Virtual Patient Modeling for Heterogeneous Intensive Care Unit Data for the Support of Artificial Intelligence

Thesis · May 2023

DOI: 10.18154/RWTH-2023-05200

1 author:

Konstantin Sharafutdinov
RWTH Aachen University
24 PUBLICATIONS **116** CITATIONS

SEE PROFILE

Virtual Patient Modeling for Heterogeneous Intensive Care Unit Data for the Support of Artificial Intelligence

---

Modellierung virtueller Patienten für heterogene Intensivstationsdaten zur Unterstützung der künstlichen Intelligenz

Von der Fakultät für Maschinenwesen der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von
Konstantin Sharafutdinov

Berichter: Universitätsprofessor Dr. rer. nat. Andreas Schuppert
Universitätsprofessor Alexander Mitsos, Ph.D.

Tag der mündlichen Prüfung: 03.05.2023

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

Abstract

Artificial intelligence (AI) and machine learning (ML) technologies have already shown their power and applicability in multiple areas of healthcare, including the intensive care unit (ICU). However, the limited generalizability of ML models developed on single-center datasets, and subsequently the impaired performance of such models in real-world settings, constitutes a significant constraint to the widespread adoption of data-based approaches in clinical practice. Furthermore, data structures and patient cohorts may significantly differ between hospitals introducing additional bias driven by data origin. These differences can be characterized as a dataset bias which is characteristic for data acquired in the ICU and represents a major challenge for the application of ML methods in the ICU setting.

In our study we propose two frameworks to address the challenge of poor generalization of ML models. The first framework enables the quantitative assessment of dataset bias based on convex hull (CH) analysis and ML methods. It allows an a priori assessment of the generalizability of ML models in a new dataset based on the CH overlaps between a dataset used for model training and the new dataset. First, CH analysis is applied to find mean CH coverage between the two datasets based on overlaps of CH projections onto subspaces spanned by all combinations of 2 features providing an upper bound for the generalization ability of a ML model. Second, 4 types of ML models are trained to classify the origin of a dataset to assess whether it is possible to distinguish between patients from different hospitals. The performance of ML models is evaluated to determine whether hospital's datasets differ in terms of underlying data distributions. Combining the results of these 2 steps, a complete vision of potential generalization issues is obtained.

The second contribution of our work is the development of a virtual patient (VP) modeling framework utilizing real-world ICU data pooled from different hospitals. VP models are computational models which simulate pathophysiological states. After being matched to real patient data, they allow to extract the core information describing a patient's status. Our VP modeling framework employs a mechanistic VP model of the cardiopulmonary system for data augmentation through identification of individualized model parameters approximating disease states of ICU patients. Parameters derived in the VP modeling framework are utilized as inputs for unsupervised ML methods which are used to characterize patient cohorts based on their mechanistic parametrization. Thus, a hybrid modeling framework for the analysis of large-scale ICU patient data is created. We show the advantages of this hybrid modeling framework in comparison to the direct utilization of original ICU data in ML algorithms. Thus, model-derived data can be utilized to reduce dataset bias and discover medically relevant patient subpopulations in heterogeneous ICU datasets. All in all, our novel frameworks integrating both mechanistic and data-driven models allow making a step towards utilization of available real-world ICU data from heterogeneous sources, which encompasses numerous benefits for healthcare.

Zusammenfassung

Die Technologien der künstlichen Intelligenz (KI) und des maschinellen Lernens (ML) haben ihre Leistungsfähigkeit bereits in vielen Bereichen des Gesundheitswesens bewiesen. Dies gilt auch für die Intensivstation (ICU). Die begrenzte Generalisierung von ML-Modellen, die auf den Datensätzen eines einzigen Zentrums entwickelt wurden stellt jedoch ein erhebliches Hindernis für die breite Anwendung der ML-Ansätze in der klinischen Praxis dar. Die Datenstrukturen und Patientenkohorten unterscheiden sich erheblich von Krankenhaus zu Krankenhaus, was zu einem zusätzlichen Bias aufgrund der Herkunft der Daten führt.

In unserer Studie schlagen wir zwei Frameworks vor, um die Herausforderung der schlechten Generalisierung von ML-Modellen im Gesundheitswesen anzugehen. Das erste Framework ermöglicht die quantitative Bewertung der Verzerrung von Datensätzen. Es ermöglicht eine a priori Bewertung der Generalisierbarkeit von ML-Modellen in einem neuen Datensatz auf der Grundlage der CH-Überschneidungen zwischen einem für das Modelltraining verwendeten Datensatz und dem neuen Datensatz. Erstens wird die CH-Analyse angewandt, um die mittlere CH-Abdeckung zwischen den beiden Datensätzen zu ermitteln, die auf den Überschneidungen der CH-Projektionen auf Unterräume basiert, die von allen Kombinationen von 2 Merkmalen aufgespannt werden. Zweitens werden 4 Arten von ML-Modellen trainiert zu beurteilen, ob es möglich ist, zwischen Patienten aus verschiedenen Krankenhäusern zu unterscheiden. Die Leistung der ML-Modelle wird bewertet, um festzustellen, ob sich die Datenverteilungen in der Krankenhäuser unterscheiden. Durch die Kombination der Ergebnisse dieser beiden Schritte erhält man einen vollständigen Überblick über potenzielle Generalisierungsprobleme.

Der zweite Beitrag unserer Arbeit ist die Entwicklung eines Frameworks für die Modellierung virtueller Patienten (VP) für reale ICU-Daten. VP-Modelle sind rechnergestützte Modelle, die pathophysiologische Zustände modellieren. Nach der Anpassung an reale Patientendaten stellt ein VP-Modell einen bestimmten pathophysiologischen Zustand eines realen Patienten dar. Unser VP Modeling Framework nutzt ein Modell des kardiopulmonalen Systems, um relevante medizinische Informationen über einzelne Patienten anhand der modellgefilterten Daten zu extrahieren, die reale physiologische Prozesse widerspiegeln. Die Ergebnisse des VP Modeling Frameworks werden als Inputs für unüberwachte ML-Methoden verwendet, die zur Charakterisierung von Patientenkohorten auf der Grundlage ihrer mechanistischen Parametrisierung verwendet werden. Wir zeigen die Vorteile solches hybriden Modellierungs-Frameworks im Vergleich zur direkten Verwendung der ICU-Daten in ML-Algorithmen. So können modellgefilterte Daten genutzt werden, um das Bias aufgrund der Herkunft der Daten zu reduzieren und medizinisch relevante Subpopulationen zu entdecken. Alles in allem ermöglichen unsere hybriden Frameworks einen Schritt in Richtung der Nutzung von realen ICU-Daten aus heterogenen Quellen, was zahlreiche Vorteile für das Gesundheitswesen mit sich bringt.

Acknowledgments

This thesis comprises the results of 5 years of work in the JRC Combine within the frames of the ASIC project. In the first place I would like to thank Prof. Schuppert for his constant and immersive supervision throughout my work in the JRC Combine. It was great to always feel the support and keen interest in my project both in times of failure and in times of success.

I would like to acknowledge Prof. Alexander Mitsos for his thorough review of this PhD thesis. His comments and suggestions have made a significant contribution to the final quality of the thesis and have helped me to gain better and deeper understanding of many concepts.

I would like to thank the whole team of the JRC for our work together. Many thanks to our JRC secretary, Mrs. Ulu-Esser, for the steady support and perfect organization of our work in the JRC Combine, for highly personal relation to students' problems and issues. I would like to thank my fellow JRC students, my colleagues, for both cheerful and intellectually intense meetings and discussions. Special thanks to Jayesh Bhat, who introduced me to the world of intensive care data and helped me a lot once I joined JRC. Many thanks to Richard Polzin, who have gone through all ups and downs of the ASIC project, but always stayed positive and has never forgotten to supply me with fresh roasted coffee beans. Thanks to Kilian Merkelbach and Moein Samadi for discussions on AI and hybrid modeling.

I would like to acknowledge the team of the ASIC project. Many thanks to Prof. Gernot Marx for organizing the project, to Prof. Johannes Bickenbach and clinical team for the support on medical questions, to Christoph Müller for intense communication and support by data delivery, to Saskia Deffge for project management, to Prof. Morris Riedel, Chadi Barakat, Hannah Mayer and Simon Fonck for the scientific discussions around the ASIC project. Special thanks to Sebastian Fritsch for his constant support from the medical side

and his thorough answers to my infinite number of emails.

I would like to thank my Alma Mater university – Moscow Institute of Physics and Technology for the possibility to study and perform research from early undergraduate years in a highly scientific and free environment. All professors and supervisors in my Bachelor years have implicitly contributed to this thesis. Many thanks to Prof. Vsevolod Sakbaev, Ilya Glukhov and Prof. Vadim Diesperov, who, unfortunately, has already passed away.

I would like to thank my family, my mother Tatiana and my father Roman, who have invested so much time to allow me to have high quality mathematical education throughout my school years. Finally, I would like to thank my wife Anastasia for the inspiration and support during my PhD studies and all my friends and relatives for all good things they have done to me.

Contents

List of Figures

List of Tables

List of Abbreviations

ADA	AdaBoost Classifier
AECC	American European Consensus Conference
AI	Artificial intelligence
anatShunt	Anatomical shunt
ARDS	Acute respiratory distress syndrome
ASIC	Algorithmic surveillance of ICU patients with acute respiratory distress syndrome
BGA	Bloodgas analysis
CH	Convex hull
CO	Cardiac output
COPD	Chronic obstructive pulmonary disease
CSV	Comma separated value
CT	Computed tomography
D	Consensus matrix
DNN	Deep neural network
ECMO	Extracorporeal membrane oxygenation
EHR	Electronic health record
f	Gas flow to the compartment
FiO_2	Fraction of inspired oxygen
Hb	Haemoglobin
HCO_3a	Bicarbonate (arterial)
ICD	International Classification of Diseases
ICU	Intensive care unit
I:E	Inspiration:expiration ratio
inR	Intercept of flow resistance curve

$inVR$	Intercept of vascular resistance curve
k	Alveolar compliance of compartment
LR	Logistic Regression
M_{ij}	Magnitude of linear dependence
$M^{(h)}(i,j)$	Connectivity matrix
MIMIC	Medical Information Mart for Intensive Care III
$m(k)$	Cluster's consensus
ML	Machine learning
MV	Mechanical ventilation
n_{cc}	Number of closed alveolar compartments
$Ncomp$	Number of alvelar compartments in the model
NPS	Nottingham Physiology Simulator
p	Pressure of alveolar compartment
PaO_2	Arterial partial pressure of oxygen
PDMS	Patient data management system
PEEP	Positive end-expiratory pressure
P_{EI}	End-inspiratory pressure
P_{ext}	External pressure on compartment
$P_{trachea}$	Tracheal pressure
PVR	Pulmonary vascular resistance
Q	Perfusion
R	Alveolar flow resistance
R^2	Coefficient of determination
RF	Random Forest
ROC AUC	Area under receiver operating characteristic curve
RQ	Respiratory quotient
RWE	Real world evidence data
SaO_2	Arterial oxygen saturation
SD	Series deadspace
S	Stiffness of alveolar compartment
SMITH	Smart Medical Information Technology for Healthcare
sR	Slope of flow resistance curve

STEP	Seeding the EuroPhysiome
SV	Stroke volume
SVM	Support Vector Machine
SvO_2	Venous oxygen saturation
sVR	Slope of vascular resistance curve
τ	Recruitment time of compartment
TOP	Threshold opening pressure of compartment
TRALI	Transfusion-related acute lung injury
$UBresist$	Upper airway resistance
v	Volume of alveolar compartment
V_c	Constant collapsing volume
$VDphys$	Anatomical deadspace volume
$VentRate$	Respiratory rate
V_{FRC}	Fractional residual capacity of the lung
VO_2	Metabolic rate of O2
VP	Virtual patient
VPH	Virtual Physiological Human
VR	Pulmonary vascular resistance of compartment
Vt	Tidal volume

List of Original Publications

This thesis includes results published throughout our work in the Joint Research Center for Computational Biomedicine (JRC-COMBINE) [1, 2, 3, 4, 5, 6]. The publications are used in the following chapters as described below including author contributions:

- Chapter 1 uses parts of the following publications [1, 2, 3, 4] to summarize the dissertation.

- Chapter 2 uses parts of the literature review on big data and artificial intelligence in medicine conducted by Pejman Farhadi, Konstantin Sharafutdinov, and Jayesh Bhat and edited Andreas Schuppert [4]: P. Farhadi, **K. Sharafutdinov**, J. S. Bhat, and A. Schuppert, "Big Data und künstliche Intelligenz in der Medizin," Telemedizin: Grundlagen und praktische Anwendung in stationären und ambulanten Einrichtungen, G. Marx, R. Rossaint, and N. Marx, Eds. Berlin, Heidelberg: Springer, 2021, pp. 423–436. doi: 10.1007/978-3-662-60611-7_37. **Author contributions**: PF, KS, JB wrote the manuscript. AS supervised the literature review.

- Chapter 3 uses the following publication to introduce the structure of the ASIC project, within which research included in this thesis was performed [5]: G. Marx, J. Bickenbach, S. Fritsch, J. Kunze, O. Maassen, S. Deffge, J. Kistermann, S. Haferkamp, I. Lutz, N. Voellm, V. Lowitsch, R. Polzin, **K. Sharafutdinov**, H. Mayer, L. Kuepfer, R. Burghaus, W. Schmitt, J. Lippert, M. Riedel, C. Barakat, A. Stollenwerk, S. Fonck, C. Putensen, S. Zenker, F. Erdfelder, D. Grigutsch, R. Kram, S. Beyer, K. Kampe, J. Gewehr, F. Salman, P. Juers, S. Kluge, D. Tiller, E. Wisotzki, S. Gross, L. Homeister, F. Bloos, A. Scherag, D. Ammon, S. Mueller, J. Palm, P. Simon, N. Jahn, M. Loeffler, T. Wendt, T. Schuerholz, P. Groeber, and A. Schuppert, "Algorithmic surveillance of ICU patients with acute respiratory distress syndrome (ASIC): protocol for a multi-

centre stepped-wedge cluster randomised quality improvement strategy," BMJ Open, vol. 11, no. 4, p. e045589, Apr. 2021, doi: 10.1136/bmjopen-2020-045589. **Author contributions**: GM and ASchu developed the concept and design of the ASIC use case. OM, SD and JK worked as project managers for the SMITH project and coordinated the use case. SJF, JBK, OM, SD and JB wrote the manuscript. GM, JB, SJF, JBK, CP, SZ, FE, RK, KK, FS, SK, SG, LH, FB, PS, NJ and TS were responsible for the extraction and summary of the guideline recommendations for the ASIC app. VL and NKV developed the technical architecture of the ASIC app and supervised the programming of the ASIC app. SH, IL, SZ, FE, DG, SB, JP, PJ, DT, EW, DA, SM, TW and PG worked on the technical implementation of the DIC at their respective centres and the data extraction for the ASIC app. ASchu, RP, KS, HM, LK, WS, RB, JL, MR, CB, ASto and SF worked on the identification of unknown risk factors with diagnostic and prognostic relevance for ARDS. ASche and JP provided feedback regarding the study design and the statistical analysis. GM and VL worked on legal issues concerning the implementation of the ASIC app. VL, NV, SH, IL worked on technical issues concerning the implementation of the ASIC app.

- Chapter 4 uses parts of the following publication [1]: **K. Sharafutdinov**, J. Bhat, S. Fritsch, K. Nikulina, M. Samadi, R. Polzin, H. Mayer, G. Marx, J. Bickenbach, and A. Schuppert, "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, vol. 5, 2022, doi: 10.3389/fdata.2022.603429. **Author contributions**: HM, SF, KS, RP, and KN worked on data acquisition and harmonization. KS, MS, and KN developed and implemented CH analysis scripts. KS, JSB, and KN developed and implemented ML routines. KS, JSB, and AS designed the research, performed analysis, analyzed the patient data, and developed the ARDS prediction model. SF gave medical advice during the development of the pipeline. SF, GM, and JB interpreted the results from a medical perspective. KS, JSB, SF, and AS wrote the manuscript.

- Chapter 5 uses parts of the literature review on the use of virtual patient modeling in intensive care conducted by Konstantin Sharafutdinov and edited by Sebastian

Fritsch and Andreas Schuppert [2]: **K. Sharafutdinov**, S. Fritsch, and A. Schuppert, "Virtuelle Patientenmodelle in der Intensivmedizin," Die digitale Intensivstation, 1st ed., G. Marx and S. Meister, Eds. MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 2022, pp. 55–71. **Author contributions**: SF and AS supervised the literature review. KS wrote the manuscript.

- Chapters 5 and 6 use parts of the following publication [3]: **K. Sharafutdinov**, S. Fritsch, M. Iravani, P. Farhadi Ghalati, S. Saffaran, D. Bates, J. Hardman, R. Polzin, H. Mayer, G. Marx, J. Bickenbach, and A. Schuppert, "Computational simulation of virtual patients reduces dataset bias and improves machine learning-based detection of ARDS from noisy heterogeneous ICU datasets," IEEE Open Journal of Engineering in Medicine and Biology, pp. 1–11, 2023, doi: 10.1109/OJEMB.2023.3243190. **Author contributions**: JGH, SS and DGB developed the original VP ARDS model (Nottingham Physiology Simulator). HM, SJF, KS, and RP worked on data acquisition and harmonization. KS and MI developed and implemented novel way of modeling ARDS virtual patients (the ARDS VP modeling framework). HM gave advice during development of the ARDS VP 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 ARDS VP modeling framework. SJF, GM and JB interpreted the results from a medical perspective. KS, SJF, and AS wrote the manuscript.

- Chapters 5 and 6 use parts of the conclusions of the following publication in the discussion on possible ways to decrease computational time of virtual patient matching [6]: C. Barakat, S. Fritsch, **K. Sharafutdinov**, G. Ingólfsson, A. Schuppert, S. Brynjólfsson, and M. Riedel, "Lessons learned on using High-Performance Computing and Data Science Methods towards understanding the Acute Respiratory Distress Syndrome (ARDS)," 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), May 2022, pp. 368–373. doi: 10.23919/MIPRO55190.2022.9803320. **Author contributions**: Conceptualization, C.B., S.F. and M.R.; methodology, C.B., S.F., M.R. and A.S.; software, C.B., K.S.; validation, C.B., A.S., S.F. and M.R.; formal analysis, A.S., S.F. and M.R; investigation, C.B.; data curation, K.S., A.S., S.F. and M.R.; writing—original draft prepara-

tion, C.B.; writing—review and editing, S.F. and M.R.; supervision, A.S., S.B., S.F. and M.R.; project administration, A.S. and M.R.; funding acquisition, A.S. and M.R.

The following bachelor thesis related to the topic of my thesis has been developed by Kateryna Nikulina under my supervision [7]:

- Chapter 4: Nikulina, K. (2021). Comparing Populations Using Convex Hull Analysis. RWTH Aachen University.

Results of this bachelor thesis have been partially used for the publication [1] and are, therefore, used in Chapter 4.

The following publications have been created in addition during the time of working on the presented thesis but did not influence the results of the thesis [8]:

- **K. Sharafutdinov**, S. J. Fritsch, G. Marx, J. Bickenbach, and A. Schuppert, "Biometric covariates and outcome in COVID-19 patients: are we looking close enough?," BMC Infectious Diseases, vol. 21, no. 1, p. 1136, Nov. 2021, doi: 10.1186/s12879-021-06823-z.

- S. Fritsch, **K. Sharafutdinov**, A. Schuppert, and J. Bickenbach, "Nutzung von künstlicher Intelligenz zur Bekämpfung der COVID-19-Pandemie," Anasthesiologie, Intensivmedizin, Notfallmedizin, Schmerztherapie, vol. 57, no. 03, pp. 185–197, Mar. 2022, doi: 10.1055/a-1423-8039.

- C. Barakat, **K. Sharafutdinov**, J. Busch, S. Saffaran, D. Bates, J. Hardman, A. Schuppert, S. Brynjólfsson, S. Fritsch, and M. Riedel, "Developing an Artificial Intelligence-Based Representation of a Virtual Patient Model for Real-Time Diagnosis of Acute Respiratory Distress Syndrome," submitted for publication.

- T. Linden, C. Ku, P. Wendland, **K. Sharafutdinov**, R. Polzin, A. Schuppert, and H. Fröhlich, on behalf of COPERIMOplus, "Survival Multi-Modal Neural Ordinary Differential Equations for Mortality Prediction of Patients with Severe Lung Disease," submitted for publication.

Parts of the following publications are included in this thesis and are reproduced with permission [2, 4]:

- Reproduced from K. Sharafutdinov, S. Fritsch, and A. Schuppert, "Virtuelle Patientenmodelle in der Intensivmedizin," Die digitale Intensivstation, 1st ed., G. Marx and S. Meister, Eds. MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 2022, pp. 55–71. Copyright ©(2023) with permission from MWV Medizinisch Wissenschaftliche Verlagsgesellschaft mbH & Co. KG.

- Reproduced from P. Farhadi, K. Sharafutdinov, J. S. Bhat, and A. Schuppert, "Big Data und künstliche Intelligenz in der Medizin," Telemedizin: Grundlagen und praktische Anwendung in stationären und ambulanten Einrichtungen, G. Marx, R. Rossaint, and N. Marx, Eds. Berlin, Heidelberg: Springer, 2021, pp. 423–436. doi: 10.1007/978-3-662-60611-7_37. Copyright ©(2023) with permission from Springer Nature.

Chapter 1

Introduction

Nowadays, with ever-increasing volumes of available data and exponentially growing computational power, the role of data-based models of artificial intelligence (AI) and machine learning (ML) becomes inevitably more important in many areas of our life. The broad term AI refers to the modeling of human intelligence processes by machines, especially computer systems. Specific AI applications include expert systems, natural language processing, speech recognition and machine vision. ML comprises a part of the broad AI research area. It summarizes a portfolio of mathematical algorithms, so-called universal machines, which allow approximating any function depending on an arbitrary number of variables, and learning the respective representation from data without any knowledge of the underlying mechanisms [9]. Hence, ML provides tools for data classification and learning of predictive models from data alone.

Despite tremendous investments in the past resulting in a great success in human average lifespan extension up to over 80 years, traditional methods of innovation in healthcare approach their saturation. Thus, complementing the conventional innovation approach by data-based approaches may be the only route to successfully tackle the complexity of diseases in aging societies [10]. Today, ML is on the top of the list of technologies that are expected to improve future healthcare. Essentially, expectations of ML in healthcare are premised on the tremendous complexity of the mechanisms driving the onset and development of complex diseases in individual patients, which significantly complicates mechanistic modeling of these phenomena. The evolution of complex diseases is affected by the molecular core processes of disease progression and by many covariates arising from a diverse genetic

background, lifestyle, exobiotic stress factors, and comorbidities [11]. Therefore, the specific feature, namely the ability to learn the behavior of a complex system from data without any explicit a priori knowledge, makes ML the method of choice for areas of healthcare and biomedicine, where the behavior of complex systems has to be learned from data as the underlying mechanistic knowledge is incomplete.

Availability of large amounts of data is a strong prerequisite for advances and successes of ML in any area of application. Fortunately, the healthcare and biomedicine fields are producing huge volumes of information from genomic sequencing, lab studies, and electronic health records (EHRs), forming one of the disciplines where so-called "Big Data" are created [12, 13, 14]. Therefore, the healthcare comprises a fruitful ground for the application of ML methods.

ML has gained a tremendous advantage by developing highly efficient algorithms for the training of so-called deep neural networks (DNN). Driven by data availability and advances of computational technology, these methods have provided new opportunities, primarily in image and time series analysis, in a broad range of application areas, including biomedical applications [15, 16], drug discovery [17], healthcare [10, 18], and precision medicine [19]. Moreover, multiple ML models have been developed for early diagnosis and prediction of diverse critical states and conditions in the intensive care unit (ICU) setting, e.g., acute respiratory distress syndrome (ARDS) [20], sepsis [21], or COVID-19 [22, 23].

However, despite some success stories mainly based on image recognition and time-series analysis, a broad breakthrough in big data analytics in medical research and healthcare could not be achieved so far. Medical data analysis provides several challenges, ranging from the tremendous complexity of the systems to be analyzed, difficulties in defining outcomes, the heterogeneity and poor quality (e.g., missingness) of the data up to an insufficient matching of the available algorithms to the intrinsic patterns of information in the data, which is far beyond the challenges of AI applications in other areas. This results in the fact that only a minor part of developed ML-based approaches are translated to the real-world healthcare setting and are used for decision support, screening, early diagnosis, and treatment of diseases [11, 24].

Furthermore, the more data-driven models are applied in a certain healthcare setting, the more the issue of impaired performance in other datasets, i.e., poor generalization of such models is becoming crucial [25, 26, 21, 27, 28]. Moreover, attempts to apply models

developed in a single hospital to patients from another hospital have also already revealed their limitations [22, 29]. In medicine, and especially in the ICU setting, there are multiple reasons why data from different hospitals significantly differ: different admission strategies, guidelines for treatment, patients' baseline values, protocols on settings of medical support devices, or definitions of cut-off values [30, 31].

On the one hand, the issue of poor generalization of developed models cannot be solved by blindly increasing the sheer size of the training dataset [28]. On the other hand, pooling data of diverse origins for developing ML tools introduces further biases driven by data origin, i.e., underlying hospital. This represents a major challenge for the application of both supervised and unsupervised ML methods, as relevant medical information is hidden behind biases introduced by different datasets [32] and multisite development of ML models often results in systems that sacrifice strong performance at a single site for systems with mediocre or poor performance at many sites [33].

In the last years, with the evolution of ML-based methods, new approaches based on deep representation learning techniques have emerged for multisite dataset adaptation [34]. They rely on the ability to extract latent shared patient features from heterogeneous clinical data by updating the parameters of underlying neural networks. Moreover, generative adversarial networks (GANs) have significantly contributed to representation learning by allowing to map the samples from heterogeneous sources to latent representations where shared information is encoded [35]. The major limitation of deep representation learning models is the lack of interpretability due to the black-box nature of representations. Several approaches have been proposed to account for this challenge. Thus, in the study by Chen et al., a deep learning representation method was augmented with a propensity score matching-related approach, namely K-nearest neighbors [36]. Furthermore, the study by Chu et al. proposed another hybrid framework supporting representation methods by an external knowledge graph capturing the center specificity [37]. It allowed exploiting the interactions between the extracted knowledge features and raw patient features. Thus, in recent years several steps have been made toward the interpretability of deep representation learning methods. Nevertheless, there is still an urgent need for interpretable methods enabling the generalization of ML models for healthcare [36].

However, models exist that allow extracting the interpretable core information describing a patient's status. Computer models complex enough to model various human pathophysi-

ological states while avoiding excessive detail are often referred to as "virtual patient (VP) models" or "in silico" patients [38]. These models allow the utilization of prior systems biology and systems medicine results implemented in explicit equation systems. Therefore, they belong to the domain of mechanistic models. They rely on real patient data, can be parametrized by adapting to data of individual patients, and, therefore, represent a specific pathophysiological state of a patient. Thus, they can be considered a "digital twin" of an actual patient at a given point in time [39].

The main property of a VP model is the ability to capture specific features of a patient's state and dynamics while relying on data collected in a real-world setting. VP models can be adapted to a patient with limited data or computational resources and can, therefore, be used at different application sites, e.g., at the bedside. Consequently, the VP models focus on healthcare applications, such as sensors, decision support systems, or alarm systems. VP models enable the use of mechanistic models and leverage data integration for further investigation or refinement of computational biomarkers [39].

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. 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 study by Dickson et al. [40] demonstrated that 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. However, previous state-of-the-art applications of VP modeling for the ICU setting were limited to the matching of stationary models to clinical trial data mainly assessed throughout dedicated studies [41, 42] or the matching of a relatively simple dynamic one-parameter VP model to the small number of real-world ICU patient datasets having high data density [40]. In all previous applications of VP modeling for the ICU, the number of created virtual patients did not exceed 100, and no studies utilized large real-world ICU databases to create VP cohorts. Moreover, to the

best of our knowledge, hybrid modeling approaches using features of virtual patients derived from routinely measured ICU data in ML models, which would allow generalization of such models, are still lacking.

Our work makes two contributions to address the challenge of dataset bias and poor generalization of ML models in ICU applications.

First, we introduce a framework for the quantitative assessment of dataset bias based on convex hull (CH) analysis and ML methods. This framework allows an a priori assessment of the generalizability of ML models in a new dataset based on the CH overlaps between a dataset used for model training and the new dataset. This framework utilizes a two-step approach. First, CH analysis is applied to find mean CH coverage between the two datasets based on overlaps of CH projections onto subspaces spanned by all combinations of 2 features. This provides an upper bound for the generalization ability of a ML model. Second, we train 4 types of ML models to classify the origin of a dataset to assess whether it is possible to distinguish between patients from different hospitals. The performance of ML models is evaluated to determine whether hospital's datasets differ in terms of underlying data distributions. Combining the results of these 2 steps, a complete vision of potential generalization issues is obtained.

Second, we propose a novel framework for individual VP modeling for real-world ICU data, which enables reduction of the impact of dataset bias in the underlying data on the downstream ML applications. We demonstrate that existing dynamic VP models can be matched to individual patient datasets using limited data routinely assessed from the individual patients in the ICU. Thus, a large ($>$ 1000 patients) cohort of virtual patients is created based on the retrospective observational ICU data pooled from different hospitals. To allow large-scale VP modeling, we introduce a number of physiologically justified assumptions to the modeling approach. Thus, we replace the original method for creating virtual ARDS patients [43] with a novel approach with significantly lower number of model parameters that have to be identified in the optimization procedure. In the VP modeling framework, patient data are mapped through an established global optimization algorithm onto individualized model parameters representing disease-driving mechanisms. These parameters comprise model-derived data.

Model-derived data are further utilized in unsupervised ML routines (clustering). The combination of the VP modeling framework with ML methods constitutes an overarching

hybrid modeling framework. We show that in comparison to the application of ML methods on the original ICU data, the hybrid framework allows to reduce dataset bias and discover medically relevant patient subpopulations. Therefore, VP modeling, in combination with ML, represents a way to address the challenge of dataset bias in real-world ICU data.

The thesis is structured as follows:

- Chapter 2 presents a review of developments, advantages, primary applications, challenges, and limitations of ML methods for the ICU and ARDS. Moreover, the challenge of impaired generalization of ML models in healthcare is discussed in detail.

- In Chapter 3, datasets used in this thesis are described, and data preparation and filtering approaches are presented.

- Chapter 4 presents a framework for comparing populations and assessing an ML model's generalization ability based on CH analysis and ML methods. This framework allows quantitative assessment of dataset bias and a priori assessment of the generalizability of ML models based on datasets, where a model is trained, and where it is intended to be used.

- Chapter 5 introduces a framework for individual virtual patient (VP) modeling for real-world ICU data. A novel way for creation of virtual ARDS patients representing real ICU patients is proposed. A complex mechanistic model representing the pulmonary system is matched to individual ICU patient data through fitting dynamic model parameters representing timely variable physiological processes to data in moving windows.

- Chapter 6 presents a use case for the application of the developed VP modeling framework. The VP modeling framework is coupled to the ML data-based methods, and the benefits of the hybrid modeling pipeline are discussed. We demonstrate, that 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.

- Chapter 7 concludes the derived results and gives an outlook over potential further applications of the VP approach for the ICU.

Chapter 2

Machine learning for intensive care

In this chapter, a review of ML methods for the ICU setting is presented. The latest developments, advantages, and main application areas of such methods are introduced. Moreover, the main challenges of the data-based modeling approach for the ICU setting, focusing on early recognition of ARDS and possible ways to address them, are discussed.

Section 2.1 presents specific properties of the ICU data that shape the scope and features of data-based approaches for the ICU. In Section 2.2 ARDS as a target for ML-based approaches is introduced. We present and discuss reasons and challenges for the limited applicability of ML models developed for early ARDS recognition. Finally, in Section 2.3, one of the main issues of the application of data-driven methods in healthcare, namely the poor generalizability of such models, is discussed.

2.1 Machine learning for intensive care

ICU data belongs to the medical big data domain and shares all its properties and challenges. However, there is a number of unique features of the ICU setting that differentiate ICU data from data from other branches of biomedicine and healthcare. ICU data are collected in the form of Real World Evidence (RWE) data collections of clinical diagnostic data or electronic medical records. The main difference of RWE data is that in contrast to clinical trials data mostly consisting of data assessed throughout dedicated studies with a strict control over potential confounders, RWE data represent observational data as assessed in clinical practice without special focus on data quality, structure, and standardization [44, 4].

Because of the properties of ICU data and their extensive size covering patient popu-

lations exceeding even large clinical studies by orders of magnitude, ML methods can be applied to the ICU datasets to:

- Find rare side effects of therapies which are not covered by designed studies [45, 46];
- Identify patient subgroups which may benefit from new treatment options or which are at increased risk for adverse effects [47, 48];
- Find subclasses of poorly characterized diseases/critical states which might support the design of follow-up clinical studies [49];
- Optimize individual therapeutic strategies [50, 51];
- Develop prediction models for disease progression in individual patients.

Especially the last application area has gained increased attention in recent years, as it can be directly utilized for early recognition of life-threatening critical conditions, such as sepsis or ARDS. So, multiple models have been developed for early diagnosis and prediction of mortality risk of diverse critical states and conditions in the ICU, including ARDS [20], sepsis [21], or COVID-19 [22, 23].

However, there are several unique properties of RWE data, which represent challenges for future research and application of ML methods. Firstly, RWE data are characterized by a high degree of uncertainty, as data collection and documentation can strongly depend on the hospital, doctors, and individual situation at data assessment and documentation. Especially manually assessed parameters are often erroneous or ambiguous [52, 53]. Thus, sophisticated statistical tests for data consistency must be applied to filter out erroneous outliers in the data. On the other side, these filters tend to eliminate all rare patterns, even if a rare pattern may indicate a new, unexpected medical feature with high relevance for the patient. So, the design of outlier detection algorithms requires a deep understanding of the patterns arising from errors in contrast to patterns with medical relevance [54, 55].

Secondly, these datasets are comprised of a very heterogeneous set of collected data for each patient [54], including:

- dynamic parameters of continuous monitoring, e.g., vital signs, blood gas analysis measurements, mechanical ventilation settings, laboratory test results;

- static parameters including patient's biometric data, previous diseases, and comorbidities in the form of codes of the international classification of diseases;
- unstructured medical reports in the form of free text;
- imaging data from radiology analysis.

These heterogeneous data require special methods for data preprocessing and merging different data types.

Thirdly, since ICU data comprise observational data assessed in clinical practice, parameters which are not assumed to be relevant for patient treatment are not or rarely assessed by the doctors. Hence, data availability varies tremendously between individual patients, and "missing data" are the standard for each patient. The availability or absence of some data types is associated with the progression of a patient in the ICU. For instance, the presence of laboratory measurements was found to be predictive for in-hospital mortality [56]. Therefore, ML analysis pipelines for ICU data must include high-performance data imputation algorithms or restrict the data analysis to the expected most relevant parameters to reduce challenges arising from "missing value" handling. This restriction, however, corrupts the optional power of ML for big data analysis, namely the chance to find novel, unexpected patterns indicating new findings with relevance for precision medicine, as data analysis must be restricted to patients where the relevant parameters were assessed. In case when "expensive" data types are required, such as MRI imaging or rarely measured parameters, this restriction often reduces the available number of patients by orders of magnitude [54]. Additionally, when only complete patient data are used, complete case analysis may reduce the power of a method and introduce selection bias [57]. In consequence, the application of powerful ML tools like deep learning, which need extensive (partially $> 10^6$) training datasets, is restricted to special application cases and cannot be used in majority of RWE data analysis tasks [4].

Finally, data which are gathered in the ICU setting consist of global indices and parameters reflecting a patient's state and are rarely directly associated with the core disease-driving mechanisms. These features rather represent surrogate markers than the patient's real pathophysiological state, 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 pa-

tients. The development of complex syndromes is influenced by various factors, including the core processes of disease progression, genetic background, lifestyle, stress factors, and comorbidities [58]. An additional crucial feature of the ICU setting is a large number of medical interventions, for instance, drug administration or mechanical ventilation. Thus, a complex feedback system is formed, in which the patient's condition causes and influences the interventions to be performed, which in turn influence the patient's condition [54, 59]. Hence, analysis of the impact of the parameters assessed in ICU data requires a mapping of the observables to the parameters controlling the core mechanisms of disease evolution. Because the mapping between the clinical diagnostic parameters and the core disease mechanisms are rarely quantified in complex diseases, utilization of ICU data for healthcare benefit today requires ML approaches.

As stated before, one of the most promising application areas of ML in the ICU setting is the early diagnosis and prediction of diverse critical states and conditions. Well-founded predictions will help physicians to choose an appropriate therapy based on predicted risk for a patient. Therefore, after translation to clinics, ML models integrated into smart alarm systems will allow improved allocation of ICU resources. Such systems, which can notify a clinician once a patient's condition begins to deteriorate, are especially of great importance for early prediction of critical states which develop on the timescale of hours and belong to the leading causes of mortality in the ICU, namely sepsis and ARDS.

In the frames of this thesis, the focus of the research will be directed onto ARDS; therefore, there is a need to describe this critical state in more detail.

2.2 Machine learning for ARDS

Acute respiratory distress syndrome (ARDS) is a potentially life-threatening condition common among ICU patients, which leads to respiratory insufficiency with relevantly impaired pulmonary gas exchange and possible multi-organ failure and fatal outcome [60, 61]. Risk factors for ARDS include sepsis, pneumonia, shock, pancreatitis, pulmonary contusion, and drowning [62]. ARDS is initiated by a lung injury followed by an inflammatory process and a diffuse damage of alveolar-capillary membrane. As a consequence, protein-rich fluid enters the alveolar space impairing gas exchange. The weight of such "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 partial collapse of the lung, i.e. formation of atelectasis, which further impairs gas exchange [63, 64, 65]. ARDS is most often associated with a high risk of rapid-onset of sepsis and life-threatening organ dysfunction [66]. Furthermore, mechanical ventilation (MV), that is necessary during ARDS treatment, can further exacerbate the pulmonary damage, causing a ventilator-induced lung injury (VILI) [67].

Incidence of ARDS worldwide remains high, with 10% of total ICU admissions and 23% of all patients requiring mechanical ventilation [68]. Despite explicit clinical definition criteria, ARDS is often under-recognized [68, 69, 70]. For instance, the large multicentre, observational 'LUNG SAFE'-trial observed that up to 39% of the ARDS cases were not diagnosed by the physicians, which suggests procedural and infrastructural deficits [68]. The fact that diagnosis is difficult and often delayed results in incomplete adherence to guideline-based therapy and high morbidity and mortality rates ranging approximately from 25% to 46% [71, 61].

ARDS often remains unrecognised until severe clinical manifestations, such as severe hypoxia, are present [69]. Early recognition and diagnosis of ARDS would allow timely application of ARDS therapy rules, including lung-protective ventilation, i.e., the use of low tidal volumes and the limitation of airway pressures. Such ventilation strategy has shown to improve outcomes compared to MV with high tidal volumes and airway pressures [72, 73]. Therefore, there is an urgent need for methods of early recognition of ARDS in the ICU setting. As ICU today represents an environment with continuous collection of data of multiple types, which are stored in the patient data management systems of hospitals, data-based ML methods can contribute to the development of diagnostic and early-warning systems for timely ARDS recognition.

Development of reliable ML prediction models for ARDS onset is based on properly identified ARDS event. Firstly, it should be found, if a patient had ARDS during his/her stay in the ICU. This information is encoded in ICU data repositories in the form of codes of the International Classification of Diseases (ICD). Starting from the ICD-10 coding system, which stays for the International Classification of Diseases, the Tenth Revision, there exists a specific code, that denotes an ARDS event: J80. J80 includes following subcategories:

- J80.01 Adult respiratory distress syndrome: Mild adult respiratory distress syndrome

- J80.02 Adult respiratory distress syndrome: Moderate adult respiratory distress syndrome
- J80.03 Adult respiratory distress syndrome: Severe adult respiratory distress syndrome
- J80.09 Adult respiratory distress syndrome: Adult respiratory distress syndrome, severity unspecified.

Thus, patients which have ICD-10 code J80 form the ARDS cohort, which can be used in further analysis (although coding is not necessarily always correct, therefore it might be possible that there are also some false-positive cases included). However, the real-world situation is much more complicated. On the one hand, patients labeled with ARDS ICD codes still represent a lower bound of the number of true ARDS cases, as a large ratio of ARDS patients is not diagnosed [68, 69]. Consequently, the large ratio of true ARDS patients is not labeled as ARDS in retrospective data. On the other hand, retrospective identification of ARDS patients is complicated by the fact, that in some freely accessible databases, which are considered to be a gold-standard of ICU databases, for instance in the American "Medical Information Mart for Intensive Care" III (MIMIC) dataset, the older ICD-9 coding system is used [74]. ICD-9 coding system does not contain a specific code for ARDS. Therefore, to retrospectively label ARDS patients in database, the ARDS label should be assigned manually based on additional criteria. For instance, in the study by Reynolds et al. following rules were used:

- ICD-9 codes for pulmonary insufficiency or respiratory failure: 518.5, 518.51, 518.52, 518.53, 518.82
- procedural codes for ventilatory support (96.70, 96.71 and 96.72) [75].

However, authors do not claim that all patients satisfying criteria mentioned above truly had ARDS. Instead, they warn that these criteria represent only one of the possible ways to retrospectively label ARDS patients in data registries, where ICD-9 coding system is used. Therefore, there is no uniform widely used accepted set of rules of how to find ARDS patients in retrospective databases with ICD coding systems older than ICD-10, which comprises a major challenge for retrospective model development on such databases. Therefore, studies on development of ML models for ARDS are utilizing diverging rules to retrospectively

label ARDS patients, including labeling by experienced physicians [47], use of ICD codes with some additional checks [76, 77], or natural language processing methods to examine radiology reports [78].

Next, even if a cohort of ARDS patients is identified, there is another significant issue on the way to ML model development for early ARDS recognition, namely retrospective identification of ARDS onset. Precise onset time of an ARDS episode has to be defined according to the Berlin criteria [79] for ARDS:

- lung injury of acute onset, within 1 week of an apparent clinical insult and with the progression of respiratory symptoms
- bilateral opacities on chest imaging (chest radiograph or CT) not explained by other lung pathology (e.g. effusion, lobar/lung collapse, or nodules)
- respiratory failure not explained by heart failure or volume overload
- decreased PaO_2/FiO_2 ratio (indicates reduced arterial oxygenation from the available inhaled gas):
  - mild ARDS: 201 – 300 mmHg ($\leq$ 39.9 kPa)
  - moderate ARDS: 101 – 200 mmHg ($\leq$ 26.6 kPa)
  - severe ARDS: $\leq$ 100 mmHg ($\leq$ 13.3 kPa)
- minimum positive end expiratory pressure (PEEP) of 5 cmH_2O.

To fully assess these criteria, following types of data are needed: arterial blood gas analysis measurements to assess PaO_2, MV settings to assess FiO_2 and PEEP, radiology data, and medical reports. However, reliable retrospective labeling of the ARDS onset constitutes a hard task due to the fact, that diagnosis according to the Berlin definition requires the clinical appraisal of certain conditions, like a hypervolemia, which is hardly to assess retrospectively. Moreover, medical imaging data is frequently lacking in retrospective databases with observational ICU data [31]. However, even if imaging data are available, reliable identification of the ARDS event still remains a challenge due to a high inter-rater variability in chest imaging [80].

Altogether, aforementioned factors form a major challenge for development of reliable ML models for early ARDS diagnosis. On the one side, a relevant number of ARDS cases stays

undiagnosed, therefore ARDS patients in retrospective cohorts represent a subpopulation of true ARDS cases. On the other side, diverging retrospective ARDS labeling criteria are used in different studies. These two main challenges can contribute to limited real-world applicability of ML models developed for early ARDS recognition.

2.3 Generalization of machine learning models and hospital bias

However, in addition to the obstacles specific to the development of models for ARDS, which were discussed in Section 2.2, there exists one generic problem for the development and application of ML methods in the ICU setting. The more data-driven models are applied in a certain healthcare setting, the more the issue of impaired performance in other datasets, i.e., poor generalization of such models, is becoming crucial [25, 26, 21, 27, 28]. If ML models are developed on one dataset, they learn data distributions, which are specific or characteristic for this particular dataset. Therefore, they perform worse on the data obtained from other sources with potentially different distributions [81, 82, 83]. Moreover, attempts to apply models developed in a single hospital to patients from another hospital have already revealed their limitations [22]. For instance, in 2019 Li Yan et al. built a simple data-driven model from electronic health records of 485 SARS-CoV2 infected patients in the region of Wuhan, China [29]. The authors claimed their model could predict the outcome for patients with $> 90\%$ accuracy using the values of only three laboratory parameters. However, the model failed to deliver the same high accuracy on patient datasets from hospitals in France, the USA, and the Netherlands [84, 85, 86]. In medicine and especially in the ICU setting, there are multiple sources of biases between different hospitals, including different admission strategies, guidelines for treatment, patients' baseline values, protocols on settings of medical support devices, or definitions of cut-off values [30, 87, 31].

An ML model effectively learns a set of rules to predict a certain outcome based on underlying data, which include markers of a patient's state, such as blood gas analysis measurements and laboratory values, and intervention data. However, the ML model inevitably relies on intervention data, either explicitly, when intervention parameters, such as MV settings or administration of medications, are included in the model as features, or implicitly, as the effects of these interventions are reflected in the patient's state parameters [33].

Protocols for interventions significantly differ between different medical centers. Moreover, protocols change with time, even within one center. Since ML models for healthcare are predominantly developed on retrospective data, it remains unclear how the performance of such models is affected by the temporal separation of the target group, even within one hospital. For instance, the study by Chen et al. showed that the relevance of clinical data decays with an effective "half-life" of about 4 months, meaning that using multiple years of historical data can be worse than one year of recent data for the development of decision-support algorithms [88].

The notion of generalizability is introduced on three levels of hierarchy, i.e., a model can possess good generalization ability: internally, if it is applicable only in the same setting where it has been developed within a short time frame; temporally, if it can be securely applied in the same setting prospectively; or externally, if it is applicable in centers different from the one, where it has been developed [89, 90].

On the one hand, the issue of poor generalization of developed models cannot be solved by blindly increasing the sheer size of the training dataset, as this does not necessarily guarantee a good performance of a model in a different setting [28]. On the other hand, pooling data of diverse origins for the development of ML tools introduces further biases driven by data origin, i.e., underlying hospital. This can represent a challenge for the application of both supervised and unsupervised ML methods, as relevant medical information is hidden behind biases introduced by different datasets [32] and multisite development of ML models often results in systems that sacrifice strong performance at a single site for systems with mediocre or poor performance at many sites [33].

In the literature, the problem of dataset bias is addressed in the field of clinical dataset adaptation [91, 92]. Clinical dataset adaptation approaches aim to allow the reliable multisite development of ML models and the adaptation of existing ML models to new datasets. Many studies on multisite dataset adaptation for ML methods are based on the so-called instance matching methods with the most popular propensity score matching (PSM) technique. PSM is a statistical matching technique that attempts to estimate the effect of an intervention by accounting for the covariates that predict receiving this intervention. The use of PSM allows the reduction of biases due to confounding variables that could be found in an estimate of the intervention effect obtained from the direct comparison of outcomes among case and control groups. In the PSM procedure, propensity scores are calculated as

coefficients of a logistic regression model predicting whether an intervention will be assigned to a patient. Then, similarities between patients can be inferred based on propensity scores. PSM allows the creation of a homogeneous matched set of patients having similar feature distribution from multiple underlying datasets, which can be used in the later analysis [93, 94]. However, instance matching methods carry several serious limitations. First, they mostly utilize linear models, which prohibits learning of underlying nonlinear relationships from heterogeneous clinical data, which are common in the ICU setting because of the complexity of patient-physician interactions, see Section 2.1 [95, 96]. Moreover, instance-matching approaches cannot model clinical setting-specific information, for instance, the differences between normal and ICU wards.

In recent years, with the evolution of ML-based methods, new approaches based on deep representation learning techniques have emerged for multisite dataset adaptation [34]. They rely on the ability to extract latent shared patient features from heterogeneous clinical data by updating the parameters of underlying neural networks. For instance, a representation learning method proposed a novel representation for information in electronic health records that explicitly models diverse factors of heterogeneity in underlying data [97]. Moreover, generative adversarial networks (GANs) have significantly contributed to the representation learning domain. In GANs, an adversarial learning strategy, which implies the learning of two neural networks, one of which learns to map the distribution of noise signal input to be identical to that of true samples, is proposed [98]. This strategy allows the mapping of the samples from heterogeneous sources to latent representations where shared information is encoded [35].

The major limitation of deep representation learning models is the lack of interpretability due to the black-box nature of representations. Several approaches have been proposed to address this challenge. Thus, in the study by Chen et al. a representation learning problem was formulated to estimate the effect of a treatment, and the deep learning representation method was augmented with PSM related approach - K-nearest neighbors [36]. This hybrid pipeline allowed to infer the unobserved (or counterfactual) outcomes in an interpretable manner. Furthermore, the study by Chu et al. proposed another hybrid framework supporting representation methods by an external knowledge graph capturing the center specificity [37]. This study proposed a mechanism to exploit the interactions between the extracted knowledge features and raw patient features. Therefore, in recent years several steps have

been made toward the interpretability of deep representation learning methods. However, there is still a considerable need for interpretable methods that allow the generalization of ML models for healthcare [36].

All in all, dataset bias introduces two challenges that must be overcome for the successful application of ML models in the real healthcare setting. First, there is a need for a method for quantitative assessment of the dataset bias. This method would allow an a priori assessment of the generalizability of ML models based on datasets, where a model is trained, and where it is intended to be used. Second, there is a need for a method to reduce biases introduced by different datasets and to extract medically relevant information from noisy heterogeneous data, for instance, data pooled from different hospitals.

To address the first challenge, a novel framework is proposed to assess the generalization capability of ML models among different datasets based on the CH overlap between multivariate datasets. To reduce dimensionality effects, we evaluate CH overlaps using the overlaps of projections onto subspaces spanned by all combinations of 2 features. In comparison to available model-based approaches [99], our CH-based method provides a model-agnostic a priori data assessment and direction of impaired generalization. To account for non-convex validity domains and areas with diverging data densities, we augment the CH approach with an ML-based approach, where 4 types of ML models are trained to classify the origin of a dataset (i.e., from which hospital) and to estimate differences in datasets with respect to underlying distributions. CH overlap of different ICU datasets and its impact on the generalization ability of ML models is investigated in detail in Chapter 4.

To address the second challenge, we utilize mechanistic virtual patient (VP) modeling as a data augmentation step before the application of ML techniques. We demonstrate how this hybrid modeling framework can be used to capture specific features of a patient's state and dynamics while reducing biases introduced by heterogeneous datasets. Fundamentals of the VP modeling approach and our novel ARDS VP modeling framework for real-world ICU data are introduced in Chapter 5. Application of our VP modeling framework to heterogeneous ICU data is presented in Chapter 6.

Chapter 3

Data

3.1 Data Description

The ICU patient data used in this study were collected by German university hospitals within the context of the use case "Algorithmic surveillance of ICU patients with acute respiratory distress syndrome" (ASIC) [5] of the "Smart Medical Information Technology for Healthcare" (SMITH) consortium which is part of the "Medical Informatics Initiative" of the German Federal Ministry of Education and Research. 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). 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. Patient inclusion criteria were age above 18 years and a cumulative duration of invasive mechanical ventilation for at least 24h. Patient datasets were collected by 8 German university hospitals in the time window between 01.03.2020 and 13.12.2021. Data collection was performed in a stepped-wedge cluster fashion, i.e., data collection duration differed among participating hospitals. Worth noticing, that the stepped-wedge cluster design of the study introduced additional temporal bias to the data.

In addition to the data collected during the ASIC project, MIMIC data, and historic retrospective datasets from participating hospitals were used as a so-called "research data repository" - an independent dataset for pipeline development. Therefore, ICU datasets which were used in this study can be split into three categories:

- MIMIC data;
- Calibration data from German hospitals - historic retrospective datasets from each of participating hospitals;
- Control data from German hospitals - data gathered during the ASIC project.

MIMIC database is one of the largest single-center ICU databases comprising deidentified, comprehensive clinical data of patients admitted to critical care units in the Beth Israel Deaconess Medical Center in Boston, Massachusetts. The database contains over 58,000 hospital admissions for 38,645 adults and 7,875 neonates and covers a time period June 2001 - October 2012. MIMIC is freely accessible to researchers worldwide under a data use agreement. Once a data use agreement is accepted, the analysis is unrestricted. This supports applications including academic and industrial research, quality improvement initiatives, and higher education coursework.

MIMIC data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more. MIMIC data were downloaded as a collection of comma separated value (CSV) files and imported into MySQL database system using scripts provided alongside with data.

Calibration data comprised fully anonymized (for anonymization rules see Section 3.2) historic retrospective datasets from each of the participating hospitals. The aim of the calibration data was to assess and quantify biases between datasets of the participating hospitals before delivery of the ASIC study data. Calibration data comprised a representative sample from the patient population of each hospital. Calibration data were chosen based on following rules:

- minimum of 1000 patient datasets;
- patients satisfying inclusion criteria for the ASIC project:
  - adult patients of the ICU ($\geq$ 18 years);
  - invasive mechanical ventilation for at least 24h;
- dataset should form a representative sample from the population, i.e., patients should be randomly selected from the cohort of patients satisfying other criteria. This criterion

should ensure, that ratio of ARDS patients in the calibration dataset is similar to the ratio of ARDS patient in the underlying population.

For calibration data delivery individual datasets were merged into single data export from each of participating hospitals, which was anonymized in the respective site. Data export from each site were transferred to the JRC Combine via the SFTP server through the data integration center of the University Hospital RWTH Aachen. All calibration datasets were checked for consistency for further use in the project.

Finally, *control data* represented retrospective, fully anonymized data of ICU patients collected during the ASIC project. Control data were gathered according to inclusion criteria of the ASIC project, were anonymized and delivered to the JRC Combine in the same fashion, as calibration data. All control datasets were checked for consistency for further use in the project.

Due to data delivery delays not all calibration and control datasets could be utilized during the project. During the development of the pipeline for the generalization assessment (CH analysis), calibration datasets from 3 hospitals were available and therefore the analysis was performed using these 3 datasets and MIMIC data. For the VP analysis and framework development the most recent data from the ASIC trial were used - control data from 4 hospitals. Thus, 8 different datasets were used in frames of the PhD project: MIMIC dataset, 3 calibration datasets, and 4 control datasets. For clarity, calibration datasets will be later referred to as Hosp A, Hosp B, and Hosp C and control datasets will be referred to as Hosp D, Hosp E, Hosp F, and Hosp G.

Patient datasets from the MIMIC database were chosen according to the inclusion criteria of the ASIC use case: invasive MV for at least 24h and patients older than 18 years. To identify the duration of invasive MV of patients from this dataset, a special MIMIC view was used[1]. The final number of patients of the analysed patient cohorts with clinical characteristics in 8 datasets under consideration are given in Table 3.1.

[1] `https://github.com/MIT-LCP/mimic-code/blob/62102b08040ac5db96af7922db8d7832ce30a813/etc/ventilation-durations.sql`

Dataset	**Total number of patients, n (%)**	**Age, years (mean ± SD)**	**Male gender, n (%)**	**Length of stay ICU, days (mean ± SD)**	**Mortality, n (%)**
Hosp A	13,067 (100)	67.3 ± 14.5	8,529 (65.3)	17.3 ± 19.4	3,742 (28.6)
Hosp B	2,976 (100)	67.3 ± 13.8	1,957 (65.8)	21.2 ± 20.1	828 (27.8)
Hosp C	1,368 (100)	68.7 ± 13.0	961 (70.2)	18.7 ± 18.1	608 (44.4)
Hosp D	3,591 (100)	67.6 ± 13.5	2,344 (65.3)	18.7 ± 19.6	1,205 (33.6)
Hosp E	1,360 (100)	69.3 ± 14.6	888 (65.3)	19.0 ± 17.7	660 (48.5)
Hosp F	2,217 (100)	68.0 ± 13.7	1,440 (65.0)	17.1 ± 18.6	729 (32.9)
Hosp G	9,040 (100)	67.3 ± 13.3	5,772 (63.8)	9.6 ± 15.5	1,587 (17.6)
MIMIC	7,683 (100)	64.1 ± 15.5	4,416 (57.5)	13.5 ± 12.4	1,277 (16.6)

Table 3.1: Clinical characteristics of the analysed patient cohorts in 8 datasets under consideration.

3.2 Data Structure

Each patients' data in all 3 types of datasets included routinely charted ICU variables collected over the whole ICU stay, biometric data and ICD codes (ICD-10 codes in German hospitals and ICD-9 in MIMIC). The full list of variables used in this study is given in Appendix A.1 (so-called *ASIC list*). This list was developed in the early phase of the ASIC project and included routinely measured ICU variables of following groups: vital signs, mechanical ventilation settings, blood gas analysis variables, lab tests, administration of medications, extracorporeal membrane oxygenation (ECMO) settings, and biometrics. For all variables required units of measurement were defined. Only variables from the ASIC list were extracted from the MIMIC database. All variables were brought to the predefined units of measurement. Moreover, a list of comorbidities which either are risk factors for ARDS or are associated with ARDS was defined (mostly based on the 'LUNG SAFE'-trial [68]). Later, COVID-19 was added to the list of comorbidities. For every comorbidity from the comorbidities list corresponding ICD-10 and ICD-9 codes were identified and corresponding dictionaries were created. The full list of comorbidities with corresponding ICD-9 and ICD-10 codes is given in Appendix A.2.

Anonymization requirement of the calibration and control data put several constraints on the data which were delivered. First, the concept of k-anonymity was applied to several variables that posed a risk to privacy including age, height, weight, and BMI. These variables were binned into groups and special criteria on the number of patients in each combination of variables were assessed. Due to this, not all datasets of patients who initially met the

Time from admission, min	**Heart rate, 1/min**	**SpO2, %**	...	**Creatinine, umol/l**
0.0	None	None		26.0
15.0	60.0	93.0		None
30.0	63.0	95.0		None
45.0	70.0	98.0		None
60.0	80.0	99.0		None
75.0	80.0	98.0		None
90.0	80.0	98.0		None
105.0	None	None		None
120.0	110.0	99.0		25.0
...	...	...	...	...

Table 3.2: Dynamic data representation for a patient. Measured variables are represented in a form of a time series starting with admission to the ICU. Patient time series are binned into 15 minutes intervals. ICU data are characterized by high missingness, i.e., not all variables are measured at all time points.

inclusion criteria could be extracted from the respective hospital and actually included in the final dataset. Second, ICD-10 codes were reduced to first 3 characters (e.g. A40). This did not apply to ICD codes of ARDS (J80.01, J80.02, J80.03, and J80.09) and COVID-19 (J07.01 and J07.02) which were delivered without any restrictions. Third, patient time series were binned into 15 minutes intervals.

From each of the patient datasets high level medical conditions (so-called *ICD blocks*) and ICD categories (so-called *ICD chapters*) were extracted based on ICD codes. For instance, ICD-10 codes from the range A00-A09 were mapped first onto an ICD block "Intestinal infectious diseases" and then onto ICD chapter "Certain infectious and parasitic diseases". Thus, all patients having ICD codes in this range were assigned to groups of "Intestinal infectious diseases" and "Certain infectious and parasitic diseases".

To sum up, each patient dataset was brought to the form of a CSV file with dynamic data of the structure given in Table 3.2. Additionally, following static features were extracted for each patient: biometric data, length of ICU stay, mortality information, comorbidities, and medical conditions.

3.3 Data Filtering

ICU data quality was described in detail in Section 2.1. In short, ICU data are characterized by high uncertainty, which is induced by large number of errors occurring during data acquisition, management, and storage. In this section we introduce filters which were

variable	Lower threshold	Upper threshold
Central venous oxygen saturation ($ScvO_2$), %	40.0	90.0
Haemoglobin (Hb), mmol/L	2.0	10.0
FiO_2, %	20.0	100.0
Horowitz index, mmHg	10.0	1500.0
PEEP, cmH_2O	2.5	30.0
P_{EI}, cmH_2O	5.0	45.0
Respiratory rate, 1/min	5.0	40.0
Body temperature, °C	30.0	45.0
Tidal volume, ml	100.0	1000.0
Base excess (arterial), mmol/l	-25.0	25.0
SaO_2, %	85.0	100.0
pH (arterial), unitless	6.0	8.0
Bicarbonate (arterial), mmol/l	10.0	50.0
$paCO_2$, mmHg	15.0	90.0
paO_2, mmHg	50.0	250.0
Inspiration : Expiration ratio (I:E), unitless	0.3	6.0
Cardiac output, l/min	1.5	20.0
Driving pressure (deltaP), cmH_2O	2.5	100.0

Table 3.3: The list of dynamic variables for filtering with corresponding thresholds. All values outside of predefined thresholds were considered unreliable and replaced with missing values.

implemented during the project and which allow to filter out some unreliable measurements in the data.

First, a list of variables needed in the virtual patient modeling procedure was created, see Appendix A.3. It included arterial blood gas variables, vital signs, MV parameters, and laboratory variables. For each of the variables from this list, lower and upper thresholds of plausible values were created based on distributions of underlying data. All values outside of predefined thresholds were considered unreliable and replaced with missing values. The full list of variables with corresponding thresholds is given in Table 3.3.

To demonstrate this approach, distribution of measured values of body temperature pooled from all underlying datasets is shown in Figure 3-1. Only body temperature measurements values between 0 °C and 45 °C are shown to suppress strong outliers (body temperatures higher than 45 °C and lower than 0 °C). Majority of measurements fall inside an interval [30 °C, 40 °C]. However, another smaller peaks are observed at around 23 °C and 0 °C. After discussions with medical doctors and employees of the IT department of University Hospital RWTH Aachen the following explanations for these two peaks were

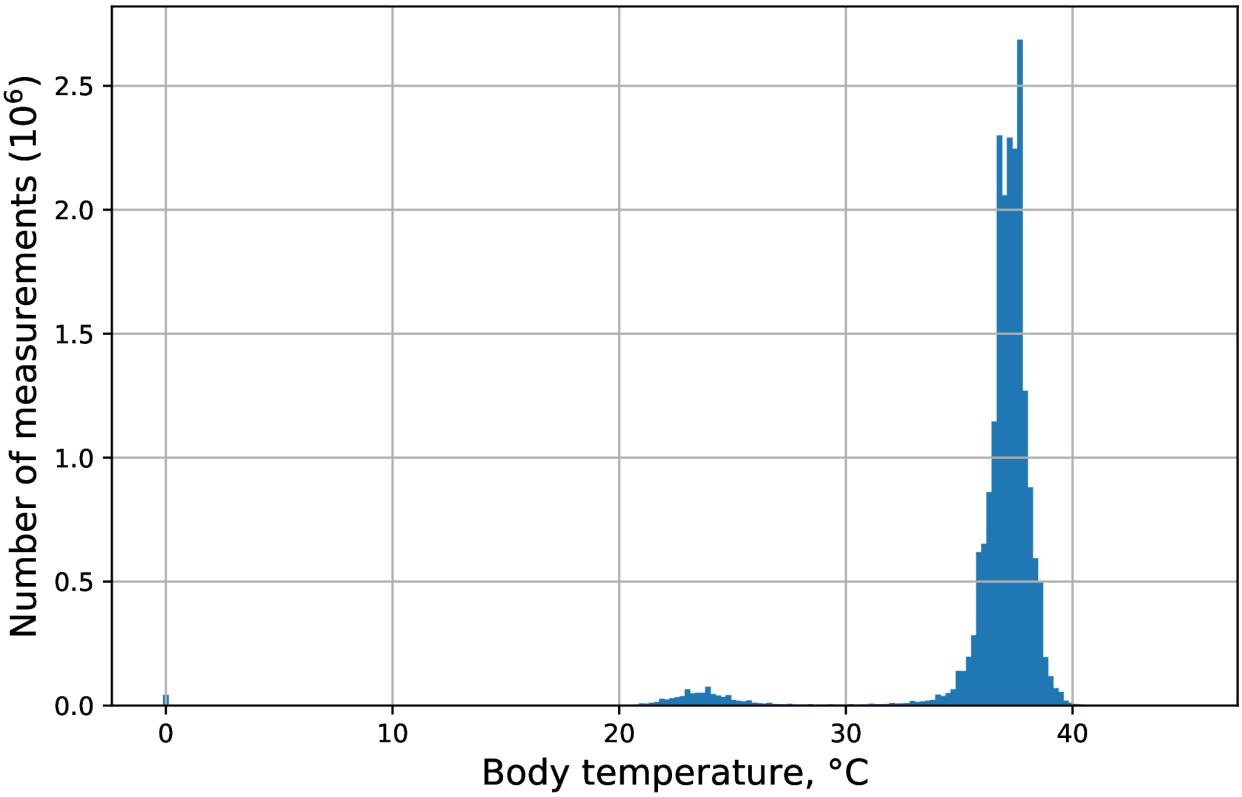


Figure 3-1: Distribution of measured values of body temperature. Three separated peaks are observed.

found: peak around 23 °C could be explained by the measurements, when body temperature sensor is disconnected from the body and measures ambient temperature. The peak around 0 °C can be explained by sensor errors. Based on the distribution of body temperature measurements it was decided to set the thresholds to the following values: 30 °C and 45 °C. Thresholds for other variables were set based on a similar approach which combines visualization and expert knowledge.

There is one variable of outstanding importance for retrospective identification of ARDS onset and classification of ARDS severity, namely the Horowitz index. The Horowitz index is calculated as PaO_2/FiO_2 and reflects arterial oxygenation relative to the composition of inhaled gas. It is used for diagnosis of ARDS and reflects the progression of this critical state. However, Horowitz index measurements were not always present in delivered datasets, even if both PaO_2 and FiO_2 measurements were present. Therefore a script for retrospective calculation of the Horowitz index based on available PaO_2 and FiO_2 measurements was developed. For each of available PaO_2 measurements the FiO_2 from the same time point

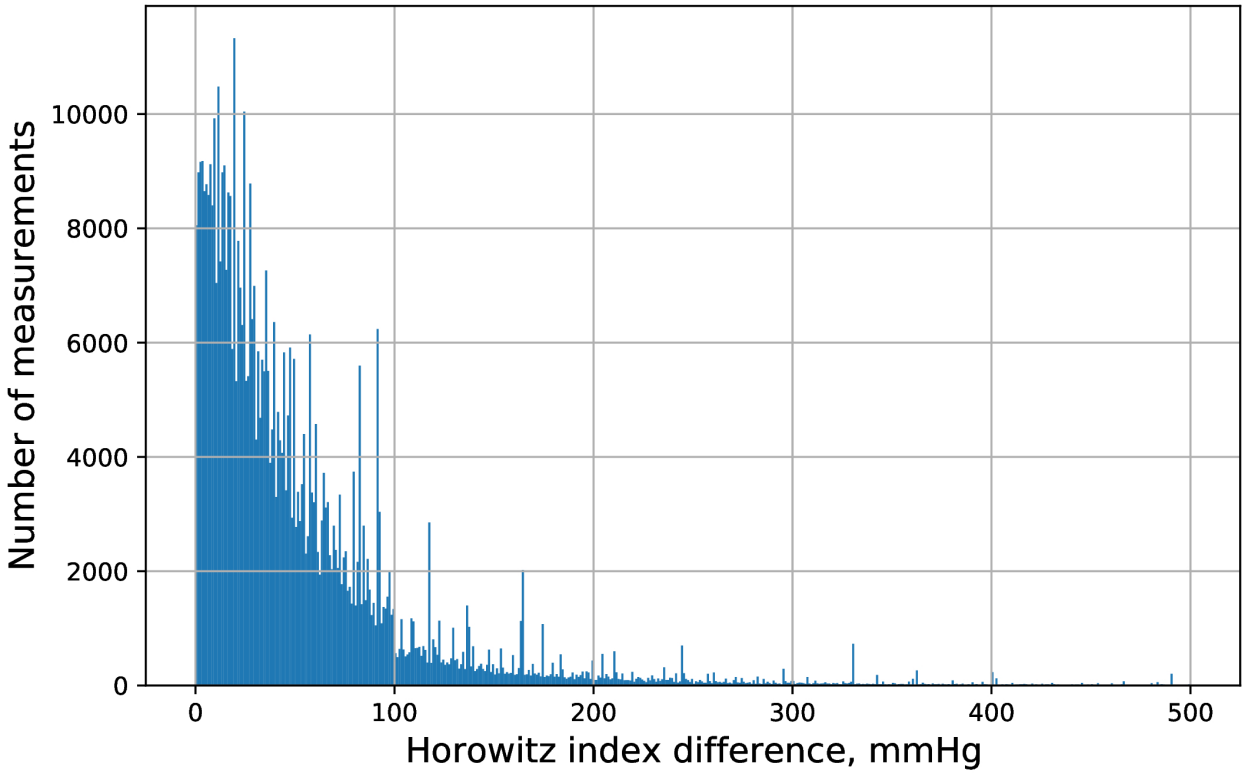


Figure 3-2: Distribution of differences between each two consecutive Horowitz index measurements.

or, if missing, the closest measurement from 8h time window before the PaO_2 measurement was used to calculate the Horowitz index for that time point. If FiO_2 measurements in the 8h window were missing, Horowitz index was not calculated. Horowitz index was calculated only for those time points, where it was missing in the delivered data, thus augmenting original Horowitz data with calculated Horowitz values. This approach was developed in a close collaboration with ICU physicians.

Next, a filter for the Horowitz index was implemented. The Horowitz index represents the main diagnostic marker for the identification of ARDS in the ICU setting and in the retrospective analysis as well. During the project we encountered, that Horowitz index measurements include multiple drops and jumps, which interrupt regions with relatively slow changes in this variable. There could be multiple reasons for such drops and jumps. The Horowitz index is a calculated variable and integrates changes of both measured values of PaO_2 and values of FiO_2, which are set by physicians. FiO_2 can be tuned freely during the ICU stay of a patient. Some treatment strategies include application of 100 % FiO_2 for short

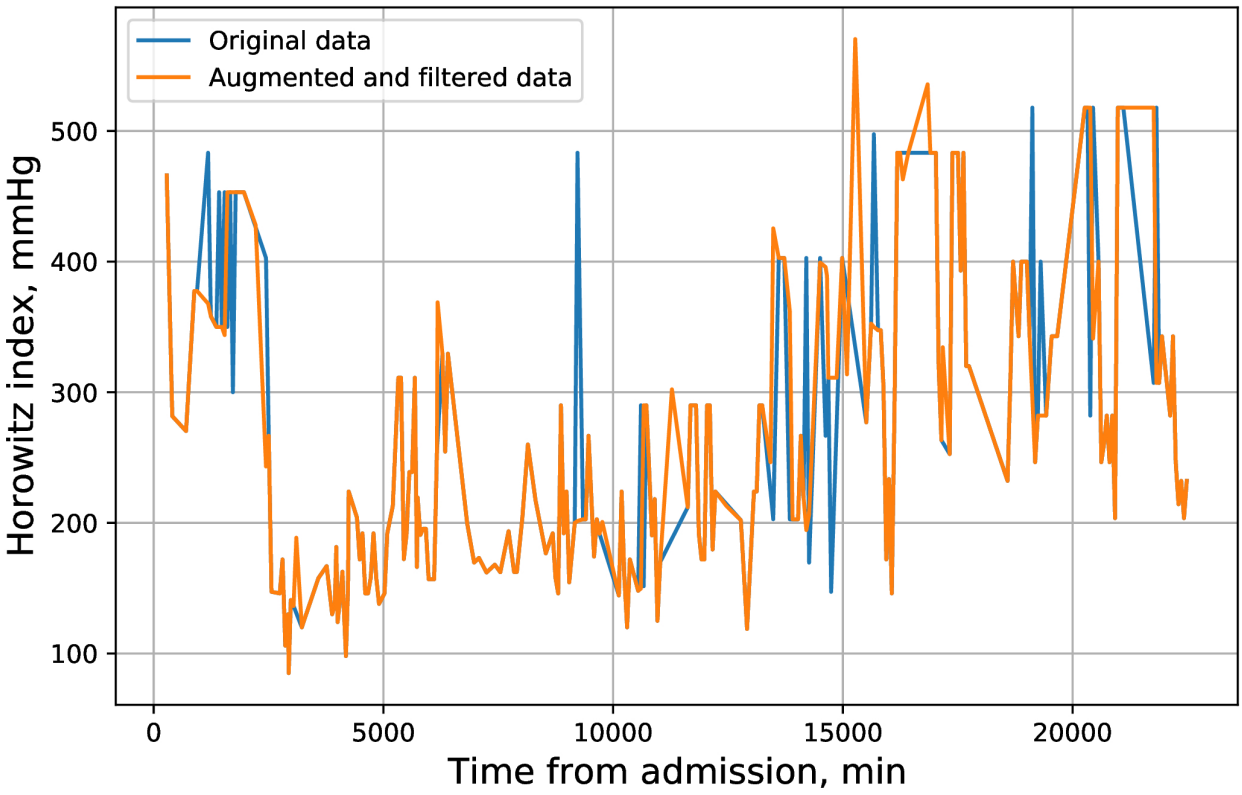


Figure 3-3: Horowitz index time series before and after data augmentation with calculated data and data filtering.

time intervals, which could cause drops in Horowitz. However, PaO_2 and FiO_2 measurement errors could also play a major role. For instance, it was discussed in collaboration with ICU physicians that PaO_2 measurements could be sometimes substituted in the patient data management system by $PaCO_2$ measurements causing the drop in the Horowitz index. Therefore, we have implemented an additional filter which removes single jumps or drops of this variable. We investigated the distribution of differences between each two consecutive Horowitz index measurements, see Figure 3-2. We decided to take 100 mmHg as a threshold to label extreme outlier drops or jumps. Filtering procedure is implemented in the way, that if one measurement differs both from the previous one and the next one more than by this threshold, it is considered as a result of measurement error and is replaced with missing value. We have calculated, that by this procedure 1% Horowitz index measurements are replaced with missing value. A Horowitz index time series of an example patient before and after data processing (augmentation with calculated values and filtering) is shown in Figure 3-3.

Therefore, data processing was performed in the following steps:

1. Checks for data consistency:
    - Units of measurement;
    - Proper data format (numeric or string);
2. Calculation of variables based on available data (Horowitz index, deltaP, etc.);
3. Filtering based on thresholds;
4. Filtering of Horowitz index drops or jumps.

Filters described in this section were implemented stepwise in response to poor matching of the VP model to some patient datasets. Therefore, these filters were not not used in the CH analysis. However, in the CH analysis a special clustering method was used to filter out outliers in the data, therefore partially compensating for threshold filtering.

3.4 Retrospective ARDS onset identification

For reliable model development for ARDS prediction or ARDS modeling it is necessary to retrospectively identify the ARDS onset time. The criteria for the diagnosis of an ARDS episode are defined in the Berlin criteria [79] and were explained in detail in Section 2.2. However, in our use case scenario, imaging data and medical notes were absent. Therefore, only the criteria for oxygenation could be assessed and were taken into account. To be able to assess these criteria, time series of Horowitz index and positive end-expiratory pressure (PEEP) were considered.

ARDS onset time is defined based on the event, when the Horowitz index drops below 300. However, there could be multiple time points with PaO_2/FiO_2 measurements ≤ 300 distributed throughout the ICU stay of a patient. Generally accepted approach is to use the first time point, when PaO_2/FiO_2 drops below 300 mmHg, as an ARDS onset [47, 78]. Nevertheless, such approach could lead to completely wrong ARDS onset labeling because of two reasons. First, there could be multiple time points with the Horowitz index ≤ 300 during the ICU stay of a patient. Second, single Horowitz index measurements can be erroneous due to the reasons described in Section 3.3. Therefore, for instance in the ASIC study [5], the definition of ARDS event was defined by group consensus of the medical professionals

Dataset	**Number of labeled ARDS patients, n (%)**	**Number of labeled ARDS patients with defined ARDS onset, n (%)**
Hosp A	999 (100)	980 (98)
Hosp B	191 (100)	97 (51)
Hosp C	162 (100)	160 (99)
Hosp D	771 (100)	732 (95)
Hosp E	455 (100)	452 (99)
Hosp G	269 (100)	234 (87)
Hosp F	566 (100)	479 (85)
MIMIC	787 (100)	571 (73)

Table 3.4: Numbers of ARDS patients in underlying datasets according to ICD labeling and the ratio of ICD labeled ARDS patients with ARDS onset, which was identified based on the criteria described in Section 3.4.

experienced in the treatment of ARDS patients in the following way: PaO_2/FiO_2 had to stay below 300 mmHg for a period of 2h or in two consecutive arterial blood gas analyses, if the time interval between them was longer than 2h. In our study we further increased the time window when the Horowitz index had to stay below 300 to 24h. By that we tried to minimize the impact of unreliable Horowitz index measurements and to ensure the presence of prolonged oxygenation impairment. This definition was approved by experienced ICU physician.

Numbers of labeled ARDS patients in underlying datasets and ratio of patients, for whom ARDS onset was identified according to criteria mentioned above are given in Table 3.4. In some datasets our approach could not reveal onset time for all patients. Especially this issue is visible in the case of Hosp B, where onset could be identified for only 51% of patients. However, there was a specific reason for that, namely measurements of PaO_2 and subsequently Horowitz index were missing for 48% of the patients of Hosp B. Therefore, it was not possible to retrospectively identify ARDS onset. Similar issue to a lesser extent was present for datasets Hosp G, Hosp F and MIMIC. Additional factor, that could cause relatively low ratio of patients with identified ARDS onset in MIMIC data, could be our ARDS labeling strategy for MIMIC data, as it does not guarantee, that labeled patients were true ARDS patients. However, for majority of German datasets our strategy for onset definition worked relatively well and onset times were identified for more than 90% of ARDS patients.

Chapter 4

Convex hull analysis for generalization assessment

In this chapter, we introduce a framework for the comparison of populations and assessment of an ML model's generalization ability. First, we apply CH analysis to find CH coverage values between datasets. Second, we train 4 types of ML models to classify the origin of a dataset. This is done to assess whether it is possible to distinguish between patients from different hospitals. The performance of ML models is evaluated to determine which hospital's datasets differ the most in terms of underlying data distributions. We apply our framework to 4 critical-care patient datasets of different origins: three datasets from German hospitals generated within the SMITH project (Hosp A, Hosp B, and Hosp C) and the MIMIC dataset.

First, the framework is applied to every pair of hospitals to find mean CH coverages and performances of ML models for classification for a data source. Second, we investigate the applicability of the developed framework using the example of ARDS. We show that drops in the performance of models developed for the classification of ARDS on the first day in the ICU are attributed to the poor intersection of CHs and to the large differences in underlying data distributions of corresponding hospitals.

Section 4.1 describes fundamentals of the CH analysis and potential limitations of this approach. In Section 4.2, materials and methods for the study are introduced including information on underlying data, CH coverage estimation procedure, and ML method for dataset comparison. Section 4.3 presents the results, followed by a discussion in Section 4.4.

4.1 Introduction to convex hull analysis

Data-driven models, such as ML methods, aim to represent systems solely from available measurement data. Hence, a critical conceptual issue of such models is their limited performance in case of extrapolation into data regions sparsely covered by the data samples used for learning the model. These models handle test data better if they come from the same dataset used for training and generalize worse on the data obtained from other sources [81, 82, 83]. Model performance drops if data used to train and test a model come from different distributions. This difference is referred to as a domain shift [83]. Unless strong assumptions are posed on the learned function, data-driven models can only be valid in regions where they have sufficiently dense coverage of training data points, which is referred to as the validity domain [100]. This can be approximated by the CH spanned by the data, which represents an upper bound of the validity domain for any ML application.

To introduce the definition of the CH, we firstly introduce the definition of a convex set. From a geometric perspective, a set $\boldsymbol{P}$ in Euclidean space $\boldsymbol{P} \subseteq \boldsymbol{R^d}$ is s convex set, if for each pair of $\boldsymbol{x_i}, \boldsymbol{x_j} \in \boldsymbol{P}$ the line segment between $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ is contained in $\boldsymbol{P}$:

$$\forall \boldsymbol{x_i}, \boldsymbol{x_j} \in \boldsymbol{P}, \forall \lambda \in [0,1] : \lambda \boldsymbol{x_i} + (1-\lambda)\boldsymbol{x_j} \in \boldsymbol{P} \tag{4.1}$$

Given a set $\boldsymbol{S}$ in a d-dimensional space $\boldsymbol{R^d}$, the CH of $\boldsymbol{S}$ is denoted as $CH(\boldsymbol{S})$ and is defined as the smallest convex set containing $\boldsymbol{S}$. For a set $\boldsymbol{S}$ with a finite number of elements n: $\{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}\}$, where $\boldsymbol{x_i} = (x_{i1}, ..., x_{id})^T$, the $CH(\boldsymbol{S})$ is given by:

$$CH(\boldsymbol{S}) = \left\{ \sum_{i=1}^{n} \alpha_i \boldsymbol{x_i} | \boldsymbol{x_i} \in \boldsymbol{S}, \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i = 1, i = 1, 2, ..., n \right\} \tag{4.2}$$

In other words, the CH of a set $\boldsymbol{S}$ is the set of all linear combinations of elements of $\boldsymbol{S}$ in which the coefficients of elements of $\boldsymbol{S}$ are nonnegative and sum to 1. Such constrained linear combinations are known as convex combinations [101, 102, 103].

CH analysis was applied and used in various areas of research. Features derived from CHs of underlying data were used to describe cellular processes [104] and evaluate similarity of protein families [105, 106]. In another study, CH approach was used to describe properties of a cohort of 103 patients with total hip replacement and distributions within the cohort. Extreme points of the CH were analyzed and considered to represent cases where a failure

is more likely to occur [107]. Moreover, the CH approach is a popular spatial metric for measuring variation. In the study by Newsome et al. this method was used to discover differences in variation among different species of sea otters [108].

Moreover, CH analysis provides a way to estimate the ability of a model to generalize. If one considers the CH of the points used in a training set generalization ability of the model tends to fail with the increase in the distance of a new point to the CH of the training set [101]. Therefore, the coverage of the CH of a test set by the CH of a training set represents an upper bound for the generalization ability of any ML-based model. In the case of learning from different populations, the mutual coverage of the CHs can serve as a measure for sufficient similarity of heterogeneous populations enabling the first estimate for the reliability of generalization of ML models. Hence, one possible approach to examine different populations for homogeneity concerning the predictive performance of ML models is to perform a CH analysis of the available data to be used for training and prediction, respectively [109, 101].

However, even if the CHs of training and test sets intersect to a large extent, there might be differences in the underlying distributions of some variables. For instance, when data of one dataset lie in a region which shows a low density of samples in the other dataset. An extreme example is a dataset consisting of two clusters of data apart from each other; the CH envelopes all dataset points, including the space between them. If the majority of samples of the second dataset fall inside the gap area between the two clusters, the generalization capacity of a model will be impaired, as there is not enough training data in that region. Although the intersection values are high, in this case, it does not allow us to judge the generalization ability of the trained model. Therefore, the CH analysis provides necessary, but not sufficient conditions for a proper generalization of ML models.

Consequently, a second step in the analysis is needed to investigate datasets for diverging underlying distributions. If there are no such differences, two datasets form a homogeneous population and are indistinguishable, otherwise, it would be possible to differentiate the datasets. Therefore, if ML classifiers can identify the origin of a drawn sample with high accuracy, we postulate that there are diverging underlying distributions of variables forming different areas with a high density of samples in two datasets. Thus, training an ML model in one dataset and applying it in the other one would mean interpolation into areas sparsely covered by training data and could impair the generalization of respective models. However,

ML methods do not provide the direction of impaired generalization (i.e., model trained on one dataset and applied in the other one and vice versa).

In contrast, CH analysis provides a model-agnostic a priori data assessment and more importantly direction of impaired generalization. The CH of one dataset may completely cover the CH of the other dataset, meaning no restrictions for generalization from the CH perspective. However, in the opposite case (the CH of the second dataset covering the first one) the CH coverage may be modest suggesting generalization issues once models developed on the second dataset will be applied to the first dataset. Furthermore, the CH analysis proposed in our framework is computationally inexpensive and is an order of magnitude faster than ML methods. Therefore, we suggest an application of the CH method for universal generalization assessment supported by the application of ML methods to reveal the scope of differences in underlying distributions. Combining the results of these 2 methods, one receives a complete vision of potential generalization issues.

4.2 Materials and methods

4.2.1 Data

In this chapter retrospective ICU data from three German datasets (datasets Hosp A, Hosp B, and Hosp C) were used. In addition, MIMIC was used as an independent dataset with different geographical origin. Initial number of patients in corresponding hospitals is given in Table 3.1. Each patient's data included routinely charted ICU variables collected over the whole ICU stay. The full list of variables is given in Appendix A.1. Data from all 4 sites were brought to the same units of measurement, checked for consistency and preprocessed according to the rules given in Section 3.3. Due to data delivery delays only calibration datasets could be utilized for the CH analysis. Additionally, filtering based on thresholds, which was explained in Section 3.3 was implemented after the development and publication of the CH analysis framework. Therefore unreliable measurement could potentially influence the results of this section. However, an outlier filtering algorithm was implemented, which was applied before every application of the CH analysis.

Data for further analysis were prepared in the following way: first, the median values of routinely charted ICU variables collected over the first day of ICU stay were extracted as features for the analysis. Features with values missing in more than 30% of patients were

Dataset	Number of patients, n (%)	Non-ARDS, n (%)	ARDS, n (%)
Hosp A	10,110 (100)	9,471 (94)	639 (6)
Hosp B	1,209 (100)	1,123 (93)	86 (7)
Hosp C	1,012 (100)	924 (91)	88 (9)
MIMIC	4,792 (100)	4,555 (95)	237 (5)

Table 4.1: Final number of patients in the datasets and number of day 1 non-ARDS/ARDS patients in datasets.

omitted. We considered features, that were present in all 4 hospitals after the data feature omission step. The final list of features (21 features overall) used in the analysis can be found in Appendix A.4. Missing values of features were filled with the hospital-wide median value for that feature.

4.2.2 Use case example: classification for ARDS on the first day of treatment in ICU

To demonstrate the applicability of the developed framework, we considered the following typical use case of the application of ML models in healthcare: classification for a critical condition based on the first-day data. We used the presence of ARDS on the first day in the ICU as an endpoint for classification. The criteria for retrospective ARDS onset identification in both German hospitals and MIMIC data were explained in Section 3.4. To be able to assess the ARDS criteria, only patients having parameters of MV [positive end-expiratory pressure (PEEP), a fraction of inspiratory oxygen (FiO_2)] and blood gas analysis measurements [partial pressure of oxygen (PaO_2)] during the first 24 h were selected. Final number of patients used in the analysis in corresponding datasets is given in Table 4.1.

To ensure that information on the ARDS/non-ARDS status of patients is present in the data, only first-day ARDS patients were chosen as a case group. The control group comprised all non-ARDS patients and patients with ARDS onset later than on the first day. A total number of day1-ARDS/non-ARDS patients in corresponding hospitals is given in Table 4.1.

In this use case, we evaluated how a ML model trained in one hospital behaves in terms of performance if it is applied in another hospital. A Random Forest Classifier was trained in each of the four hospitals separately to classify ARDS and non-ARDS patients and tested in the other unseen hospitals. Performance in all datasets was assessed with area under

receiver operating characteristic curve (ROC AUC).

4.2.3 Convex hull coverage estimation

CH coverage for a new dataset was defined as the ratio of data points of a new dataset that lie inside of the CH of the initial dataset in the pair. An example of CH intersections for hospitals (Hosp B, Hosp C) and for the pair of features, arterial oxygen saturation (SaO_2) and arterial bicarbonate, is shown in Figure 4-1. It should be noted that CH coverage is not a symmetric measure, i.e., CH coverage of Hosp A by Hosp B can differ from CH coverage of Hosp B by Hosp A. CH coverage for each feature combination was assessed in 2 dimensions, i.e., for each combination of pair of features the coverage of CH of one hospital was calculated for all other hospitals. For instance, if hospitals Hosp A and Hosp B were considered, for each pair of features, CH coverage of Hosp A by Hosp B and CH coverage of Hosp B by Hosp A were calculated. CH coverages were assessed using bootstrapping of underlying data (100 times).

Let denote as $\boldsymbol{Hosp\ A}$ a dataset of Hosp A, which contains n_A elements $\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_{n_A}}$, where $\boldsymbol{x_k} = (x_{ki}, x_{kj})^T$ is a two-dimensional vector in a feature space (i, j). Analogously, $\boldsymbol{Hosp\ B}$ is a dataset of Hosp B with n_B elements $\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_{n_B}}$, where $\boldsymbol{x_l} = (x_{li}, x_{lj})^T$ is a two-dimensional vector in a feature space (i, j). Let denote as $CH^{i,j}(\boldsymbol{Hosp\ B})$ a CH of the dataset $\boldsymbol{Hosp\ B}$ in a feature space (i, j). Then we introduce the equation for a CH coverage for Hosp A by Hosp B in a two-dimensional feature space (i, j), where i and j denote feature indices:

$$CH^{i,j}_{cov}(Hosp\ A, Hosp\ B) = \frac{\sum_{\boldsymbol{x_k} \in \boldsymbol{Hosp\ A}} 1[\boldsymbol{x_k} \in CH^{i,j}(\boldsymbol{Hosp\ B})]}{\sum_{\boldsymbol{x_k} \in \boldsymbol{Hosp\ A}} 1}. \tag{4.3}$$

In higher dimensions intersections of CHs identified from datasets of sizes, which are usually available in single hospitals, tend to shrink even for datasets drawn from the same distribution due to the curse of dimensionality. Hence, we tested overlapping data by means of the overlaps of projections onto subspaces spanned by all combinations of 2 features. In case of overlapping CHs, the CHs of all projections will overlap as well. The opposite holds only in the case of homogeneous data distributions within the box in full data space spanned by the intersection of all projections. We assume that this is the case for real-world data available in healthcare and our approach delivers an acceptable approximation for the

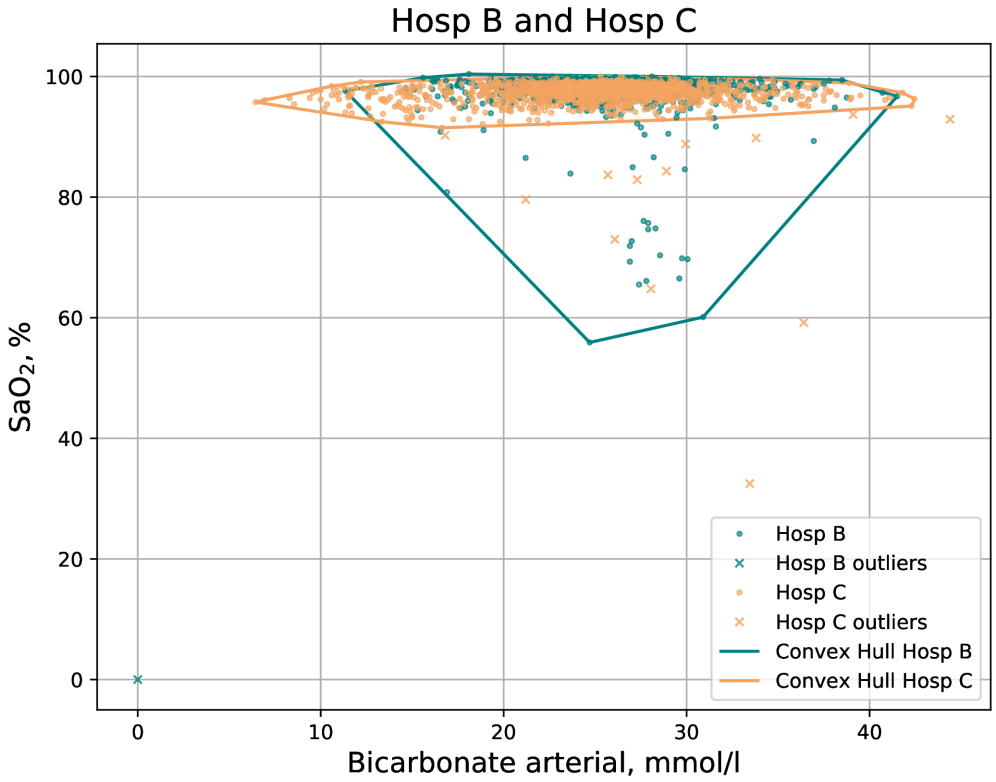


Figure 4-1: Example of CH intersection for the pair of hospitals (Hosp A, Hosp B) and the pair of features: SaO_2 and arterial bicarbonate. Overlaps of underlying data were inspected by means of the overlaps of projections onto subspaces spanned by all combinations of 2 features. Some data points are filtered out by the DBSCAN method prior to the construction of the CH. Outliers were identified in each 2D subspace before the calculation of CH overlaps.

estimation of translational predictivity for practical use.

CH coverage with respect to a single feature i was calculated as the median CH coverage value of all feature pairs that contain this feature:

$$CH^{i}_{cov}(Hosp\ A,Hosp\ B) = med(CH^{i,1}_{cov}(Hosp\ A, Hosp\ B), \\ CH^{i,2}_{cov}(Hosp\ A, Hosp\ B), ..., CH^{i,n}_{cov}(Hosp\ A, Hosp\ B)). \tag{4.4}$$

Next, the distribution of CH coverages for all features was computed. Finally, mean CH coverages for each pair of hospitals were calculated as the mean CH coverage among all features:

$$CH_{cov}(Hosp\ A, Hosp\ B) = \frac{\sum_{i \in n} CH^{i}_{cov}(Hosp\ A, Hosp\ B)}{n}, \tag{4.5}$$

where n is the number of features. Additionally, we specified the value of the first quartile minus 1.5 $\times$ interquartile range of the distribution as a threshold for low-coverage features. A low-coverage feature was defined as a feature with a CH coverage value that lies below the threshold. Such features were identified for each pair of datasets.

To eliminate the influence of outliers on the CH analysis, a density-based data clustering algorithm DBSCAN[1] was applied to the data. Before each run of the CH algorithm, outliers were removed using the DBSCAN method. The DBSCAN algorithm allows finding the densest areas of points that are considered to form clusters, whereas points, which cannot be assigned to any cluster are labeled as outliers. There are two core parameters for the DBSCAN algorithm that influence the resulting number of clusters and outliers. The first one - ϵ, determines the considered neighborhood of a point. The second parameter, minimal samples, defines the minimal number of points within an ϵ-neighborhood of a point to form a cluster. In our analysis, two DBSCAN parameters are tuned so that only one resulting cluster is allowed (which corresponds to the core dataset of the underlying hospital) and the number of outliers does not exceed 10%. Algorithm for outlier detection and filtering using the DBSCAN clustering algorithm was developed by Kateryna Nikulina in frames of the bachelor thesis [7].

4.2.4 Machine learning method for classification of a dataset

The classification task was defined to distinguish patients between two hospitals. Four classifiers, namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and AdaBoost (ADA) were used. Since the target label (hospital source identifier) was imbalanced due to different number of patients in underlying datasets, the "class weight" hyperparameter for LR, RF, and SVM was set to the "balanced" option. The prepared dataset was split into the train/validation (80%) and test (20%) sets. An optimal set of model hyperparameters was found using grid search with stratified 5-fold cross-validation on the train/validation set. A ROC AUC score was used to evaluate the performance of the chosen model. Predictions on the test set were evaluated with ROC AUC, precision, recall,

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html`

and F1 score metrics.

ML methods were trained two times. First, all features were used to train ML models. Second, features with low CH coverage were omitted from the analysis and ML models were retrained. This allowed to judge, whether the discriminating ability of ML models was predominantly caused by different CHs of underlying data or by differences in underlying data distributions of corresponding hospitals.

4.2.5 Python 3 modules used in this study and system requirements

In this study, the SciPy [110] Python 3 [111] spatial library with the Quickhull algorithm and the Delaunay class was used for CH analysis. The Scikit-learn [112] implementations of ML classification methods were svm.SVC, linear_model.LogisticRegression, ensemble.RandomForestClassifier and ensemble.AdaBoostClassifier. CH and ML analysis was performed on the computational cluster of the RWTH Aachen University using 1 node with 40 cores, 2.66 GHz, 4 GB RAM. The longest runtime for the CH analysis was 16 min. The runtime for the ML script comprised 24 h. Analysis was tested as well on the 2018 quadcore laptop i7-8565U CPU @ 1.80 GHz × 8.

CH and ML methods used in this study were developed by Kateryna Nikulina and Konstantin Sharafutdinov. They are available as a python package "chgen". Example scripts on how to use this package are available in the repository of the JRC Combine RWTH Aachen[2].

4.3 Results

4.3.1 Application of CH analysis to each pair of datasets

Figure 4-2 shows the mean CH coverage for each pair of hospitals. For each German hospital, minimum coverage was found when data of the corresponding hospital were covering the MIMIC dataset (last column in Figure 4-2). However, that was not the case for the opposite situation. Maximum mean CH coverage was found for cases when MIMIC data covers data from German hospitals (last row in Figure 4-2).

Features with low CH coverage values were identified for each pair of hospitals. These features are shown in Table 4.2. The table is not symmetric since features with low coverage

[2] `https://git.rwth-aachen.de/jrc-combine/chgen`

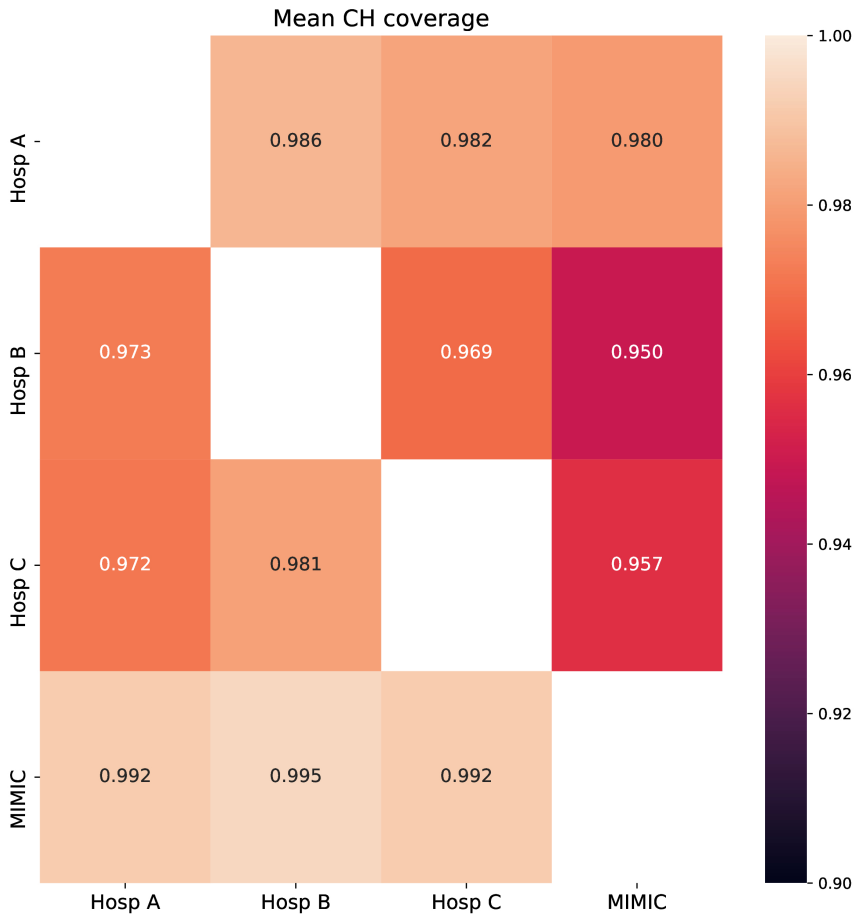


Figure 4-2: CH analysis results for data from four hospitals. Mean CH coverage over all features is shown. Rows - initial population, columns - population which CH is covered by the CH of the initial population.

values when the first hospital's data cloud is covering the second one may be different from features in the opposite coverage situation. Results of the mean CH coverage are accompanied by the number of features with low CH coverage in each case of the datasets' comparison. For each German hospital, a maximum number of such features was found when data of German hospitals were covering the MIMIC dataset (3 or 2 features correspondingly, last column in Table 4.2). Distributions of values for the features found in the case when data of German hospitals were covering the MIMIC dataset for all four datasets are shown in Figure 4-3. CH coverages for all features in the case of MIMIC coverage are given in Appendix A.5.

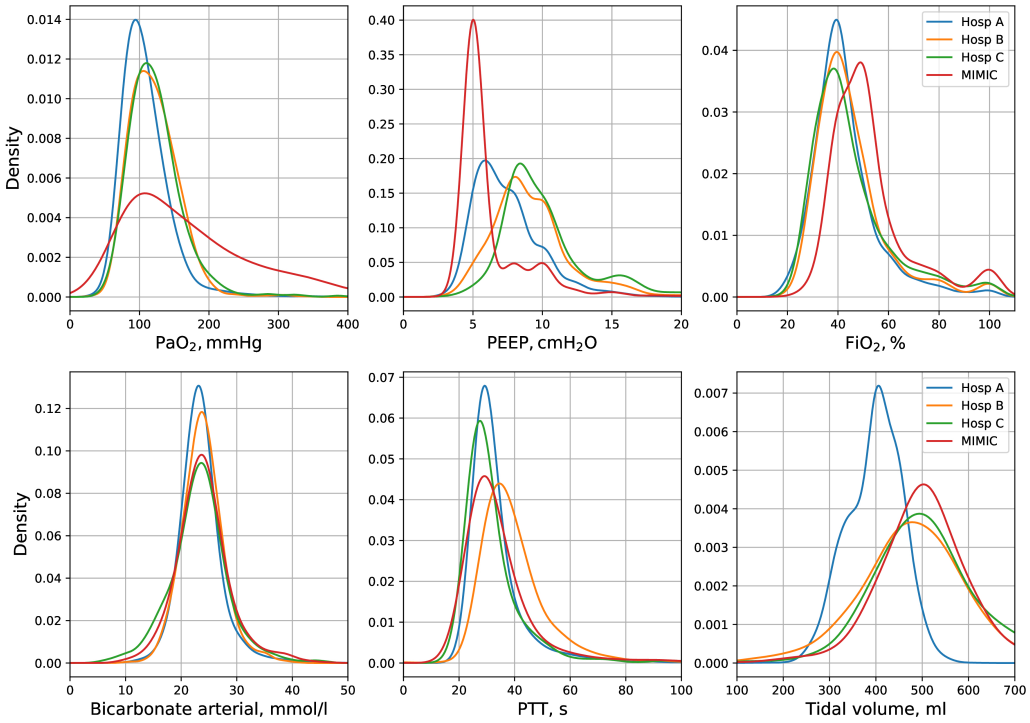


Figure 4-3: Distributions of values for the features with low CH coverage found in the case when data of German hospitals were covering the MIMIC dataset.

4.3.2 Application of ML routines for classification of the hospital

Results of the application of ML routines to classify the hospital for every pair of hospitals are shown in Figure 4-4 (A). Results of the ADA method are shown, as it gained the highest performance in terms of ROC AUC in all cases. In each pair of hospitals, the hospital where the patient samples were derived from could be almost perfectly classified (ROC AUC ≥ 0.94). The best separation was obtained between the MIMIC cohort and German hospitals. German hospitals looked more alike to classifiers. The worst separation was observed between Hosp B and Hosp C.

After the exclusion of the features with low CH coverage values, and retraining with the best-performing ML classifiers, the largest ROC AUCs were still observed between the MIMIC cohort and German hospitals (see Figure 4-4 (B)).

	Hosp A	**Hosp B**	**Hosp C**	**MIMIC**
Hosp A	-	Tidal volume, PEEP	Tidal volume	PaO_2, Tidal volume, PTT
Hosp B	PaO_2, Respiratory rate	-	Bicarbonate arterial, Respiratory rate, PTT	PaO2, Bicarbonate arterial, PTT
Hosp C	FiO_2, PEEP	-	-	PaO2, PEEP
MIMIC	FiO_2, Lactate arterial	Lactate arterial	Lactate arterial	-

Table 4.2: Lists of variables with low CH intersections for all pairs of hospitals. Rows - initial population, columns - population, CH of which is covered by the CH of the initial population.

4.3.3 Use case example: classification for ARDS on the first day of treatment in ICU

The results of the classification task are shown in Figure 4-5. Diagonal cells represent the performance of a specialized model which was trained and tested in the same hospital. The performance of specialized models strongly differed among hospitals under consideration, with the lowest ROC AUC of 0.79 in MIMIC and the highest of 0.94 in Hosp B. To test the generalization ability of developed models, they were tested on other unseen datasets, i.e., other hospitals (non-diagonal cells).

If the population of the new hospital is similar to or more homogeneous than the one of the original hospitals concerning the condition under consideration, the performance of the model will stay on a similar level or can be even higher than in the original hospital. However, if the population differs from the original one, performance will be impaired. For each specialized model trained in German hospitals the largest drop in performance was observed when the respective model was applied in the MIMIC dataset with the strongest drop of 0.26 for a model trained in Hosp B. Overall, models developed in Germany, showed impaired performance compared to the specialized MIMIC model. The opposite was not the case, as the MIMIC model showed similar performance in German hospitals to the performance in the original cohort.

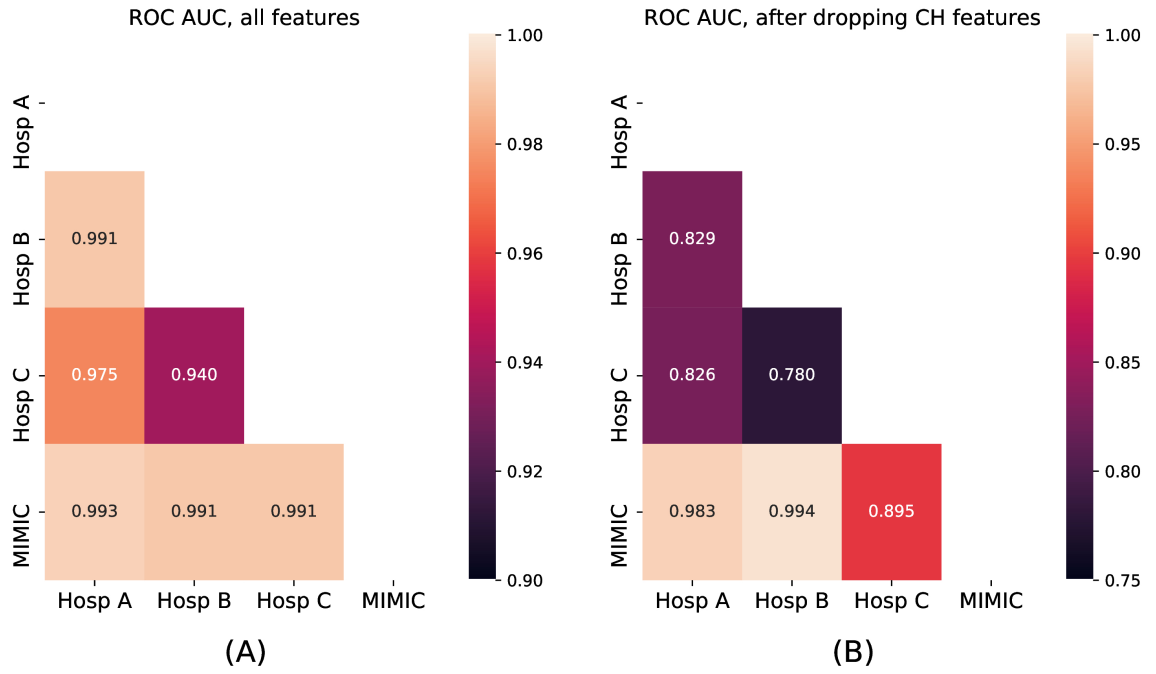


Figure 4-4: ROC AUC for classification for a hospital. (A): Performance of an ML learning algorithm for classification for a hospital. (B): Performance of an ML learning algorithm for classification for a hospital after removal of features with low CH coverage values. Numbers in cells reflect the ROC AUC of the classifier trained to separate between hospital 1 (row name) and hospital 2 (column name).

4.4 Discussion

"Internal" model performance on structurally similar, previously unseen data, gathered from the same source used for model training, can be contrasted with "external" model performance on new, previously unseen data from other sources. ML models perform worse in external cohorts due to several reasons such as different protocols, confounding variables, or heterogeneous populations [25, 26, 27]. Moreover, medical data can be biased by a variety of factors such as admission policies, hospital treatment protocols, country-specific guidelines, clinician discretion, healthcare economy, etc. Furthermore, labeling or coding criteria of a certain disease or syndrome and treatment guidelines evolve with time [113]. Since ML models for healthcare are predominantly developed on retrospective data, it remains unclear how the performance of such models is affected by the temporal separation of the target group even within one hospital.

Similarly, model reproducibility and model transportability have distinct objectives [89]. While reproducibility focuses on the performance of the model in the same target popu-

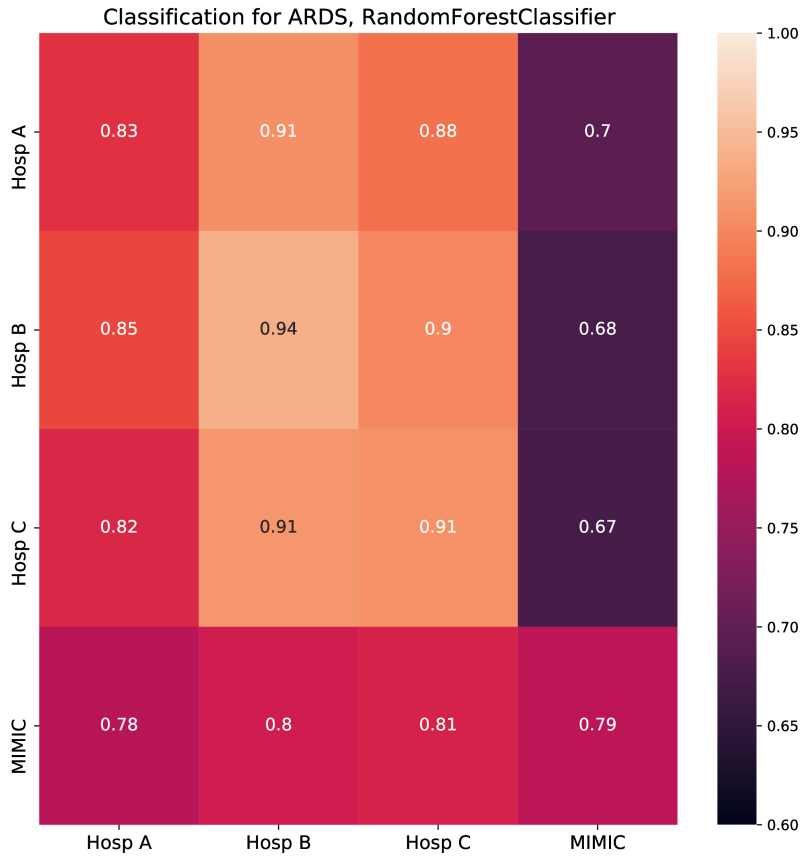


Figure 4-5: Random forest classifier classification results (cross-prediction matrix) for ARDS on the first day in ICU. RF trained in each of the four hospitals (row name) and applied in each of the four hospitals (column name). Diagonal cells represent the performance of specialized models which were trained and tested in the same hospital. Non-diagonal cells represent the performance of such models once they are applied in other hospitals and reflect ability of a model to generalize to the unseen population of another hospital. Twenty-one features common for all four hospitals were used to build corresponding RF models. Performance is depicted in terms of ROC AUC.

lation, transportability refers to performance in different but related source populations. Nevertheless, the closeness of this relationship between populations must be ascertained to achieve valid results of external validation. The performance will be poor in a sample that is too different from the data used for development. Conversely, a test sample that is too similar will overestimate the predictive performance showing reproducibility rather than transportability. To address these different aspects, an elaborate validation approach as described in the study by Debray et al. [99] seems necessary. They recommend the examination of the validation datasets in the first step to ensure adequate relatedness using a case-mix of a dataset and subsequent evaluation of the model with respect to the perceived

relatedness. The first step in their analysis is very similar to the ML step of our framework, as it evaluates whether it is possible to identify the dataset where a sample belongs to using the logistic regression model. Therefore it shares all the limitations of ML approaches for the evaluation of differences in distributions which are described below. The second step in the analysis pipeline of Debray et al. is also model-dependent, as it is based on the performance of a pretrained model in other datasets.

In this study, we have introduced another method for population comparison and assessment of a model's generalizability. First, it estimates the similarity of the underlying populations in terms of mean CH coverage. Second, it estimates the differences in datasets in terms of underlying data distributions. These two tasks are accomplished by the application of 2 methods–first the CH analysis and followed by the ML classifiers.

During the application of the framework on the datasets obtained from 4 hospitals, we found that there were significant differences in CH coverage among pairs of hospitals (see Figure 4-2). The lowest CH coverages for each of the German hospitals were observed when the MIMIC dataset was covered by data obtained from the corresponding hospital. However, in the opposite case i.e., Hosp A/Hosp B/Hosp C covered by MIMIC, the coverages were large. This shows that Hosp B/Hosp C and to a lesser extent Hosp A represented a part of the data space, spanned by data of MIMIC. In other words, data from German hospitals comprised, in greater or lesser proportions, parts of the MIMIC data cloud.

All four datasets exhibited differences in underlying data distributions. Once trained, ML classifiers were able to distinguish data coming from different sources with ROC AUC larger than 0.94, suggesting nearly perfect identification of the hospital from where the patient data originated from (see Figure 4-4). After the omission of features with low CH coverages, the performance of retrained models dropped. However, the performance of models distinguishing MIMIC from German hospitals was still largely supporting the finding, that the MIMIC dataset significantly differed from German hospitals.

To demonstrate that our framework can be used to assess the generalization ability of ML models, we considered a use case of classification for the first day of ARDS data. A specialized model was trained for each of the four hospitals' data. Then it was applied to unseen hospital data and the performance of the model on the original data was compared to those of the new data. We observed 2 clusters of datasets, namely German hospitals and MIMIC (see Figure 4-5). Models developed for German hospitals' data exhibited the

largest drop in performance once applied to MIMIC. That was not the case in the opposite situation, i.e., application of the MIMIC model to German hospitals data, where almost no drops were observed. CH analysis fully supported these findings. First, for each of the German hospitals, the lowest CH coverages were observed when the MIMIC dataset was covered by data from corresponding hospitals suggesting the impaired performance of models developed in German hospitals and applied in MIMIC. Second, mean CH coverages of German datasets by MIMIC data were found close to 1, suggesting full CH coverage and thus, the absence of limitations for generalization.

Moreover, smaller drops in performance were observed when models developed on data from Hosp B or Hosp C were applied to data from Hosp A. This is in line with corresponding CH coverages (Hosp A by Hosp B/Hosp C), which are in the medium range. Interestingly, when models, developed in Hosp A or MIMIC were applied in Hosp B or Hosp C we did not observe any drop in performance, but rather a slight increase. It could be the case if the population of the new hospital is similar to or more homogeneous than the one of the original hospitals concerning the condition under consideration. In our case, it would mean, that fewer non-ARDS patients with low Horowitz index are present in Hosp B/C compared to Hosp A/MIMIC. On the other hand, the necessary condition for the proper generalization, in this case, is satisfied by the fact, that CH coverages of Hosp B/C by Hosp A/MIMIC are among the largest in our study. Overall, the results of cross-prediction for ARDS were found to be in accordance with the results of the CH analysis of corresponding datasets.

Application of ML routines for classification for a hospital also supported the finding, which suggests that the MIMIC data significantly differed from German datasets, as the best separation with ROC AUCs > 0.99 was obtained between the MIMIC cohort and German hospitals. Nearly perfect separation was still possible after the exclusion of features with low CH coverage. This result indicated that the MIMIC cohort is not only less covered by German data, but exhibits diverging underlying data distributions once compared to German hospitals. However, while ML methods indicated, that there were significant differences in underlying distributions and performance of a model could be impaired, they did not point in the direction of proper or poor generalization, i.e., models trained in dataset A and applied in dataset B and vice versa. This constitutes an advantage of the CH method, as it is originally asymmetric and allows to assessment direction of impaired generalization. Moreover, the CH assessment is universal and does not depend on the particular ML classification method.

However, there could be multiple other reasons for such strong discrepancies in models' performance. First, some of the features with low CH coverage (PEEP, FiO_2, tidal volume) belong to parameters, which are set by physicians in the ICU, thus suggesting different treatment strategies in underlying hospitals. Distributions of values for these features also support this hypothesis. In Figure 4-3 there are clear differences in distributions of PEEP settings among all underlying hospitals. Moreover, tidal volume and FiO_2 settings are different in one of the 4 datasets, namely in Hosp A and MIMIC respectively. This clearly indicates diverging treatment guidelines among hospitals for patient populations of comparable severity, as all patients (except for MIMIC dataset) satisfied the same inclusion criteria of the ASIC use case. Moreover, distribution of PaO_2 values for MIMIC is skewed towards larger values, suggesting less severe state of MIMIC patients and subsequently different admission policies in the underlying hospital for MIMIC. Most probably, patients in much milder pulmonary states were admitted to the ICU at Beth Israel hospital.

Second, diverging ARDS labeling criteria (ICD-10 in Germany vs. ICD-9 in MIMIC) might contribute to label uncertainty in ARDS classification. Issues with proper ARDS labeling were described in Section 2.2 and represent a relevant issue on the way to translation of models developed on MIMIC data to healthcare setting, as MIMIC ARDS patients found with the rules described in Section 2.2 can represent a subpopulation of true ARDS patients and also include some false-positives.

Finally, the timespans of data collection overlap only partially. MIMIC data were collected between 2001 and 2012, Hosp A data between 2009 and 2019. Data from Hosp B and Hosp C were collected after 2012. This is relevant since in 2012 the American European Consensus Conference (AECC) definition of ARDS changed to the currently accepted Berlin definition [113]. This limitation might be crucial based on results of the previous study [88], which shows, that timespans of data collection play an outstanding role in model development and the relevance of clinical data decays with an effective "half-life" of about 4 months.

Nevertheless, the main observation is valid regardless of particular ARDS labeling: MIMIC data do significantly differ from all three other hospitals in this study. Given that this database is considered nearly a gold standard of open ICU databases, an external validation for models developed on this database is absolutely necessary. In the best case, a special pipeline for the assessment of the transferability of trained models should be included

in the data preparation step before a model development, so that generated models might exhibit significantly better performance.

Our study has other limitations that have to be considered. The geometric nature of the CH method implies that the CH method is sensitive to outliers in the data. One outlier point, that is located far from the main data cloud, will significantly increase the size of the CH leading to the overestimation of the CH, once it is calculated for the underlying dataset [114]. To eliminate the influence of noisy data on the CH analysis, a density-based data clustering algorithm DBSCAN [115] was applied to the data so that during each run of the CH algorithm, outliers were removed using the DBSCAN method. Additionally, to increase the robustness of the CH analysis results, each CH analysis execution was averaged over 100 runs with bootstrapped data.

Moreover, the geometric nature of the CH analysis implies an approximation of the validity domain as a hyperrectangular space spanned by extreme data points, which comprise the boundary of the corresponding CH. This represents a significant limitation of our CH-based method in case of application on non-convex data distributions, for instance those which can be thought of as having holes, separated clusters, or being banana-shaped [116]. Thus, to uncover non-convex validity domains the CH approach can be augmented by an application of one-class support vector machines (SVMs) for modeling complex validity domains [117, 118, 119]. For this purpose a so-called Support Vector Domain Description (SVDD) have been used to model validity domains of data-driven models [120]. In case of real-world data available in the ICU such non-convex validity domains can be potentially induced by parameters set by physicians. For instance, oxygen therapy with 100% oxygen in inhaled air is a usual practice in the ICU, see Figure 4-3. Then, 2-dimensional dataset which includes FiO_2 as a feature will potentially have a separate cluster with values of FiO_2=100% and the CH of this dataset will be artificially enlarged. Therefore, CH assessment provides an upper bound of the validity domain for ML applications and, thus, the necessary, but not sufficient condition for a proper generalization of ML models. Potentially, such non-convex structures in underlying data distributions can be revealed using the second step in the framework, namely the application of ML models.

Another potential weakness of our study design is that imputation was done using median values of the feature for underlying hospital without taking into account multidimensional feature distribution. However, this could not influence the results of the CH analysis, as this

Hospital	**CH coverage**
Hosp A	0.987
Hosp B	0.916
Hosp C	0.886
MIMIC	0.972

Table 4.3: CH coverage of the test set by the train set in the same hospital, where ML models were developed.

imputation did not influence the CHs of the underlying datasets. Moreover, the influence of this type of imputation on the results of cross-prediction of models for ARDS prediction was minimized by the fact, that we specifically have chosen patients with charted data of the main variables important for the identification of the ARDS state: PaO_2, FiO_2, and PEEP.

Another important question is how to define cutoff values between good and bad performance for both CH and ML analysis. We estimated CH coverages between train and test sets for the same hospital (see Table 4.3). These can be used as benchmarks for CH intersections for reasonable generalization. However, these also differed among hospitals, but here a clear correlation with the sample size of the cohort was observed. For instance, in Hosp C a test set of 202 patients was covered by a train set of 810 patients. It represents a real-world situation, as usually one deals with datasets, which represent a sample from original population. Such datasets are of smaller size, than original population. Therefore, a CH which is built based on such subsample represents an approximation of a CH of the true CH of the whole underlying dataset and depends on sample size. The size of the CH is sensitive to the size of a subsample [121, 7]. Therefore, the estimates for proper CH coverage should also depend on the sample size under consideration. For large datasets (Hosp A/MIMIC), where test set sizes were comparable to the sizes of smaller datasets in the study (Hosp C) they comprise 0.987/0.972. For ML routines, there is no rule of thumb to define minimum ROC AUC to judge whether hospitals cannot be distinguished. Usually, values of ROC AUC < 0.7 are considered to indicate poor discrimination performance.

Additionally, sample size can potentially be a factor, while considering CH intersections and machine learning results. However, there are some pieces of evidence, that this is at least not a dominant factor for generalization differences. First, models developed in small cohorts of Hosp B/Hosp C for ARDS classification deliver similar performance in Hosp A, as a specialized model of that hospital. Second, the model developed in Hosp A has a high generalization error in MIMIC (0.13), but a model developed in MIMIC shows the opposite

behavior in Hosp A having a small generalization error (0.01). Third, a model developed in the smallest Hosp C does not exhibit any generalization error in a dataset of completely different Hosp B. Therefore, we are of the strong opinion, that different sample sizes in underlying hospitals cannot explain such strong discrepancies in models' performance in different hospitals.

Finally, another important remark is that as the dimension of a dataset grows, then a trained ML model will almost always lie in the extrapolation range once applied to unseen data [122]. This is a consequence of the curse of dimensionality and has to be considered in all ML applications and especially in deep learning where models are dealing with hundreds or thousands of features. A theorem 4.6 states [123], that given a d-dimensional dataset $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}\}$ with i.i.d. samples uniformly drawn from an hyperball, the probability that a new sample $\boldsymbol{x}$ is in interpolation regime has the following asymptotic behavior:

$$\lim_{d\to\infty} p(\boldsymbol{x} \in CH(\boldsymbol{X})) = \begin{cases} 1 & \text{if } N > d^{-1}2^{d/2} \\ 0 & \text{if } N < d^{-1}2^{d/2}. \end{cases} \tag{4.6}$$

However, ML models that are intended for the real healthcare setting usually require a high degree of interpretability and therefore (mostly) contain a limited number of features [36]. In our study minimum N comprised 1012 patients in a dataset of Hosp C, therefore yielding the limiting dimension of the feature space to 30 according to the theorem 4.6. Thus, results of our study on interpolation and extrapolation regimes remain valid. Given that for the development of ML tools for the ICU larger datasets are usually used including tens of thousands of patients, our CH analysis approach provides a useful framework for evaluation of generalization ability of ML models with relatively small number of features.

Chapter 5

Novel ARDS virtual patient modeling framework for real-world ICU data

This chapter introduces a framework for individual ARDS virtual patient (VP) modeling for real-world ICU data. This framework is based on the original pulmonary model (Nottingham Physiology Simulator (NPS)), which was developed by Jonathan Hardman and Declan Bates [124]. In our study, we propose a novel way how virtual ARDS patients representing real ICU patients can be created in the matching procedure of the simulator to original patient data.

In Section 5.1, a historical perspective of the development of mechanistic virtual patient modeling approach starting from the foundation of the discipline called "Computational Physiology" through models of the Virtual Physiological Human project to the approaches of the virtual patient modeling is given. Section 5.2 introduces the main features of the Nottingham Physiology Simulator (NPS). In Section 5.3, our ARDS virtual patient modeling framework is presented, including our approach for ARDS modeling, sensitivity analysis, and optimization procedure. Finally, in Section 5.4, the limitations of the VP modeling approach are discussed with a focus on the limitations of the proposed VP modeling framework.

5.1 Introduction

5.1.1 History and foundations of virtual patient modeling

In 1993, the so-called "Physiome project" was launched by the International Union of Physiological Sciences (IUPS) with the main goal of adding the new discipline called "Computational Physiology" to the area of physiology. In this rapidly developing field of research, mathematicians, physicists, and bioengineers are collaborating with physiologists and molecular biologists to quantitatively link behavior among the various heterogeneous levels of organization in the human body: from genes and proteins to cells, tissues, organs, and whole interacting organ systems [125, 126, 38].

Living systems in many respects share similar characteristics with other complex systems from fields of physics and engineering: enormous complexity of interaction between respective entities, high redundancy, inhomogeneity, anisotropic and nonlinear behavior. In frames of the systems medicine, living, medically relevant systems are described, modelled and simulated using methods similar to those used for complex technical processes. However, a key distinctive feature of biological systems is their ability to adapt themselves in response to changing environmental conditions, which is determined by the complex interaction between the regulation of gene activity and the physical environment of the system, the details of which are often still not understood. The structure and function of such systems are thus coupled. Therefore, the main goal of computational physiology is to adequately describe these relationships in a computationally efficient manner and to develop models and software systems that account for this unique feature of the living organism. Studies of computational physiology have addressed this challenge by gradually linking organ physiology to tissues, tissues to cells, and ultimately to genomic and proteomic data [125].

To completely describe a living system in the computational physiology modeling framework, it is considered at three modeling levels:

- Biological - biological problem in terms of the biological domain
- Mathematical - mathematical formulation of the biological problem
- Computation representation languages - representations of models in a formal language that is machine interpretable and that represents computational abstractions of entities, mathematical relationships, and rules for their interpretation.

The path from biological formulation of a phenomenon through mathematical description to model implementation as software forms a general pipeline of computational physiology at each level of biological organization [125]. Different modeling approaches exhibit profiles of specific strengths and weaknesses; hence the choice of a particular technique should be based on the problem under consideration as well as data quality and availability [127, 128].

The Physiome project was intended to convey a quantitative description of the physiological dynamics and functional behavior of the intact organism. Therefore, the models of computational physiology initially targeted the general or "reference" human organism, which was described by a series of models at different levels of organization. However, later it became clear that phenotypic heterogeneity between individuals, especially when affected by pathology, required more specific models rather than general tools. Therefore, the focus shifted to models capable of describing a specific pathological condition, thus creating a "virtual" or "in silico" patient. In 2007, the Virtual Physiological Human Institute for Integrative Biomedical Research (VPH Institute) was created with the aim to complement the physiological multiscale modeling focus of the Physiome project with greater clinical relevance and industrial opportunities, including the development of patient-specific computational models for personalized medicine and the simulation of pathophysiological processes [126].

The three main goals of the VPH Institute were:

- Digital Patient: patient-specific modeling to support medical decision-making.
- In silico Clinical Trials: models to improve preclinical and clinical evaluation of new biomedical products; technologies to reduce, refine, and partially replace animal testing and clinical trials.
- Personalized Health Predictions: models based on data collected from mobile sensors, wearables and environmental sensors to advise individual patients.

VPH approaches have now been applied to many subsystems of the human body and to a wide range of domains [129, 130, 131]. Traditionally, however, computational physiology has focused on those organ systems which have been the most studied biophysically: the cardiovascular, neuromotor, musculoskeletal, and respiratory systems.

One of the oldest areas of research in computational physiology is modeling of the lung. Traditionally, physicians have used global indices and variables that reflect the state of

the lung, such as blood gas analysis values or volumes measured by spirometry. However, this represents a significant simplification. Real lungs exhibit inhomogeneous characteristics, such as structural asymmetries and regional variations in ventilation and perfusion, which are not captured by standard diagnostic methods. Computational modeling combined with state-of-the-art medical imaging provides a framework to predict lung function based on individual structure so that global measurements can be interpreted in terms of regional function. Computational physiology of the lung relies on several main pillars such as anatomically based models of the airway, vasculature, airway resistance, and tissue elasticity [132]. When combined with vascular models, they can provide a relationship between ventilation and perfusion throughout the lung [133]. When these models are complemented by a gas exchange model, the distribution of oxygen and carbon dioxide in the lung can also be predicted [134]. If this complex modeling framework is further augmented with computed tomography data from a patient, it becomes possible for physicians to understand the overall picture of gas exchange in the lung, but also, for example, the distribution of blood clots in pulmonary artery during pulmonary embolism. This approach can later be applied to relevant lung diseases such as asthma [134, 135] or COPD [134, 136, 137].

However, comprehensive physiological models that aim to represent physiological states as accurately as possible often require large amounts of data and computing power, which limits real-world applicability of such models. Therefore, for the utilization in the real-world healthcare setting, approaches of virtual patient (VP) modeling have the highest potential. VP models are physiological models which are complex enough to describe an essential pathophysiological mechanism for a condition under consideration, but at the same time not over-complicated such that for parameterization of VP models limited data and computational resources are required [38, 39, 138].

5.1.2 Virtual patient modeling for critical care medicine

Critical care medicine is a significant application area for innovations in computational medicine. Large amounts of data are routinely generated during the treatment of ICU patients, such as blood gas analysis (BGA), respirator settings, laboratory, and vital signs data. To process and store these huge amounts of data, a large number of hospitals are already equipped with electronic health record or patient data management systems (PDMS). Critical care datasets are therefore very large, but at the same time very heterogeneous.

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 often molecular core processes of disease progression, 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 [54, 59].

Therefore, data gathered in the ICU setting are surrogate markers representing a tremendous simplification of clinical reality. Subsequently, relevant medical signals about a patient state are often disturbed by noise or missing completely. For instance, the real 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 extract relevant patient information, approaches of VP modeling can be used. 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 rarely measured, such as cardiac output, anatomical shunt, or distribution of biophysical characteristics of alveolar compartments across the lungs. These parameters can be inferred 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 contain important information about the patient's health status, which cannot be directly extracted from routinely measured ICU data due to aforementioned reasons [39]. In the area of pulmonary research, the key feature of this approach resides in the ability to "look inside" the lung by investigating the relationship between treatment strategies, identified physiological parameters, and manifestations of a particular critical condition. By studying these relationships possible physiological mechanisms that drive particular diseases could be uncovered [139].

Another application of VP models in intensive care medicine manifests itself in the area of clinical trials. Prospective randomized clinical trials are mostly associated with specific difficulties in the ICU, e.g., identification of the right patient population or large number of potentially beneficial treatment strategies, especially with respect to MV settings [140, 139]. Considering VP's as digital twins of real ICU patients, they can assist in the design, planning

and execution of the study. The use of such individualized computer simulations is also called "in silico clinical study" [141]. The basic idea is to initialize a large number of digital twins of individual patients and model a specific disease/intervention. It has already been shown to be useful in the field of mechanical ventilation [142, 143, 139] and medical device optimization [144]. The advantages of in silico studies in this regard are the control of patient variability, as VP's can account for multiple dynamic conditions and model different external stimuli without affecting other patient characteristics.

There are numerous examples of VP models for the ICU that vary both in terms of the pathophysiological state under consideration and the complexity of the underlying model. These range from simple one-compartment models with few parameters [145] to comprehensive models with several hundred degrees of freedom [132]. In general, as complexity increases, more data are needed to fit real patients, either in the form of a longer time interval for fitting or additional parameters, sometimes rarely measured. An excellent example of a VP model that possesses the sufficient complexity to correctly capture multiple physiological states while using primarily routine variables is the Nottingham Physiology Simulator (NPS).

5.2 Nottingham Physiology Simulator as virtual patient model

5.2.1 Introduction to the model

The Nottingham Physiology Simulator (NPS) includes a comprehensive simulation model of the pulmonary system based on mechanistic models of ventilation and gas exchange [124], which was later extended to include cardiovascular components [146]. The simulator has already been validated in multiple studies using real patient data [124, 147, 148].

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, see Figure 5-1. Dynamics of the model components (transport of air from mouth to airway and alveoli, the gas exchange between alveoli and their corresponding capillaries, and the gas exchange between blood and peripheral tissue compartment) is discretized and implemented in MATLAB programming language [149] as a system of algebraic equations, obtained or approximated from the published literature, experimental data and clinical observations. These equations are solved in a series in a

naive iterative manner in a loop for the predefined desired number of iterations.

The simulator represents a dynamic cardiopulmonary state in vivo that is initialized with a list of input parameters. Some of these parameters are routinely measured in intensive care setting, such as certain blood test measurements or respirator settings. Others, however, are rarely measured, such as cardiac output or anatomical shunt. Finally, such parameters as lung properties across different alveolar compartments can be measured under experimental conditions or corresponding methods are lacking completely. They must be estimated in the matching procedure of the VP to real patient data, which is performed using optimization methods.

Once the simulator is initialized and the simulation starts, variables describing the state of a simulated patient (i.e., blood gas analysis variables) change as a response to the external stimulus by mechanical ventilation (implemented via parameters of respirator settings). If the model parameters do not change, after a number of iterations state variables equilibrate reaching a steady state of the system ventilator-patient. Equilibration time comprises 2 hours, i.e., patient state variables reach the steady state after 2 hours of simulation. Thus, the simulator can be used to simulate pathophysiological state and responses of the patient on the time scales up to two hours, given constant values of the model parameters

During the simulation, the model also calculates variables that are not directly measurable but may carry important information about the patient's health status, for example lung perfusion. This information can either be used directly by physicians or integrated into machine learning systems for critical condition prediction, classification tasks, or subcohort identification and characterization. Finally, VP also enables model-based imputation of missing or infrequently recorded variables, as they can be identified in the optimization procedure.

Model components have been extensively described elsewhere, see Supplementary file for the study by Das et al. [142]. The unique property of the NPS is that the lung is modeled using 100 alveolar compartments (see Figure 5-1), each of which may have different properties. Each alveolar compartment i has 6 unique and configurable biophysical characteristics:

- alveolar compliance k_i
- alveolar inlet resistance R_i
- vascular resistance VR_i of corresponding capillary unit

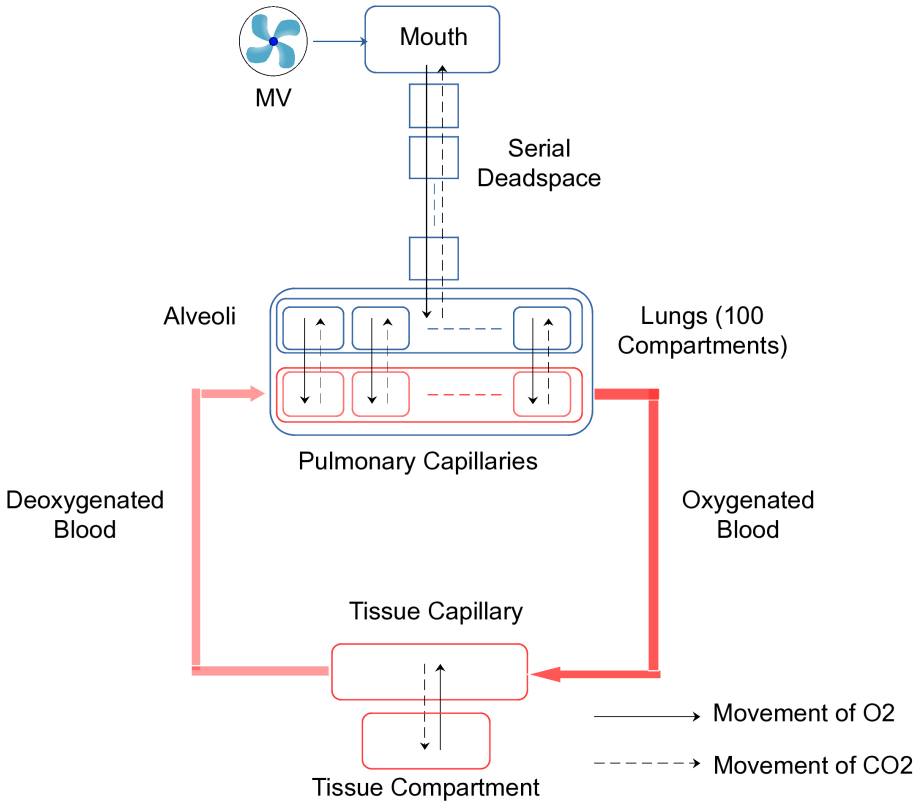


Figure 5-1: Structure of the physiological model implemented in the Nottingham Physiology Simulator. Model includes following components: mechanical ventilator (MV), serial deadspace, alveolar-capillary interface, arterial and venous blood, tissue compartment and tissue-capillary interface.

- extrinsic (interstitial) pressure $P_{ext,i}$
- threshold opening pressure TOP_i
- recruitment time τ_i.

It allows modeling of variable gas and blood flows in the lung. Thus, ventilation-perfusion mismatch can be modeled, allowing the simulation of conditions such as chronic obstructive pulmonary disease (COPD) [137], acute hypoxaemic respiratory failure in COVID-19 patients [150] and ARDS [43, 143, 139, 151, 152, 41].

Here we will summarize main components of the model and provide corresponding model equations for model components which are relevant for the further analysis, namely calcula-

tion of pressures of alveolar compartments, calculation of air flows and volumes of alveolar compartments, and cardiovascular calculations. This section is based on the description of the components of the NPS as they are given in the supplementary material of the study by Das et al. [142].

5.2.2 Calculation of pressures of alveolar compartments

First, given the volume of the lung and $Ncomp = 100$ alveolar compartments, the pressure in each of the alveolar compartments is calculated based on volume-pressure compliance curve. For each compartment i the alveolar pressure p_i (as the pressure above atmospheric in cmH_2O) for the given volume of alveolar compartment v_i in milliliters and given time point t_k is determined by:

$$p_i(t_k) = \begin{cases} S_i(v_i(t_k) - V_c)^2 - P_{ext,i} & \text{if } v_i(t_k) > 0 \\ 0 & \text{if } v_i(t_k) \leq 0. \end{cases} \tag{5.1}$$

where

$$S_i = k_i Ncomp^2/200000 \text{ and } V_c = 0.2V_{FRC}/Ncomp. \tag{5.2}$$

The alveolar compartments are arranged in parallel (see in Figure 5-1) and interact with the series deadspace with respect to the movement of gases. The flow of air into the alveolar compartments is achieved by a positive pressure provided by the ventilator (either P_{EI} or PEEP) and the air moves along the pressure gradient. The equation models the behavior of the intact lung / chest-wall complex. The use of the square of the difference between v_i and V_c causes alveolar pressure to increase at volumes below V_c, leading to exhalation and collapse of alveolar compartment.

$P_{ext,i}$ represents the effective net pressure generated by the sum of the effects of factors outside each alveolus that act to distend that alveolus. Positive components of $P_{ext,i}$ include the outward pull of the chest wall, and negative effects include the compressive effect of interstitial fluid in the alveolar wall. Decreasing $P_{ext,i}$ increases the alveolar pressure such that the pressure gradient forces the air out of the alveolar compartment until the compartment collapses. Therefore, large negative external pressure models a scenario where there is compression from outside the alveolus, for instance through interstitial edema, causing

collapse.

The parameter S_i is a scalar that determines the intra-alveolar pressure for a given volume (with respect to a constant collapsing volume V_c) and is dependent on the alveolar compliance k_i. The units of S_i are $cmH_2O \times ml^{-2}$ and it can be described as *stiffness of alveolar compartment*, as for the fixed volume v_i increase of S_i increases the corresponding alveolar pressure of the alveolar compartment. Thus, a larger pressure of mechanical ventilator is needed to drive air into the compartment. Therefore, the compartment will be behaving as a stiffer lung unit.

Finally, V_c is defined as a "constant collapsing volume" at which the alveolus tends to empty and represents a fundamental mechanical property of tissue and surfactant. V_{FRC} represents fractional residual capacity or the resting volume of the lung.

5.2.3 Calculation of gas flow and volumes of compartments

Once pressures in alveolar compartments are determined, the gas flow f_i to the compartment or out of the compartment is calculated based on pressure gradient:

$$f_i = \frac{P_{trachea} - p_i}{UB_{resist} + R_i}, \tag{5.3}$$

where $P_{trachea}$ is the tracheal pressure (equal to P_{EI} during inhalation and PEEP during exhalation), p_i is the pressure in the compartment, $UBresist$ is the upper airway resistance and R_i is the alveolar flow resistance. Further in the manuscript we will use $R_i \equiv R_i + UB_{resist}$ as effective flow resistance of the compartment.

f_i defines the volume of gas per time slice that should be moving in or out of the lung. A positive value implies that tracheal pressure is higher than alveolar pressure and hence due to the pressure gradient there is flow waiting to move into the alveolus (inflow). Negative f_i value indicates that the pressure gradient is reversed due to the decrease in the tracheal pressure, and thus the flows out of the different alveoli need to be added (outflow). The difference between inflow and outflow determines the volume of gas waiting to move up or down the anatomical deadspace. Once the gases have been transported, alveolar variables such as v_i and p_i are recalculated using the new alveolar volumes.

The pressure differential created by the mechanical ventilator drives the gas through the system. The series deadspace (SD) is located between the mouth and the alveolar

compartments and represents trachea, bronchi, and the bronchioles where no gas exchange occurs. Inhaled gases pass through the SD during inspiration and alveolar gases pass through the SD during expiration.

The new volume of gas $v_i(t_{k+1})$ in the alveolar compartment i is then given by:

$$v_i(t_{k+1}) = v_i(t_k) + f_i(t_k). \tag{5.4}$$

In the model, the inhaled air is assumed to consist of five gases: oxygen, nitrogen, carbon dioxide, water vapour and an additional gas used to model additives such as helium or other anaesthetic gases. Finally, volumes and partial pressures of each gas in each alveolar compartment are calculated.

Moreover, to be able to model reopening or recruitment of collapsed alveolar units threshold opening pressure TOP_i and recruitment time τ_i of alveolar compartments were introduced to the model.

5.2.4 Cardiovascular calculations

Each of the alveolar compartments i has corresponding pulmonary capillary with pulmonary vascular resistance VR_i, which determines the blood flow through this compartment (later on in the thesis VR will be referred to as vascular resistance of the compartment, measured in dynes/sec/cm^{-5}). 100 pulmonary capillaries can be modeled as 100 resistors based on the analogy with electrical circuits. Then, blood flow (cardiac output) is analogous to electric current and vascular resistance is analogous to resistance of a resistor in a circuit. To calculate the blood flow through each of the compartments, firstly we have to calculate the overall resistance of the system (pulmonary vascular resistance - PVR):

$$\frac{1}{PVR} = \frac{1}{R_1} + \frac{1}{R_2} + ... + \frac{1}{R_{100}}. \tag{5.5}$$

Blood flow Q through the lung comprises a part of the cardiac output CO, as some blood is bypassing the lung due to anatomical shunt:

$$Q = (1 - anatShunt)CO. \tag{5.6}$$

Then, blood flow Q_i through each compartment i is given by:

$$Q_i = \frac{PVR \times Q}{VR_i}. \tag{5.7}$$

Once blood flows through each of capillary units and volumes of each of alveolar compartments are determined, equilibration of the gaseous content of the alveoli with the blood flowing in the pulmonary capillaries occurs. Equilibrium is achieved by the iterative movement of the gases between a capillary and an alveolar unit until a set equilibrium point is reached. This process is influenced by haemoglobin content/saturation, pH level of blood, etc. These effects are incorporated into the model by the use of established blood gas laws. Thus, the actual behaviour of gas transfer across the pulmonary alveolar-capillary barrier is approximated as closely as possible. This process is repeated for all alveolar units every sampling interval.

5.3 The virtual patient modeling framework for ICU data

5.3.1 Creation of virtual ARDS patients

The main aim of our study was to model ARDS development. The earlier approach which was used by Das et al. to create virtual ARDS patients [43], however, did not suit our setting due to several reasons. The main limitation of the old approach was the requirement to identify the large number of model parameters, including properties of single alveolar compartments. For instance, in the original study model calibration to data required identification of more than 200 model parameters. They were identified in a global optimization procedure, which required large computational resources. Genetic algorithms (GA) were used for virtual patient matching in the original studies utilizing the NPS [43, 153]. GA belong to a family of stochastic global optimisation algorithms suitable for black-box optimization, i.e., algorithms that assume the objective and/or constraint functions are given by black-boxes (in this case by computer simulation). However, GA usually require larger numbers of function evaluations than gradient-based techniques [154]. To speed up the optimization process, a parallelized implementation of a genetic algorithm was employed in the original study. The matching process was performed under Matlab 2015a using the Global Optimization Toolbox and Parallel Computing Toolbox. Nevertheless, optimization procedure still required the use of the 'Minerva' high performance computing cluster provided

by the University of Warwick with 396 nodes (2×hexa-core 2.66 GHz 24 GB RAM). This limited applicability of such approaches to a small number of patient datasets, as creation of each virtual patient required extensive computational resources.

Moreover, for each of the patients hundreds of parameters were identified in the optimization procedure, whereas objective function includes only several measurements of limited number of variables (usually blood gas analysis measurements). It comprised the second limitation of such approach which layed in a concern about parameter identifiability , i.e., whether it is theoretically possible to estimate unique parameter values from data, given the quantities measured.

On contrary, our VP modeling framework was initially intended to be used on large databases with real-world ICU data including thousands of patients. Therefore, the initial approach for the creation of virtual ARDS patients, which was used in the earlier studies, was not suitable for our setting due to aforementioned reasons. Furthermore, initial modeling approach did not integrate any a priori information about pathophysiology of ARDS patients, i.e., no specific assumptions were made about the distributions of values of compartmental parameters during the ARDS state of a patient.

Thus, to fill this gap and to address the issue of impaired identifiability we tried to decrease the number of parameters for optimization to a smaller number by integrating a number of assumptions based on the a priori medical knowledge about the evolution of the ARDS state of an ICU patient. The initial aim of our VP modeling framework was to model the transition from non-ARDS state to the ARDS state and an early stage of the ARDS development (till 1 day after the ARDS onset). Information extracted from the VP modeling was intended to be later used in ML tools for the early ARDS detection. This comprised the main justification for a number of assumptions used in the approach for the reduction of the number of optimization parameters. The reduction of the number of optimization parameters was particularly needed for compartmental parameters, which were introduced in Section 5.2, as together they comprise 600 degrees of freedom which have to be identified for each of the patients.

Therefore, firstly fixed values were assigned for parameters which are responsible for alveolar recruitment and these parameters were excluded from the optimization procedure. Thus, threshold opening pressure TOP_i and recruitment time τ_i of all alveolar compartments were set to large values (40 cmH_2O and 10 sec correspondingly) modeling the situation,

when no recruitment is possible after alveolar collapse. There are three main approaches for alveolar recruitment, namely application of high PEEP, recruitment maneuvers, and prone positioning [155]. Events of the application of high enough PEEP levels for alveolar recruitment (>16 cmH_2O [156]) are rare in German hospitals (see Figure 4-3 for the distribution of applied PEEP values on the first day in the ICU) and PEEP is rather used to prevent the collapse and not to reopen already collapsed alveoli. Furthermore, recruitment maneuvers have not been proven to provide mortality benefit in ARDS patients. Therefore, despite relatively high reported application rates of recruitment maneuvers (the LUNG SAFE study reported the rate of 21% of all ARDS patients [68]), true application rates in hospitals under consideration remain unclear. Based on the experience of the physician consulting our research, recruitment maneuvers are rarely used in German hospitals. Moreover, our approach ignores potential application of prone positioning, which is applied in 8% to 14% of all ARDS patients [68, 157]. However, physicians are usually hesitating to apply it, as proning is very laborious and can also cause a number of complications. Therefore, usually there is a time gap between the drop in oxygenation (reflected by the Horowitz index) and application of prone positioning. All this partially justifies the assumption that collapsed alveoli are assumed to stay collapsed in our modeling framework. Nevertheless, these limitations have to be considered once the framework is applied. Finally, for the modeling of recruitment maneuvers these parameters are, however, essential [142].

Secondly, instead of the identification of stiffness of the alveolar compartment S_i and external pressure on the compartment $P_{ext,i}$ separately for each individual compartment, which would require the identification of 200 parameters, the ARDS was modeled through the introduction of closed alveolar compartments. Closed alveolar compartments model the formation of atelectases, which are an integral part of the pathological picture of ARDS. To account for the spatial distribution of atelectases and lung properties over the lung, we defined the index of compartment to represent the distance to the compartment from the lower part of the lung in the supine position of the patient. Thus, compartments with the indices close to 0 represent lower (or dorsal/gravity-dependent) parts of the lung, whereas compartments with the indices close to 100 represent upper (or ventral/nondependent) parts of the lung.

As described in Section 2.2, the spatial distribution of atelectases in the lung is controlled by two mechanisms. The first one are the infiltrates which are caused by the factors that

trigger the ARDS and which affect the whole lung, for instance sepsis, pancreatitis, or inhalation of toxic gases. In these cases atelectases are distributed uniformly throughout the lung. The second mechanism starts to be dominant in the later stages of the ARDS. 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 [158, 159, 160].

To model the formation of atelectases we introduced a single parameter controlling the number of closed alveolar compartments (n_{cc}). As our VP modeling framework was intended to model the early stage of the ARDS development, we decided to use the uniform distribution of atelectases over the lung and modeled it through closing of n_{cc} out of 100 compartments on random. Closed compartments were introduced through setting values of external pressure of n_{cc} random compartments to large negative value (-30 cmH_2O). Large negative external pressure models a scenario where there is compression from outside the alveolus, for instance through interstitial edema, causing collapse and complete alveolar shunt, see Figure 5-2. Additionally, stiffness of alveoli of closed compartments was set to large value (0.16 $cmH_2O \times ml^{-2}$), modeling stiffer lung units. Minimum value of n_{cc} was set to 0 and maximum value to 80.

The notion of collapsed alveolar compartments was already used in the work by Das et al. [142]. However, in that study the assessment of the number of closed compartments was performed after the identification of 3 model parameters ($P_{ext,i}$, S_i and TOP_i) for each of the alveolar compartments in the optimization process. Such approach again required identification of 300 parameters in the optimization procedure carrying all limitations which were mentioned above. Our approach allowed to model the building of atelectases by introduction of one parameter only, given the limitation of the early phase of the ARDS development.

Next, instead of fitting flow resistance R_i and vascular resistance VR_i values for each of the compartments separately, we introduced parameters defining distributions of R_i and VR_i across alveolar compartments. Vascular and flow resistance of compartments were each introduced through two parameters (sVR and $inVR$ staying for slope and intercept of VR and sR and inR staying for slope and intercept of R respectively) defining distribution of

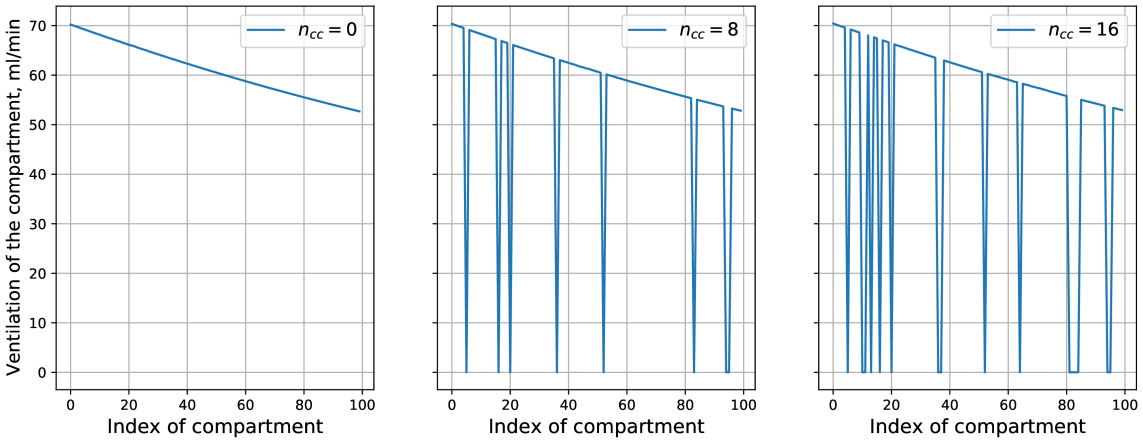


Figure 5-2: Distribution of the ventilation for the patient for different number of closed compartments (n_{cc}). Ventilation of an alveolar compartment is calculated as the air flow through the compartment in a minute. Ventilation of closed compartments is equal to 0 causing complete shunt.

VR and R values over 100 alveolar compartments based on the VR and R values of a healthy patient (see Figures 5-3 and 5-5):

$$\begin{aligned} VR_i &= VR_{healthy} + (i-49)sVR + inVR \\ R_{100-i} &= R_{healthy} + (i-49)sR + inR. \end{aligned} \tag{5.8}$$

When sVR and $inVR$ are equal to 0, vascular resistance of compartments VR_i resembles vascular resistance of the healthy patient, see Figure 5-3 and is equal to the default vascular resistance for the compartment (1.6×10^4 dynes/sec/cm^{-5}/min), as in the study by Das et al. [142]. On the other hand, maximum values of the sVR and $inVR$ were defined in the way, that final distribution of VR resembles that of the ARDS patient (patient A from Das et al. [142]). Therefore, values of sVR and $inVR$ within these limits can model all intermediate states between distribution of VR of the healthy patient and the ARDS patient.

Moreover, we examined, whether the final distribution of VR resembles the real distribution of perfusion in the lung of a patient in a supine position. In order to do that, we defined index of compartment to represent the distance to the compartment from the lower part of the lung, as described above. It is known, that lower parts of the lung are both perfused and ventilated better, than upper parts of the lung due to gravitational effects [161, 162]. From available literature we extracted relative ratios of ventilation and perfusion

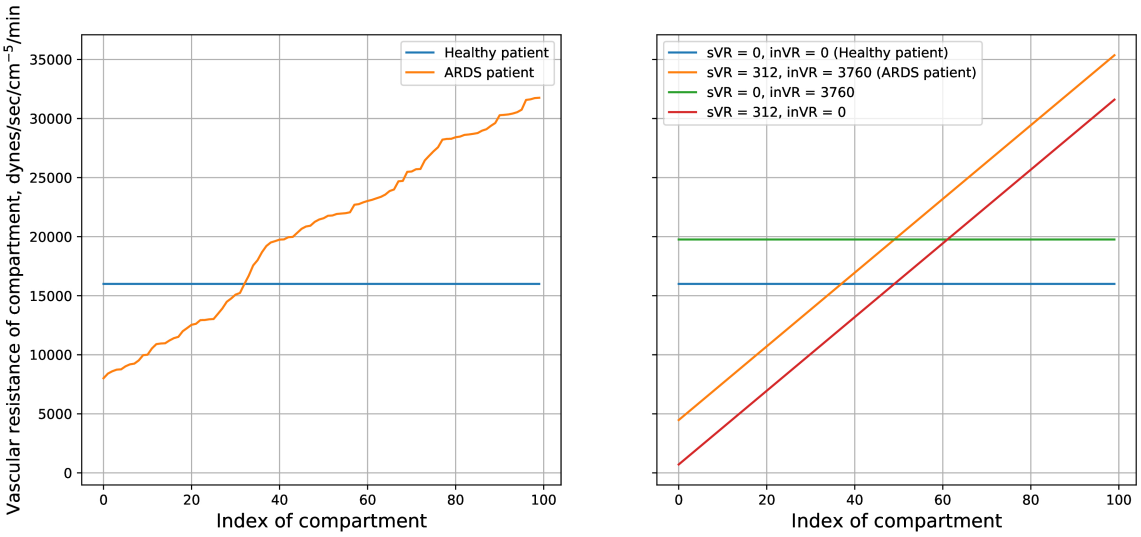


Figure 5-3: Left figure: distributions of VR in predefined healthy and ARDS patients. Right figure: distributions of VR for extreme values of sVR and $inVR$.

in the upper and lower parts of the lung and defined sVR and sR accordingly. For instance, ventilation in the lower part of the lung is on average 2-2.5 times larger than in the upper part of the lung, whereas perfusion in gravity-dependent parts is on average up to 3 times larger than in ventral parts [161, 162].

Using different values of sVR and $inVR$ and constant cardiac output volume, one can calculate distribution of the perfusion through the lung and verify that the difference in perfusion between upper and lower parts of the lung approximates that from the literature. For instance, in Figure 5-4 distribution of the perfusion for the patient with extreme values of $sVR = 312$ dynes/sec/cm^{-5}/min and $inVR = 3760$ dynes/sec/cm^{-5}/min and $CO = 8000$ ml is shown. The sum of perfusion of the lower 30 compartments is 4377 ml and the sum of perfusion of the upper 30 compartments is 1206 ml. The ratio is then 3.62, which is close to real physiological value of the patients (3).

Analogously, when sR and inR are equal to 0, flow resistance of compartments resembles flow resistance of compartments of the healthy patient, see Figure 5-5 and is equal to the default flow resistance for the compartment (5×10^{-3} kPa/ml/min). Upper limit for inR was determined as the intercept of the linear fitting of the distribution for R for the predefined ARDS patient. Lower limit for sR was determined in the way to model reasonable physiological difference in air flow through lower and upper parts of the lung, as described above. Finally, distribution for R was inverted in order to provide a physiological picture of

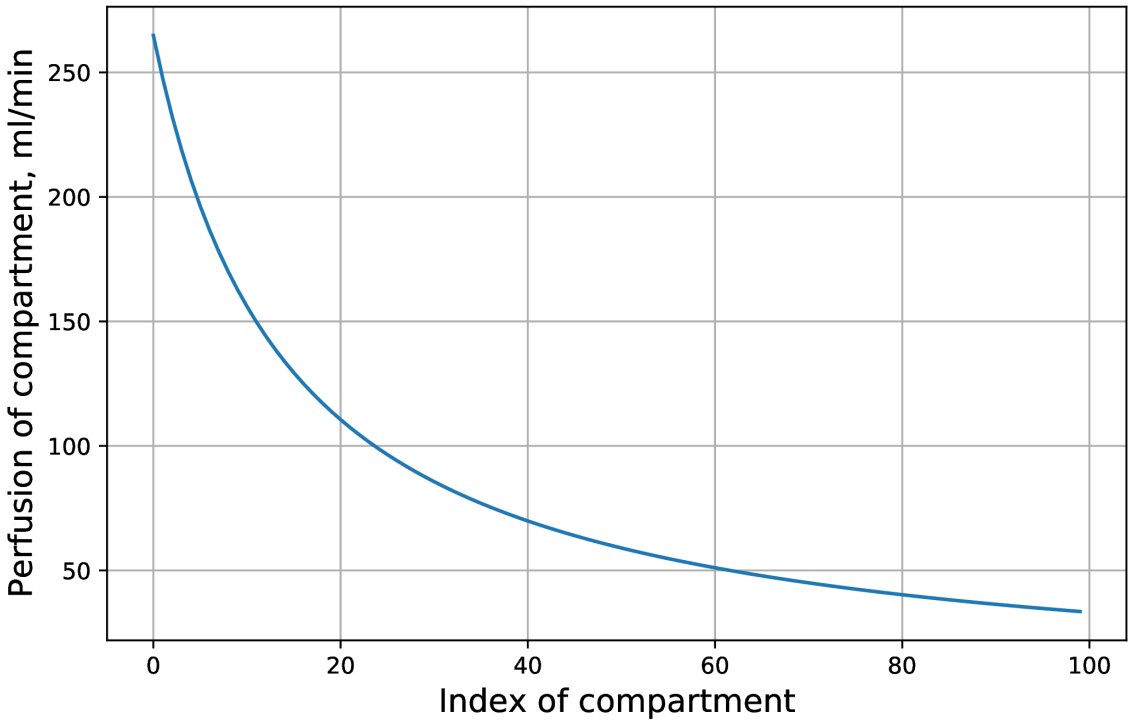


Figure 5-4: Distribution of the perfusion for the patient with $sVR = 312$ dynes/sec/cm^{-5}/min and $inVR = 3760$ dynes/sec/cm^{-5}/min and $CO = 8000$ ml. Strong difference between perfusion of lower parts of the lung (indexes close to 0) and upper parts of the lung (indexes close to 100) is observed.

the lung (therefore R_{100-i} in the equation 5.8), where lower lung parts are ventilated more, than upper lung parts.

Finally, to fully define a virtual patient (parameterize the simulator) a number of other important parameters describing the state of the patient had to be identified. In contrast to compartmental parameters, which had to be defined for each of the compartments separately, these parameters can be considered as global parameters, as they describe the physiological state of the patient as whole. To these parameters belonged 3 groups of parameters: those parameters which are used in the NPS and are present in the underlying ICU data (tidal volume, base excess arterial, etc.); those parameters which were used to parameterize the simulator in previous studies [124, 163, 148, 137, 142, 43], but are fully or partially absent in the underlying ICU data (cardiac output, inspiration : expiration ratio, central venous oxygen saturation); those parameters which were identified in the optimization procedure in previous studies using the NPS (anatomical shunt, respiratory quotient, anatomical deadspace volume, metabolic rate of O_2) [124, 163, 148, 137, 142, 43]. The full

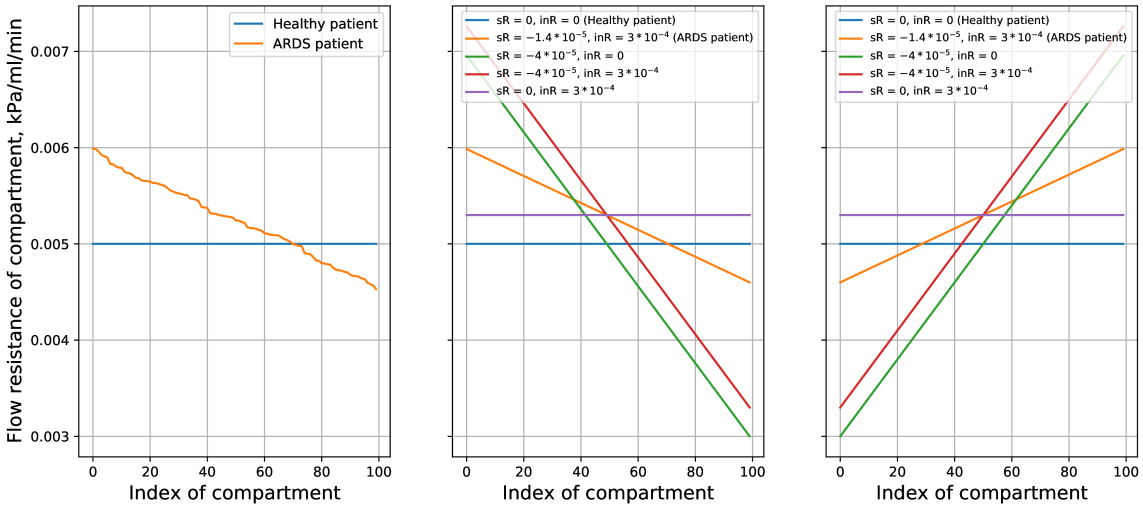


Figure 5-5: Left figure: distributions of R in predefined healthy and ARDS patients. Middle figure: distributions of R for extreme values of sR and inR and for values of sR and inR, which resemble distribution of R of the ARDS patient. Right figure: inverted distributions of R for extreme values of sR and inR and for values of sR and inR, which resemble distribution of R of the ARDS patient. Inversion was needed to model physiological distributions of ventilation over the lung, when lower parts of the lung are better ventilated, than upper parts of the lung.

list of global parameters is given in Appendix A.3.

For each of the global parameters physiologically meaningful ranges were identified based on existing literature in case of parameters missing in the data and on distributions of values in underlying data for parameters present in the data. These ranges for both compartmental and global parameters are given in Appendix A.6.

5.3.2 Sensitivity analysis

To identify which model parameters are predominantly responsible for the model responses and have to be identified, we performed a sensitivity analysis. Sensitivity analysis was performed for every parameter in corresponding ranges defined in Appendix A.6 with respect to simulator outputs, which are of high relevance for assessment of pulmonary conditions, namely arterial blood gas variables: PaO_2, $PaCO_2$, and HCO_3a. Firstly, the baseline values for all model parameters were defined reflecting the "average" patient. Baseline values were calculated as the mean value between lower and upper threshold of the corresponding parameter. The only exception was the parameter defining the number of closed alveolar compartments (n_{cc}), as the middle value of 40 closed compartments was already causing

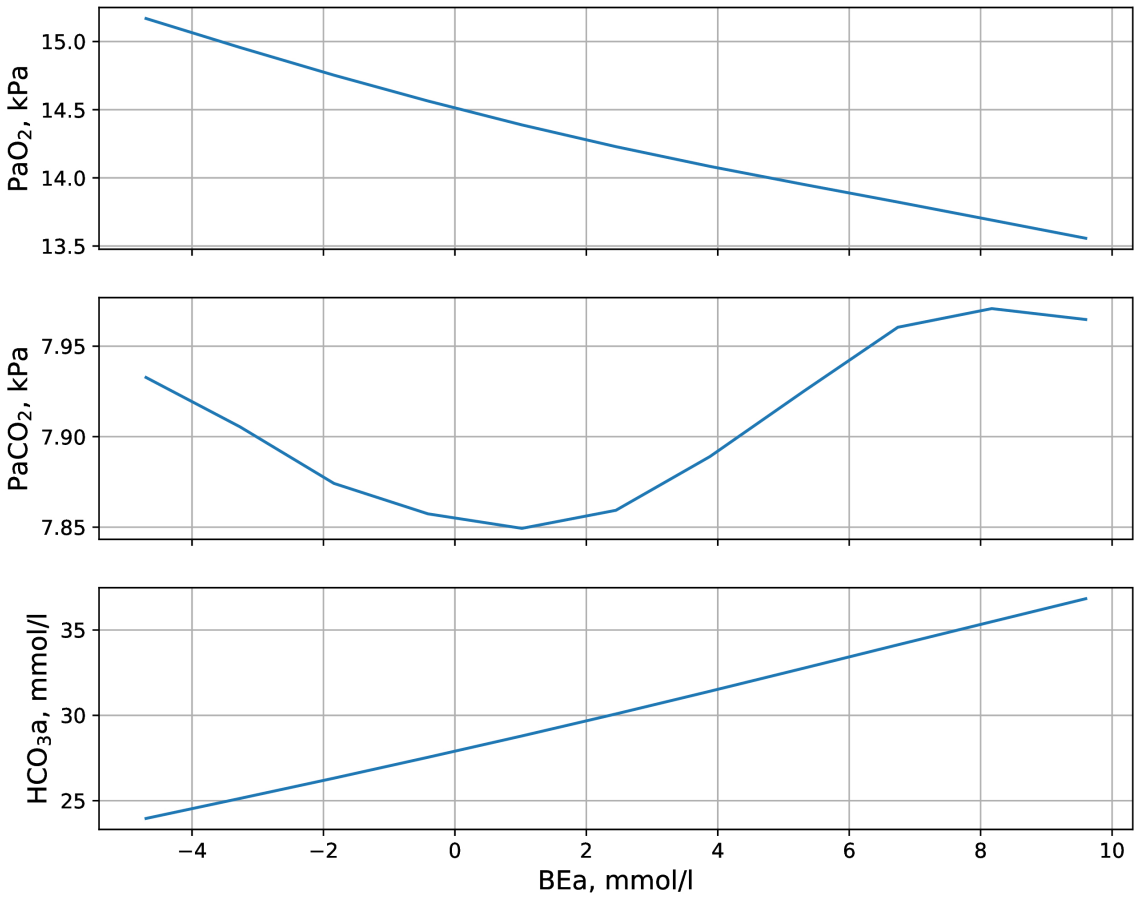


Figure 5-6: Results of the univariate local sensitivity analysis varying BEa on a grid of 11 values between lower threshold and upper threshold for this parameter. Linear dependence of PaO_2 and HCO_3a and non-linear dependence of $PaCO_2$ values with respect to BEa are observed.

very low PaO_2 values and high $PaCO_2$ values meaning completely impaired oxygenation. Therefore, the baseline value for the n_{cc} was set to 0, simulating the intact lung.

We performed univariate local sensitivity analysis varying one parameter at a time on a grid of 11 values between lower threshold and upper threshold, see Figure 5-6 for the example of sensitivity analysis with respect to BEa. Again, n_{cc} represented an exception and was varied between 0 and 40 closed compartments, because of the reasoning discussed in the previous paragraph. Then, following 2 metrics were calculated for each of the parameters with respect to 2 outputs (PaO_2 and $PaCO_2$):

- the coefficient of determination R^2 of the linear fit
- the magnitude M of the linear dependence.

R^2 was used to judge the linearity of the following dependence:

$$\boldsymbol{y_{ij}} = s \times \boldsymbol{x_j} + k, \tag{5.9}$$

where $\boldsymbol{y_{ij}}$ is a vector of simulator outputs of a variable $i \in$ [PaO_2, $PaCO_2$, HCO_3a] in response to the corresponding values of parameter j, $\boldsymbol{x_j}$ represents values of the corresponding parameter j of the sensitivity analysis, s is a slope and k is an intercept.

M was used to judge on the magnitude or strength of the dependence, i.e., how much do changes in independent variable affect the dependent variable. It and was introduced through:

$$M_{ij} = \frac{(\max(\boldsymbol{y_{ij}}) - \min(\boldsymbol{y_{ij}}))/\overline{\boldsymbol{y_{ij}}}}{(\max(\boldsymbol{x_i}) - \min(\boldsymbol{x_i}))/\overline{\boldsymbol{x_i}}}, \tag{5.10}$$

where $i \in$ [PaO_2, $PaCO_2$, HCO_3a] and j stays for every parameter of the sensitivity analysis.

Using two predefined metrics we were able firstly to assess, if a linear dependence takes place and then evaluate the magnitude of such dependence. These two metrics were not suited for nonlinear dependencies, though. Therefore, if such ones were present, they were inspected visually.

Results of the sensitivity analysis are given in Table 5.1. Parameters could be split into three groups, namely those which are linearly influencing all three outputs (PaO_2, $PaCO_2$, and HCO_3a), those which are non-linearly influencing at least one of the dependent variables and those which are not influencing outputs of the simulator. To those linearly influencing both outputs among others belong body temperature, anatomical shunt, cardiac output and number of closed compartments. Non-linear input-output relationships were observed for instance, for P_{EI} and sVR. pHa, tidal volume, SvO_2, PaO_2, $PaCO_2$, HCO_3a, and SaO_2 did not have impact on outputs, indicating that initial values of blood gas variables are changed in the simulation procedure and do not influence equilibrium values obtained in the end of the simulation.

Based on the results of the sensitivity analysis following 11 parameters have to be identified in the matching procedure of the simulator to real patient data (optimization procedure): anatomical shunt (anatShunt), respiratory quotient (RQ), anatomical deadspace volume (VDphys), metabolic rate of O2 (VO2), cardiac output (CO), inspiration : expira-

Parameter	$\mathbf{R^2(PaO_2)}$	$\mathbf{R^2(PaCO_2)}$	$\mathbf{R^2(HCO_3a)}$	$\mathbf{M(PaO_2)}$	$\mathbf{M(PaCO_2)}$	$\mathbf{M(HCO_3a)}$
P_{EI}**	3.65×10^{-1}	6.12×10^{-1}	6.12×10^{-1}	1.28	4.61	9.73×10^{-1}
VO_2**	8.40×10^{-1}	1.00	1.00	8.51×10^{-1}	9.91×10^{-1}	3.44×10^{-1}
T*	1.00	1.00	1.00	1.42	1.47×10^{-1}	4.49×10^{-2}
RQ*	9.96×10^{-1}	1.00	1.00	1.63×10^{-1}	9.60×10^{-1}	2.99×10^{-1}
PEEP**	9.98×10^{-1}	9.38×10^{-1}	9.38×10^{-1}	5.62×10^{-2}	1.01	2.90×10^{-1}
FiO_2**	9.37×10^{-1}	6.38×10^{-1}	6.38×10^{-1}	1.10	4.00×10^{-2}	1.24×10^{-2}
I:E**	8.63×10^{-1}	7.23×10^{-1}	7.23×10^{-1}	2.68×10^{-2}	8.08×10^{-1}	2.09×10^{-1}
n_{cc}*	9.03×10^{-1}	9.57×10^{-1}	9.57×10^{-1}	3.80×10^{-1}	4.69×10^{-1}	1.27×10^{-1}
anatShunt*	8.92×10^{-1}	9.93×10^{-1}	9.93×10^{-1}	8.15×10^{-1}	2.86×10^{-2}	8.71×10^{-3}
VDphys*	9.98×10^{-1}	9.80×10^{-1}	9.80×10^{-1}	3.10×10^{-2}	5.04×10^{-1}	1.51×10^{-1}
CO**	9.99×10^{-1}	9.63×10^{-1}	9.63×10^{-1}	6.56×10^{-1}	1.28×10^{-2}	3.90×10^{-3}
Hb**	9.99×10^{-1}	7.54×10^{-1}	7.54×10^{-1}	2.13×10^{-1}	6.25×10^{-2}	4.42×10^{-2}
VentRate**	9.01×10^{-3}	9.59×10^{-2}	9.59×10^{-2}	1.14×10^{-2}	2.28×10^{-1}	6.73×10^{-2}
BEa**	9.91×10^{-1}	3.21×10^{-1}	3.21×10^{-1}	1.93×10^{-2}	2.63×10^{-3}	7.30×10^{-2}
sVR**	6.13×10^{-1}	7.84×10^{-1}	7.84×10^{-1}	1.24×10^{-2}	2.93×10^{-2}	8.82×10^{-3}
inR*	9.98×10^{-1}	1.00	1.00	2.00×10^{-3}	3.20×10^{-2}	9.78×10^{-3}
sR**	7.62×10^{-1}	9.80×10^{-1}	9.80×10^{-1}	7.78×10^{-4}	1.95×10^{-2}	5.98×10^{-3}
SaO_2†	9.00×10^{-2}	1.38×10^{-2}	1.38×10^{-2}	1.30×10^{-3}	1.75×10^{-3}	8.39×10^{-4}
inVR*	9.52×10^{-1}	9.87×10^{-1}	9.87×10^{-1}	5.23×10^{-4}	1.78×10^{-3}	5.27×10^{-4}
HCO_3a†	1.51×10^{-1}	2.87×10^{-1}	2.87×10^{-1}	3.36×10^{-4}	4.07×10^{-4}	2.03×10^{-4}
$PaCO_2$†	2.90×10^{-1}	4.20×10^{-1}	4.20×10^{-1}	2.20×10^{-4}	3.07×10^{-4}	1.47×10^{-4}
SvO_2†	3.62×10^{-2}	2.35×10^{-2}	2.35×10^{-2}	1.50×10^{-4}	3.69×10^{-4}	1.53×10^{-4}
PaO_2†	3.51×10^{-2}	4.52×10^{-1}	4.52×10^{-1}	1.76×10^{-4}	1.76×10^{-4}	8.46×10^{-5}
Vt†	-	-	-	0.00	0.00	0.00
pHa†	-	-	-	0.00	0.00	0.00

Table 5.1: Results of the sensitivity analysis for the input parameters of the simulator. Sensitivity was evaluated based on two metrics: R^2 reflecting the linearity of input-output relationship and M measuring the strength of the response. Sensitivity was calculated with respect to the outputs of the simulator: PaO_2, $PaCO_2$, and HCO_3a. For Vt and pHa coefficient of determination could not be calculated, as no changes in simulator outputs with respect to changes in these parameters were observed. Parameters which are linearly influencing all 3 outputs are labeled with *; those which are non-linearly influencing at least one of the outputs are labeled with **; those which are not influencing outputs of the simulator are labeled with †.

tion ratio (I:E), sVR, $inVR$, sR, inR, and n_{cc}. All other parameters are either present in the patient data in majority of patients or do not have impact on simulator outputs.

We have to point at the limitations of such sensitivity analysis. The analysis we performed comprises so-called *local* sensitivity analysis and can tell the behaviour of the outputs only in the vicinity of one fixed point (in this case "average" patient, as was defined before). Therefore, there is no guarantee, that these dependencies will hold in other parts of the hyperparameter space. However, as input values of blood gas variables do not influence the outputs of the simulator, this rule should hold in other parts of the hyperparameter space. This holds true for the Vt and pHa, which are not used in the simulator. Therefore, all

parameters which outputs of the simulator are sensitive to were either taken from underlying data or added to the list of parameters that have to be identified in the optimization procedure.

5.3.3 Optimization procedure

Objective function

Parameters that were selected in Section 5.3.2 have to be identified in the optimization procedure, which provides a parametrization of the simulator given underlying data of a real patient. Thus, the digital twin or the virtual patient of the real patient is created. The objective function includes simulator outputs, which are of high relevance for assessment of pulmonary conditions, namely arterial blood gas variables: PaO_2, $PaCO_2$, and HCO_3a. As equilibration of the simulator occurs after 2h time interval from the start of the simulation, the patient data were binned in the 2h bins so that the values of each bin can be used as inputs to the simulator and the values of the next bin can be compared to the outputs of the simulator.

Suitable time points for the simulation from patient datasets were chosen as those, where measurements of the variables needed for the calculation of the objective function (PaO_2, $PaCO_2$, HCO_3a) and the parameters with largest magnitudes found in the sensitivity analysis (P_{EI}, PEEP, FiO_2) were charted. Body temperature was also among the most important parameters found in the sensitivity analysis. However, body temperature measurements were present in most patient datasets, therefore it was not used to prefilter patients. All other parameters needed to parametrize the simulator, but which were not included in the list of parameters for the optimization, were taken from the data of the patient for whom optimization was performed using measurements of a corresponding variable from the time point of the start of simulation. If value on that time point was missing, padding of the last available value before the time point under consideration was used. If values of that variable were missing completely - the mean value of that variable of the population was used.

Only patient datasets with charted values of the aforementioned variables on at least 3 time bins in a time window of interest (24h) were selected for further analysis. Therefore, number of suitable time points within the time window of interest, which were used in the optimization, varied from 3 to 12. The simulator was initialized at each bin and the patient

data of the next bin were compared to the outputs of the simulator.

The minimization problem for the objective function Y is given by:

$$\min_{p_1,\cdots,p_j}(Y) = \min_{p_1,\cdots,p_j}(\frac{\sum_{time\ points}\sqrt{\sum_{i=1}^{3} r_i^2}}{n_{time\ points}}),\ r_i = \frac{y_i^{meas} - y_i^{sim}}{\sigma_i}. \tag{5.11}$$

where y_i are arterial blood gas analysis variables $y \in$ [PaO_2, $PaCO_2$, HCO_3a] and p are optimization parameters. Here, the difference between original measured values y^{meas} and simulator outputs y^{sim} is taken and scaled with the standard deviation of the corresponding variable of a patient. Then, this difference is summed over all blood gas analysis variables. Furthermore, the residuals are summed over all time points in the time window under consideration to account for longitudinal fitting of the simulator to the real patient data. Dividing the objective function by the number of time points does not influence the optimization result. However, it is required to compare optimization residuals for different patients, as the number of time points within the time interval for optimization may vary.

Optimization method

This information can be used to determine a performance measure, i.e., an objective function, but the analytical expression is not available due to the complexity of the simulations.

Next, we had to choose a suitable optimization method. In our case, the analytical calculation of the gradients of the objective function with respect to optimization parameters was not possible due to the complexity of underlying simulations. Moreover, values of the objective function are only available through a time-consuming computer simulation, i.e., calculation of objective function incorporates results of several runs of the simulator (up to 12 runs of the simulator for optimization window of 1 day). One run of the simulator (2h simulation) takes up to 3 minutes of computational time on a 2018 quadcore laptop (i7-8565U CPU @ 1.80 GHz × 8) [6]. Such a long computational time of one simulation is determined by the MATLAB implementation of the simulator. The simulator was provided to us as a proprietary implementation by Prof. Declan Bates for the exclusive use in the ASIC project. In frames of the ASIC project our aim was to provide a proof of concept of the VP modeling pipeline for the ICU data. Therefore, reimplementation of the simulator in another programming language for the runtime optimization was not feasible within the scope of the ASIC project. However, possible ways to address this issue are discussed in

Section 5.4.

As the objective function includes outputs of multiple simulations, one function evaluation may take up to 36 minutes of computational time (given simulations for different time points are run serially). Estimation of derivatives is therefore also computationally expensive, as it would require 11 additional objective function evaluations at each optimization step. Hence, we considered the objective function as a black-box. In the literature, such type of optimization problems is typically called a black-box optimization problem with costly evaluation [164, 165], where the aim is to optimize the objective function within a small budget of function evaluations. After comprehensive search for a suitable optimization method we ended up with the RBFOpt library - an open-source library for black-box optimization with costly function evaluations [164]. The RBFOpt library provides an implementation of an established Radial Basis Function method originally proposed by Gutmann in 2001 [166]. The main feature of the algorithm is that it builds and iteratively refines a surrogate model of the objective function approximating the original objective function. Despite the fact that the original algorithm is not new, numerical comparisons for the current RBFOpt implementation performed by Costa et al. on a test set of problems taken from the literature revealed that the RBFOpt is one of the best algorithms for costly black-box optimization: it outperformed the open-source solvers included in the comparison in the original publication and performed slightly better than a commercial solver [164].

The RBFOpt library constituted two other advantages for our use case. First, in the original publication it was tested on a set of low-dimensional problems (up to 15 optimization parameters) with a small budget of function evaluation and outperformed other methods [164]. Therefore, the RBFOpt method was suitable for our problem with 11 optimization parameters. Second, the RBFOpt library handles mixed-integer problems, i.e. optimization parameters can include both continuous and integer variables. It was the case in our study, as n_{cc} was an integer variable and other 10 parameters were continuous variables. All in all, the RBFOpt library was chosen as an optimization tool for our study.

Stopping criteria

The limiting factor for the optimization was computational time, as each run of the simulator took on average 3 minutes of computational time. As the simulator was implemented in the MATLAB language, but optimization routine was implemented in Python, there was a need

for Python-MATLAB interface. In our trials we ended up with the following approach: optimization for each patient dataset was performed on one core of the compute cluster. Parallelization was therefore performed on the patient level, i.e., matching for 400 patients was performed in parallel using 400 cores of the compute cluster. This approach allowed to get results for the cohort of patients in a reasonable time (several days). However, we did not parallelize simulation for multiple time points in the time interval for a single patient due to issues in parallelization of the Python-MATLAB interface. Therefore, matching for individual patient was performed in a serial way and took up to 3 days of compute time on a single core.

Therefore, a crucial issue was the identification of the early stopping criteria for optimization to reduce the computational time of the optimization routine, as it was intended to be used for thousands of ICU patients. We performed a trial optimization run with 300 iterations for 39 patients to identify which number of iterations provides the best trade-off between compute time and residual of the optimization. For each of the patients, optimization was performed 10 times with different starting random configurations (390 optimization runs overall). Then, for each of the runs we investigated, how the minimum residuum till iteration n deviates from the minimum residuum of the whole optimization run (of 300 iterations).

Results of this analysis are shown in Figure 5-7. Median of the deviation of minimum residuum till iteration n from the minimum residuum of the whole optimization run (of 300 iterations) is shown for each iteration n. Fast convergence is observed for the first 50 iterations changing to slow convergence after 50 optimization iterations. At 100 iterations median deviation from the minimum residuum comprises less than 1.5%, see right figure. Moreover, 95 % of the optimization runs had deviation not larger than 20% from the minimum residuum within first 100 iterations. Therefore, number of iterations equal to 100 was chosen as a stopping criterion for the matching procedure of the simulator to real patient data.

5.4 Limitations of the virtual patient modeling

Starting from simple models with only a few free parameters each, VP modeling has developed into a large research and development area with simulators for various organ subsystems

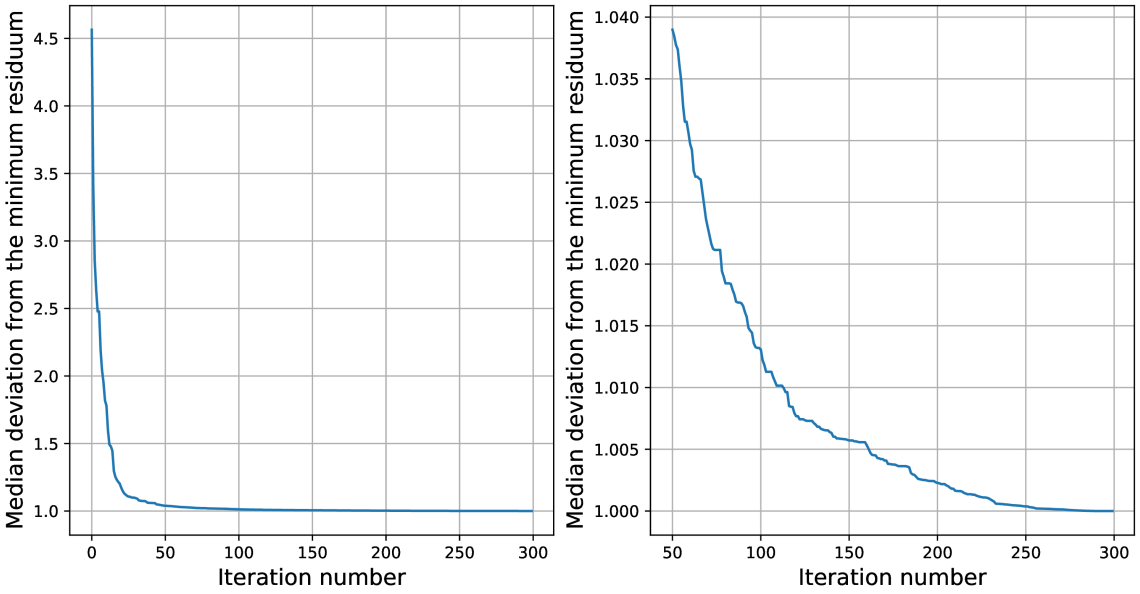


Figure 5-7: Median of the deviation of minimum residuum till iteration n from the minimum residuum of the whole optimization run (of 300 iterations) is shown for each iteration number n. Fast convergence is observed for the first 50 iterations changing to slow convergence after 50 optimization iterations.

encompassing thousands of degrees of freedom. Appropriate model systems are therefore available for many questions, and their selection is based on the scientific question to be investigated and the available data [126, 38].

VP models can be useful for healthcare in several ways. By fitting a model to individual patients, one can access unmeasurable parameters that contain potentially important information about the patient's health status. This information can be used directly to investigate relationships between treatment strategies, identified physiological variables, and manifestations of a particular critical condition uncovering possible pathophysiological mechanisms that drive particular diseases. Moreover, this information can be integrated into machine learning systems for critical condition prediction, classification tasks, or subcohort identification and characterization.

Fitting the VP model to a real patient creates a digital twin that can be used to study the patient's responses to external stimuli, such as changes in mechanical ventilation settings or drug administration. This enables in silico clinical trials. As prospective randomized clinical trials in the ICU are challenging, in silico clinical trials may assist in the design, planning and execution of real clinical studies by investigating effects of different external stimuli on the cohort of digital twins, keeping other patient characteristics intact.

Limiting the applicability of VP models in real-world settings is the complex validation of the developed models. Validation of VP models remains a critical factor, as it must be ensured that a model responds correctly, i.e., similar to real patient's responses, to changes in clinical inputs. On the individual patient level, one compares predicted outcomes with actual measured data from single real patients. Cohort-level validation is performed with "before-after validation" or "cross-validation". In the former case, a specific intervention is applied on a VP cohort and outcomes are compared to the application of the protocol in clinical use. Then, statistics of the VP cohort are compared to those of the original cohort assessing weather VP model accurately resembles cohort dynamics. The goal of the latter method is to decouple the effects of the clinical inputs from the data used to create the virtual patients (called training data). In this way, it can be assessed whether the VP cohorts accurately model patient dynamics independent of the training cohort [39].

Another validation approach is described by the concept of "robustness" [135, 167] where one examines the responses of a model to external or internal stimuli under uncertainty of the model parameters. If the model matches patient behavior and exhibits the same degree of variation as a living system (a patient), then the model is considered to be validated. The opposite case indicates either insufficient parameter adaptation of the model to the patient or structural errors of the model with respect to a particular use case. Moreover, such analysis provides a tool to evaluate the cumulative effect of model parameter uncertainty on model results [148]. Unfortunately, VP models are often applied on and validated by only a limited number of patients, in best cases reaching hundreds of patients [151]. The reason is that clinical trial data needed for validation is mostly hardly accessible. However, with the emerging open-access medical databases, more data become available for the validation and application of VP models [31].

The original NPS was validated against real patient data using the data from clinical trials [124, 148]. We have not explicitly validated our modeling approach against real patient data due to the challenges discussed above. However, our ARDS VP modeling framework was implicitly validated by two observations discovered in the use case presented in Chapter 6, where we modeled ARDS development with the increase of the number of closed alveolar compartments. First, on a cohort of 1007 patients with suspected ARDS acceptable quality of fitting (simulator outputs within 2 standard deviations of measured data) was observed for 96% patients in the time window before suspected ARDS onset (when all 11 parameters

were identified in the optimization procedure) and for 85% patients in the time window after suspected ARDS onset (when only a new value for the number of closed alveolar compartments n_{cc} was identified). Acceptable quality of fitting in both windows was observed for 82% of patients. It implicitly proves the structural validity of the model and ability to model development of ARDS by varying one parameter only. Second, clustering performed on features of virtual patients allowed to discover a cluster enriched with diagnosed ARDS patients. This cluster had the largest increase in the number of closed compartments n_{cc}, which fully supported our approach of modeling ARDS by introducing closed alveolar compartments. More details on this study are given in Chapter 6.

However, there are limitations of our modeling approach that has to be considered. First, we integrated an assumption of uniform distribution of atelectases over the lung, which is based on the initial intention to model an early development stage of ARDS. Second, our approach does not account for alveolar recruitment through recruitment maneuvers or prone positioning, which seems to be a reasonable assumption in the early stages of the ARDS development. Moreover, we model ARDS development with changes in one parameter (n_{cc}) only, keeping other identified parameters constant. Thus, VP modeling has the potential to extract a lot of 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.

A crucial limitation of VP modeling is that it still requires relatively large number of physiological variables measured over a time when a VP model is fitted to real data. Despite that, data requirements in VP modeling are lower than in more complex models of computational physiology, where imaging data play a dominant role. However, this requirement significantly reduces the amount of patient datasets, which can be used in VP modeling [39]. To address this issue, instead of using only patient datasets with complete data, missing parameters or physiological variables can be identified in the optimization procedure, as it is done for I:E and CO in the proposed modeling framework. This, however, again increases the complexity of the optimization procedure. Therefore, a balance between the number of optimization parameters, availability of data, and computational resources should be maintained.

Two most important challenges in the matching procedure of the simulator to real patient

data are the identifiability of the parameters given underlying data and uncertainty of the defined parameters of single virtual patients. Identifiability of parameters in the global optimization algorithm means the identification of the unique parameter vector, which yields the global minimum of the objective function. However, a unique global minimum can be missing. Moreover, the identification of the global minimum may require inacceptable computation times. Therefore, in the case of the optimization of the black-box function with costly evaluations other criteria, such as fitting quality may be applied. For instance, optimization algorithm is stopped once satisfactory quality of fitting, which is accepted in practice, is achieved [148]. In our modeling framework we run optimization procedure for predefined number of iterations and then quality of fitting is evaluated. Later, assessment of fitting quality with respect to underlying patient data can be performed to identify potential modeling limitations.

To address the issue of uncertainty in identified parameters, techniques of uncertainty quantification (UQ) should be applied to patient-specific models. Using these methods, parameters can be described with distributions and not with single values. The uncertainty of the identified value for the parameter can be quantified based on the distribution for that parameter for the underlying patient. Thus, the UQ approach provides a natural framework for predictions, where for each single real patient multiple corresponding virtual patients can be sampled based on the distributions for underlying parameters. Thus, uncertainties in outputs of the simulations for these virtual cohort can be assessed [138].

A solution for the UQ is offered by Bayesian framework. In fully Bayesian approaches, a prior distribution on the parameters is modified to a posterior distribution based on the observed data. Such inference provides the ability to sample from the posterior distribution of the parameters, which for instance allows to calculate descriptive statistics, i.e., the posterior mean. Markov chain Monte Carlo (MCMC) methods are the main class of algorithms for full posterior sampling based on a principled accept–reject scheme. However, both the identification of global minimum of the objective function and application of MCMC methods require large number of iterations what is unacceptable in the case of costly function evaluations [168, 148].

Another important limitation is the complexity of the fitting process, which usually requires the use of high-performance computing for the optimization procedure. For instance, in the study by Das et al. model calibration to data was performed using the ‘Minerva’

high performance computing cluster provided by the University of Warwick with 396 nodes (2×hexa-core 2.66 GHz 24 GB RAM) [169]. Due to the implementation of the number of physiologically-based assumptions, we managed to significantly decrease the number of parameters that have to be identified in the optimization procedure. However, matching of the simulator to individual patient data still required the use of the computational cluster of the RWTH Aachen University with 10 nodes (40 cores each, 2.66 GHz, 4 GB RAM) and matching procedure took on average several days of computational time. Therefore, the VP modeling framework introduced in this thesis still requires significant computational resources, which prohibits a straightforward implementation of such method at the bedside.

Long running time of the simulator is the reason for the aforementioned limitations including UQ and challenges by implementation of the VP methods at the bedside. As mentioned earlier, reimplementation of the simulator in another programming language for the runtime optimization was not feasible within the scope of the ASIC project. However, the first step to shorten the simulation time was done in the study by Barakat et al. [6], where the original MATLAB implementation of the simulator was converted into a compilable and portable version in C code. It allowed to scale up the simulation both in terms of speed of execution of individual tasks and in terms of the number of tasks that can be executed concurrently. For instance, the execution time of one simulation was reduced from 259.1 seconds for original MATLAB simulation to 100.8 seconds for the C version running in parallel on 48 CPUs. However, such decreased run time still is not sufficiently low for large-scale use of VP modeling in the real ICU setting, as matching procedure will still take hours for a single patient.

A potential approach to significantly reduce the run time of the simulation is to replace the mechanistic VP simulator with a data-based black box model. With the rapid growth of performance of data-driven approaches, which have been discussed in Introduction, ML models can be trained to directly approximate the simulation model for the purpose of accelerating matching and data assimilation of virtual patient models. The fundamental idea is to learn to approximate the simulation-based outputs and then use these computationally efficient surrogates (for instance a DL model which outputs reproduce the outputs of the simulator for the same values of input variables) in later tasks such as VP matching and extraction of model parameters. This approach can enable creation and further analysis of large-scale virtual patient cohorts with minimal computational requirements, additionally al-

lowing application of tools for UQ, for instance MCMC methods. Moreover, such pretrained portable ML model can be used at the bedside with limited computational resources. However, several challenges remain to be addressed. First, to build the training database from simulation remains time-consuming requiring large number of simulation runs. It is unclear how exhaustive the range of inputs needs to be in order for the machine learning surrogate to be able to mimic the simulation model over a wide range of parameter values [138]. Secondly, surrogate ML model replacing a simulator will lack such key advantages of mechanistic modeling as explainability and extrapolability (ability to match a patient different from all patients, that have been matched before).

Chapter 6

Virtual patient modeling reduces dataset bias and improves machine learning-based detection of ARDS from noisy heterogeneous ICU datasets

The issue of impaired performance of data-driven models on datasets different from those used for training, i.e. poor generalizability of ML models in healthcare settings, has been introduced in Section 2.3. This issue can be addressed by pooling of data from diverse origins, i.e. hospitals. However, pooling of data for development of ML tools introduces further biases driven by data origin. Such differences were uncovered in Section 4.4, where it was shown, that both treatment strategies and admission policies differ among hospitals which are used in this thesis, which might represent a challenge for the application of ML methods.

In this chapter, 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. The framework for individual VP modeling for real-world ICU data which was introduced in Section 5.3 is used to create a large ($>$ 1000 patients) cohort of virtual patients based on the retrospective observational ICU data of patients with

suspected ARDS pooled from different hospitals. In the optimization procedure patient data are mapped onto individualized model parameters approximating disease states of patients. These parameters comprise data derived in the matching procedure of the VP model to real patient data (later referred to as model-derived data).

Next, we compare the results of an unsupervised ML method (clustering) in two cases: where the learning is based on original patient data and on model-derived data. More robust cluster configurations are observed in clustering using the VP model-derived data. VP model-based clustering also reduces biases introduced by the inclusion of data from different hospitals. Moreover, methodology introduced in this chapter provide a way to address the challenge of the under-recognition of ARDS by physicians. VP model-based clustering allowed to discover an additional cluster with significant enrichment of patients with diagnosed ARDS. Thus, model-derived data can be potentially used to identify non-diagnosed ARDS patients (which belong to the same cluster as diagnosed ARDS patients), providing a route to improved ML model development for early ARDS recognition.

Overall analysis pipeline is shown in Figure 6-1. Our results indicate that mechanistic VP 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 subpopulations.

Section 6.1 introduces materials and methods for the study including information on the approach for modeling of ARDS development, underlying data, clustering, and enrichment analysis. Section 6.2 presents the results, followed by a discussion in Section 6.3.

6.1 Methodology

6.1.1 Creation of a virtual patient cohort

To fully define each of the virtual patients, the simulator was fitted to individual patient data using established global optimization algorithm as described in Section 5.3.3. The model parameters that were identified in the optimization procedure included 2 groups of parameters, as described in Section 5.3: rarely measured physiological variable if missing in patient data (anatomical shunt, respiratory quotient, anatomical deadspace volume, metabolic rate of O2, cardiac output, and inspiration to expiration ratio) and parameters defining distributions of properties of alveolar compartments (vascular resistance, flow resistance of compartments

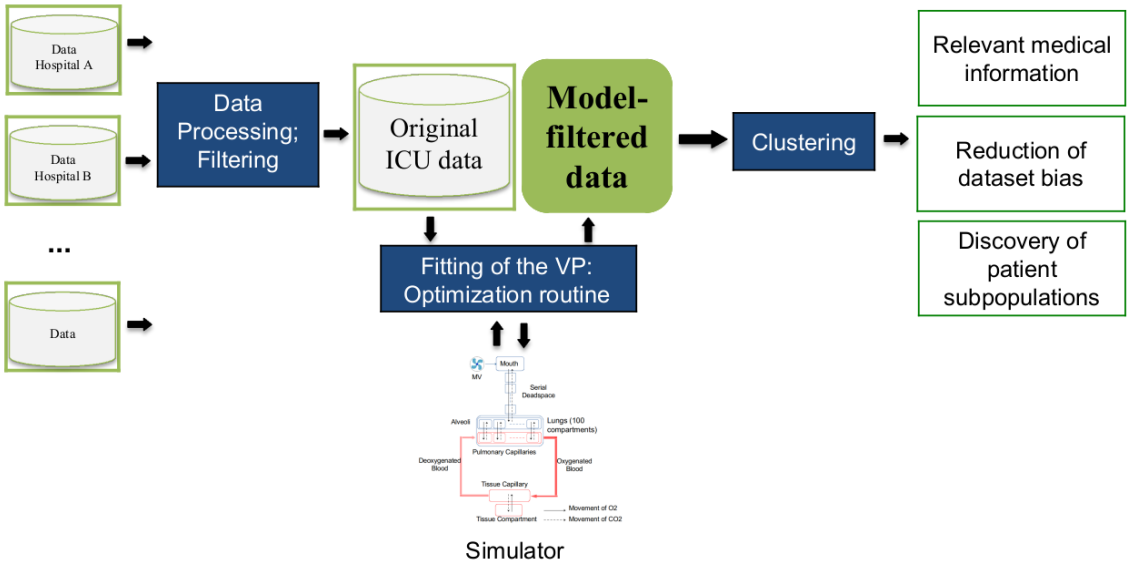


Figure 6-1: Scheme of the VP modeling framework. Heterogeneous patient ICU data are processed and filtered in the first step. Then, optimization routine is utilized to match a VP model to single ICU patient datasets. Thus, model-derived data of single patients are extracted. These data are further used in clustering methods. Discovered clusters are further analyzed.

and number of closed alveolar compartments). 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 Section 5.3.3.

We introduced one change to the parameters, that have to be identified in the optimization procedure to further improve the modeling. In the sensitivity analysis introduced in Section 5.3.3 cardiac output (CO) values were found to have highest influence on the outputs of the simulator. However, CO is known to change with time as a response to changes in heart rate and stroke volume (SV), where heart rate can vary significantly within short time periods (minutes and hours). As heart rate data were almost always present in the patient data, we decided to replace CO with the SV multiplied by the heart rate extracted from the real data. Thus, SV was the parameter, that was identified in the optimization procedure.

Final bounds for the optimization parameters are given in Table 6.1.

The optimization procedure was performed in two time windows relative to the onset of ARDS (t_0): from t_0 - 2d to t - 1d (window 1) and from t_0 to t + 1d (window 2), where d stands for 1 day.

Parameter	Lower bound	Upper bound
SV, ml	30.00	150.00
I:E, unitless	0.17	0.50
sR, kPa×min/ml	4.00×10^{-5}	0.00
inR, kPa×min/ml	0.00	3.00×10^{-4}
sVR, dynes/sec/cm^{-5}	0.00	312.00
inVR, dynes/sec/cm^{-5}	0.00	3760.00
n_{cc}, unitless	0	80
anatShunt, unitless	0.01	0.30
RQ, unitless	0.60	1.00
VO_2, ml/min	150.00	550.00
VDphys, ml	50.00	150.00

Table 6.1: Lower and upper bounds for parameters that have to be identified in the optimization procedure.

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. VP fitting scheme is given in Figure 6-2.

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 parameters and simulation outputs used in the further analysis is given in Appendix A.7). For each of the patients, these parameters comprised model-derived data consisting of 18 features.

6.1.2 Data

In this chapter fully depersonalized data on ICU patients from four German hospitals (datasets Hosp D, Hosp E, Hosp F, and Hosp G) collected during the project "Algorithmic surveillance of ICU patients with acute respiratory distress syndrome" (ASIC) were used. Additionally, a historic retrospective dataset from one of the participating hospitals was included into the analysis (Hosp A). It comprised fully depersonalized data of ICU patients that were extracted according to the same rules as within the ASIC project. The time

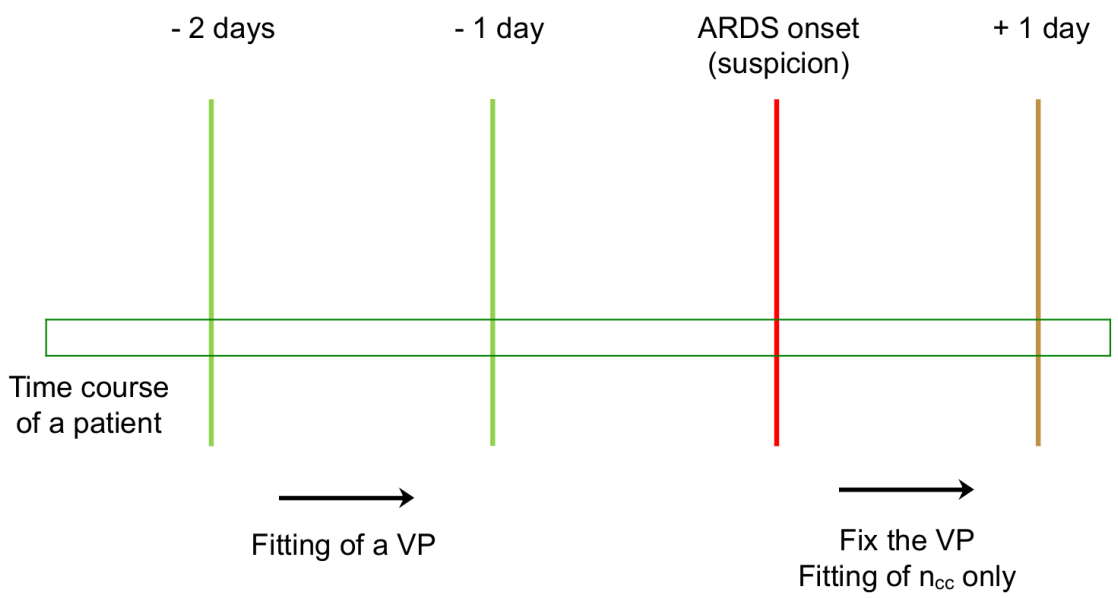


Figure 6-2: VP fitting scheme. Firstly, VP is parameterized by fitting the model to the real patient in the time window between 2 to 1 days before ARDS onset. Next, the VP configuration is fixed and only the number of closed alveolar compartments n_{cc} is fitted after suspected ARDS onset.

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. Initial number and clinical characteristics of patients in corresponding hospitals is given in Table 3.1. The overall number of patients in 5 datasets comprised 29,275 patients.

Each patient's data included routinely charted ICU variables collected over the whole ICU stay, biometric data and ICD-10 codes. The full list of variables is given in Appendix A.1. Data from all five datasets were brought to the same units of measurement, checked for consistency, and preprocessed according to the rules given in Section 3.3.

The criteria for the diagnosis of ARDS were introduced in Section 2.2. As medical imaging data were missing in our dataset, only suspected ARDS onset time could be determined according to rules introduced in Section 3.4. It was defined as the timepoint when the Horowitz 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 and blood gas analysis (see Section 5.3.3) charted both before and after the suspected ARDS onset were selected. The

Dataset	Initial number of patients, n (%)	Final number of patients, n (%)
Hosp A	13,067 (100)	467 (4)
Hosp D	3,591 (100)	127 (4)
Hosp E	1,360 (100)	110 (8)
Hosp G	2,217 (100)	114 (5)
Hosp F	9,040 (100)	189 (2)

Table 6.2: Initial and final number of patients in the datasets under consideration.

final number of patients fulfilling these criteria comprised 1,007 patients. The initial and final number of patients for the analysis in corresponding hospitals is given in Table 6.2.

6.1.3 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 original data, which were used as inputs to the simulator. For each of the variables, mean values in windows 1 and 2 (before and after suspected ARDS onset respectively) were calculated and used as features (see Appendix A.8). Additionally, the mean Alveolar–arterial gradient (A-aO_2 or A–a gradient) in each of the time windows was calculated and added to the list of features:

$$A - aO_2 = F_iO_2(P_{atm} - P_{H_2O}) - PaCO_2/0.8 - PaO_2, \tag{6.1}$$

where at sea level $P_{atm} = 760$ mmHg and $P_{H_2O} = 47$ mmHg. Increased A–a gradient representing the large difference between the alveolar and arterial concentrations of oxygen implies ventilation-perfusion mismatch, therefore reflecting impaired integrity of the alveolar capillary unit. In states of ventilation-perfusion mismatch, such as pulmonary embolism or ARDS, oxygen is not effectively transferred from the alveoli to the blood which results in an elevated A-a gradient [170]. Finally, difference in Horowitz index between window 1 and window 2 was included to the list of features and used as a feature reflecting the severity of development of oxygenation impairment. Next, highly correlated features with a Pearson correlation coefficient between corresponding measurements larger than 0.9 were omitted from the analysis.

The second dataset comprised model-derived data: simulator outputs and parameters found in the optimization procedure (see Appendix A.7). The former dataset thus repre-

sented data from the cohort of original patients, while the latter represented the model-derived data or data from the virtual patient cohort.

Consensus k-means clustering was performed for different number of clusters (from 2 to 8 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 [171]. Based on produced clusters a consensus matrix D is formed, which reflects how many times two items of the dataset occurred in the same cluster:

$$D(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)}, \tag{6.2}$$

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, hierarchical clustering is performed on the consensus matrix to produce final clusters. 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 the observed clusters based on the predefined threshold. The threshold for the final clustering configuration was chosen in the way, that at most 20% of patients were labeled as outliers.

Quality of clustering was assessed using mean cluster's consensus, as described in the study by Monti et al. [171]. Firstly a cluster's consensus $m(k)$ is defined as the average consensus index between all pairs of items belonging to the same cluster k:

$$m(k) = \frac{1}{\frac{N_k(N_k-1)}{2}} \sum_{i,j \in I_k, i<j} D(i,j), \tag{6.3}$$

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. Then, 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.

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 $\alpha = 0.05$ [172]. 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 testing using Benjamini-Hochberg correction [173].

Clinical conditions which were used in the enrichment analysis were defined based on ICD-10 codes of underlying patients. Following clinical conditions were extracted from the codes: all ICD-10 chapters (A00-B99: Certain infectious and parasitic diseases, C00-D48: Neoplasms, etc.), all ICD-10 blocks (A00-A09: Intestinal infectious diseases, A15-A19: Tuberculosis, etc..), and comorbidities related to ARDS, which are given in Appendix A.2. Only clinical conditions with prevalence of more than 1% in the underlying population were used in the analysis.

6.1.4 Modules used in this study and system requirements

Original simulator requires the installation of MATLAB [149] 2009 or above, with a suitable C compiler installed. Simulations in frames of this thesis were conducted using MATLAB 2020a. The RBFOpt library [164] was used for fitting the VP model to real patient data in the optimization procedure, as described in Section 5.3.3. The following Python 3 programming language [111] implementations were used in this chapter: scikit-learn [112] implementation of k-means clustering was used in the consensus clustering algorithm (sklearn.cluster.KMeans); scipy [110] 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 $\alpha = 0.05$.

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

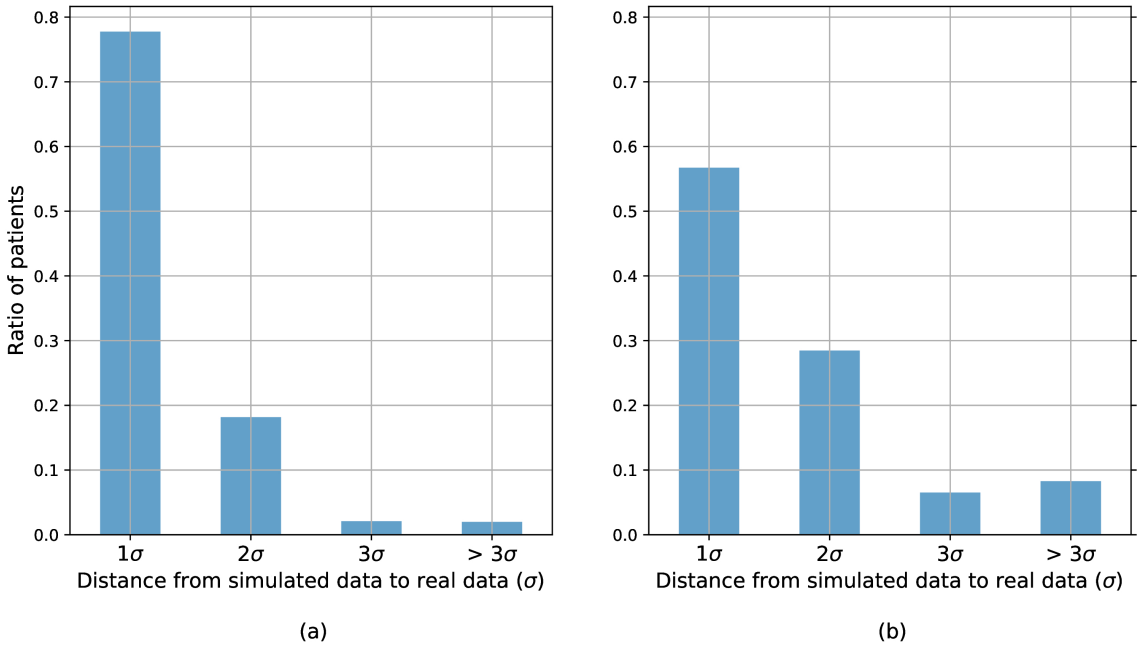


Figure 6-3: Quality of fitting the simulator to real patient. Cohort of 1007 patients with suspected ARDS. Acceptable quality of fitting (simulator outputs within 2 standard deviations of measured data) was observed for 96% patients in the window before suspected ARDS onset and for 85% patients in the time window after suspected ARDS onset.

time points in each of the 2 windows. Therefore, matching procedure took on average several days of computational time. Optimization routine was tested as well on the 2018 quadcore laptop i7-8565U CPU @ 1.80 GHz × 8.

6.2 Results

6.2.1 Optimization results

Fitting quality of the optimization procedure for all patients is shown in Figure 6-3. Acceptable quality of fitting (simulator outputs within 2 standard deviations of measured data) was observed for 96% patients in the window before suspected ARDS onset and for 85% patients in the time window after suspected ARDS onset. Acceptable quality of fitting in both windows was observed for 82% or 823 patients. which were used in the subsequent analysis. Thus, reliable model-derived data were obtained for 823 patients, which were used in the subsequent analysis.

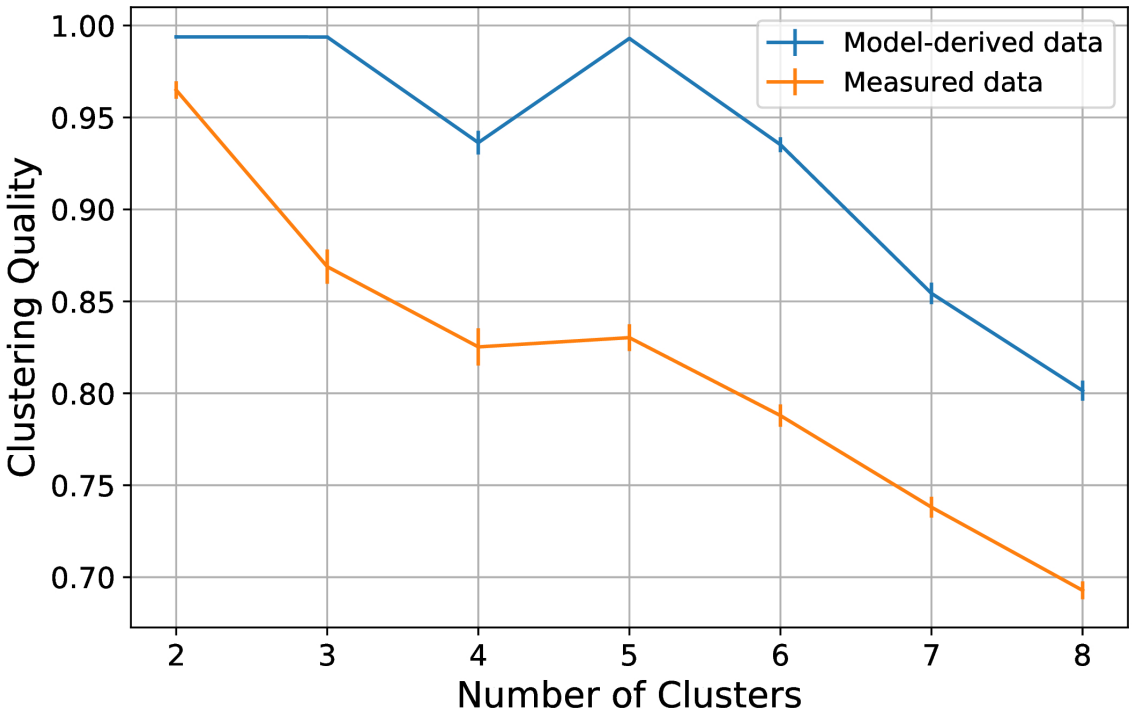
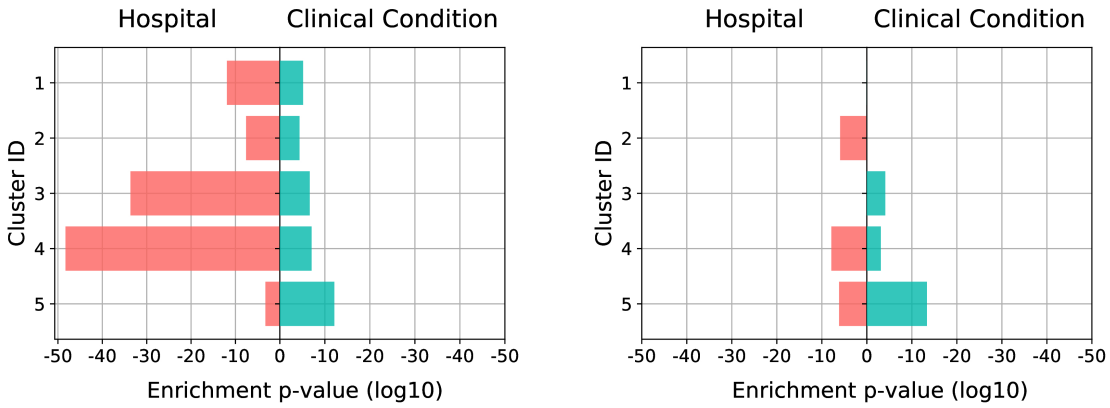


Figure 6-4: Clustering quality for different numbers of clusters for clustering on original measured data (orange line) and model-derived data (blue line) data. Mean clustering quality with 95% confidence intervals over repeated (100 times) clustering on subsample (80%) of dataset is shown. Mean clustering quality and results of a two-tailed Student's t-test for mean quality of clustering are given in Table 6.3.

6.2.2 Clustering on original measured data

Clustering quality for different configurations of the number of clusters is shown in Figure 6-4. 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. Therefore, the number of clusters for further investigation was fixed to 5.

Each of the 5 discovered clusters had characteristic 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 Figure 6-5 (a). Enrichment results are given in Appendix A.9. Finally, none of the discovered clusters had significant



(a) (b)

Figure 6-5: Significance of enriched clinical conditions and 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 hospital (red bar). In clustering on original data, all 5 discovered clusters are significantly enriched with data from some hospitals. In clustering on simulation data, 2 clusters without enrichment for a hospital are observed and overall magnitude of enrichment with respect to a hospital is decreased.

enrichment of diagnosed ARDS patients (according to ICD-10 code J80.x).

6.2.3 Clustering on model-derived data

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 Figure 6-4 for the results of clustering, Table 6.3 for the results of the t-test, and Appendix A.10 for enrichment results). While on the original measured data, the quality decreased significantly already after increasing the number of clusters to 3, in 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, and thus the number of clusters for further investigation was fixed to 5, similarly to the case of clustering on the original data.

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 Figure 6-5 (b) (significance of 5.0×10^{-49}, 2.2×10^{-34}, 1.2×10^{-12}, 2.5×10^{-8}, 5.8×10^{-5} in measured data vs. 1.3×10^{-8}, 6.9×10^{-7}, 1.2×10^{-6} in model-derived

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	9.481×10^{-21}
3	0.869 (0.860, 0.878)	0.994 (0.993, 0.995)	26.205	1.103×10^{-46}
4	0.825 (0.815, 0.835)	0.936 (0.930, 0.942)	18.137	2.373×10^{-41}
5	0.830 (0.823, 0.837)	0.993 (0.992, 0.994)	43.713	3.222×10^{-67}
6	0.788 (0.782, 0.794)	0.935 (0.931, 0.939)	39.515	2.879×10^{-88}
7	0.738 (0.732, 0.744)	0.854 (0.848, 0.860)	28.077	6.148×10^{-71}
8	0.693 (0.688, 0.698)	0.801 (0.796, 0.806)	29.223	2.781×10^{-73}

Table 6.3: 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.

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 Horowitz 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.

6.3 Discussion

As described in Section 2.1, ICU data consist of global indices comprising the high-level observations of the real pathophysiological state of the patient, which do not directly infer the core disease state of the patient. Moreover, medical interventions in the ICU setting can differ significantly among diverse hospitals introducing additional hospital bias to the datasets [54, 8]. Thus, relevant medical signals about a patient's state are often disturbed or missing completely. However, the ability of VP models, when appropriately adapted, to create a digital twin for a real patient enables assessment of patient-specific variables that are not readily measurable (e.g., vascular resistances, transpulmonary pressures, anatomic shunt, etc.) and 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 [39].

In the current work, for the first time, the simulator was 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 82% 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 extracted for 823 patients, supporting the structural and methodological validity of the proposed ARDS modeling framework.

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, indicating better separation and more robust clustering configurations.

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. Moreover, 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 clinical 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, i.e. MIMIC, for the study presented in this chapter, which could have covered different patient populations. All patients satisfied the same strict inclusion criteria and patient datasets were chosen and filtered according to uniform rules.

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 lower than in the clustering on original data (see Figure 6-5). 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 diagnosed ARDS patients. 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 of 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, which was discussed in detail in Section 2.2. Insufficient quality of labeling represents an additional factor that contributes to impaired generalization of 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 [68, 69, 70]. On the other hand, reliable retrospective labeling constitutes a challenging task, due to the fact that diagnosis according to the Berlin definition requires the clinical appraisal of certain conditions, such as hypervolemia, which are not assessable retrospectively. 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 still remains a challenge due to a high inter-rater variability in chest imaging [80]. Finally, studies on the development of ML models for ARDS are utilizing diverging rules to retrospectively label ARDS patients [47, 78, 77].

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 Horowitz 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 [174]. 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.

The study presented in this chapter 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 Horowitz 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 were introduced as fundamental limitations of the VP modeling approach in Section 5.4. They 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 of number of patients in the final analysis cohort compared to the initial cohort (see Table 6.2). 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 older datasets (Hosp A) 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 this dataset and to 4 datasets from other hospitals. 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 significant computational resources, as described in Section 5.3.3. This limitation also did not allow us to perform comprehensive uncertainty quantification analysis, which would require application of MCMC routines and would significantly increase already long running times. All this still tremendously complicates a straightforward implementation of such approaches at the bedside.

Chapter 7

Conclusion

In conclusion, this work proposes the way on how to address the challenge of poor generalization of AI methods, such as ML models, developed for the use in healthcare in order to allow application of these models in heterogeneous ICU datasets. This work comprises two main contributions. First, a framework for the quantitative assessment of differences between datasets is proposed. This framework investigates whether an ML model trained on one dataset will lay in the extrapolation regime once applied on the other dataset based on the CH intersections of the underlying datasets. Moreover, differences in distributions of variables between underlying datasets are evaluated using ML routines.

This framework enables the quantification of dataset bias and validation of developed models before a potential application at the bedside. Given the training data and a retrospective dataset from a hospital where the model is intended to be used, we can judge the generalization ability in another hospital. On the use case of classification for the first day ARDS, we showed that the strongest drop in performance is associated with the poor intersection of CHs of corresponding hospitals and with differences in underlying data distributions. Therefore, we suggest the application of this framework as a first tool to assess the transferability of trained models and quantify the differences between datasets, i.e., dataset bias [1].

The second contribution of our work is the development of the VP modeling framework for real-world ICU data. It utilizes a sophisticated mechanistic VP model and advanced optimization methods to extract relevant medical information on individual ICU patients from observational data of mixed origin. To allow large-scale VP modeling for real-world

ICU data we:

- implemented ICU data preprocessing scripts that bring ICU data from multiple sites to the same format, check the data for consistency, and perform data filtering (Section 3.3),
- implemented a number of assumptions based on the ARDS physiology to the modeling approach, i.e., the original algorithm for the creation of virtual ARDS patients [43] was replaced with a novel approach, which allowed to reduce the number of model parameters, which have to be identified in the optimization procedure from hundreds to eleven (Section 5.3.1),
- used an established optimization algorithm suitable for our setting for matching of virtual patients to original ICU patient data (Section 5.3.3).

To demonstrate the utility of patient-specific features extracted from real ICU patient data, they were utilized as inputs for unsupervised ML methods. Such hybrid modeling framework, where ML models can be trained either on the model-derived data only or on a combination of original data with patient-specific parameters found in the VP modeling framework, provides many benefits which allow tackling important problems limiting the use of ML in the ICU setting.

Initially, hybrid modeling was developed in the context of modeling of complex chemical processes. Hybrid models provide a continuous transition between pure mechanistic models and pure black-box ML, allowing an optimal balance between existing prior knowledge, data availability, and predictive ability [175, 176, 177, 178]. Data-based approaches can be utilized to infer relationships where mechanistic understanding is partially or completely lacking, i.e., between underlying measured or model-derived data and patient outcomes. Moreover, VP models are usually limited to one subsystem or 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. Here, ML methods again can be utilized to infer these influences from available data. Thus, combination of VP modeling with ML models comprises a hybrid modeling approach, which provides a route to address conceptual problems of both pure black-box modeling approaches and pure mechanistic approaches and to contribute to transition to personalized medicine.

Previous applications of hybrid approaches incorporating both mechanistic and data-based modeling have already resulted in successes in other areas of research. For instance, model-derived parameters of individual patients were used to infer important clinical covariates for a patient state [179] or stratify patients [180]. However, to the best of our knowledge, for the first time, we created a hybrid modeling framework for the analysis of large-scale ICU patient data, which couples the complex simulator to real-world data and utilizes the outputs in the ML routines.

In our study, we have highlighted the advantages of the hybrid modeling approach in comparison to the direct utilization of original ICU data in ML algorithms. First, our results support the hypothesis that mechanistic modeling can be used to infer individualized parameters approximating disease states of patients. These parameters, while being used as inputs to ML routines, enable significant reduction of biases introduced by learning from heterogeneous datasets. This result logically follows the observation of Dickson et al. [40], who demonstrated that 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. Our finding allows to draw the conclusion, that parameters of the matched VPs represent a feature space, where shared information of patient dynamics is encoded. Compared to other approaches for multisite dataset adaptation, which build black-box latent representation spaces based on GANs [35, 98], the VP approach provides a high degree of interpretability, as VP parameters reflect feature of defined pathophysiological state. In healthcare interpretability of the results of predictive models plays a major role, as for physicians, practitioners, and policy makers it is crucial to trust the validity and accuracy of the model, as well as understand how the model works, what recommendation has been made by the model, and why [181].

Second, the hybrid modeling framework allowed the discovery of patient cohorts driven exclusively by medical conditions. Moreover, the hybrid pipeline allowed to find a cluster with significant enrichment of diagnosed ARDS patients. This finding supports our ARDS modeling approach, as features of the patients in this cluster resemble those expected for ARDS patients in our ARDS modeling approach. However, the second and more important consequence of this finding is the potential of our approach for the retrospective identification of non-diagnosed ARDS patients, which would enable the training of ML models for early ARDS recognition on reliably labeled data. Nevertheless, further research and retrospective

validation by physicians is needed to prove this hypothesis.

There are multiple further application directions for the virtual patient modeling in the ICU. First, the VP modeling framework can be applied at the bedside to immediately create a digital twin of the patient once they are admitted to the ICU (after the calibration time needed to collect data for the VP model matching) and continuously update the VP parameters along the stay of the ICU patient. Then, physiological parameters continuously extracted from the digital twin can either be directly used as biomarkers, i.e., number of closed alveolar compartments can indicate impaired oxygenation before it is reflected in measured variables, or embedded into ML routines for early detection of critical states, as the information content in the parameters of the VP can comprise added value for the ML routines. Our approach for modeling ARDS development and state evolution of the patient in the ICU with changes in only one parameter of the virtual patient, namely the number of closed alveolar compartments, comprised a significant simplification of clinical reality. To continuously extract other important parameters, one can increase the modeling detail by allowing other VP parameters to vary within physiological limits on the characteristic time scales.

Second, digital twins created for real ICU patients can be utilized to suggest optimal treatment strategies. The guidelines for mechanical ventilation in the ICU for similar patient populations significantly vary among different hospitals, which was uncovered and discussed in Section 4.4 [1]. Therefore, there are no generally accepted rules for the choice of MV settings. Thus, VP modeling provides a way to apply simultaneously numerous MV strategies to the digital twin and choose an optimal one based on predefined criteria, i.e., constant minute ventilation or constant alveolar ventilation [143, 41].

Third, the virtual patient modeling framework presented in this thesis still reflects only pulmonary and (significantly simplified) cardiovascular components of the patient pathophysiological state. This framework can be coupled to similar frameworks developed for other organ systems and pathophysiological conditions, for instance, models for the heart [138] or glycaemic control models [40]. The overarching framework of the virtual ICU human could then be used for all purposes described in this thesis. Here, however, the trade-off between the complexity of the model and real-world data availability must be maintained.

Nevertheless, to realize the benefits of VP modeling mentioned above, the expensive computational requirements of the matching procedure should be overcome. We have discussed

possible ways to address this challenge in Section 5.4. The most promising approach is to replace a VP simulator with a surrogate ML model mimicking the outputs of the simulator for the same inputs. Such an approach would enable online learning of the VP parameters in the case of limited computational resources, i.e., at the bedside. Thus, VP modeling approaches will be applicable in every hospital with minimal computational requirements.

All in all, our developed frameworks allow the utilization of available real-world ICU databases for virtual patient modeling and hybrid modeling, which encompasses numerous benefits for healthcare. 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 virtual patient modeling in combination with machine learning in the future will allow a significant contribution to "Precision Medicine" supporting the transition from standardized, one-size-fits-all care to personalized, one-method-fits-all care with benefits for patients, physicians, and the healthcare system.

References

[1] K. Sharafutdinov, J. S. Bhat, S. J. Fritsch, K. Nikulina, M. E. Samadi, R. Polzin, H. Mayer, G. Marx, J. Bickenbach, and A. Schuppert, "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*, vol. 5, 2022.

[2] K. Sharafutdinov, S. Fritsch, and A. Schuppert, "Virtuelle Patientenmodelle in der Intensivmedizin," in *Die digitale Intensivstation* (G. Marx and S. Meister, eds.), pp. 55–71, MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 1 ed., 2022.

[3] K. Sharafutdinov, S. J. Fritsch, M. Iravani, P. F. Ghalati, S. Saffaran, D. G. Bates, J. G. Hardman, R. Polzin, H. Mayer, G. Marx, J. Bickenbach, and A. Schuppert, "Computational simulation of virtual patients reduces dataset bias and improves machine learning-based detection of ARDS from noisy heterogeneous ICU datasets," *IEEE Open Journal of Engineering in Medicine and Biology*, pp. 1–11, 2023.

[4] P. Farhadi, K. Sharafutdinov, J. S. Bhat, and A. Schuppert, "Big Data und künstliche Intelligenz in der Medizin," in *Telemedizin: Grundlagen und praktische Anwendung in stationären und ambulanten Einrichtungen* (G. Marx, R. Rossaint, and N. Marx, eds.), pp. 423–436, Berlin, Heidelberg: Springer, 2021.

[5] G. Marx, J. Bickenbach, S. J. Fritsch, J. B. Kunze, O. Maassen, S. Deffge, J. Kistermann, S. Haferkamp, I. Lutz, N. K. Voellm, V. Lowitsch, R. Polzin, K. Sharafutdinov, H. Mayer, L. Kuepfer, R. Burghaus, W. Schmitt, J. Lippert, M. Riedel, C. Barakat, A. Stollenwerk, S. Fonck, C. Putensen, S. Zenker, F. Erdfelder, D. Grigutsch, R. Kram, S. Beyer, K. Kampe, J. E. Gewehr, F. Salman, P. Juers, S. Kluge, D. Tiller, E. Wisotzki, S. Gross, L. Homeister, F. Bloos, A. Scherag, D. Ammon, S. Mueller, J. Palm, P. Simon, N. Jahn, M. Loeffler, T. Wendt, T. Schuerholz, P. Groeber, and A. Schuppert, "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, Apr. 2021. Publisher: British Medical Journal Publishing Group Section: Intensive care.

[6] C. Barakat, S. Fritsch, K. Sharafutdinov, G. Ingólfsson, A. Schuppert, S. Brynjólfsson, and M. Riedel, "Lessons learned on using High-Performance Computing and Data Science Methods towards understanding the Acute Respiratory Distress Syndrome (ARDS)," in *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pp. 368–373, May 2022. ISSN: 2623-8764.

[7] K. Nikulina, *Comparing Populations Using Convex Hull Analysis*. Bachelor's Thesis, RWTH Aachen University, Aachen, Dec. 2021.

[8] K. Sharafutdinov, S. J. Fritsch, G. Marx, J. Bickenbach, and A. Schuppert, "Biometric covariates and outcome in COVID-19 patients: are we looking close enough?," *BMC Infectious Diseases*, vol. 21, p. 1136, Nov. 2021.

[9] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press, 2014.

[10] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44–56, Jan. 2019. Number: 1 Publisher: Nature Publishing Group.

[11] H. Fröhlich, R. Balling, N. Beerenwinkel, O. Kohlbacher, S. Kumar, T. Lengauer, M. H. Maathuis, Y. Moreau, S. A. Murphy, T. M. Przytycka, M. Rebhan, H. Röst, A. Schuppert, M. Schwab, R. Spang, D. Stekhoven, J. Sun, A. Weber, D. Ziemek, and B. Zupan, "From hype to reality: data science enabling personalized medicine," *BMC Medicine*, vol. 16, p. 150, Aug. 2018.

[12] R. Bellazzi, "Big data and biomedical informatics: a challenging opportunity," *Yearbook of Medical Informatics*, vol. 9, pp. 8–13, May 2014.

[13] J. S. Beckmann and D. Lew, "Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities," *Genome Medicine*, vol. 8, p. 134, Dec. 2016.

[14] C. H. Lee and H.-J. Yoon, "Medical big data: promise and challenges," *Kidney Research and Clinical Practice*, vol. 36, pp. 3–11, Mar. 2017. Publisher: Korean Society of Nephrology.

[15] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, vol. 18, pp. 851–869, Sept. 2017.

[16] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Molecular Systems Biology*, vol. 12, p. 878, July 2016. Publisher: John Wiley & Sons, Ltd.

[17] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep Learning in Drug Discovery," *Molecular Informatics*, vol. 35, no. 1, pp. 3–14, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201501008.

[18] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, Jan. 2019. Number: 1 Publisher: Nature Publishing Group.

[19] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M.

Boca, S. J. Swamidass, A. Huang, A. Gitter, and C. S. Greene, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, p. 20170387, Apr. 2018. Publisher: Royal Society.

[20] M. Rashid, M. Ramakrishnan, V. P. Chandran, S. Nandish, S. Nair, V. Shanbhag, and G. Thunga, "Artificial intelligence in acute respiratory distress syndrome: A systematic review," *Artificial Intelligence in Medicine*, vol. 131, p. 102361, Sept. 2022.

[21] M. Schinkel, K. Paranjape, R. S. Nannan Panday, N. Skyttberg, and P. W. B. Nanayakkara, "Clinical applications of artificial intelligence in sepsis: A narrative review," *Computers in Biology and Medicine*, vol. 115, p. 103488, Dec. 2019.

[22] L. Wynants, B. V. Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, E. Albu, B. Arshi, V. Bellou, M. M. J. Bonten, D. L. Dahly, J. A. Damen, T. P. A. Debray, V. M. T. d. Jong, M. D. Vos, P. Dhiman, J. Ensor, S. Gao, M. C. Haller, M. O. Harhay, L. Henckaerts, P. Heus, J. Hoogland, M. Hudda, K. Jenniskens, M. Kammer, N. Kreuzberger, A. Lohmann, B. Levis, K. Luijken, J. Ma, G. P. Martin, D. J. McLernon, C. L. A. Navarro, J. B. Reitsma, J. C. Sergeant, C. Shi, N. Skoetz, L. J. M. Smits, K. I. E. Snell, M. Sperrin, R. Spijker, E. W. Steyerberg, T. Takada, I. Tzoulaki, S. M. J. v. Kuijk, B. C. T. v. Bussel, I. C. C. v. d. Horst, K. Reeve, F. S. v. Royen, J. Y. Verbakel, C. Wallisch, J. Wilkinson, R. Wolff, L. Hooft, K. G. M. Moons, and M. v. Smeden, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, vol. 369, p. m1328, Apr. 2020. Publisher: British Medical Journal Publishing Group Section: Research.

[23] M. L. Chee, M. E. H. Ong, F. J. Siddiqui, Z. Zhang, S. L. Lim, A. F. W. Ho, and N. Liu, "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, p. 4749, Jan. 2021. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[24] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A Review of Challenges and Opportunities in Machine Learning for Health," *AMIA Summits on Translational Science Proceedings*, vol. 2020, pp. 191–200, May 2020.

[25] F. Cabitza, R. Rasoini, and G. F. Gensini, "Unintended Consequences of Machine Learning in Medicine," *JAMA*, vol. 318, pp. 517–518, Aug. 2017.

[26] 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 Medicine*, vol. 15, p. e1002683, Nov. 2018. Publisher: Public Library of Science.

[27] G. Mårtensson, D. Ferreira, T. Granberg, L. Cavallin, K. Oppedal, A. Padovani, I. Rektorova, L. Bonanni, M. Pardini, M. G. Kramberger, J.-P. Taylor, J. Hort, J. Snædal, J. Kulisevsky, F. Blanc, A. Antonini, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, S. Lovestone, A. Simmons, D. Aarsland, and E. Westman, "The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study," *Medical Image Analysis*, vol. 66, p. 101714, Dec. 2020.

[28] A. Wong, E. Otles, J. P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penoza, M. Ghous, and K. Singh, "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients," *JAMA internal medicine*, vol. 181, pp. 1065–1070, Aug. 2021.

[29] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, X. Huang, Y. Xiao, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, B. Jiao, C. Cheng, Y. Zhang, A. Luo, L. Mombaerts, J. Jin, Z. Cao, S. Li, H. Xu, and Y. Yuan, "An interpretable mortality prediction model for COVID-19 patients," *Nature Machine Intelligence*, vol. 2, pp. 283–288, May 2020. Number: 5 Publisher: Nature Publishing Group.

[30] C. Kelliny, J. William, W. Riesen, F. Paccaud, and P. Bovet, "Metabolic syndrome according to different definitions in a rapidly developing country of the African region," *Cardiovascular Diabetology*, vol. 7, p. 27, Sept. 2008.

[31] C. M. Sauer, T. A. Dam, L. A. Celi, M. Faltys, M. A. A. de la Hoz, L. Adhikari, K. A. Ziesemer, A. Girbes, P. J. Thoral, and P. Elbers, "Systematic Review and Comparison of Publicly Available ICU Data Sets—A Decision Guide for Clinicians and Data Scientists," *Critical Care Medicine*, vol. 50, p. e581, June 2022.

[32] C. Sáez, N. Romero, J. A. Conejero, and J. M. García-Gómez, "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, pp. 360–364, Feb. 2021.

[33] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, "The myth of generalisability in clinical research and machine learning in health care," *The Lancet Digital Health*, vol. 2, pp. e489–e492, Sept. 2020. Publisher: Elsevier.

[34] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, Aug. 2013. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[35] Z. Huang and W. Dong, "Adversarial MACE Prediction After Acute Coronary Syndrome Using Electronic Health Records," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, pp. 2117–2126, Sept. 2019. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[36] D. Chen, S. Liu, P. Kingsbury, S. Sohn, C. B. Storlie, E. B. Habermann, J. M. Naessens, D. W. Larson, and H. Liu, "Deep learning and alternative learning strategies for retrospective real-world clinical data," *npj Digital Medicine*, vol. 2, pp. 1–5, May 2019. Number: 1 Publisher: Nature Publishing Group.

[37] J. Chu, J. Chen, X. Chen, W. Dong, J. Shi, and Z. Huang, "Knowledge-aware multicenter clinical dataset adaptation: Problem, method, and application," *Journal of Biomedical Informatics*, vol. 115, p. 103710, Mar. 2021.

[38] M. Viceconti and P. Hunter, "The Virtual Physiological Human: Ten Years After," *Annual Review of Biomedical Engineering*, vol. 18, pp. 103–123, July 2016.

[39] J. G. Chase, J.-C. Preiser, J. L. Dickson, A. Pironet, Y. S. Chiew, C. G. Pretty, G. M. Shaw, B. Benyo, K. Moeller, S. Safaei, M. Tawhai, P. Hunter, and T. Desaive, "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, p. 24, Dec. 2018.

[40] J. L. Dickson, K. W. Stewart, C. G. Pretty, M. Flechet, T. Desaive, S. Penning, B. C. Lambermont, B. Benyó, G. M. Shaw, and J. G. Chase, "Generalisability of a Virtual Trials Method for Glycaemic Control in Intensive Care," *IEEE Transactions on Biomedical Engineering*, vol. 65, pp. 1543–1553, July 2018. Conference Name: IEEE Transactions on Biomedical Engineering.

[41] S. Mistry, A. Das, S. Saffaran, N. Yehya, T. E. Scott, M. Chikhani, J. G. Laffey, J. G. Hardman, L. Camporota, and D. G. Bates, "Validation of at-the-bedside formulae for estimating ventilator driving pressure during airway pressure release ventilation using computer simulation," *Respiratory Research*, vol. 23, p. 101, Dec. 2022.

[42] S. Mistry, B. S. Brook, S. Saffaran, M. Chikhani, D. M. Hannon, J. G. Laffey, T. E. Scott, L. Camporota, J. G. Hardman, and D. G. Bates, "A computational cardiopulmonary physiology simulator accurately predicts individual patient responses to changes in mechanical ventilator settings," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2022, pp. 3261–3264, July 2022.

[43] A. Das, M. Haque, M. Chikhani, W. Wang, J. G. Hardman, and D. G. Bates, "Creating virtual ARDS patients," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2729–2732, Aug. 2016. ISSN: 1558-4615.

[44] R. E. Sherman, S. A. Anderson, G. J. Dal Pan, G. W. Gray, T. Gross, N. L. Hunter, L. LaVange, D. Marinac-Dabic, P. W. Marks, M. A. Robb, J. Shuren, R. Temple, J. Woodcock, L. Q. Yue, and R. M. Califf, "Real-World Evidence — What Is It and What Can It Tell Us?," *New England Journal of Medicine*, vol. 375, pp. 2293–2297, Dec. 2016. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMsb1609216.

[45] M. Neuville, N. El-Helali, E. Magalhaes, A. Radjou, R. Smonig, J.-F. Soubirou, G. Voiriot, A. Le Monnier, S. Ruckly, L. Bouadma, R. Sonneville, J.-F. Timsit, and B. Mourvillier, "Systematic overdosing of oxa- and cloxacillin in severe infections treated in ICU: risk factors and side effects," *Annals of Intensive Care*, vol. 7, p. 34, Mar. 2017.

[46] F. Gao and Y. Zhang, "Inotrope Use and Intensive Care Unit Mortality in Patients With Cardiogenic Shock: An Analysis of a Large Electronic Intensive Care Unit Database," *Frontiers in Cardiovascular Medicine*, vol. 8, 2021.

[47] D. Zeiberg, T. Prahlad, B. K. Nallamothu, T. J. Iwashyna, J. Wiens, and M. W. Sjoding, "Machine learning for patient risk stratification for acute respiratory distress syndrome," *PLOS ONE*, vol. 14, p. e0214465, Mar. 2019. Publisher: Public Library of Science.

[48] A. Dejam, B. E. Malley, M. Feng, F. Cismondi, S. Park, S. Samani, Z. A. Samani, D. S. Pinto, and L. A. Celi, "The effect of age and clinical circumstances on the outcome of red blood cell transfusion in critically ill patients," *Critical Care*, vol. 18, p. 487, Aug. 2014.

[49] L. D. Bos, L. R. Schouten, L. A. van Vught, M. A. Wiewel, D. S. Y. Ong, O. Cremer, A. Artigas, I. Martin-Loeches, A. J. Hoogendijk, T. van der Poll, J. Horn, N. Juffermans, C. S. Calfee, and M. J. Schultz, "Identification and validation of distinct biological phenotypes in patients with acute respiratory distress syndrome by cluster analysis," *Thorax*, vol. 72, pp. 876–883, Oct. 2017.

[50] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care," *Nature Medicine*, vol. 24, pp. 1716–1720, Nov. 2018. Number: 11 Publisher: Nature Publishing Group.

[51] M. M. Ghassemi, S. E. Richter, I. M. Eche, T. W. Chen, J. Danziger, and L. A. Celi, "A data-driven approach to optimized medication dosing: a focus on heparin," *Intensive Care Medicine*, vol. 40, pp. 1332–1339, Sept. 2014.

[52] B. L. Taylor, S. M. Selbst, and A. E. C. Shah, "Prescription writing errors in the pediatric emergency department," *Pediatric Emergency Care*, vol. 21, pp. 822–827, Dec. 2005.

[53] K. M. Hirata, A. H. Kang, G. V. Ramirez, C. Kimata, and L. G. Yamamoto, "Pediatric Weight Errors and Resultant Medication Dosing Errors in the Emergency Department," *Pediatric Emergency Care*, vol. 35, pp. 637–642, Sept. 2019.

[54] M. Ghassemi, L. A. Celi, and D. J. Stone, "State of the art review: the data revolution in critical care," *Critical Care*, vol. 19, p. 118, Dec. 2015.

[55] M. Mamdani and A. S. Slutsky, "Artificial intelligence in intensive care medicine," *Intensive Care Medicine*, vol. 47, pp. 147–149, Feb. 2021.

[56] A. Sharafoddini, J. A. Dubin, D. M. Maslove, and J. Lee, "A New Insight Into Missing Data in Intensive Care Unit Patient Profiles: Observational Study," *JMIR Medical Informatics*, vol. 7, p. e11605, Jan. 2019.

[57] A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen, "Missing data and multiple imputation in clinical epidemiological research," *Clinical Epidemiology*, vol. 9, pp. 157–166, Mar. 2017.

[58] J. H. Chen and S. M. Asch, "Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations," *New England Journal of Medicine*, vol. 376, pp. 2507–2509, June 2017. Publisher: Massachusetts Medical Society.

[59] F. Michard and J. L. Teboul, "Predictive analytics: beyond the buzz," *Annals of Intensive Care*, vol. 9, p. 46, Apr. 2019.

[60] S. E. Cochi, J. A. Kempker, S. Annangi, M. R. Kramer, and G. S. Martin, "Mortality Trends of Acute Respiratory Distress Syndrome in the United States from 1999 to 2013," *Annals of the American Thoracic Society*, vol. 13, pp. 1742–1751, Oct. 2016.

[61] K. Raymondos, T. Dirks, M. Quintel, U. Molitoris, J. Ahrens, T. Dieck, K. Johanning, D. Henzler, R. Rossaint, C. Putensen, H. Wrigge, R. Wittich, M. Ragaller, T. Bein, M. Beiderlinden, M. Sanmann, C. Rabe, J. Schlechtweg, M. Holler, F. Frutos-Vivar, A. Esteban, H. Hecker, S. Rosseau, V. von Dossow, C. Spies, T. Welte, S. Piepenbrock, and S. Weber-Carstens, "Outcome of acute respiratory distress syndrome in university and non-university hospitals in Germany," *Critical Care (London, England)*, vol. 21, p. 122, May 2017.

[62] E. Eworuke, J. M. Major, and L. I. Gilbert McClain, "National incidence rates for Acute Respiratory Distress Syndrome (ARDS) and ARDS cause-specific factors in the United States (2006–2014)," *Journal of Critical Care*, vol. 47, pp. 192–197, Oct. 2018.

[63] L. A. Huppert, M. A. Matthay, and L. B. Ware, "Pathogenesis of Acute Respiratory Distress Syndrome," *Seminars in Respiratory and Critical Care Medicine*, vol. 40, pp. 31–39, Feb. 2019.

[64] L. K. Reiss, A. Schuppert, and S. Uhlig, "Inflammatory processes during acute respiratory distress syndrome: a complex system," *Current Opinion in Critical Care*, vol. 24, pp. 1–9, Feb. 2018.

[65] P. van der Zee and D. Gommers, "Recruitment Maneuvers and Higher PEEP, the So-Called Open Lung Concept, in Patients with ARDS," *Critical Care*, vol. 23, p. 73, Mar. 2019.

[66] K. Kambas, M. M. Markiewski, I. A. Pneumatikos, S. S. Rafail, V. Theodorou, D. Konstantonis, I. Kourtzelis, M. N. Doumas, P. Magotti, R. A. DeAngelis, J. D. Lambris, and K. D. Ritis, "C5a and TNF- Up-Regulate the Expression of Tissue Factor in Intra-Alveolar Neutrophils of Patients with the Acute Respiratory Distress Syndrome," *The Journal of Immunology*, vol. 180, pp. 7368–7375, June 2008. Publisher: American Association of Immunologists Section: CELLULAR IMMUNOLOGY AND IMMUNE REGULATION.

[67] J. H. T. Bates and B. J. Smith, "Ventilator-induced lung injury and lung mechanics," *Annals of Translational Medicine*, vol. 6, p. 378, Oct. 2018.

[68] G. Bellani, J. G. Laffey, T. Pham, E. Fan, L. Brochard, A. Esteban, L. Gattinoni, F. van Haren, A. Larsson, D. F. McAuley, M. Ranieri, G. Rubenfeld, B. T. Thompson, H. Wrigge, A. S. Slutsky, A. Pesenti, LUNG SAFE Investigators, and ESICM Trials Group, "Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries," *JAMA*, vol. 315, pp. 788–800, Feb. 2016.

[69] S. Fröhlich, N. Murphy, A. Doolan, O. Ryan, and J. Boylan, "Acute respiratory distress syndrome: underrecognition by clinicians," *Journal of Critical Care*, vol. 28, pp. 663–668, Oct. 2013.

[70] G. Bellani, T. Pham, and J. G. Laffey, "Missed or delayed diagnosis of ARDS: a common and serious problem," *Intensive Care Medicine*, vol. 46, pp. 1180–1183, June 2020.

[71] J. Phua, J. R. Badia, N. K. J. Adhikari, J. O. Friedrich, R. A. Fowler, J. M. Singh, D. C. Scales, D. R. Stather, A. Li, A. Jones, D. J. Gattas, D. Hallett, G. Tomlinson, T. E. Stewart, and N. D. Ferguson, "Has Mortality from Acute Respiratory Distress Syndrome Decreased over Time?," *American Journal of Respiratory and Critical Care Medicine*, vol. 179, pp. 220–227, Feb. 2009. Publisher: American Thoracic Society - AJRCCM.

[72] N. Petrucci and C. De Feo, "Lung protective ventilation strategy for the acute respiratory distress syndrome," *The Cochrane Database of Systematic Reviews*, p. CD003844, Feb. 2013.

[73] M. B. P. Amato, M. O. Meade, A. S. Slutsky, L. Brochard, E. L. V. Costa, D. A. Schoenfeld, T. E. Stewart, M. Briel, D. Talmor, A. Mercat, J.-C. M. Richard, C. R. R. Carvalho, and R. G. Brower, "Driving pressure and survival in the acute respiratory distress syndrome," *The New England Journal of Medicine*, vol. 372, pp. 747–755, Feb. 2015.

[74] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, May 2016.

[75] H. N. Reynolds, M. McCunn, U. Borg, N. Habashi, C. Cottingham, and Y. Bar-Lavi, "Acute respiratory distress syndrome: estimated incidence and mortality rate in a 5 million-person population base," *Critical Care*, vol. 2, p. 29, Mar. 1998.

[76] C. Lam, C. F. Tso, A. Green-Saxena, E. Pellegrini, Z. Iqbal, D. Evans, J. Hoffman, J. Calvert, Q. Mao, and R. Das, "Semisupervised Deep Learning Techniques for Predicting Acute Respiratory Distress Syndrome From Time-Series Clinical Data: Model Development and Validation Study," *JMIR formative research*, vol. 5, p. e28028, Sept. 2021.

[77] C. Lam, R. Thapa, J. Maharjan, K. Rahmani, C. F. Tso, N. P. Singh, S. C. Chetty, and Q. Mao, "Multitask Learning With Recurrent Neural Networks for Acute Respiratory Distress Syndrome Prediction Using Only Electronic Health Record Data: Model Development and Validation Study," *JMIR Medical Informatics*, vol. 10, p. e36202, June 2022. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

[78] S. Le, E. Pellegrini, A. Green-Saxena, C. Summers, J. Hoffman, J. Calvert, and R. Das, "Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS)," *Journal of Critical Care*, vol. 60, pp. 96–102, Dec. 2020.

[79] ARDS Definition Task Force, V. M. Ranieri, G. D. Rubenfeld, B. T. Thompson, N. D. Ferguson, E. Caldwell, E. Fan, L. Camporota, and A. S. Slutsky, "Acute respiratory distress syndrome: the Berlin Definition," *JAMA*, vol. 307, pp. 2526–2533, June 2012.

[80] M. W. Sjoding, T. P. Hofer, I. Co, A. Courey, C. R. Cooke, and T. J. Iwashyna, "Interobserver Reliability of the Berlin ARDS Definition and Strategies to Improve the Reliability of ARDS Diagnosis," *Chest*, vol. 153, pp. 361–367, Feb. 2018.

[81] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, pp. 1521–1528, June 2011. ISSN: 1063-6919.

[82] E. A. AlBadawy, A. Saha, and M. A. Mazurowski, "Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing," *Medical Physics*, vol. 45, pp. 1150–1158, Mar. 2018.

[83] E. H. P. Pooch, P. L. Ballester, and R. C. Barros, "Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification," June 2020. arXiv:1909.01940 [cs, eess, stat].

[84] M. Barish, S. Bolourani, L. F. Lau, S. Shah, and T. P. Zanos, "External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19," *Nature Machine Intelligence*, vol. 3, pp. 25–27, Jan. 2021. Number: 1 Publisher: Nature Publishing Group.

[85] C. Dupuis, E. De Montmollin, M. Neuville, B. Mourvillier, S. Ruckly, and J. F. Timsit, "Limited applicability of a COVID-19 specific mortality prediction rule to the intensive care setting," *Nature Machine Intelligence*, vol. 3, pp. 20–22, Jan. 2021. Number: 1 Publisher: Nature Publishing Group.

[86] M. J. R. Quanjel, T. C. van Holten, P. C. Gunst-van der Vliet, J. Wielaard, B. Karakaya, M. Söhne, H. S. Moeniralam, and J. C. Grutters, "Replication of a mortality prediction model in Dutch patients with COVID-19," *Nature Machine Intelligence*, vol. 3, pp. 23–24, Jan. 2021. Number: 1 Publisher: Nature Publishing Group.

[87] J. Gallifant, J. Zhang, M. d. P. A. Lopez, T. Zhu, L. Camporota, L. A. Celi, and F. Formenti, "Artificial intelligence for mechanical ventilation: systematic review of design, reporting standards, and bias," *British Journal of Anaesthesia*, vol. 128, pp. 343–351, Feb. 2022. Publisher: Elsevier.

[88] J. H. Chen, M. Alagappan, M. K. Goldstein, S. M. Asch, and R. B. Altman, "Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets," *International Journal of Medical Informatics*, vol. 102, pp. 71–79, June 2017.

[89] A. C. Justice, K. E. Covinsky, and J. A. Berlin, "Assessing the Generalizability of Prognostic Information," *Annals of Internal Medicine*, vol. 130, pp. 515–524, Mar. 1999. Publisher: American College of Physicians.

[90] D. G. Altman and P. Royston, "What do we mean by validating a prognostic model?," *Statistics in Medicine*, vol. 19, no. 4, pp. 453–473, 2000.

[91] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Norén, Y.-C. Li, P. E. Stang, D. Madigan, and P. B. Ryan, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers," *Studies in Health Technology and Informatics*, vol. 216, pp. 574–578, 2015.

[92] J. Chen, L. Sun, C. Guo, and Y. Xie, "A fusion framework to extract typical treatment patterns from electronic medical records," *Artificial Intelligence in Medicine*, vol. 103, p. 101782, Mar. 2020.

[93] R. H. Dehejia and S. Wahba, "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *The Review of Economics and Statistics*, vol. 84, pp. 151–161, Feb. 2002.

[94] P. C. Austin, "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioral Research*, vol. 46, pp. 399–424, May 2011. Publisher: Routledge _eprint: https://doi.org/10.1080/00273171.2011.568786.

[95] D. L. Streiner and G. R. Norman, "The Pros and Cons of Propensity Scores," *Chest*, vol. 142, pp. 1380–1382, Dec. 2012.

[96] G. N. Okoli, R. D. Sanders, and P. Myles, "Demystifying propensity scores," *British Journal of Anaesthesia*, vol. 112, pp. 13–15, Jan. 2014.

[97] S. Pokharel, G. Zuccon, X. Li, C. P. Utomo, and Y. Li, "Temporal tree representation for similarity computation between medical patients," *Artificial Intelligence in Medicine*, vol. 108, p. 101900, Aug. 2020.

[98] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, pp. 139–144, Oct. 2020.

[99] T. P. A. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E. W. Steyerberg, and K. G. M. Moons, "A new framework to enhance the interpretation of external validation studies of clinical prediction models," *Journal of Clinical Epidemiology*, vol. 68, pp. 279–289, Mar. 2015. Publisher: Elsevier.

[100] P. Courrieu, "Three algorithms for estimating the domain of validity of feedforward neural networks," *Neural Networks*, vol. 7, pp. 169–174, Jan. 1994.

[101] X. Zhou and Y. Shi, "Nearest Neighbor Convex Hull Classification Method for Face Recognition," in *Computational Science – ICCS 2009* (G. Allen, J. Nabrzyski, E. Seidel, G. D. van Albada, J. Dongarra, and P. M. A. Sloot, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 570–577, Springer, 2009.

[102] R. L. Graham, "An efficient algorith for determining the convex hull of a finite planar set," *Information Processing Letters*, vol. 1, pp. 132–133, June 1972.

[103] R. V. Shesu, T. V. S. U. Bhaskar, E. P. R. Rao, M. Ravichandran, and B. V. Rao, "An improved method for quality control of in situ data from Argo floats using convex hulls," *MethodsX*, vol. 8, p. 101337, Jan. 2021.

[104] A. Kawa, M. Stahlhut, A. Berezin, E. Bock, and V. Berezin, "A simple procedure for morphometric analysis of processes and growth cones of neurons in culture using parameters derived from the contour and convex hull of the object," *Journal of Neuroscience Methods*, vol. 79, pp. 53–64, Jan. 1998.

[105] K. Tian, X. Zhao, and S. S. T. Yau, "Convex hull analysis of evolutionary and phylogenetic relationships between biological groups," *Journal of Theoretical Biology*, vol. 456, pp. 34–40, Nov. 2018.

[106] X. Zhao, K. Tian, R. L. He, and S. S. T. Yau, "Convex hull principle for classification and phylogeny of eukaryotic proteins," *Genomics*, vol. 111, pp. 1777–1784, Dec. 2019.

[107] D. O'Rourke, M. Bottema, and M. Taylor, "Sampling strategies for approximating patient variability in population-based finite element studies of total hip replacement," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 35, no. 3, p. e3168, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cnm.3168.

[108] S. D. Newsome, J. D. Yeakel, P. V. Wheatley, and M. T. Tinker, "Tools for quantifying isotopic niche space and dietary variation at the individual and population level," *Journal of Mammalogy*, vol. 93, pp. 329–341, Apr. 2012.

[109] G. Ostrouchov and N. Samatova, "On FastMap and the convex hull of multivariate data: toward fast and robust dimension reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1340–1343, Aug. 2005. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[110] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Mar. 2020. Number: 3 Publisher: Nature Publishing Group.

[111] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[112] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, "Scikit-learn: Machine Learning in Python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Nov. 2011.

[113] J. Kunze, S. Fritsch, A. Peine, O. Maaßen, G. Marx, and J. Bickenbach, "Management of ARDS: From ventilation strategies to intelligent technical support – Connecting the dots," *Trends in Anaesthesia and Critical Care*, vol. 34, pp. 50–58, Oct. 2020.

[114] B. J. Worton, "A Convex Hull-Based Estimator of Home-Range Size," *Biometrics*, vol. 51, no. 4, pp. 1206–1215, 1995. Publisher: [Wiley, International Biometric Society].

[115] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN," *ACM Transactions on Database Systems*, vol. 42, pp. 19:1–19:21, July 2017.

[116] A. M. Schweidtmann, J. M. Weber, C. Wende, L. Netze, and A. Mitsos, "Obey validity limits of data-driven models through topological data analysis and one-class classification," *Optimization and Engineering*, vol. 23, pp. 855–876, June 2022.

[117] R. J. Malak, Jr. and C. J. J. Paredis, "Using Support Vector Machines to Formalize the Valid Input Domain of Predictive Models in Systems Design Problems," *Journal of Mechanical Design*, vol. 132, Sept. 2010.

[118] E. Roach, R. R. Parker, and R. J. Malak, "An Improved Support Vector Domain Description Method for Modeling Valid Search Domains in Engineering Design Problems," pp. 741–751, American Society of Mechanical Engineers Digital Collection, June 2012.

[119] M. Quaglio, E. S. Fraga, E. Cao, A. Gavriilidis, and F. Galvanin, "A model-based data mining approach for determining the domain of validity of approximated models," *Chemometrics and Intelligent Laboratory Systems*, vol. 172, pp. 58–67, Jan. 2018.

[120] D. Tax and R. Duin, *Data domain description using support vectors.* Jan. 1999. Pages: 256.

[121] J. Syväranta, A. Lensu, T. J. Marjomäki, S. Oksanen, and R. I. Jones, "An Empirical Evaluation of the Utility of Convex Hull and Standard Ellipse Areas for Assessing Population Niche Widths from Stable Isotope Data," *PLOS ONE*, vol. 8, p. e56094, Feb. 2013. Publisher: Public Library of Science.

[122] R. Balestriero, J. Pesenti, and Y. LeCun, "Learning in High Dimension Always Amounts to Extrapolation," Oct. 2021. arXiv:2110.09485 [cs].

[123] I. Bárány and Z. Füredi, "On the shape of the convex hull of random points," *Probability Theory and Related Fields*, vol. 77, pp. 231–240, Feb. 1988.

[124] J. G. Hardman, N. M. Bedforth, A. B. Ahmed, R. P. Mahajan, and A. R. Aitkenhead, "A physiology simulator: validation of its respiratory components and its ability to predict the patient's response to changes in mechanical ventilation.," *BJA: British Journal of Anaesthesia*, vol. 81, pp. 327–332, Sept. 1998.

[125] J. B. Bassingthwaighte, "Strategies for the Physiome Project," *Annals of Biomedical Engineering*, vol. 28, pp. 1043–1058, Aug. 2000.

[126] E. J. Crampin, M. Halstead, P. Hunter, P. Nielsen, D. Noble, N. Smith, and M. Tawhai, "Computational physiology and the physiome project: Computational physiology and the physiome project," *Experimental Physiology*, vol. 89, pp. 1–26, Jan. 2004.

[127] N. Kusch, L. Turnhoff, and A. Schuppert, "Modeling from Molecule to Disease and Personalized Medicine," in *Handbook of Biomarkers and precision Medicine*, pp. 245–250, CRC Press, 2019.

[128] L. Turnhoff, N. Kusch, and A. Schuppert, ""Big Data and Dynamics"—The Mathematical Toolkit Towards Personalized Medicine," in *Patterns of Dynamics* (P. Gurevich, J. Hell, B. Sandstede, and A. Scheel, eds.), Springer Proceedings in Mathematics & Statistics, (Cham), pp. 338–369, Springer International Publishing, 2017.

[129] J. A. Weis, M. I. Miga, L. R. Arlinghaus, X. Li, V. Abramson, A. B. Chakravarthy, P. Pendyala, and T. E. Yankeelov, "Predicting the Response of Breast Cancer to Neoadjuvant Therapy Using a Mechanically Coupled Reaction-Diffusion Model," *Cancer Research*, vol. 75, pp. 4697–4707, Nov. 2015.

[130] V. B. Shim, P. J. Hunter, P. Pivonka, and J. W. Fernandez, "A Multiscale Framework Based on the Physiome Markup Languages for Exploring the Initiation of Osteoarthritis at the Bone–Cartilage Interface," *IEEE Transactions on Biomedical Engineering*,

vol. 58, pp. 3532–3536, Dec. 2011. Conference Name: IEEE Transactions on Biomedical Engineering.

[131] J. W. Fernandez, V. B. Shim, and P. J. Hunter, "Integrating degenerative mechanisms in bone and cartilage: A multiscale approach," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6616–6619, Aug. 2012. ISSN: 1558-4615.

[132] M. Tawhai, A. R. Clark, G. M. Donovan, and K. S. Burrowes, "Computational Modeling of Airway and Pulmonary Vascular Structure and Function: Development of a "Lung Physiome"," *Critical Reviews™ in Biomedical Engineering*, vol. 39, no. 4, 2011. Publisher: Begel House Inc.

[133] M. Kim, R. Bordas, W. Vos, R. A. Hartley, C. E. Brightling, D. Kay, V. Grau, and K. S. Burrowes, "Dynamic flow characteristics in normal and asthmatic lungs," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 31, no. 12, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cnm.2730.

[134] K. S. Burrowes, J. De Backer, R. Smallwood, P. J. Sterk, I. Gut, R. Wirix-Speetjens, S. Siddiqui, J. Owers-Bradley, J. Wild, D. Maier, and C. Brightling, "Multi-scale computational models of the airways to unravel the pathophysiological mechanisms in asthma and chronic obstructive pulmonary disease (AirPROM)," *Interface Focus*, vol. 3, p. 20120057, Apr. 2013. Publisher: Royal Society.

[135] J.-S. Kim, N. V. Valeyev, I. Postlethwaite, P. Heslop-Harrison, K.-H. Cho, and D. G. Bates, "Analysis and extension of a biochemical network model using robust control theory," *International Journal of Robust and Nonlinear Control*, vol. 20, pp. 1017–1026, Nov. 2009.

[136] K. S. Burrowes, T. Doel, and C. Brightling, "Computational modeling of the obstructive lung diseases asthma and COPD," *Journal of Translational Medicine*, vol. 12, p. S5, Nov. 2014.

[137] W. Wang, A. Das, T. Ali, O. Cole, M. Chikhani, M. Haque, J. G. Hardman, and D. G. Bates, "Can computer simulators accurately represent the pathophysiology of individual COPD patients?," *Intensive Care Medicine Experimental*, vol. 2, p. 23, Dec. 2014.

[138] S. A. Niederer, Y. Aboelkassem, C. D. Cantwell, C. Corrado, S. Coveney, E. M. Cherry, T. Delhaas, F. H. Fenton, A. V. Panfilov, P. Pathmanathan, G. Plank, M. Riabiz, C. H. Roney, R. W. dos Santos, and L. Wang, "Creation and application of virtual patient cohorts of heart models," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 378, p. 20190558, June 2020. Publisher: Royal Society.

[139] A. Das, L. Camporota, J. G. Hardman, and D. G. Bates, "What links ventilator driving pressure with survival in the acute respiratory distress syndrome? A computational study," *Respiratory Research*, vol. 20, p. 29, Dec. 2019.

[140] J.-L. Vincent, J. B. Hall, and A. S. Slutsky, "Ten big mistakes in intensive care medicine," *Intensive Care Medicine*, vol. 41, pp. 505–507, Mar. 2015.

[141] M. Viceconti, A. Henney, and E. Morley-Fletcher, "In silico clinical trials: how computer simulation will transform the biomedical industry," *International Journal of Clinical Trials*, vol. 3, pp. 37–46, May 2016.

[142] A. Das, O. Cole, M. Chikhani, W. Wang, T. Ali, M. Haque, D. G. Bates, and J. G. Hardman, "Evaluation of lung recruitment maneuvers in acute respiratory distress syndrome using computer simulation," *Critical Care*, vol. 19, p. 8, Dec. 2015.

[143] S. Saffaran, A. Das, J. G. Hardman, N. Yehya, and D. G. Bates, "High-fidelity computational simulation to refine strategies for lung-protective ventilation in paediatric acute respiratory distress syndrome," *Intensive Care Medicine*, vol. 45, pp. 1055–1057, July 2019.

[144] M. Viceconti, C. Cobelli, T. Haddad, A. Himes, B. Kovatchev, and M. Palmer, "In silico assessment of biomedical products: The conundrum of rare but not so rare events in two case studies," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 231, pp. 455–466, May 2017. Publisher: IMECHE.

[145] A. Ben-Tal, "Simplified models for gas exchange in the human lungs," *Journal of Theoretical Biology*, vol. 238, pp. 474–495, Jan. 2006.

[146] A. Das, M. Haque, M. Chikhani, W. Wang, T. Ali, O. Cole, J. G. Hardman, and D. G. Bates, "Development of an integrated model of cardiovascular and pulmonary physiology for the evaluation of mechanical ventilation strategies," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, (Milan), pp. 5319–5322, IEEE, Aug. 2015.

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

[148] A. Das, Z. Gao, P. P. Menon, J. G. Hardman, and D. G. Bates, "A systems engineering approach to validation of a pulmonary physiology simulator for clinical applications," *Journal of The Royal Society Interface*, vol. 8, pp. 44–55, Jan. 2011.

[149] *MATLAB version 9.10.0.1613233 (R2021a).* Natick, Massachusetts: The Mathworks, Inc., 2021.

[150] L. Weaver, A. Das, S. Saffaran, N. Yehya, T. E. Scott, M. Chikhani, J. G. Laffey, J. G. Hardman, L. Camporota, and D. G. Bates, "High risk of patient self-inflicted lung injury in COVID-19 with frequently encountered spontaneous breathing patterns: a computational modelling study," *Annals of Intensive Care*, vol. 11, p. 109, Dec. 2021.

[151] S. Saffaran, A. Das, J. G. Laffey, J. G. Hardman, N. Yehya, and D. G. Bates, "Utility of Driving Pressure and Mechanical Power to Guide Protective Ventilator Settings in Two Cohorts of Adult and Pediatric Patients With Acute Respiratory Distress Syndrome: A Computational Investigation," *Critical Care Medicine*, vol. Publish Ahead of Print, Apr. 2020.

[152] A. Das, S. Saffaran, M. Chikhani, T. E. Scott, M. Laviola, N. Yehya, J. G. Laffey, J. G. Hardman, and D. G. Bates, "In Silico Modeling of Coronavirus Disease 2019 Acute

Respiratory Distress Syndrome: Pathophysiologic Insights and Potential Management Implications," *Critical Care Explorations*, vol. 2, p. e0202, Sept. 2020.

[153] W. Wang, "Computational simulation indicates that moderately high-frequency ventilation can allow safe reduction of tidal volumes and airway pressures in ARDS patients," p. 12, 2015.

[154] P. BAJPAI and M. Kumar, "Genetic Algorithm - an Approach to Solve Global Optimization Problems," *Indian Journal of Computer Science and Engineering*, vol. 1, pp. 199–206, Oct. 2010.

[155] G. Mols, H.-J. Priebe, and J. Guttmann, "Alveolar recruitment in acute lung injury," *BJA: British Journal of Anaesthesia*, vol. 96, pp. 156–166, Feb. 2006.

[156] A. Jabbari, E. Alijanpour, P. Amri Maleh, and B. Heidari, "Lung protection strategy as an effective treatment in acute respiratory distress syndrome," *Caspian Journal of Internal Medicine*, vol. 4, no. 1, pp. 560–563, 2013.

[157] C. Guérin, P. Beuret, J. M. Constantin, G. Bellani, P. Garcia-Olivares, O. Roca, J. H. Meertens, P. A. Maia, T. Becher, J. Peterson, A. Larsson, M. Gurjar, Z. Hajjej, F. Kovari, A. H. Assiri, E. Mainas, M. S. Hasan, D. R. Morocho-Tutillo, L. Baboi, J. M. Chrétien, G. François, L. Ayzac, L. Chen, L. Brochard, A. Mercat, and investigators of the APRONET Study Group, the REVA Network, the Réseau recherche de la Société Française d'Anesthésie-Réanimation (SFAR-recherche) and the ESICM Trials Group, "A prospective international observational prevalence study on prone positioning of ARDS patients: the APRONET (ARDS Prone Position Network) study," *Intensive Care Medicine*, vol. 44, pp. 22–37, Jan. 2018.

[158] L. Puybasset, P. Cluzel, P. Gusman, P. Grenier, F. Preteux, and J. J. Rouby, "Regional distribution of gas and tissue in acute respiratory distress syndrome. I. Consequences for lung morphology. CT Scan ARDS Study Group," *Intensive Care Medicine*, vol. 26, pp. 857–869, July 2000.

[159] J. J. Rouby, L. Puybasset, P. Cluzel, J. Richecoeur, Q. Lu, and P. Grenier, "Regional distribution of gas and tissue in acute respiratory distress syndrome. II. Physiological correlations and definition of an ARDS Severity Score. CT Scan ARDS Study Group," *Intensive Care Medicine*, vol. 26, pp. 1046–1056, Aug. 2000.

[160] L. Puybasset, P. Gusman, J.-C. Muller, P. Cluzel, P. Coriat, J.-J. Rouby, and a. C. S. A. S. G. the Group, "Regional distribution of gas and tissue in acute respiratory distress syndrome. III. Consequences for the effects of positive end-expiratory pressure," *Intensive Care Medicine*, vol. 26, pp. 1215–1227, Sept. 2000.

[161] S. Nyrén, P. Radell, S. E. Lindahl, M. Mure, J. Petersson, S. Larsson, H. Jacobsson, and A. Sánchez-Crespo, "Lung Ventilation and Perfusion in Prone and Supine Postures with Reference to Anesthetized and Mechanically Ventilated Healthy Volunteers," *Anesthesiology*, vol. 112, pp. 682–687, Mar. 2010.

[162] A. C. Henderson, R. C. Sá, R. J. Theilmann, R. B. Buxton, G. K. Prisk, and S. R. Hopkins, "The gravitational distribution of ventilation-perfusion ratio is more uniform in prone than supine posture in the normal human lung," *Journal of Applied Physiology*, vol. 115, pp. 313–324, Aug. 2013. Publisher: American Physiological Society.

[163] R. A. McCahon, M. O. Columb, R. P. Mahajan, and J. G. Hardman, "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," *British Journal of Anaesthesia*, vol. 101, pp. 358–365, Sept. 2008.

[164] A. Costa and G. Nannicini, "RBFOpt: an open-source library for black-box optimization with costly function evaluations," *Mathematical Programming Computation*, vol. 10, pp. 597–629, Dec. 2018.

[165] A. Charles and H. Warren, *Derivative-Free and Blackbox Optimization*. No. 2197-1773, Springer Cham, 1 ed., Dec. 2017.

[166] H.-M. Gutmann, "A Radial Basis Function Method for Global Optimization," *Journal of Global Optimization*, vol. 19, pp. 201–227, Mar. 2001.

[167] P. P. Menon, I. Postlethwaite, S. Bennani, A. Marcos, and D. G. Bates, "Robustness analysis of a reusable launch vehicle flight control law," *Control Engineering Practice*, vol. 17, pp. 751–765, July 2009.

[168] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer Texts in Statistics, New York, NY: Springer, 2004.

[169] A. Das, M. Haque, M. Chikhani, O. Cole, W. Wang, J. G. Hardman, and D. G. Bates, "Hemodynamic effects of lung recruitment maneuvers in acute respiratory distress syndrome," *BMC Pulmonary Medicine*, vol. 17, p. 34, Feb. 2017.

[170] P. J. Hantzidiamantis and E. Amaro, "Physiology, Alveolar to Arterial Oxygen Gradient," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2022.

[171] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," *Machine Learning*, vol. 52, pp. 91–118, July 2003.

[172] I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier, "Enrichment or depletion of a GO category within a class of genes: which test?," *Bioinformatics (Oxford, England)*, vol. 23, pp. 401–407, Feb. 2007.

[173] 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.

[174] K. Komiya, T. Akaba, Y. Kozaki, J.-i. Kadota, and B. K. Rubin, "A systematic review of diagnostic methods to differentiate acute lung injury/acute respiratory distress syndrome from cardiogenic pulmonary edema," *Critical Care*, vol. 21, p. 228, Aug. 2017.

[175] B. Fiedler and A. Schuppert, "Local identification of scalar hybrid models with tree structure," *IMA Journal of Applied Mathematics*, vol. 73, pp. 449–476, June 2008.

[176] A. Schuppert and T. Mrziglod, "Hybrid Model Identification and Discrimination with Practical Examples from the Chemical Industry," in *Hybrid Modeling in Process Industries*, CRC Press, 2018. Num Pages: 26.

[177] M. E. Samadi, S. Kiefer, S. J. Fritsch, J. Bickenbach, and A. Schuppert, "A training strategy for hybrid models to break the curse of dimensionality," *PLOS ONE*, vol. 17, p. e0274569, Sept. 2022. Publisher: Public Library of Science.

[178] K. Merkelbach, A. M. Schweidtmann, Y. Müller, P. Schwoebel, A. Mhamdi, A. Mitsos, A. Schuppert, T. Mrziglod, and S. Schneckener, "HybridML: Open source platform for hybrid modeling," *Computers & Chemical Engineering*, vol. 160, p. 107736, Apr. 2022.

[179] A. Procopio, S. De Rosa, F. Montefusco, G. Canino, A. Merola, J. Sabatino, C. Critelli, C. Indolfi, F. Amato, and C. Cosentino, "Analysis of a Cardiac-Necrosis-Biomarker Release in Patients with Acute Myocardial Infarction via Nonlinear Mixed-Effects Models," *Applied Sciences*, vol. 12, p. 13038, Jan. 2022. Number: 24 Publisher: Multidisciplinary Digital Publishing Institute.

[180] D. Fey, M. Halasz, D. Dreidax, S. P. Kennedy, J. F. Hastings, N. Rauch, A. G. Munoz, R. Pilkington, M. Fischer, F. Westermann, W. Kolch, B. N. Kholodenko, and D. R. Croucher, "Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients," *Science Signaling*, vol. 8, pp. ra130–ra130, Dec. 2015. Publisher: American Association for the Advancement of Science.

[181] I. Kolyshkina and S. Simoff, "Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach," *Frontiers in Big Data*, vol. 4, 2021.

Appendix

A.1 List of variables assessed in the ICU which were used in this study.

Overall, 83 diagnostic variables routinely assessed in the ICU including 7 biometric variables were used in this study. Additionally, data on drug administration of 23 drugs were used.

Vital signs:

- Heart rate
- Peripheral oxygen saturation (SpO_2)
- Systolic arterial pressure (SAP)
- Mean arterial pressure (MAP)
- Diastolic arterial pressure (DAP)
- Central venous pressure
- Systolic pulmonary arterial pressure
- Mean pulmonary arterial pressure
- Diastolic pulmonary arterial pressure
- Body temperature
- 24h urine output
- 24h fluid balance
- Pulmonary artery occluded pressure
- Extravascular lung water index
- Global end-diastolic volume index
- Cardiac output (bolus)
- Cardiac index (bolus)
- Cardiac output (continuous)

- Cardiac index (continuous)
- Systemic vascular resistance index
- Pulmonary vascular resistance index
- Stroke volume (bolus)
- Stroke volume index (bolus)
- Stroke volume (continuous)
- Stroke volume index (continuous)
- Extracorporeal blood flow
- Extracorporeal gas flow (O_2)
- Extracorporeal gas composition

Ventilatory settings:

- Respiratory rate
- Respiratory rate (spontaneous)
- Tidal volume
- Tidal volume per ideal body weight
- Tidal volume (spontaneous)
- End-inspiratory pressure (P_{EI})
- Positive end-expiratory pressure (PEEP)
- Driving pressure (deltaP)
- Fraction of inspired oxygen (FiO_2)
- Fraction of expired oxygen (FeO_2)
- Inspiration : Expiration ratio (I:E)
- Pulmonary compliance

- Inhaled nitric oxide
- End-tidal carbon dioxide ($etCO_2$)

Blood gas analysis variables:

- pH (arterial)
- $PaCO_2$
- PaO_2
- SaO_2
- PaO_2/FiO_2 ratio (P/F ratio; Horowitz index)
- Base excess (arterial)
- Bicarbonate (arterial)
- Lactate (arterial)
- Central venous oxygen saturation ($ScvO_2$)

Laboratory variables:

- Albumin
- Alanine transaminase (ALT)
- Amylase
- Aspartate transaminase (AST)
- Bilirubin
- Brain natriuretic peptide (BNP)
- C-Reactive Protein (CRP)
- Creatine kinase
- Creatine kinase-MB
- Creatinine

- D-dimers
- Haematocrit
- Haemoglobin (Hb)
- International normalized ratio (INR)
- Interleukin-6
- Lactate dehydrogenase (LDH)
- Leukocytes
- Lymphocytes
- Lipase
- NT-pro brain natriuretic peptide (NT-pro BNP)
- Procalcitonin (PCT)
- Platelets
- Partial thromboplastin time (PTT)
- Troponin
- Urea

Biometrics:

- Height
- Weight
- Age
- Gender
- ARDSnet ideal body weight (StdWeightARDS)
- StdWeightARDS (Female) = 45.5 + 0.91 (Height [cm] - 152.4)
- StdWeightARDS (Male) = 50.0 + 0.91 (Height [cm] - 152.4)

Medications:

- Dobutamine intravenous continuous
- Epinephrine intravenous continuous
- Norepinephrine intravenous continuous
- Vasopressin intravenous continuous
- Milrinone intravenously continuous
- Levosimendan intravenously continuous
- Propofol intravenously continuous
- Midazolam intravenous continuous
- Clonidine intravenously continuous
- Dexmedetomidine intravenous continuous
- Ketanest intravenous continuous
- Isoflurane inhalation
- Sevoflurane inhalation
- Sufentanil intravenous continuous
- Fentanyl intravenous continuous
- Morphine intravenous continuous
- Rocuronium intravenous bolus
- Furosemide intravenous continuous
- Hydrocortisone intravenously by bolus
- Prednisolone intravenously by bolus
- Dexamethasone intravenous by bolus
- Terlipressin intravenous bolus
- Fludrocortisone peroral boluswise

A.2 List of comorbidities associated with ARDS with dictionaries of correpsonding ICD-9 and ICD-10 codes.

Overall, 22 comorbidities were assessed using predefined dictionaries of ICD-10 and ICD-9 codes.

Comorbidity	**ICD-9 Codes**	**ICD-10 Codes**
Aspiration	E8794, E8705, 99732	T17, T17.8, T17.9, J95.4, O89.0
Drowning	9941	T75.1
Burn trauma	941, 942, 943, 944, 945, 946, 947, 948, 949, E89, E98	T20, T21, T22, T23, T24, T25, T26, T27, T28, T30, T31, T32
Chronic heart failure	428	I50, I25, I27, I11.0, I13.0, I05, I06, I07, I08, I09, I31.0, I31.1, I34, I35, I36, I37, I42, I43, I51.5
Chronic liver failure	572	K72.1, K74.6, K70.3, K72.0, K71.7, K74.3, K74.4, K74.5, K76.1, K73.0, K73.2, K73.9, K70.4, K71.0, K74.0, K72.9, B18.0, B18.1, B18.2
Chronic renal failure	585, 586	N18
COVID-19	NaN	U07.1
Diabetes mellitus	250, 249	E10, E11, E13, E14
Drug overdose	969, 971, 972, 973, 974, 975, 976, 977, 978, 979, 980, 981, 982	T40.0, T40.1, T40.4, T43, T44, T45, T46, T48, T50, T51, T52, F11.0, F11.1, F11.2, T40.2, T40.3, T50.9
Fettembolie	9581	T79.1
Hematologic neoplasm	200, 205, 206, 206, 207, 201	C81, C82, C83, C84, C85, C86, C88, C90, C91, C92, C93, C94, C95, C96
Immunosuppression	7953, 279	D80, D81.1, D82.0, D83.8, D84.1, D86, D89, D90
Inhalation	E911, E912, 506, 507, 508, E869	J68, T27.1, T27.3, T27.5, T27.7, T59
Liver failure	NaN	K72, K70, K71
Pankreatitis	5770	K85, K85.0, K85.00, K85.01, K85.1, K85.10, K85.11, K85.2, K85.20, K85.21, K85.3, K85.30, K85.31, K85.8, K85.80, K85.81, K85.9, K85.90, K85.91, K86.1

Pneumonia	322, 3220, 3221, 3222, 3229, 11505, 11515, 11595, 480, 4800, 4801, 4802, 4803, 4808, 4809, 481, 482, 4820, 4821, 4822, 4823, 48230, 48231, 48232, 48239, 4824, 48240, 48241, 48242, 48249, 4828, 48281, 48282, 48283, 48284, 48289, 4829, 483, 4830, 4831, 4838, 484, 4841, 4843, 4845, 4846, 4847, 4848, 485, 486, 5171, 99731, 99732	J10.0, J11.0, J12.0, J12.1, J12.2, J12.3, J12.8, J12.9, J13, J14, J15.0, J15.1, J15.2, J15.3, J15.4, J15.5, J15.6, J15.7, J15.8, J15.9, J16.0, J16.8, J17, J17.0, J17.1, J17.2, J17.3, J17.8, J18.0, J18.1, J18.2, J18.8, J18.9, J69.0, J69.1, J69.8, A48.1
Pulmonary vasculitis	4464	M30.1, M31.3, M31.7, D89.1, D69.0
Renal failure	NaN	N19
Sepsis	9959, 99590, 99591, 99592, 99593, 99590, 99594	A32.7, A39.1, A39.2, A40, A41, R65, R65.0, R65.1, R65.2, R65.3, R65.9, R57.2
Transfusion-related acute lung injury (TRALI)	5187	T80, T80.8, J95.88

Thoraxtrauma	861, 8610, 86100, 86101, 86102, 86103, 8611, 86110, 86111, 86112, 86113, 8612, 86120, 86121, 86122, 8613, 86130, 86131, 86132	S20, S20.0, S20.1, S20.11, S20.14, S20.2, S20.3, S20.31, S20.4, S20.41, S20.8, S20.81, S20.85, S21, S21.0, S21.1, S21.2, S21.8, S21.80, S21.86, S21.87, S21.83, S21.84, S21.85, S21.9, S22, S22.0, S22.00, S22.01, S22.02, S22.03, S22.04, S22.05, S22.06, S22.1, S22.3, S22.31, S22.32, S22.4, S22.40, S22.41, S22.42, S22.43, S22.44, S22.5, S22.8, S23, S23.0, S23.1, S23.11, S23.12, S23.16, S23.2, S23.3, S24, S24.0, S24.1, S24.10, S24.11, S24.12, S24.2, S24.7, S24.71, S24.72, S24.73, S24.74, S24.75, S24.76, S24.77, S25, S25.0, S25.1, S25.2, S25.3, S25.4, S25.5, S25.8, S25.88, S26, S26.0, S26.81, S26.82, S26.88, S26.9, S27, S27.0, S27.1, S27.2, S27.3, S27.31, S27.32, S27.38, S27.4, S27.5, S27.6, S27.8, S27.81, S27.82, S27.88, S29, S29.0, S29.7, S29.8, S29.9
Transfusion	99961, 99962, 99963, 99971, 99972, 99973, 99976, 99977, 99978, 99989, 99983, 99980, 99985, 99984	NaN

Table A.1: List of comorbidities associated with ARDS with dictionaries of correpsonding ICD-9 and ICD-10 codes. Overall, 22 comorbidities were assessed using predefined dictionaries of ICD-10 and ICD-9 codes. NaN - no suitable ICD codes could be found or comorbidity was not assessed in all datasets.

A.3 List of ICU variables which are needed to fully parameterize the simulator and define a virtual patient.

- Base excess arterial (BEa)
- Cardiac output (CO)
- Fraction of inspired oxygen (FiO_2)
- Bicarbonate (arterial) (HCO_3a)
- Haemoglobin (Hb)
- Inspiration : Expiration ratio (I:E)
- Positive end-expiratory pressure (PEEP)
- $PaCO_2$
- PaO_2
- SaO_2
- SvO_2
- Body temperature (T)
- Respiratory rate (VentRate)
- Tidal volume (Vt)
- pH arterial (pHa)
- End-inspiratory pressure (P_{EI})
- Anatomical shunt (anatShunt)
- Respiratory quotient (RQ)
- Metabolic rate of O_2 (VO_2)
- Anatomical deadspace volume (VDphys)

A.4 List of variables used for the CH analysis and for classification of ARDS on the first day in ICU.

Overall, medians of 16 diagnostic variables routinely assessed in the ICU over the first day were used. Additionally, biometrics were used (5 variables).

- Horowitz index
- Respiratory rate
- Tidal volume
- FiO_2
- PEEP
- PaO_2
- Lactate arterial
- Bicarbonate arterial
- SpO_2
- Heart rate
- Leukocytes
- Platelets
- Urea
- Creatinine
- Haemoglobin
- Partial thromboplastin time (PTT)
- Height
- Weight
- Age

- StdWeightARDS
- Gender

A.5 CH coverages for all features.

Feature	**MIMIC covered by Hosp A**	**MIMIC covered by Hosp B**	**MIMIC covered by Hosp C**
PaO_2	0.974	0.690	0.794
Platelets	0.998	0.975	0.979
FiO_2	0.995	0.977	0.963
SpO_2	0.998	0.976	0.977
Tidal volume	0.787	0.982	0.973
Bicarbonate arterial	0.994	0.939	0.972
Horowitz index	0.992	0.963	0.952
Urea	0.993	0.971	0.971
Respiratory rate	0.996	0.963	0.978
Heart rate	0.996	0.978	0.983
Haemoglobin	0.995	0.967	0.971
Lactate arterial	0.999	0.991	0.992
Leukocytes	0.997	0.983	0.977
Creatinine	0.990	0.969	0.959
PEEP	0.992	0.966	0.907
PTT	0.978	0.903	0.959

Table A.2: CH coverages for all features. MIMIC data covered by other hospitals.

A.6 Physiologically meaningful ranges for parameters of the simulator.

Parameter	Lower threshold	Upper threshold
CO, ml/min	3400.00	11800.00
FiO_2, unitless	0.25	0.71
Haemoglobin, g/l	73.00	125.00
I:E, unitless	0.17	0.50
$paCO_2$, kPa	3.93	8.01
paO_2, kPa	7.89	20.30
PEEP, cmH_2O	4.46	12.24
T, °C	35.80	38.40
VentRate, unitless	10.00	30.00
Vt, ml	287.20	488.92
HCO_3a, mmol/l	20.00	34.90
BEa, mmol/l	-4.70	9.60
P_{EI}, cmH2O	13.25	31.60
SaO_2, unitless	0.90	0.99
pHa, unitless	7.25	7.53
SvO_2, unitless	0.40	0.90
sR, kPa×min/ml	-40.00×10^{-6}	0.00
inR, kPa×min/ml	0.00	3.00×10^{-4}
sVR, dynes/sec/cm^{-5}	0.00	312.00
inVR, dynes/sec/cm^{-5}	0.00	3760.00
n_{cc}, unitless	0.00	80.00
anatShunt, unitless	0.01	0.30
RQ, unitless	0.60	1.00
VO_2, ml/min	150.00	550.00
VDphys, ml	50.00	150.00

Table A.3: Physiologically meaningful ranges for parameters of the simulator. Lower and upper thresholds were identified based on existing literature in case of parameters missing in the data and on distributions of values in underlying data for parameters present in the data. Parameters fully or partially missing in the data are given in bold.

A.7 Parameters used as model-based filtered data.

Parameters used as model-based filtered data. These were found in the optimization procedure in both time windows (labeled with *) or calculated based on simulator outputs (labeled with †) for each of the patients. For each of the patients these parameters comprised model-based filtered data consisting of 18 features. Values of these parameters were used as features in the clustering procedure.

- anatomical shunt (anatShunt)*
- respiratory quotient (RQ)*
- anatomical deadspace volume (VDphys)*
- metabolic rate of O_2 (VO_2)*
- slope of vascular resistance (sVR)*
- intercept of vascular resistance (inVR)*
- slope of flow resistance (sR)*
- intercept of flow resistance (inR)*
- number of closed compartments in window 1 (n_{cc1})*
- number of closed compartments in window 2 (n_{cc2})*
- fitting residual in window 1†
- fitting residual in window 2†
- lung ventilation in window 1†
- ratio of shunted blood in window 1†
- lung ventilation in window 2†
- ratio of shunted blood in window 2†
- increase in number of closed compartments (n_{cc2} - n_{cc1})†
- volume of shunted blood through closed compartments in window 2†

A.8 Features extracted from original measured data.

Mean values of following variables in time windows 1 and 2 (before and after suspected ARDS onset respectively) were calculated and used as features in the clustering on original measured data. Additionally difference in Horowitz index between window 1 and window 2 was used as a feature (Horowitz drop).

- Base excess arterial (BEa)
- Fraction of inspired oxygen (FiO_2)
- Bicarbonate (arterial) (HCO_3a)
- Haemoglobin (Hb)
- Positive end-expiratory pressure (PEEP)
- $PaCO_2$
- PaO_2
- SaO_2
- Body temperature (T)
- Respiratory rate (VentRate)
- Tidal volume (Vt)
- pH arterial (pHa)
- End-inspiratory pressure (P_{EI})
- PaO_2/FiO_2 ratio (P/F ratio; Horowitz index)
- Alveolar–arterial gradient (A-aO_2; A–a gradient)
- Horowitz drop

A.9 Results of enrichment analysis in clusters discovered in clustering on original measured data.

Cluster ID	Condition/Hospital	p-value	Prevalence cluster	Prevalence population
1	Hosp A	1.164×10^{-12}	0.746	0.464
1	Other disorders of the nervous system	6.961×10^{-06}	0.500	0.328
1	Cerebrovascular diseases	5.595×10^{-04}	0.523	0.390
1	Other disorders of eye and adnexa	1.212×10^{-03}	0.138	0.068
2	Hosp A	2.529×10^{-08}	0.679	0.464
2	Persons encountering health services for specific procedures and health care	4.250×10^{-05}	0.723	0.570
2	Hosp D	5.811×10^{-05}	0.234	0.124
3	Hosp F	2.190×10^{-34}	0.466	0.123
3	Hosp E	3.234×10^{-12}	0.291	0.109
3	Occupied and unoccupied key numbers	2.223×10^{-07}	0.500	0.317
3	Functional impairment	3.545×10^{-07}	0.291	0.147
3	Persons encountering health services for examination and investigation	2.650×10^{-05}	0.514	0.363
3	Codes for special purposes	5.548×10^{-05}	0.878	0.758
3	Injuries to the shoulder and upper arm	3.947×10^{-04}	0.088	0.034
3	Obesity and other hyperalimentation	1.275×10^{-03}	0.149	0.080
3	Thoraxtrauma	2.002×10^{-03}	0.189	0.114
3	Injuries to the thorax	2.002×10^{-03}	0.189	0.114
3	Injuries to the head	2.940×10^{-03}	0.243	0.162
4	Hosp G	5.043×10^{-49}	0.636	0.180
4	Persons encountering health services for examination and investigation	8.975×10^{-08}	0.552	0.363
4	Occupied and unoccupied key numbers	1.033×10^{-05}	0.468	0.317
4	Other diseases of intestines	3.647×10^{-05}	0.318	0.196
4	Chronic lower respiratory diseases	7.238×10^{-05}	0.201	0.107
5	Renal failure	7.680×10^{-13}	0.792	0.454
5	Aplastic and other anaemias	1.696×10^{-11}	0.958	0.697
5	Diseases of liver	3.956×10^{-09}	0.396	0.164
5	Liver Failure	5.741×10^{-09}	0.333	0.124
5	Diseases of the genitourinary system	2.450×10^{-08}	0.865	0.621

5	Diseases of the blood and blood-forming organs	2.526×10^{-08}	0.969	0.775
5	Sepsis	4.334×10^{-08}	0.802	0.550
5	General symptoms and signs	5.499×10^{-08}	0.917	0.699
5	Diseases of the digestive system	9.884×10^{-07}	0.698	0.467
5	Coagulation defects, purpura and other haemorrhagic conditions	3.379×10^{-06}	0.760	0.546
5	Other bacterial diseases	4.122×10^{-06}	0.740	0.525
5	Other diseases of the digestive system	1.149×10^{-05}	0.250	0.106
5	Mycoses	7.332×10^{-05}	0.354	0.196
5	Complications of surgical and medical care, not elsewhere classified	8.169×10^{-05}	0.552	0.369
5	Suppurative and necrotic conditions of lower respiratory tract	3.792×10^{-04}	0.094	0.027
5	Hosp A	5.660×10^{-04}	0.625	0.464
5	Diseases of oesophagus, stomach and duodenum	8.531×10^{-04}	0.281	0.159
5	Soft tissue disorders	1.449×10^{-03}	0.156	0.070
5	Diseases of the skin and subcutaneous tissue	1.532×10^{-03}	0.438	0.299
5	Diseases of peritoneum	1.996×10^{-03}	0.188	0.095
5	Other respiratory diseases principally affecting the interstitium	2.685×10^{-03}	0.229	0.129
5	Certain infectious and parasitic diseases	4.320×10^{-03}	0.885	0.781
5	Other disorders of the skin and subcutaneous tissue	4.929×10^{-03}	0.365	0.249
5	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	5.172×10^{-03}	0.948	0.865
5	Infectious agents with resistance to certain antibiotics or chemotherapeutic agents	6.836×10^{-03}	0.333	0.226
5	Glomerular diseases	7.804×10^{-03}	0.094	0.039

Table A.4: Hospitals and clinical conditions which are over-represented in the discovered clusters. Results for clustering on original measured data are shown.

A.10 Results of enrichment analysis in clusters discovered in clustering on model-based filtered data.

Cluster ID	Condition/Hospital	p-value	Prevalence cluster	Prevalance population
1	-	-	-	-
2	Hosp A	1.212×10^{-06}	0.632	0.464
3	Other respiratory diseases principally affecting the interstitium	8.030×10^{-05}	0.269	0.129
3	ARDS	1.549×10^{-04}	0.247	0.118
4	Hosp A	1.308×10^{-08}	0.721	0.464
4	Other bacterial diseases	7.838×10^{-04}	0.673	0.525
4	Aplastic and other anaemias	1.040×10^{-03}	0.827	0.697
4	Sepsis	1.165×10^{-03}	0.692	0.550
5	Occupied and unoccupied key numbers	4.514×10^{-14}	0.640	0.317
5	Persons encountering health services for examination and investigation	1.129×10^{-11}	0.658	0.363
5	Hosp G	6.872×10^{-07}	0.360	0.180
5	Codes for special purposes	1.235×10^{-05}	0.910	0.758
5	Hosp D	4.379×10^{-05}	0.252	0.124

Table A.5: Hospitals and clinical conditions which are over-represented in the discovered clusters. Results for clustering on model-based filtered data shown. In cluster 1 no significant enrichment was found.