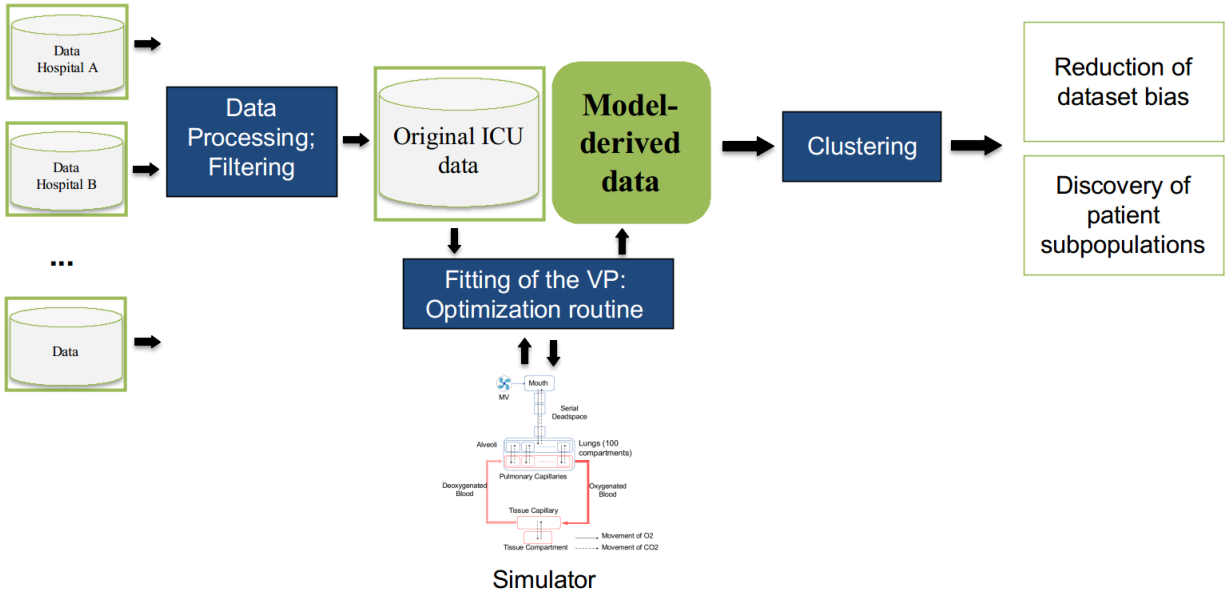
This article has been accepted for publication in IEEE Open Journal of Engineering in Medicine and Biology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJEMB.2023.3243190



Computational simulation of virtual patients reduces dataset bias and improves machine learning-based detection of ARDS from noisy heterogeneous ICU datasets. Mechanistic virtual patient modeling is used to infer model-derived parameters on individual patients, significantly reducing biases introduced by learning from heterogeneous datasets and allowing improved discovery of patient cohorts driven exclusively by medical conditions.

This article has been accepted for publication in IEEE Open Journal of Engineering in Medicine and Biology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJEMB.2023.3243190

Computational simulation of virtual patients reduces dataset bias and improves machine learning-based detection of ARDS from noisy heterogeneous ICU datasets

Konstantin Sharafutdinov[1,2,3], Sebastian Johannes Fritsch[3,4,5], Mina Iravani[1,2,3], Pejman Farhadi Ghalati[1,2], Sina Saffaran[6], Declan G. Bates[6], Jonathan G. Hardman[7], Richard Polzin[1,2,3], Hannah Mayer[3,8], Gernot Marx[3,4], Johannes Bickenbach[3,4], Andreas Schuppert[1,2,3]

[1] Institute for Computational Biomedicine, RWTH Aachen University, Aachen, Germany
[2] Joint Research Center for Computational Biomedicine, RWTH Aachen University, Aachen, Germany
[3] SMITH Consortium of the German Medical Informatics Initiative, Leipzig, Germany
[4] Department of Intensive Care Medicine, University Hospital RWTH Aachen, Aachen, Germany
[5] Juelich Supercomputing Centre, Forschungszentrum Juelich, Juelich, Germany
[6] School of Engineering, University of Warwick, Coventry, UK
[7] School of Medicine, University of Nottingham, Nottingham, UK
[8] Systems Pharmacology & Medicine, Bayer AG, Leverkusen, Germany

CORRESPONDING AUTHOR: Konstantin Sharafutdinov (e-mail: ksharafutdin@ukaachen.de)

This publication of the SMITH consortium was supported by the German Federal Ministry of Education and Research (Grant Nos. 01ZZ1803B and 01ZZ1803M)
This article has supplementary downloadable material

ABSTRACT Goal: Machine learning (ML) technologies that leverage large-scale patient data are promising tools predicting disease evolution in individual patients. However, the limited generalizability of ML models developed on single-center datasets, and their unproven performance in real-world settings, remain significant constraints to their widespread adoption in clinical practice. One approach to tackle this issue is to base learning on large multi-center datasets. However, such heterogeneous datasets can introduce further biases driven by data origin, as data structures and patient cohorts may differ between hospitals. Methods: In this paper, we demonstrate how mechanistic virtual patient (VP) modeling can be used to capture specific features of patients' states and dynamics, while reducing biases introduced by heterogeneous datasets. We show how VP modeling can be used for data augmentation through identification of individualized model parameters approximating disease states of patients with suspected acute respiratory distress syndrome (ARDS) from observational data of mixed origin. We compare the results of an unsupervised learning method (clustering) in two cases: where the learning is based on original patient data and on data derived in the matching procedure of the VP model to real patient data. Results: More robust cluster configurations were observed in clustering using the model-derived data. VP model-based clustering also reduced biases introduced by the inclusion of data from different hospitals and was able to discover an additional cluster with significant ARDS enrichment. Conclusions: Our results indicate that mechanistic VP modeling can be used to significantly reduce biases introduced by learning from heterogeneous datasets and to allow improved discovery of patient cohorts driven exclusively by medical conditions.

INDEX TERMS ARDS, Computational Simulation, Dataset Bias, Machine Learning, Virtual Patients

IMPACT STATEMENT Mechanistic virtual patient modeling can be used to infer individualized parameters approximating disease states of patients, significantly reducing biases introduced by learning from heterogeneous datasets and allowing improved discovery of patient cohorts driven exclusively by medical conditions.

I. Introduction

Artificial intelligence (AI) and machine learning (ML) models have already shown their potential applicability in diverse areas of healthcare [1-3]. Several models have been developed for the early diagnosis and prediction of critical states and conditions in the ICU, e.g., ARDS [4], sepsis [5] and COVID-19 [6-9].

However, the more data-driven models are applied in healthcare settings, the more the issue of impaired performance on different datasets, i.e. poor generalizability of such models, is becoming apparent [5, 10-13]. If ML models are developed on one dataset, they learn data distributions which are specific or characteristic for this

particular dataset and perform worse on data obtained from other sources with potentially different distributions [14-16]. Moreover, attempts to apply models developed in a single hospital to patients from another hospital have also already revealed significant limitations [17, 18]. In medicine generally, but particularly in the ICU setting, there are multiple reasons why data from different hospitals can differ significantly, e.g., different admission strategies, guidelines for treatment, patients' baseline values, protocols on settings of medical support devices or definitions of cut-off values [19-21].

On the one hand, the issue of poor generalizability of developed models cannot be solved by blindly increasing the size of the training dataset, as this does not necessarily guarantee a good performance of a model on another dataset [10]. On the other hand, pooling of data from diverse origins for development of AI/ML tools introduces further biases driven by data origin. This can represent a challenge for the application of both supervised and unsupervised AI/ML methods, as relevant medical information can be hidden behind biases introduced by different datasets [22].

A potential solution to these challenges is to exploit models that allow to infer the core information approximating a patient's status. Such computer models, which are complex enough to model heterogeneous human pathophysiological states, are often referred to as "virtual patient (VP) models" or "in silico" patients [23]. These mechanistic models rely on real patient data and represent a specific pathophysiological state of a patient. Therefore, they can be considered a "digital twin" of a real patient at a given point in time. VP models aim to capture specific features of patient dynamics while avoiding excessive detail. They are based on well accepted and understood physiological principles and can be adapted to represent individual patients [24]. VP modeling, therefore, enables data augmentation through identification of individualized model parameters in the matching procedure of the VP model to real patient data. These model-derived parameters represent an approximation of a disease state of a patient and potentially should not depend on the assessment protocols of the underlying dataset. Therefore, models integrating these parameters are expected to be generalizable across different application sites. In the area of in silico clinical trials encouraging results support this hypothesis. Thus, the responses of the matched VP cohorts to the insulin therapy were generalizable across different hospitals once they were compared to the responses of original cohorts in corresponding hospitals [25]. Moreover, previous applications of hybrid approaches incorporating both mechanistic and data-based modeling have already resulted in successes in other areas of research. Thus, model-derived parameters of individual patients were used to infer important clinical covariates for a patient state [26] or stratify patients [27].

In this paper, we investigate how a mechanistic VP model can be employed to infer model-derived individualized parameters from ICU data pooled from diverse hospitals. We show that such data augmentation allows a reduction in the bias introduced by diverse datasets, and provides clinically meaningful information from noisy heterogeneous data, for instance from data pooled from different hospitals, which allows improved discovery of patient subpopulations through clustering. We demonstrate our approach on a cohort of patients with suspected acute respiratory distress syndrome (ARDS) - a potentially life-threatening condition assessed from multiple hospitals in Germany as part of the ASIC project [28].

During the development of ARDS, due to an inflammatory process and a diffuse damage of alveolar-capillary membrane, protein-rich fluid enters the alveolar space impairing gas exchange. The weight of such a "wet lung" leads to an increased gravitational pressure on the lower, dependent lung compartments. This pressure in combination with the already present edema leads to the formation of atelectases, especially under mechanical ventilation (MV) with inadequate settings [29-31]. This leads to respiratory insufficiency with relevantly impaired pulmonary gas exchange and possible multi-organ failure and fatal outcomes [32, 33]. Despite the existence of an explicit clinical definition (the Berlin definition [34]), significant numbers of patients with ARDS are unrecognized or recognized late by clinicians [35-37]. Thus, diagnosis is difficult and often delayed resulting in incomplete adherence to guideline-based therapy and high morbidity and mortality rates [32, 33]. Failure to recognize ARDS in a timely fashion leads to failure to use strategies that improve survival [37]. Early diagnosis of ARDS may facilitate measures to avoid progression of the lung injury, including protective mechanical ventilation, fluid restriction, and adjunctive measures proven to improve survival such as prone positioning.

Therefore, there is an urgent need for methods that could assist clinicians in early recognition of ARDS in the ICU setting. Several ML models have been developed for the early diagnosis of ARDS in the ICU [4]. However, insufficient quality of ARDS labeling in retrospective datasets, which is caused by under-recognition of ARDS by clinicians [35-37] and by the ambiguities in the use of the Berlin definition [4], represents an important challenge for successful development of applicable ML models, as they must be trained on properly labeled ARDS events. In this paper we provide a way to address this issue. We show that a mechanistic VP model can be used to infer a set of model-derived parameters approximating disease states of individual patients from raw data, which can be used to identify non-diagnosed ARDS patients, providing a route to improved ML model development for early ARDS recognition.

II. Materials and Methods

A. Computational model

The simulator used in this study includes a comprehensive model of the pulmonary system based on mechanistic models of ventilation and gas exchange [38]. It was later extended to include cardiovascular components [39]. The simulator has

already been validated using real patient data [40, 41]. Internally, the model is constructed as a system of differential algebraic equations obtained from published literature, experimental data, and observational studies, that quantitatively represent established physiological processes. The equations are solved iteratively, with the solutions of one iteration at a time point used as inputs to the iteration at the next time step. This allows accurate representation and observation of gradual changes in several parameters that are otherwise difficult to estimate. The simulator consists of different modules representing the airways, the lung as a collection of ventilated alveolar compartments coupled to mechanical ventilator, anatomical shunt, dead space and the tissue compartment. The lung is modeled using 100 alveolar compartments, each of which may have different properties such as flow resistance, vascular resistance, compliance, etc. Thus, ventilation-perfusion mismatch can be modeled, allowing the simulation of conditions such as ARDS [42-44].

The simulator represents a dynamic cardiopulmonary state in vivo that is initialized with numerous input parameters. Some of these parameters are routinely measured in intensive care setting, such as blood gas analysis (BGA) measurements or respirator settings (the full list of parameters used as inputs for the model is given in the Supplementary List I). Others, however, are rarely measured, such as cardiac output, anatomical shunt or biophysical characteristics of individual alveolar compartments, and thus these must be estimated using optimization procedures.

B. Creation of a virtual patient cohort

To fully define each of the virtual patients, the simulator was fitted to individual patient data using advanced global optimization algorithm [45-47]. The model parameters that were identified in the optimization procedure included 2 groups of parameters. Firstly, rarely measured physiological parameters (anatomical shunt, respiratory quotient, anatomical dead space volume, metabolic rate of O2, cardiac stroke volume, and inspiration to expiration ratio), were determined through optimization if they were missing in patient data. Parameters defining distributions of properties of alveolar compartmental parameters (vascular resistance and flow resistance of compartments) were also identified in the optimization process. To model ARDS development, another main parameter was introduced to the optimization procedure – the number of closed alveolar compartments (n_{cc}), accounting for the formation of atelectases and modeled through increased external pressure on the compartment leading to no ventilation and complete alveolar shunt. The optimization problem was formulated to find a configuration of model parameters that minimizes the difference between the model outputs and the observed patient data (arterial blood gas values at all time points in a window). Further details on the optimization procedure are given in the Supplementary File.

The optimization procedure was performed in two time windows relative to the onset of ARDS (t_0): from t_0 - 2d to t_0 - 1d (window 1) and from t_0 to t_0 + 1d (window 2), where d stands for 1 day. We assumed a patient to be in a steady non-ARDS state in the window 1 and in a steady ARDS state in the window 2. The one day interval between the two windows was assumed to represent a transient state and was excluded from the optimization. The optimal parameterization of the simulator for each patient in the window 1 comprised a VP configuration. To model ARDS development, in the window 2 optimization was performed exclusively for the n_{cc} keeping the VP configuration found in the first window intact.

TABLE I
INITIAL AND FINAL NUMBER OF PATIENTS IN THE HOSPITALS UNDER CONSIDERATION.

Hospital	Initial number of patients	Final number of patients
Hosp A	3,591	127
Hosp B	13,067	467
Hosp C	1,360	110
Hosp D	2,217	114
Hosp E	9,040	189

After fitting the simulator to individual patients, a list of parameters was calculated based on simulator outputs and parameters found in the optimization procedure in both time windows for each of the patients. These parameters, among others, included n_{cc}, ventilation and shunted blood fraction (the full list of optimized and simulation output parameters is given in the Supplementary List II). For each of the patients, these parameters comprised model-derived data consisting of 18 features.

C. Data

Four German hospitals (later referred to as Hosp A, Hosp C, Hosp D and Hosp E) provided retrospective, fully depersonalized data on ICU patients collected during the project "Algorithmic surveillance of ICU patients with acute respiratory distress syndrome" (ASIC) [28] of the SMITH consortium, which is part of the German Medical Informatics Initiative. The ASIC project was approved by the independent Ethics Committee (EC) at the RWTH Aachen Faculty of Medicine (local EC reference number: EK 102/19, date of approval: 26.03.2019). The ASIC project was registered at the German Clinical Trials Register (Registration Number: DRKS00014330). The Ethics Committee waived the need to obtain Informed consent for the collection and retrospective analysis of the de-identified data as well as the publication of the results of the analysis. Additionally, a historical dataset from one of the participating hospitals was included into the analysis (Hosp B). It comprised fully depersonalized data of ICU patients that were extracted according to the same rules as within the ASIC project. The time period for the historical dataset started with the introduction of the patient data management system in the ICU of the respective hospital and ended with the start of the ASIC project and covered a period of 10 years. Patient inclusion criteria were age above 18 years and a cumulative duration of invasive MV of at least 24 hours. There were no explicit exclusion criteria. Each patient's data included routinely charted ICU parameters collected over the whole ICU stay, biometric data and ICD-10 codes. The full list of parameters used in this study is given in

TABLE II
CLUSTERING QUALITY FOR CONFIGURATIONS WITH DIFFERENT NUMBER OF CLUSTERS IN CASE OF CLUSTERING ON ORIGINAL MEASURED DATA AND MODEL-DERIVED DATA. MEAN CLUSTERING QUALITY WITH 95 % CONFIDENCE INTERVAL AND RESULTS OF A TWO-TAILED STUDENT'S T-TEST FOR MEAN QUALITY OF CLUSTERING ARE SHOWN.

Number of Clusters	Mean Quality Measured	Mean Quality Simulated	Statistic	p-value
2	0.965 (0.960, 0.970)	0.994 (0.993, 0.995)	11.726	9.481E-21
3	0.869 (0.860, 0.878)	0.994 (0.993, 0.995)	26.205	1.103E-46
4	0.825 (0.815, 0.835)	0.936 (0.930, 0.942)	18.137	2.373E-41
5	0.830 (0.823, 0.837)	0.993 (0.992, 0.994)	43.713	3.222E-67
6	0.788 (0.782, 0.794)	0.935 (0.931, 0.939)	39.515	2.879E-88
7	0.738 (0.732, 0.744)	0.854 (0.848, 0.860)	28.077	6.148E-71
8	0.693 (0.688, 0.698)	0.801 (0.796, 0.806)	29.223	2.781E-73

Supplementary List I. Data from all five datasets were brought to the same units of measurement and were checked for consistency. During depersonalization, the concept of k-anonymity was applied to several parameters that posed a risk to privacy including age, height, weight, and BMI. These parameters were binned into intervals and the number of patients in each interval and in each combination of intervals of 4 parameters was assessed. If there were less than 8 patients in a particular interval or less than 10 patients in any combination of intervals including this interval, all patients of this interval were excluded from the analysis. Due to this, not all datasets of patients who initially met the inclusion criteria could be extracted from the respective hospital and included in the final dataset. The overall number of patients in the final dataset comprised 29,275 patients.

The criteria for the diagnosis of ARDS are defined in the Berlin criteria [34]. As medical imaging data were missing in our dataset, only suspected ARDS onset time could be determined according to the Berlin criteria. It was defined as the timepoint when the ratio of arterial partial pressure of oxygen (PaO_2) and the inspired fraction of oxygen (FiO_2), also known as P/F ratio or Horovitz index, dropped below 300 mmHg for the first time and stayed below this threshold for at least 24 hours. Moreover, to be able to fit a simulator to the ICU data and create a cohort of virtual patients, only patients having specific MV, blood gas analysis and other parameters charted both before and after the suspected ARDS onset were selected. The final number of patients fulfilling these criteria comprised 1,007 patients. The initial and final number of patients in corresponding hospitals is given in Table I. A full description of data preparation is given in the Supplementary File.

D. Consensus clustering and enrichment analysis

We generated two datasets from the patient data representing the individual disease status to be used in the clustering algorithm. The first dataset comprised mean values of original measured parameters, which were used as inputs to the simulator, calculated on time windows 1 and 2 (before and after suspected ARDS onset respectively, see Supplementary List III). The second dataset comprised model-derived data: simulator outputs and parameters found in the optimization procedure (see Supplementary List II). The former dataset thus represented data from the cohort of original patients, while the latter represented the model-derived data, i.e. data from the virtual patient cohort.

Consensus k-means clustering was performed for different number of clusters in each of the cases. Consensus clustering is based on repeated multiple times (1000 times) clustering of the sampled data from the original dataset and is known to produce robust clusters [48]. To further increase robustness of discovered clusters, another step was introduced to the clustering procedure. It was allowed to assign an outlier label to some patients, if they could not be securely assigned to any of observed clusters. In the clustering procedure, quality of clustering was assessed using mean cluster's consensus, as described in [48]. This metric is introduced based on consensus matrix D:

$$D(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)} \quad (1),$$

where $M^{(h)}$ is a connectivity matrix of the perturbed dataset obtained in the h-th resampling of the original dataset and $M^{(h)}(i, j)$ is equal to 1, if items i and j belong to the same cluster in h-th clustering repetition and 0 otherwise. $I^{(h)}$ is the $(N \times N)$ indicator matrix such that its (i, j)-th entry is equal to 1 if both items i and j are present in the perturbed dataset and 0 otherwise. Then, a cluster's consensus m(k) is defined as the average consensus index between all pairs of items belonging to the same cluster k:

$$m(k) = \frac{1}{\frac{N_k(N_k-1)}{2}} \sum_{i,\, j \in I_k,\, i<j} D(i,\, j) \quad (2),$$

where I_k is the set of indices of items belonging to cluster k and N_k is a number of items in cluster k. Finally, the mean cluster's consensus is the cluster's consensus averaged over all clusters. This metric is a summary statistic which reflects the mean stability of clusters discovered in the consensus clustering algorithm and represents the overall robustness of discovered configuration of clusters. Mean clustering quality with 95 % confidence intervals was calculated by repeated (100 times) clustering on subsamples (80%) of dataset. A full description of the clustering procedure is given in the Supplementary File.

For each of the discovered clusters, enrichment with respect to clinical conditions and to each of the 5 underlying hospitals was evaluated using one-sided hypergeometric test for enrichment with a significance level of $\alpha = 0.05$ [49]. Analogously to gene set enrichment analysis, this method allows to identify clinical conditions (or hospitals) that are over-represented in a particular cohort (cluster) of patients compared to the whole population. For instance, if patients of Hosp A are encountered in a particular cluster more frequently than in the overall patient population formed of 5 hospitals, then that cluster is enriched with patients of Hosp A. Observed statistical significance values for each of conditions under consideration were corrected for multiple

This article has been accepted for publication in IEEE Open Journal of Engineering in Medicine and Biology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJEMB.2023.3243190

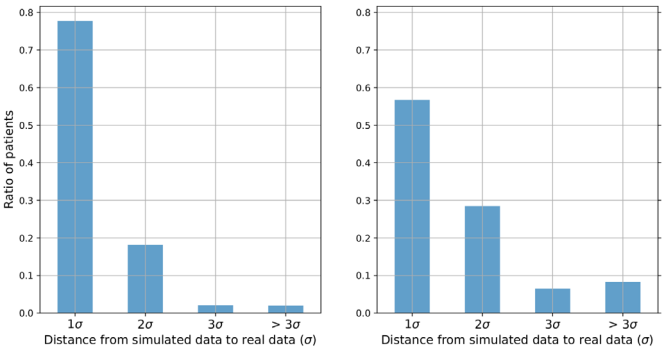


Fig. 1. Quality of fitting the simulator to real patient in the time window before suspected ARDS onset (a) and after suspected ARDS onset (b). Cohort of 1007 patients with suspected ARDS. Acceptable quality of fitting (simulator outputs within 2 standard deviations of measured data) was observed for 95.9% patients in the window before suspected ARDS onset and for 84.5% patients in the time window after suspected ARDS onset.

testing using Benjamini-Hochberg correction [50].

E. Modules used in the study

In this study, the RBFOpt package [39] was used for fitting the VP model to real patient data in the optimization procedure. The following Python programming language [47] implementations were used in the study: scikit-learn [48] implementation of k-means clustering was used in the consensus clustering algorithm (sklearn.cluster.KMeans); scipy [49] implementations of hierarchical clustering were used in the consensus clustering algorithm (scipy.cluster.hierarchy, scipy.spatial.distance); statistical analysis was performed with scipy library (scipy.stats.hypergeom, scipy.stats.ttest_ind). Clustering results were compared using a two-tailed Student's t-test with a significance level of $\alpha = 0.05$.

III. Results

A. Creation of a virtual patient cohort

Fitting quality of the optimization procedure for all patients is shown in Fig. 1. Acceptable quality of fitting (simulator outputs within 2 standard deviations of measured data) was observed for 95.9% patients in the window before suspected ARDS onset and for 84.5% patients in the time window after suspected ARDS onset. Acceptable quality of fitting in both windows was observed for 81.7% or 823 patients. Thus, reliable model-derived data were obtained for 823 patients, which were used in the subsequent analysis.

B. Clustering results

Clustering quality for different configurations of the number of clusters is shown in Fig. 2. For original measured data the best clustering quality was observed for 2 clusters, followed by a steep decrease in clustering quality for 3 clusters and gradual decrease of clustering quality for clustering configurations with a cluster number larger than 5.

In contrast to the clustering on the original measured data, the clustering quality on model-derived data was found to be significantly higher for all configurations of number of clusters (see Fig. 2 for the results of clustering and Table II for the results of the t-test). While on the original measured data, the quality decreased significantly already after

Fig. 2. Clustering quality for different numbers of clusters for clustering on original measured data (orange line) and model-derived data (blue line) data. Mean clustering quality with 95 % confidence intervals over repeated (100 times) clustering on subsample (80%) of dataset is shown. Mean clustering quality and results of a two-tailed Student's t-test for mean quality of clustering are given in Table II.

increasing the number of clusters to 3, in the case of the model-derived data, the quality remained high for 2, 3 and 5 clusters. However, a cluster number above 5 also resulted in a steep decrease in clustering quality in this dataset. Thus, the number of clusters for further investigation was fixed to 5 for both clustering on original and model-derived data.

In case of clustering on original data each of the 5 discovered clusters had certain clinical conditions, which were over-represented in the respective clusters. However, all clusters were found to be driven by data from one or several particular hospitals, i.e. significant enrichment with respect to the hospital was found. Furthermore, 4 out of 5 clusters were dominated by significant over-representation of underlying hospitals, i.e. the highest enrichment was observed with respect to the hospital and not to the clinical condition, see Fig. 3 (a). Enrichment results are given in Supplementary Table I. Finally, none of the discovered clusters had significant enrichment of diagnosed ARDS patients (according to ICD-10 code J80.x).

In contrast, clustering on model-derived data revealed 2 mixed clusters, i.e. clusters without over-representation of any underlying hospital. In the remaining 3 clusters, although such an over-representation could be observed, it was significantly lower than in the clustering on measured original data, see Fig. 3 (b) and Supplementary Table II (significance of 5.0E-49, 2.2E-34, 1.2E-12, 2.5E-8, 5.8E-5 in measured data vs. 1.3E-8, 6.9E-7, 1.2E-6 in model-derived data).

Additionally, clustering on model-derived data was able to discover a cluster with significant ARDS over-representation of diagnosed ARDS patients. This group of patients exhibited multiple properties which are specific for ARDS patients. These encompass the lowest Horovitz index among all clusters, the lowest number of ventilation-free days and the highest mortality. Finally, this cluster showed the largest increase in number of closed alveolar compartments (n_{cc}) among all clusters.

This article has been accepted for publication in IEEE Open Journal of Engineering in Medicine and Biology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJEMB.2023.3243190

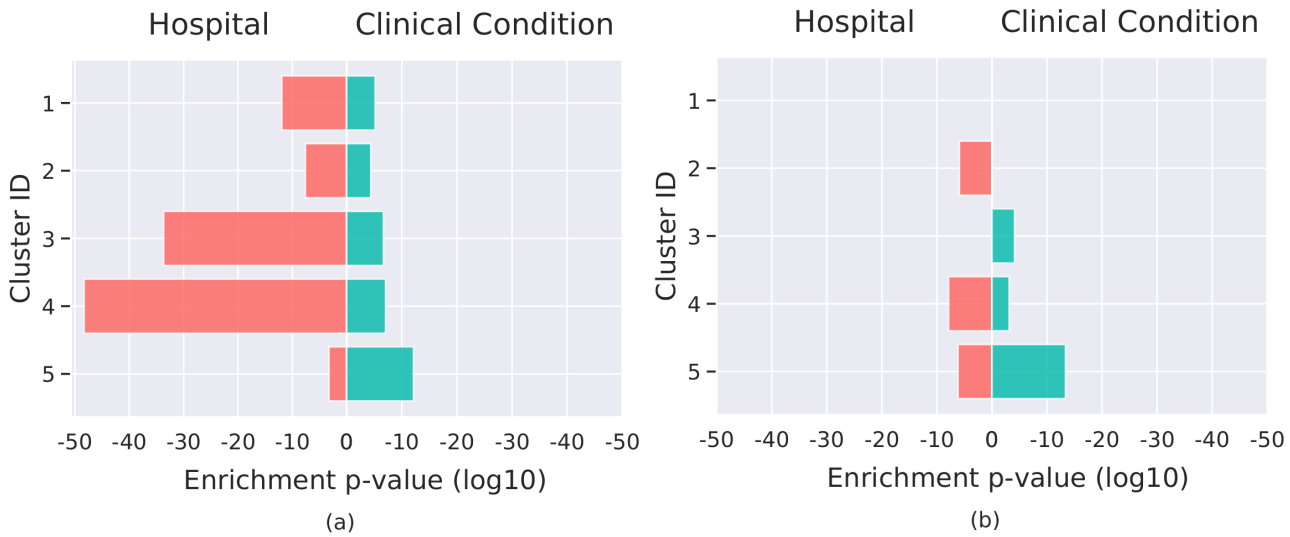


Fig. 3. Significance of enrichment of clinical conditions and underlying hospitals in discovered clusters for clustering on original measured data (a) and model-derived data (b). The highest enrichment in each of the clusters is shown both for enrichment of clinical conditions (green bar) and for enrichment with respect to a hospital (red bar). In clustering on original data, all 5 discovered clusters are significantly enriched with data from some hospitals. In clustering on model-derived data, 2 clusters without enrichment for a hospital are observed and overall magnitude of enrichment with respect to a hospital is decreased.

IV. Discussion

Data which are gathered in the ICU setting consist of global indices and parameters that reflect the state of the lung, such as BGA values or MV settings. However, these features in reality represent surrogate markers for the real pathophysiological state of the patient, leading to a significant simplification of clinical reality. In essence, ICU data are based on systematic monitoring of the enormous complexity of mechanisms accompanying the occurrence and progression of acute syndromes in individual patients. The development of complex syndromes is controlled not only by the core processes of disease progression (often molecular), but also by a large number of covariates arising from a diverse genetic background, lifestyle, exobiotic stress factors, and comorbidities. Another important factor is the large number of medical interventions in the context of intensive care, such as drug administration or MV. All these factors form highly complex feedback systems, in which the patient's condition causes and influences the interventions to be performed, which in turn influence the patient's condition. Such interventions can differ significantly among diverse hospitals introducing additional bias to the datasets [54, 55]. Subsequently, relevant medical signals about a patient's state are often disturbed by noise or are missing completely. For instance, the human lung has inhomogeneous characteristics such as structural asymmetries and regional variations in ventilation and perfusion that cannot be captured by standard diagnostic methods.

To be able to infer relevant patient information, approaches of systems medicine and computational physiology can be used. Systems medicine aims to describe, model, and simulate living, medically relevant systems using methods similar to those used for complex technical processes. The main goal of computational physiology as a part of systems medicine is the adequate description of these relationships in a computationally efficient manner and the development of models that consider unique properties of the living organisms in response to their environment [23, 56]. One of the pillars of computational physiology is VP modeling. The overall VP approach relies on the ability to determine parameters from data that are both patient-specific and time-varying, accounting for variability within and between patients. The ability of VP models, when appropriately adapted, to create a digital twin for a real patient also enables assessment of patient-specific parameters that are not readily measurable (e.g., vascular resistances, transpulmonary pressures, anatomic shunt, etc.). These unmeasurable parameters contain potentially important information about the patient's health status, which cannot be extracted from routinely measured ICU data due to the previously mentioned reasons [24].

In this paper, we demonstrate how a VP modeling framework can be applied to large ICU patient cohorts pooled from different hospitals to reduce dataset bias and to infer parameters approximating patients' disease states. First, we show how a mechanistic VP model can be used to derive model parameters of individual patients with suspected ARDS, which comprise model-derived data. Secondly, we show how these data can be further utilized to improve clustering quality and discover medically relevant patient subpopulations.

A comprehensive physiological model, that was used in this study was already validated against real patient data [40, 41]. However, in the current study, the simulator was firstly used to create a large (>1000 patients) cohort of virtual patients based on the retrospective observational data pooled from different hospitals. VP model fitting to real ICU patients showed a reasonable fitting quality. Acceptable fit in both time windows was observed for 81.7% of the patients in the cohort. The larger ratio of patients with acceptable quality of fitting in the first window can be explained by the

fact that 11 parameters were optimized in the window 1, whereas only 1 parameter, namely n_{cc}, was determined in the window 2. Therefore, reliable model-derived data were obtained for 823 patients. The optimization was performed separately for 2 time windows, which allowed to parameterize a patient in a steady non-ARDS state (window 1), and then track the ARDS development by changes in the number of closed compartments. The optimization using the data from both time windows together to parametrize a VP would potentially enable a better average fit in the time windows. However, this parametrization would correspond to an "average" state and would not allow to follow the progression of the ARDS. Moreover, the optimization of VP parameters other than n_{cc} in the window 2 would potentially allow a better fitting quality in that window. Thus, in the future studies our modeling approach can be improved by allowing other VP parameters to vary within physiologically meaningful ranges during ARDS development, which might improve quality of ARDS modeling. The cohort of patients for whom acceptable fitting quality could not be achieved is of particular interest for further research. On the one hand, our approach for ARDS simulation integrates several assumptions and cannot guarantee an accurate approximation of all pathophysiological processes of ICU patients. On the other hand, the virtual patient model itself may be limited and fail in modeling certain states of ICU patients. For instance, we found that the cohort of patients with low fitting quality is characterized by significantly lower end-inspiratory pressures in the window 2. However, no clinical condition was found to be enriched in this cohort. Nevertheless, further research is needed to fully inspect reasons for low fitting quality.

To demonstrate the utility of the obtained model-derived data, we used a classic unsupervised learning approach, namely clustering. We compared the clustering on original data vs. clustering on inferred model-derived data. Intermediate clustering quality was observed in the clustering on original data, meaning that the consensus clustering method was struggling to split a full cohort into homogeneous groups and find a stable configuration of clusters. In contrast, clustering on model-derived data revealed significantly better clustering quality for all configurations of number of clusters.

More importantly, clustering based on the original data was strongly affected by the diversity of underlying hospitals. In all discovered clusters, patients from a particular hospital were significantly over-represented. In 4 out of 5 clusters, such enrichment was found to be the most significant for that cluster. These observations indicate that clustering on observed data is dominated more by the hospital source and much less by underlying medical conditions. Therefore, clustering on the pooled data is biased by the data source and does not allow to find mixed subgroups of patients. This finding is even more striking given the fact that we did not use external ICU datasets, e.g., MIMIC, HiRID, or AmsterdamUMCdb, for this study, which could have covered different patient populations. All patients in this study satisfied the same strict inclusion criteria and were later filtered and chosen according to uniform rules. For instance, chest X-ray data were not available during the study, which represented the main limitation for the retrospective ARDS diagnosis in the cohort. However, clustering on model-derived data obtained from each of the virtual patients allowed us to find 2 clusters of mixed hospital origin, i.e. clusters without over-representation of any underlying hospital. Moreover, although significant enrichment with respect to the hospital was still present in 3 out of 5 clusters, its magnitude was much less than in the clustering on original data (see Fig. 3).

These findings support the main characteristic of the VP models, namely the ability to identify relevant data patterns and infer individualized model parameters approximating the disease state from underlying data by leveraging mechanistic physiological principles while simultaneously avoiding an excessive level of detail.

Another interesting observation was that clustering on original measured data was not able to find a subgroup of "true" diagnosed ARDS patients. Partially, these patients were uniformly distributed among discovered clusters and did not form a separate group with typical ARDS properties, e.g., an impaired oxygenation or high driving pressures for MV. In contrast, clustering on model-derived data was able to discover a cluster with significant ARDS over-representation and clinical properties, which resemble those of ARDS patients.

This finding is especially important in the context of unreliable ARDS labeling in retrospective data. Insufficient quality of labeling represents an additional factor that contributes to impaired generalization of AI/ML models developed on retrospective ICU data. For the proper development of ML models for ARDS diagnosis and prediction, such models have to be trained on reliably labeled data. On the one hand, patients labeled with ARDS ICD codes still represent a lower bound on the number of true ARDS cases, as large numbers of ARDS patients are not diagnosed [35-37]. On the other hand, reliable retrospective labeling constitutes a challenging task, since diagnosis according to the Berlin definition requires the clinical appraisal of certain conditions, such as hypervolemia, which are not assessable retrospectively. This lack of data is a critical point also for the future work on ARDS. A formalization of fluid overload is a challenging task, since there is no metric which is measured routinely to classify the fluid status of a patient. For instance, a cumulative fluid balance is not suitable to conclude on a hypervolemia. Thus, it remains a clinical appraisal which needs to be assessed at the bedside. Datasets containing this information are highly desirable for the future work on ARDS. However, they are not available yet and their generation would be quite laborious. Thus, it is questionable if they will ever reach the required size to be used in ML algorithms. Moreover, medical imaging data are frequently lacking in retrospective databases with observational ICU data. However, even if imaging data are available, reliable identification of the

ARDS event remains a challenge due to a high interrater variability in chest imaging [57]. Finally, studies on the development of AI models for ARDS are utilizing diverging rules to retrospectively label ARDS patients [58-60].

All patients in the cohort under consideration had a time point (suspected ARDS onset), when a part of the Berlin definition which accounts for the impaired oxygenation was satisfied. Presence of "true" ARDS patients in the cohort was guaranteed by the fact, that some patients had ICD-10 code for diagnosed ARDS. However, some of the patients might have had ARDS, but were not diagnosed and therefore lacked the ICD-10 code for ARDS, since it is known that a relevant number of ARDS cases stays undiagnosed. Therefore, the "true" ARDS cohort would have consisted of these two groups of patients: the "true positives" and "false negatives". Our hypothesis was that the patients from these two groups would be similar to each other and form a shared cluster in the clustering procedure. However, that was not the case for the clustering on original measured data, as none of the discovered clusters was enriched with diagnosed ARDS patients. Clustering on measured data was therefore not able to differentiate between ARDS patients and patients with other conditions, that could have led to decreased Horovitz index. In contrast, through clustering on model-derived data we were able to discover a cluster with significant ARDS over-representation and clinical properties, which resemble those of ARDS patients. At the same time this cluster was not enriched with other pathological conditions, which often have similar clinical picture, such as for instance Heart Failure [61]. Furthermore, this ARDS cluster had the largest increase in the number of closed compartments (n_{cc}) in the model, which fully supports our approach of modeling ARDS by introducing closed alveolar compartments. Our findings suggest that the identified ARDS cluster might also include those ARDS patients which were not diagnosed by the ICU staff. Therefore, this approach could be additionally used to identify non-diagnosed ARDS patients, although further research and retrospective validation is needed to prove this hypothesis.

Our study has some limitations that have to be considered. First, as the actual ARDS clinical diagnosis time was not present in underlying data, the ARDS onset was identified retrospectively based on the Horovitz index. Potential availability of the ARDS diagnosis time would allow precise identification of the time windows for fitting of the VP model (at least for the diagnosed ARDS patients) enabling identification of more reliable VP configurations in future studies. However, to the best of our knowledge, no available database of clinical data contains clinical diagnosis timestamps. Therefore, datasets containing this information will have to be created from the ground up. Second, parameters of the virtual patients that were identified in the window before suspected ARDS onset were assumed to stay constant in the observation window of 2 days. This is only partially true, as most of the identified parameters are changing with time. Therefore, our approach to model ARDS development represents a significant simplification of the complex pathophysiological processes, which are happening during this critical condition. However, in our opinion, it covers the most important clinical manifestation of ARDS and can be used as the first approximation for the modeling. Moreover, our ARDS modeling approach was validated by the fact that the ARDS cluster, which was discovered in the data, had the largest increase in number of closed compartments, as expected. Nevertheless, VP modeling has the potential to infer additional information about the patient status which was not used in this study. For instance, by introducing physiologically meaningful changes in other VP parameters during ARDS development, one might significantly improve quality of ARDS modeling. However, it should be noted that model-derived parameters represent a virtual entity. Therefore, detailed clinical evaluation and validation should be performed before they are used in any support systems at the bedside.

Extensive data requirements and complexity of the fitting process of the VP model constituted additional limitations of the study. The former did not allow us to use all available patient data and was the reason for the significantly lower number of patients in the final analysis cohort compared to the initial cohort (see Table I). It must be considered that to reach the aim to create a sufficiently large dataset for the analysis, not only data collected during the current project but also a historical dataset (Hosp B) were included. It cannot be ruled out that patient populations or therapeutic concepts have changed over the years introducing additional bias into the analysis. However, this limitation reflects the real-world situation, as ML models are mostly developed on retrospective datasets with some temporal separation from datasets, where such models are intended to be used. Furthermore, this limitation does not influence the overall conclusions of the study, as enrichment of a similar magnitude was observed with respect to the Hosp B and the other 4 hospitals (see Supplementary Table I). The latter limitation required the use of the computing cluster for the optimization procedure. Although our approach was limited only to the identification of at most 11 parameters for each of the virtual patients, it required the use of advanced global optimization algorithm and significant computational resources. Matching of the simulator to individual patient data and further analysis was performed on the computational cluster of the RWTH Aachen University using 10 nodes with 40 cores each, 2.66 GHz, 4 GB RAM. The longest runtime for one simulation comprised 5 min. Optimization for each patient required repetitive (100 iterations) simulation for multiple time points in each of the 2 windows. Therefore, the overall matching procedure took on average several days of computational time. All this still tremendously complicates a straightforward implementation of such methods at the bedside.

In general, VP modeling possesses further limitations, restraining its applicability in real-world setting. First, it requires complex validation of the developed models. Second, VP models are usually limited to an organizational level of the human body and do not consider the influence of

exogenous covariates, e.g., preexisting diseases, lifestyle, genetic predispositions, or environmental influences [24].

V. Conclusions

In this study we have shown how a mechanistic VP model can be used to infer parameters approximating disease states of individual patients with suspected ARDS from observational data of mixed origin. Our results support the hypothesis that mechanistic modeling can be used to significantly reduce biases in data, introduced by pooling of data from different hospitals and to allow a discovery of patient cohorts driven exclusively by medical conditions. Overall, the continuous development of hybrid modeling approaches integrating diverse computational technologies, continuing increases in computational power, and ever-growing numbers of available datasets leads to the expectation that these technologies will make a significant contribution to precision medicine, with benefits for patients, physicians, and the healthcare system as a whole.

Supplementary Materials

Supplementary materials include description of data preparation, optimization, and clustering approaches used in the study. Supplementary List I contains the full list of parameters used as inputs for the model. Supplementary List II contains the full list of optimized and simulation output parameters which comprise model-derived data. Supplementary List III contains the full list of features which were extracted from original measured data and used in the clustering procedure. Enrichment analysis results for each of discovered clusters in case of clustering on original measured data are given in Supplementary Table I. Finally, enrichment analysis results for each of discovered clusters in case of clustering on model-derived data are given in Supplementary Table II.

Acknowledgment

This publication of the SMITH consortium was supported by the German Federal Ministry of Education and Research (Grant Nos. 01ZZ1803B and 01ZZ1803M).

Conflicts of interests

All authors declare no conflicts of interest in this paper. HM is an employee of Bayer AG, Germany. HM has stock ownership with Bayer AG, Germany.

Author contributions

JGH, SS and DGB developed the VP ARDS model. HM, SJF, KS, and RP worked on data acquisition and harmonization. KS and MI developed and implemented VP ARDS modeling pipeline. HM gave advice during development of the VP ARDS modeling framework. KS and PFG developed and implemented clustering routines. KS and AS designed the research and performed analysis of the patient data. SJF gave medical advice during the development of the pipeline. SJF, GM and JB interpreted the results from a medical perspective. KS, SJF, and AS wrote the manuscript. All authors read and approved the final manuscript.

References

[1] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat Med,* vol. 25, no. 1, pp. 44-56, Jan 2019, doi: 10.1038/s41591-018-0300-7.

[2] M. Ghassemi *et al.*, "A Review of Challenges and Opportunities in Machine Learning for Health," *AMIA Jt Summits Transl Sci Proc,* vol. 2020, pp. 191-200, 2020.

[3] H. Fröhlich *et al.*, "From hype to reality: data science enabling personalized medicine," *BMC Med,* vol. 16, no. 1, p. 150, Aug 27 2018, doi: 10.1186/s12916-018-1122-7.

[4] A.-K. I. Wong *et al.*, "Machine Learning Methods to Predict Acute Respiratory Failure and Acute Respiratory Distress Syndrome," *Frontiers in Big Data*, vol. 3, 2020, doi: 10.3389/fdata.2020.579774.

[5] M. Schinkel *et al.*, "Clinical applications of artificial intelligence in sepsis: A narrative review," *Computers in Biology and Medicine,* vol. 115, p. 103488, 2019/12/01/ 2019, doi: .

[6] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ,* vol. 369, p. m1328, 2020, doi: 10.1136/bmj.m1328.

[7] L. Wynants *et al.*, "Update to living systematic review on prediction models for diagnosis and prognosis of covid-19," *BMJ,* vol. 372, p. n236, 2021, doi: 10.1136/bmj.n236.

[8] R. Gomes *et al.*, "A Comprehensive Review of Machine Learning Used to Combat COVID-19," *Diagnostics*, vol. 12, no. 8*,* doi: 10.3390/diagnostics12081853.

[9] M. L. Chee *et al.*, "Artificial Intelligence Applications for COVID-19 in Intensive Care and Emergency Settings: A Systematic Review," *International Journal of Environmental Research and Public Health*, vol. 18, no. 9*,* doi: 10.3390/ijerph18094749.

[10] A. Wong *et al.*, "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients," *JAMA Intern Med,* vol. 181, no. 8, pp. 1065-1070, Aug 1 2021, doi: 10.1001/jamainternmed.2021.2626.

[11] F. Cabitza *et al.*, "Unintended Consequences of Machine Learning in Medicine," *Jama,* vol. 318, no. 6, pp. 517-518, Aug 8 2017, doi: 10.1001/jama.2017.7797.

[12] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Med,* vol. 15, no. 11, p. e1002683, Nov 2018, doi: 10.1371/journal.pmed.1002683.

[13] G. Mårtensson *et al.*, "The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study," *Medical Image Analysis,* vol. 66, p. 101714, 2020/12/01/ 2020, doi: .

[14] E. A. AlBadawy *et al.*, "Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing," *Medical physics,* vol. 45, no. 3, pp. 1150-1158, 2018.

[15] E. H. P. Pooch *et al.*, "Can We Trust Deep Learning Based Diagnosis? The Impact of Domain Shift in Chest Radiograph Classification," in *Thoracic Image Analysis*, Cham, J. Petersen *et al.*, Eds., 2020// 2020: Springer International Publishing, pp. 74-83.

[16] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, 2011: IEEE, pp. 1521-1528.

[17] J. Goncalves *et al.*, "Li Yan et al. reply," *Nature Machine Intelligence,* vol. 3, no. 1, pp. 28-32, 2021.

[18] K. Sharafutdinov *et al.*, "Application of convex hull analysis for the evaluation of data heterogeneity between patient populations of different origin and implications of hospital bias in downstream machine-learning-based data processing: A comparison of 4 critical-care patient datasets," *Frontiers in Big Data,* Original Research vol. 5, 2022, doi: 10.3389/fdata.2022.603429.

[19] J. Gallifant *et al.*, "Artificial intelligence for mechanical ventilation: systematic review of design, reporting standards, and bias," *British Journal of Anaesthesia,* 2021.

[20] C. M. Sauer *et al.*, "Systematic Review and Comparison of Publicly Available ICU Data Sets-A Decision Guide for Clinicians and Data Scientists," *Crit Care Med,* vol. 50, no. 6, pp. e581-e588, Jun 1 2022, doi: 10.1097/ccm.0000000000005517.

[21] C. Kelliny *et al.*, "Metabolic syndrome according to different definitions in a rapidly developing country of the African region," *Cardiovascular Diabetology,* vol. 7, no. 1, p. 27, 2008/09/18 2008, doi: 10.1186/1475-2840-7-27.

[22] C. Sáez *et al.*, "Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset," Journal of the American Medical Informatics Association, vol. 28, no. 2, pp. 360–364, Feb. 2021, doi: 10.1093/jamia/ocaa258.

[23] M. Viceconti and P. Hunter, "The Virtual Physiological Human: Ten Years After," *Annual Review of Biomedical Engineering,* vol. 18, no. 1, pp. 103-123, 2016/07/11 2016, doi: 10.1146/annurev-bioeng-110915-114742.

[24] J. G. Chase *et al.*, "Next-generation, personalised, model-based critical care medicine: a state-of-the art review of in silico virtual patient models, methods, and cohorts, and how to validation them," *BioMedical Engineering OnLine,* vol. 17, no. 1, p. 24, 2018/02/20 2018, doi: 10.1186/s12938-018-0455-y.

[25] D. Fey *et al.*, "Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients," *Science Signaling*, vol. 8, no. 408, pp. ra130–ra130, Dec. 2015, doi: 10.1126/scisignal.aab0990.

[26] A. Procopio *et al.*, "Analysis of a Cardiac-Necrosis-Biomarker Release in Patients with Acute Myocardial Infarction via Nonlinear Mixed-Effects Models," *Applied Sciences*, vol. 12, no. 24, Art. no. 24, Jan. 2022, doi: 10.3390/app122413038.

[27] J. L. Dickson *et al.*, "Generalisability of a Virtual Trials Method for Glycaemic Control in Intensive Care," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 7, pp. 1543–1553, Jul. 2018, doi: 10.1109/TBME.2017.2686432.

[28] G. Marx *et al.*, "Algorithmic surveillance of ICU patients with acute respiratory distress syndrome (ASIC): protocol for a multicentre stepped-wedge cluster randomised quality improvement strategy," *BMJ Open,* vol. 11, no. 4, p. e045589, Apr 8 2021, doi: 10.1136/bmjopen-2020-045589.

[29] L. A. Huppert *et al.*, "Pathogenesis of Acute Respiratory Distress Syndrome," *Semin Respir Crit Care Med,* vol. 40, no. 1, pp. 31-39, Feb 2019, doi: 10.1055/s-0039-1683996.

[30] P. van der Zee and D. Gommers, "Recruitment Maneuvers and Higher PEEP, the So-Called Open Lung Concept, in Patients with ARDS," *Crit Care,* vol. 23, no. 1, p. 73, Mar 9 2019, doi: 10.1186/s13054-019-2365-1.

[31] L. K. Reiss *et al.*, "Inflammatory processes during acute respiratory distress syndrome: a complex system," *Curr Opin Crit Care,* vol. 24, no. 1, pp. 1-9, Feb 2018, doi: 10.1097/mcc.0000000000000472.

[32] S. E. Cochi *et al.*, "Mortality trends of acute respiratory distress syndrome in the United States from 1999 to 2013," *Annals of the American Thoracic Society,* vol. 13, no. 10, pp. 1742-1751, 2016.

[33] K. Raymondos *et al.*, "Outcome of acute respiratory distress syndrome in university and non-university hospitals in Germany," *Critical Care,* vol. 21, no. 1, pp. 1-17, 2017.

[34] V. M. Ranieri *et al.*, "Acute respiratory distress syndrome: the Berlin Definition," *Jama,* vol. 307, no. 23, pp. 2526-33, Jun 20 2012, doi: 10.1001/jama.2012.5669.

[35] G. Bellani *et al.*, "Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries," *Jama,* vol. 315, no. 8, pp. 788-800, Feb 23 2016, doi: 10.1001/jama.2016.0291.

[36] S. Fröhlich *et al.*, "Acute respiratory distress syndrome: Underrecognition by clinicians," *Journal of Critical Care,* vol. 28, no. 5, pp. 663-668, 2013/10/01/ 2013, doi: .

[37] G. Bellani *et al.*, "Missed or delayed diagnosis of ARDS: a common and serious problem," *Intensive Care Med,* vol. 46, no. 6, pp. 1180-1183, Jun 2020, doi: 10.1007/s00134-020-06035-0.

[38] J. G. Hardman, "Respiratory physiological modelling—the design, construction, validation and application of a set of original respiratory physiological models.," PhD thesis, Division of Anaesthesia and Intensive Care, University of Nottingham, 2001.

[39] S. Mistry *et al.*, "A computational cardiopulmonary physiology simulator accurately predicts individual patient responses to changes in mechanical ventilator settings," Annu Int Conf IEEE Eng Med Biol Soc, vol. 2022, pp. 3261–3264, Jul. 2022, doi: 10.1109/EMBC48229.2022.9871182.

[40] J. G. Hardman *et al.*, "A physiology simulator: validation of its respiratory components and its ability to predict the patient's response to changes in mechanical ventilation," *Br J Anaesth,* vol. 81, no. 3, pp. 327-32, Sep 1998, doi: 10.1093/bja/81.3.327.

[41] A. Das *et al.*, "A systems engineering approach to validation of a pulmonary physiology simulator for clinical applications," *J R Soc Interface,* vol. 8, no. 54, pp. 44-55, Jan 6 2011, doi: 10.1098/rsif.2010.0224.

[42] R. A. McCahon *et al.*, "Validation and application of a high-fidelity, computational model of acute respiratory distress syndrome to the examination of the indices of oxygenation at constant lung-state," *Br J Anaesth,* vol. 101, no. 3, pp. 358-65, Sep 2008, doi: 10.1093/bja/aen181.

[43] A. Das *et al.*, "What links ventilator driving pressure with survival in the acute respiratory distress syndrome? A computational study," *Respir Res,* vol. 20, no. 1, p. 29, Feb 11 2019, doi: 10.1186/s12931-019-0990-5.

[44] A. Das *et al.*, "Creating virtual ARDS patients," *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2016, pp. 2729-2732, Aug 2016, doi: 10.1109/embc.2016.7591294.

[45] H. M. Gutmann, "A Radial Basis Function Method for Global Optimization," *Journal of Global Optimization,* vol. 19, no. 3, pp. 201-227, 2001/03/01 2001, doi: 10.1023/A:1011255519438.

[46] A. Costa and G. Nannicini, "RBFOpt: an open-source library for black-box optimization with costly function evaluations," *Mathematical Programming Computation,* vol. 10, no. 4, pp. 597-629, 2018/12/01 2018, doi: 10.1007/s12532-018-0144-7.

[47] G. Nannicini, "On the implementation of a global optimization method for mixed-variable problems," *Open Journal of Mathematical Optimization,* vol. 2, pp. 1-25, 2021.

[48] S. Monti *et al.*, "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," *Machine Learning,* vol. 52, no. 1, pp. 91-118, 2003/07/01 2003, doi: 10.1023/A:1023949509487.

[49] I. Rivals *et al.*, "Enrichment or depletion of a GO category within a class of genes: which test?," *Bioinformatics*, vol. 23, no. 4, pp. 401–407, Feb. 2007, doi: 10.1093/bioinformatics/btl633.

[50] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 57, no. 1, pp. 289-300, 1995/01/01 1995, doi: .

[51] G. van Rossum, "Python reference manual," *Department of Computer Science [CS],* no. R 9525, 1995.

[52] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research,* vol. 12, pp. 2825-2830, 2011.

[53] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods,* vol. 17, no. 3, pp. 261-272, 2020/03/01 2020, doi: 10.1038/s41592-019-0686-2.

[54] M. Ghassemi *et al.*, "State of the art review: the data revolution in critical care," Critical Care, vol. 19, no. 1, p. 118, Dec. 2015, doi: 10.1186/s13054-015-0801-4.

[55] K. Sharafutdinov *et al.*, "Biometric covariates and outcome in COVID-19 patients: are we looking close enough?," *BMC Infectious Diseases,* vol. 21, no. 1, p. 1136, 2021/11/04 2021, doi: 10.1186/s12879-021-06823-z.

[56] E. J. Crampin *et al.*, "Computational physiology and the physiome project," *Experimental Physiology,* vol. 89, no. 1, pp. 1-26, 2004/01/01 2004, doi: .

[57] M. W. Sjoding *et al.*, "Interobserver Reliability of the Berlin ARDS Definition and Strategies to Improve the Reliability of ARDS Diagnosis," *CHEST,* vol. 153, no. 2, pp. 361-367, 2018, doi: 10.1016/j.chest.2017.11.037.

[58] D. Zeiberg *et al.*, "Machine learning for patient risk stratification for acute respiratory distress syndrome," *PLOS ONE,* vol. 14, no. 3, p. e0214465, 2019, doi: 10.1371/journal.pone.0214465.

[59] S. Le *et al.*, "Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS)," *Journal of Critical Care,* vol. 60, pp. 96-102, 2020/12/01/ 2020, doi: .

[60] C. Lam *et al.*, "Multitask Learning With Recurrent Neural Networks for Acute Respiratory Distress Syndrome Prediction Using Only Electronic Health Record Data: Model Development and Validation Study," *JMIR Med Inform,* vol. 10, no. 6, p. e36202, 2022/6/15 2022, doi: 10.2196/36202.

[61] K. Komiya *et al.*, "A systematic review of diagnostic methods to differentiate acute lung injury/acute respiratory distress syndrome from cardiogenic pulmonary edema," *Crit Care,* vol. 21, no. 1, p. 228, Aug 25 2017, doi: 10.1186/s13054-017-1809-8.