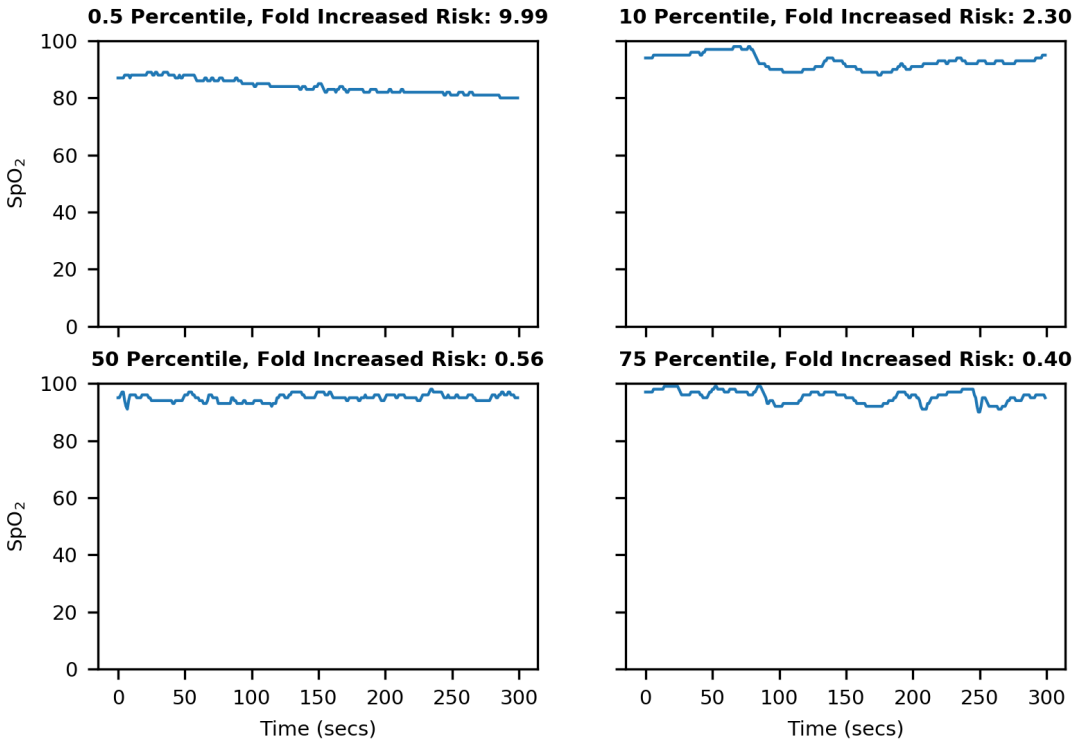

Figure 7. Lower moving threshold in SpO_2 predicts increased risk of death. Four 10-minute heart rate records that correspond to increasing death risk with a decreasing percentile value for SpO_2 moving threshold.

Bibliography

1 Moss TJ, Lake DE, Calland JF, *et al.* Signatures of subacute potentially catastrophic illness in the ICU: model development and validation. *Crit Care Med* 2016; **44**: 1639–48.

2 Griffin MP, O'Shea TM, Bissonette EA, Harrell FE, Lake DE, Moorman JR. Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness. *Pediatr Res* 2003; **53**: 920–6.

3 Moorman JR, Carlo WA, Kattwinkel J, *et al.* Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial. *J Pediatr* 2011; **159**: 900-6.e1.

4 Fairchild KD, Schelonka RL, Kaufman DA, *et al.* Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial. *Pediatr Res* 2013; **74**: 570–5.

5 Schelonka RL, Carlo WA, Bauer CR, *et al.* Mortality and neurodevelopmental outcomes in the heart rate characteristics monitoring randomized controlled trial. *J Pediatr* 2020; **219**: 48–53.

6 Kovatchev BP, Farhy LS, Cao H, Griffin MP, Lake DE, Moorman JR. Sample asymmetry analysis of heart rate characteristics with application to neonatal sepsis and systemic inflammatory response syndrome. *Pediatr Res* 2003; **54**: 892–8.

7 Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol* 2000; **278**: H2039-49.

8 Lake DE, Richman JS, Griffin MP, Moorman JR. Sample entropy analysis of neonatal heart rate variability. *Am J Physiol Regul Integr Comp Physiol* 2002; **283**: R789-97.

9 Richman JS, Lake DE, Moorman JR. Sample Entropy. In: Numerical computer methods, part E. Elsevier, 2004: 172–84.

10 Fulcher BD, Little MA, Jones NS. Highly comparative time-series analysis: the empirical structure of time series and their methods. *J R Soc Interface* 2013; **10**: 20130048.

11 Fulcher BD, Jones NS. hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. *Cell Syst* 2017; **5**: 527-531.e3.

12 Fulcher BD, Lubba CH, Sethi SS, Jones NS. A self-organizing, living library of time-series data. *Sci Data* 2020; **7**: 213.

13 Wang X, Smith K, Hyndman R. Characteristic-Based Clustering for Time Series Data. *Data Min Knowl Discov* 2006; **13**: 335–64.

14 Lubba CH, Sethi SS, Knaute P, Schultz SR, Fulcher BD, Jones NS. catch22: CAnonical Time-series CHaracteristics. *Data Min Knowl Discov* 2019; **33**: 1821–52.

15 Griffin MP, O'Shea TM, Bissonette EA, Harrell FE, Lake DE, Moorman JR. Abnormal heart rate characteristics are associated with neonatal mortality. *Pediatr Res* 2004; **55**: 782–8.

16 Levinson MA, Niestroy J, Al Manir S, *et al.* FAIRSCAPE: A framework for FAIR and

reproducible biomedical analytics. *BioRxiv* 2020; published online Aug 11. DOI:10.1101/2020.08.10.244947.

17 Kunze J, Rogers R. The ARK Identifier Scheme. UC Office of the President: California Digital Library (2008). 2008; published online May 22. https://escholarship.org/uc/item/9p9863nc (accessed June 11, 2020).

18 Gil Y, Miles S, Belhajjame K, et al. PROV Model Primer: W3C Working Group Note 30 April 2013. 2013. https://www.w3.org/TR/prov-primer (accessed Feb 5, 2021).

19 Bishop CM. Pattern Recognition and Machine Learning. Singapore: Springer, 2006.

20 Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004; **69**: 066138.

21 Kraskov A, Stögbauer H, Andrzejak RG, Grassberger P. Hierarchical clustering using mutual information. *Europhys Lett* 2005; **70**: 278.

22 Maechler M, Rousseeuw P, Struyf A, *et al.* CRAN - Package 'Cluster.' Comprehensive R Archive Network, 2019 https://CRAN.R-project.org/package=cluster (accessed Dec 29, 2020).

23 Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; **49**: 1373–9.

24 Leisman DE, Harhay MO, Lederer DJ, *et al.* Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. *Crit Care Med* 2020; **48**: 623–33.

25 Sauerbrei W, Perperoglou A, Schmid M, *et al.* State of the art in selection of variables and functional forms in multivariable analysis-outstanding issues. *Diagn Progn Res* 2020; **4**: 3.

26 Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 1996; **58**: 267–88.

27 Altmann EG, Hallerberg S, Kantz H. Reactions to extreme events: Moving threshold model. *Physica A: Statistical Mechanics and its Applications* 2006; **364**: 435–44.

28 Sun L, Joshi M, Khan SN, Ashrafian H, Darzi A. Clinical impact of multi-parameter continuous non-invasive monitoring in hospital wards: a systematic review and meta-analysis. *J R Soc Med* 2020; **113**: 217–24.

29 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; **1**: 206–15.

30 Guzzetti S, Borroni E, Garbelli PE, *et al.* Symbolic dynamics of heart rate variability: a probe to investigate cardiac autonomic modulation. *Circulation* 2005; **112**: 465–70.

31 Azar Y, Broder AZ, Karlin AR, Linial N, Phillips S. Biased random walks. *Combinatorica* 1996; **16**: 1–18.

32 Weaver W. Probability, rarity, interest, and surprise. *Sci Mon* 1948; **67**: 390–2.

33 Griffin MP, Moorman JR. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. *Pediatrics* 2001; **107**: 97–104.

34 Richardson DK, Gray JE, McCormick MC, Workman K, Goldmann DA. Score for Neonatal Acute Physiology: a physiologic severity index for neonatal intensive care. *Pediatrics* 1993; **91**: 617–23.

35 Gray JE, Richardson DK, McCormick MC, Workman-Daniels K, Goldmann DA. Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index. *Pediatrics* 1992; **90**: 561–7.

36 Griffin MP, Lake DE, Bissonette EA, Harrell FE, O'Shea TM, Moorman JR. Heart rate characteristics: novel physiomarkers to predict neonatal infection and death. *Pediatrics* 2005; **116**: 1070–4.

37 Sullivan BA, McClure C, Hicks J, Lake DE, Moorman JR, Fairchild KD. Early heart rate characteristics predict death and morbidities in preterm infants. *J Pediatr* 2016; **174**: 57–62.

38 Pollack MM, Koch MA, Bartel DA, *et al.* A comparison of neonatal mortality risk prediction models in very low birth weight infants. *Pediatrics* 2000; **105**: 1051–7.

39 Richardson DK, Corcoran JD, Escobar GJ, Lee SK. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. *J Pediatr* 2001; **138**: 92–100.

40 Sullivan BA, Wallman-Stokes A, Isler J, *et al.* Early pulse oximetry data improves prediction of death and adverse outcomes in a two-center cohort of very low birth weight infants. *Am J Perinatol* 2018; **35**: 1331–8.

41 Saria S, Rajani AK, Gould J, Koller D, Penn AA. Integration of early physiological responses predicts later illness severity in preterm infants. *Sci Transl Med* 2010; **2**: 48ra65.

42 Khera R, Haimovich J, Hurley NC, *et al.* Use of machine learning models to predict death after acute myocardial infarction. *JAMA Cardiol* 2021; published online March 10. DOI:10.1001/jamacardio.2021.0122.

43 Engelhard MM, Navar AM, Pencina MJ. Incremental Benefits of Machine Learning-When Do We Need a Better Mousetrap? *JAMA Cardiol* 2021; published online March 10. DOI:10.1001/jamacardio.2021.0139.