UNIVERSITY OF CALIFORNIA SAN DIEGO

Using Machine Learning to Elucidate Disease Heterogeneity and Improve Prognostication in Patients with COPD

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics

by

Nancy (Fang) Yuan

Committee in charge:

Professor Theresa Gaasterland, Chair
Professor Tsung-Ting (Tim) Kuo, Co-Chair
Professor Laura Crotty Alexander
Professor Michael Hogarth
Professor Venktesh Ramnath

2023

The Dissertation of Nancy (Fang) Yuan is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

TO MY DEAR PARENTS AND ANCESTORS, MY SISTER FRIENDS, AND MY BELOVED FUR BABIES, MAX AND LEO

EPIGRAPH

“Live, then, and be happy, beloved children of my heart, and never forget, that until the day God will deign to reveal the future to man, all human wisdom is contained in these two words, ‘Wait and Hope.’” - Alexandre Duma

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude and love for my parents for their continual support, encouragement, and love throughout my life. Likewise, I would like to thank my late great grandmother, my dear grandparents, and my godmother, for raising me into late toddlerhood after my parents immigrated to the United States. Additionally, I would like to thank my dear auntie, and my incredibly talented cousin for inspiring me to travel and pursue creativity and self-expression. I would also like to thank my uncle for challenging me to think outside the box; my other dear auntie for spoiling me with beautiful gifts; and my other brilliant cousin for showing me how to play harder after working hard. I would also like to thank my late great auntie, a prolific playwright who gave the best hugs. Lastly, I would like to convey my love and gratitude for my late great grandfather, and my late grandmother, both of whom had passed before I was born but whose spirits have been with me always. Thank you all for lovingly putting up with my shenanigans over the years. As a family we have not had the luxury of being physically close to each other on a daily basis. Thank you all for helping me appreciate the beauty and meaning behind the sacrifices we make for each other to achieve our dreams.

Similarly, I would like to thank the many incredible mentors that I have had along this long academic journey, with special thanks to my thesis committee members.

Chapter 2, in full, is adapted from a paper that appeared in Annals of the American Thoracic Society, Volume 19, Issue 2, Pages 1993-2002, December 2022. The dissertation author was the primary author of this paper, which was coauthored by Kyle Hasenstab, Tara Retson, David Lynch, Douglas Conrad, and Albert Hsiao. The dissertation author thanks Kyle

Hasenstab, Tara Retson, David Lynch, Douglas Conrad, and Albert Hsiao for providing meaningful feedback and discussion.

Chapter 3, in full, is adapted from a paper that appeared in Radiology: Cardiothoracic Imaging, Volume 3, Issue 2, e200477, April 8, 2021. The dissertation author was one of the primary authors of this paper, which was coauthored by Kyle Hasenstab, Tara Retson, David Lynch, Douglas Conrad, Seth Kligerman, and Albert Hsiao. The dissertation author thanks Kyle Hasenstab, Tara Retson, David Lynch, Douglas Conrad, Seth Kligerman, and Albert Hsiao for providing meaningful feedback and discussion.

Chapter 4 presents unpublished material, and is coauthored with Theresa Gaasterland. The dissertation author was the primary author of this chapter. The dissertation author thanks Theresa Gaasterland, Albert Hsiao, and Shamim Nemati for providing meaningful feedback and discussion on the computational experiments performed in this chapter.

Chapter 5 presents unpublished material, and is coauthored with Theresa Gaasterland. The dissertation author was the primary author of this chapter. The dissertation author thanks Theresa Gaasterland for providing meaningful feedback and discussion on the computational experiments performed in this chapter.

Lastly, I would like to thank my dearest sister friends for all the support, care, and love that you have given me, near and far. Much thanks to Meredith, Elizabeth, Priscilla, Disha, Anh, and Jen. I give special thanks to my late friend, Sarah Rose, who appeared in my dream after she passed in Jan 2021 and gave me a glimpse into a world beyond the physical plane of this existence. Sarah remains a beacon of light in many people's memories, including mine.

VITA

2014 Bachelor of Arts in History, University of Texas at Austin

2015 Bachelor of Science in Cell & Molecular Biology, University of Texas at Austin

2017 Master of Science in Biological Sciences, University of California San Diego

2023 Doctor of Philosophy in Bioinformatics and Systems Biology with a Specialization in Biomedical Informatics, University of California San Diego

PUBLICATIONS

Yuan, N. F., Hasenstab, K., Retson, T., Conrad, D. J., Lynch, D. A., & Hsiao, A. (2022). Unsupervised Learning Identifies Computed Tomographic Measurements as Primary Drivers of Progression, Exacerbation, and Mortality in Chronic Obstructive Pulmonary Disease. *Annals of the American Thoracic Society*, *19*(12), 1993–2002. https://doi.org/10.1513/AnnalsATS.202110-1127OC

Hasenstab, K. A., **Yuan, N.**, Retson, T., Conrad, D. J., Kligerman, S., Lynch, D. A., Hsiao, A., & COPDGene Investigators (2021). Automated CT Staging of Chronic Obstructive Pulmonary Disease Severity for Predicting Disease Progression and Mortality with a Deep Learning Convolutional Neural Network. *Radiology. Cardiothoracic imaging*, *3*(2), e200477. https://doi.org/10.1148/ryct.2021200477

Hasenstab, K. A., Tabalon, J., **Yuan, N.**, Retson, T., & Hsiao, A. (2021). CNN-based Deformable Registration Facilitates Fast and Accurate Air Trapping Measurements at Inspiratory and Expiratory CT. *Radiology. Artificial intelligence*, *4*(1), e210211. https://doi.org/10.1148/ryai.2021210211

Humphries, J., Xiong, L., Liu, J., Prindle, A., **Yuan, F.**, Arjes, H. A., Tsimring, L., & Süel, G. M. (2017). Species-Independent Attraction to Biofilms through Electrical Signaling. *Cell*, *168*(1-2), 200–209.e12. https://doi.org/10.1016/j.cell.2016.12.014

Wagner, D. T., Zeng, J., Bailey, C. B., Gay, D. C., **Yuan, F.**, Manion, H. R., & Keatinge-Clay, A. T. (2017). Structural and Functional Trends in Dehydrating Bimodules from trans-Acyltransferase Polyketide Synthases. *Structure (London, England : 1993)*, *25*(7), 1045–1055.e2. https://doi.org/10.1016/j.str.2017.05.011

FIELD OF STUDY

APPLICATIONS OF MACHINE LEARNING AND DEEP LEARNING IN HEALTH INFORMATICS

ABSTRACT OF THE DISSERTATION

Using Machine Learning to Elucidate Disease Heterogeneity and Improve Prognostication in Patients with COPD

by

Nancy (Fang) Yuan

Doctor of Philosophy in Bioinformatics and Systems Biology with a Specialization in

Biomedical Informatics

University of California San Diego, 2023

Professor Theresa Gaasterland, Chair
Professor Tsung-Ting (Tim) Kuo, Co-Chair

Chapter 2 work:

Chronic obstructive pulmonary disease (COPD) is a heterogeneous syndrome, with phenotypic manifestations that tend to be distributed along a continuum. Unsupervised machine

learning based on broad selection of imaging and clinical phenotypes may be used to identify primary variables that define disease axes and stratify patients with COPD.

To identify primary variables driving COPD heterogeneity using principal component analysis (PCA), and to define disease axes and assess the prognostic value of these axes across three outcomes: progression, exacerbation, and mortality.

We included 7331 patients between 39 and 85 years, of which 40.3% are black and 45.8% are female smokers with a mean of 44.6 pack years from the COPDGene Phase 1 cohort (2008-2011) in our analysis. Out of a total of 916 phenotypes, 147 continuous clinical, spirometric, and CT features were selected. For each component (PC), we computed a principal component score (PCS) based on feature weights. We used PCS distributions to define disease axes along which we divided the patients into quartiles. To assess the prognostic value of these axes, we applied logistic regression analyses to estimate 5-year (n=4159) and 10-year (n=1487) odds of progression. Cox regression and Kaplan-Meier analyses were performed to estimate 5-year and 10-year risk of exacerbation (n=6532) and all-cause mortality (n=7331).

The first PC, accounting for 43.7% of variance, was defined by CT measures of air trapping and emphysema. The second PC, accounting for 13.7% of variance, was defined by spirometric and CT measures of vital capacity and lung volume. The third PC, accounting for 7.9% of the variance, was defined by CT measures of lung mass, airway thickening, and body habitus. Stratification of patients across each disease axis revealed up to 3.2-fold [2.4, 4.3] greater odds of 5-year progression, 5.4-fold [4.6, 6.3] greater risk of 5-year exacerbation, and 5.0-fold [4.2, 6.0] greater risk of 10-year mortality between the highest and lowest quartiles.

Unsupervised learning analysis of the COPDGene cohort reveals CT measurements may bolster patient stratification along the continuum of COPD phenotypes. Each of the disease axe

also individually demonstrate prognostic potential, predictive of future FEV1 decline, exacerbation, and mortality.

Chapter 3 work:

To define clinically meaningful stages of disease using CT imaging for patients with COPD, we developed a deep learning-based algorithm to stage severity of COPD through quantification of emphysema and air trapping on CT and assessed the ability of proposed stages to prognosticate 5-year progression and mortality.

In this retrospective study, an algorithm using co-registration and lung segmentation was developed in-house to automate quantification of emphysema and air trapping from inspiratory and expiratory CT images. The algorithm was then tested on a separate group of 8951 patients from the COPDGene study (2007-2017). With measurements of emphysema and air trapping, bivariable thresholds were determined to define CT stages of severity (mild, moderate, severe, and very severe), and evaluated for their ability to prognosticate disease progression and mortality using logistic regression and Cox regression.

Based on CT stages, odds of disease progression were greatest among patients with very severe disease (odds ratio [OR], 2.67 [95% CI: 2.02, 3.53; P<.001), and elevated in patients with moderate disease (1.50 [1.22, 1.84]; P = .001). Hazard ratio of mortality for very severe disease on CT was 2.23 times normal ([1.93, 2.58]; P<.001). When combined with GOLD staging, patients with GOLD 2 disease had the greatest odds of disease progression when CT stage was severe (4.48 [3.18, 6.31]; P<.001) or very severe (4.72 [3.13, 7.13]; P<.001).

Automated CT algorithms can facilitate staging of COPD severity, which has comparable diagnostic performance to spirometric GOLD staging, and provides further prognostic value when used in conjunction with GOLD stage. CT-based severity stratification of patients with COPD can

prognosticate disease progression and mortality and may allow for improved diagnosis and management.

Chapter 4 work:

The prevalence of COVID-19 has placed undue burden on the healthcare system, signaling a critical need to develop innovative machine learning (ML) strategies to improve triage and care for patients who are hospitalized with COVID-19. Recent developments in artificial intelligence have shown that deep learning algorithms such as convolutional neural networks (CNN) are effective in automating various tasks in medical imaging such as pneumonia detection [1]. The challenge remains in implementing these algorithms efficiently in the clinic where there may be insufficient hardware to support CNN with hundreds of hidden layers and millions of parameters [2].

To address this challenge, we developed a shallow 3-layer CNN classifier, SimplePNUnet, to detect the presence of pneumonia using a subset of 2000 frontal chest X-rays from the publicly accessible dataset released as part of the 2018 RSNA Pneumonia Detection Challenge. We fine-tuned model hyperparameters using Bayesian optimization. Additionally, we assessed the robustness of SimplePNUnet across different downsampling techniques and batch sizes. Performance of the algorithm was evaluated using the area under receiver operating characteristic curve (AUROC).

The performance of the best model across the training (n=1000), validation (n=500), and test (n=500) sets was comparable (AUC, 0.870 vs 0.812 vs 0.834, respectively) and within range of deeper U-Net based pneumonia detection algorithms (AUC, 0.80-0.90). SimplePNUnet achieved good performance on an external validation set (AUC, 0.712), demonstrating its generalizability to evaluate other COVID-19 patient cohorts.

Chapter 5 work:

COPD exacerbations are characterized by sudden worsening of symptoms that often result in increased future exacerbation and mortality risk. Identifying clinical and sociodemographic characteristics that are strongly predictive of severe exacerbations may help screen for patients at high risk for experiencing an exacerbation that may significantly increase their burden of healthcare costs. Here, we built a novel, interpretable machine learning framework to predict severe exacerbation in patients with COPD.

To curate the data, a total of 118 demographic, socioeconomic, laboratory testing, pulmonary function testing, and CT imaging features available for 7781 patients were curated from the Phase 1 COPDGene data. Next, for each data type, we applied the Boruta feature selection method to identify only the features that may be relevant for predicting severe exacerbation. After performing Boruta feature selection, we trained a total of five unimodal models using data from only a single data type. We further trained four multimodal models using different combinations of data types. For each model, the data was split 70% for training and 30% for testing, using a total of four different model architectures to predict severe exacerbation: logistic regression, K-nearest neighbors, random forest classifier, and lightGBM. The hyperparameters of each model were optimized using Bayesian optimization. Each model was evaluated using the area under the receiver operating characteristic curve (AUROC), among other performance metrics. Model interpretability was achieved using SHAP global and local interpretation plots that summarize the average marginal contribution of each feature to severe exacerbation.

Unimodal models using lab testing data-only, pulmonary function test data-only, or CT imaging data-only had strong performance (highest AUC, 0.77, 0.81, 0.77, respectively).

Combining lab testing data with social determinants of health data improved performance significantly (highest AUC, 0.62 to 0.78). The highest performing multimodal model was trained using a combination of demographic, lab testing, social determinants of health, and pulmonary function data (highest AUC, 0.82). Model interpretability analysis performed on the multimodal model trained with all the features revealed BODE, baseline age, and $FEF_{25-75\%}$ to be among the top contributing factors toward a positive prediction of severe exacerbation.

Integrating additional data types into multimodal models improved performance for predicting severe exacerbation. Unimodal models tuned using Bayesian optimization likewise achieved similarly strong performance when trained with lab test data, pulmonary function test data, or CT imaging data. SHAP global and local interpretation plots offer significant insights into explaining how the models arrived at their predictions. Identifying BODE, baseline age, and $FEF_{25-75\%}$ as the most influential features toward a positive prediction of severe exacerbation may help physicians in screening patients at high-risk for experiencing future exacerbations.

CHAPTER 1: INTRODUCTION

An unconventional source of motivation for this body of work.

"Do you ever study the Chinese zodiac to try to matchmake people?" His bewitching olive green eyes sparkled sardonically. "Of course not," I quipped. I went home after the date with my head spinning with memories that his question triggered. *That sunset evening against a crowded backdrop, a glance of eternity, a shared kiss before everything dissolved into a flash of white.* What would I do to relive that moment with my long-ago soulmate? What wouldn't I do? His archly posed question catalyzed my metaphysical dive into eastern and western astrological systems.

Readers of this thesis who insist on seeing and understanding this world solely through scientific lens may dismiss astrology as pseudoscience. To this I say, I agree in the sense that we have not yet formalized a rigorous scientific framework to test astrological principles. Readers may even harbor prejudices against me for taking astrology seriously enough to study it. To that I say, spirituality and science can be pursued in parallel, as both attempt to shed light into the unknown by helping us better understand ourselves and by extension, our world. What started for me as a spiritual journey on finding true love actually led to the formation of some of the main scientific questions addressed in the body of my thesis. Before diving into those questions, let me first walk you through an ultra-brief history of predictive science.

An ultra-brief history of predictive science.

Deeply ingrained in the human psyche is the curiosity of the unknown future. Early societies attempted to satisfy this curiosity by devising creative strategies to predict future events. Astrological systems emerged as a product of meticulously recorded observations across millennia of how celestial bodies moved in the heavenly skies. Astrologers recognized the

predictive power behind the cyclical patterns of movement, inferred spiritual and practical meaning from them, and shared their interpretations to their contemporaries in the form of mythological stories. The night sky was not the only medium the ancients used in their predictive frameworks. The ancient Chinese had pyromancy, a form of divination whereby diviners carved questions into tortoise shells, applied heat, and then interpreted the crack patterns to make predictions about war, health, agriculture, and other impactful events during that time.[1,2,3,4] In a similar vein, the Greeks had a chosen priestess at Delphi who would fall into a trance-like state after breathing in gaseous fumes emanating from a natural chasm and then would start channeling cryptic messages from the Oracle.[5]

As fantastical as these early endeavors may seem, they highlight our inherent drive to use the most precise and accurate instruments we have available to foretell the future. Since the late 1950s, technological advancements in artificial intelligence (AI) have vastly improved and refined the ability to predict future events. Computers became faster, cheaper, and more effective at data storage. At the same time, machine learning algorithms improved significantly.

Machine learning (ML) methods train computers to learn patterns in the data and make predictions based on these patterns. The ML models use digitized data (input features) to make predictions of specified outcomes (or output labels). The goal of 'learning' is to derive features that are most predictive of the specific labels under consideration. Deep Learning (DL) models, a branch of ML, achieve this through multiple layers of nonlinear processing (known as *artificial neural networks*), which produce a distributed and hierarchical representation of the input features.[6,7] The idea of hierarchical representation was inspired by the organization of the human visual object recognition system into different processing layers.[6,8]

By the late 1980s, Yann LeCun integrated back propagation with deep convolutional neural networks to recognize handwritten zip code digits provided by the U.S. Postal Service.[9] The next significant evolutionary step for deep learning took place a decade later, when computers started becoming faster at data processing, as graphics processing units (GPUs) were developed to run more computationally expensive algorithms designed for image recognition.

We now live in the age of "Big Data". In the healthcare domain, deep learning has successfully automated various classification and segmentation tasks in medical imaging.[13,16] Specifically, deep learning algorithms have accurately detected the presence of common diseases such as pneumonia, coronary artery disease, and cancer.[23] Algorithmic performance is commonly compared to that of healthcare professionals to assess clinical utility. Among the challenges of integrating new AI algorithms into clinical practice is interpretability. What information does the algorithm "look at" when it is making predictions? Developing strategies to identify the salient features that influence predictions and to quantify the extent of influence is an ongoing area of research.

Applications of machine learning in healthcare.

Over the past decade, ML and DL methods have gained increasing attention and application in clinical medicine.[17,18,19,20,21,22,23,24] Clinical data from the Electronic Health Record (EHR) can be collected either at varying time intervals (e.g., medication and laboratory data or radiology imaging) or continuously (e.g., vital sign data from remote monitoring sensors).[6] Given the breadth and richness of patient data currently available, ML and DL can compute complex interactions or temporal relationships among multivariate risk factors that bedside clinicians and patients commonly overlook.[6,18]

Developing robust and interpretable models for predicting longitudinal health outcomes for patients remains one of the top priorities for AI scientists working in healthcare. Studies have shown among the challenges of integrating new AI algorithms into clinical practice is inadequate interpretability, which refers to the ability of users to understand the decision-making process of the model.[19,20,21] The level of trust a clinician has toward an AI decision support system may be affected by organizational policies, workplace culture, and inherent biases.[22] As the integration of AI systems continues to impact the healthcare system, a relationship of trust must be made between clinicians and AI. Arguably, the burden lies on AI developers to integrate interpretability dimensions into the models such that it is transparent for end-users to see how much each of the input features contribute to the overall prediction. This is especially true for predicting adverse outcomes for patients with complex, systemic diseases such as chronic obstructive pulmonary disease (COPD).

Developing AI-driven techniques to elucidate COPD heterogeneity and improve prognostication.

COPD affects over 200 million people worldwide and remains the fourth leading cause of death by disease in the United States.[10,11,12] COPD is predominantly caused by tobacco smoking, and is characterized by chronic airway inflammation, destruction of downstream alveoli and vasculature which leads to emphysema, and air trapping due to airflow obstruction.[13] While there is no cure for COPD, much of the ongoing research is focused on improving patient management. It has been challenging to identify the relative importance of specific clinical and social determinants of health (SDOH) predictors and drivers of COPD disease progression, exacerbation, and mortality over time.

Given that COPD may be a risk factor for more severe COVID-19 disease, it is natural to fear for patients with underlying COPD.[14] Furthermore, the pandemic has placed additional

burden on the workload of radiologists who read chest radiographs. There is a critical need to develop innovative machine learning (ML) and deep learning (DL) strategies to improve triage and care for patients who are hospitalized with COVID-19.

The heterogeneity of medical data makes it challenging to develop robust predictive models that contribute to the understanding of a patient's disease trajectory. In the context of COPD research, several ML and DL models utilizing either clinical, imaging, or administrative health data have been built to predict exacerbation and mortality.[15,16,17] While these methods have shown moderate success in predicting such adverse outcomes, a common limitation of these approaches is that they only use one type of data at the exclusion of other types and thus cannot contextualize other clinical information as would be done in medical practice, which may limit clinical translation.[18,19,20,21,23,24]

Data fusion refers to the process of joining data from multiple modalities in order to extract complementary and more complete information for better performing ML models as opposed to using a single data modality. For AI (artificial intelligence) applications to be successfully applied to COPD prognostication, an integrative approach combining clinical (e.g. pulmonary function tests), social determinants of health (SDOH), and imaging data may be necessary.

SDOH data and medical imaging contain valuable information that has been underutilized in COPD disease outcome prediction. Previous studies have observed a social class gradient in the prevalence of respiratory symptoms in adults with COPD.[35] Specifically, low socioeconomic status is a known risk factor for various adverse COPD health outcomes, including acute exacerbations.[36] More recent studies have shown that quantitative measurements from computed tomography (CT) correlate with spirometry and are useful in COPD patient

stratification and prognostication.[13] As such, CT measurements of emphysema and air trapping have been included in the recently updated diagnostic criteria for COPD from investigators of the NIH-supported COPDGene study.[37]

Multimodal DL models that can incorporate clinical data and pixel data from images have been successfully implemented in applications outside of medicine, such as autonomous driving and video classification.[38,39] Recent medical literature shows a growing trend in the use of fusion models leveraging both EHR and imaging data for solving complex tasks that cannot be readily addressed by a single data modality. In one study, Lungren *et al*. developed and compared different multimodal fusion model architectures that use both pixel data from volumetric CT pulmonary angiography scans and patient data from EMR to successfully automate detection of pulmonary embolisms.[39] Multimodal DL architectures can also be applied to COPD prognostication.

We had a unique opportunity to leverage the longitudinal cohort from the COPDGene study, the electronic health record (EHR) from UCSD Health, and the well-annotated publicly available chest radiograph datasets on Kaggle. We developed a multimodal approach by combining clinical, sociodemographic, and imaging data from these rich datasets to perform our data-driven analyses. Our overarching goal was to integrate different data types and build a framework that identifies mechanisms underlying heterogeneous forms of COPD, as well as to automate the diagnosis and improve prognostication for both COPD and COVID-19 patients.

In Chapter 2, we designed and implemented a data-driven approach to identify primary drivers of disease progression, exacerbation, and mortality in COPD patients. We hypothesized that CT imaging features will be among the primary drivers of adverse patient outcomes. We built an unsupervised machine learning framework, using principal component analysis (PCA) to

stratify patients with COPD along principal axes of disease. To confirm the clinical value of these disease axes, we applied logistic regression and Cox proportional hazards analysis to evaluate their association with downstream clinical outcomes, such as disease progression, exacerbation, and mortality.

In Chapter 3, we developed efficient computational methods to define clinically meaningful stages of disease using computed tomography (CT) for patients with COPD. We hypothesized that CT imaging significantly captures the underlying pathophysiology of COPD and contains valuable information for accurate staging of the disease. We developed a deep learning-based algorithm to stage the severity of COPD through quantification of emphysema and air trapping on CT and assessed the ability of proposed stages to prognosticate progression, exacerbation, and all-cause mortality.

In Chapter 4, we developed a computationally tractable algorithm to automate pneumonia detection for patients with COVID-19. We hypothesized that a shallow convolutional neural network (CNN) performs as well as the deeper CNNs in detecting pneumonia. We developed and validated a 3-layer CNN that can be retrained and evaluated for patient cohorts in medically underserved areas with limited computing capabilities.

In Chapter 5, we developed and evaluated multimodal machine learning approaches that integrate clinical, imaging, and sociodemographic data to predict clinically relevant outcomes in patients with COPD. We hypothesized that the multimodal model that uses a fusion of different types of data will outperform at least one single-modality model in predicting exacerbation risk. We built single-modality baseline models (e.g. “clinical data only”, “sociodemographic data only”, and “imaging data only”) using deep neural networks. Additionally, we developed

multimodal fusion models using different combinations of these features and compared the performance of the fusion models against baseline models.

The analytic work in chapters 2 through 5 use the following cohorts containing retrospective, deidentified patient data:

COPDGene cohort: We leveraged the COPDGene Phase 1 (study enrollment) and Phase 2 (5-10 year follow up) longitudinal dataset. COPDGene is a 10-year cross sectional prospective cohort of over 10,000 adult current and former smokers enrolled between January 2008 and June 2011 at 21 clinical centers across the United States.[37] The primary goals of this study are to characterize smokers with and without COPD, using spirometry, medical history, respiratory symptoms, and inspiratory and expiratory chest CT scans.[37]

2018 RSNA pneumonia challenge chest radiographs: The 2018 RSNA pneumonia challenge included 25,684 chest radiographs spanning the range of pathologies from this data set, distributed as DICOM images.[40] The chest X-rays in this dataset are a subset of a larger NIH database comprising 112,120 frontal chest radiographs, each of which was labeled with findings and diagnoses extracted from radiology text reports using natural language processing.[40] These pathologies included atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia.[40]

COVID-19 Radiography Dataset (COV-RAD): The COVID-19 Radiography Dataset is a publicly accessible Kaggle dataset compiled by a team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh along with their collaborators from Pakistan and Malaysia.[41,42] Collaborating with physicians at these sites, they have created a database of chest X-ray images for COVID-19 positive cases along with Normal and Viral Pneumonia images. Over the past two years, they have increased the database to 3616 COVID-

19 positive cases along with 10,192 Normal, 6012 Lung Opacity (Non-COVID lung infection), and 1345 Viral Pneumonia images and corresponding lung masks.[41,42]

Chapters 2-5 report on the ML approaches that we developed to leverage these cohorts for deeper understanding and more effective data-driven diagnostic and prognostic strategies for patients with COPD and COVID-19. Chapter 6 will contain concluding remarks on our findings from each chapter and a brief discussion on future research directions.

REFERENCES

1. Nylan, M. (2001). *The Five "confucian" Classics*. New Haven: Yale University Press.
2. O´Brien, P. (2014). *Divination Sacred Tools for Reading the Mind of God*. U.S. Games Systems.
3. Redmond, G.P., & Hon, T. (2014). *Teaching the I Ching (book of Changes)*. Aar Teaching Religious Studies. Oxford: Oxford University Press.
4. Wilhelm, R., & Baynes, C.F. (1985). *The I Ching : Or, Book of Changes*. [3d Ed.] ed. Bollingen Series, 19. Princeton, N.J.: Princeton University Press.
5. Mellenthin, J., & Shapiro, S.O. *Mythology Unbound: An Online Textbook for Classical Mythology*. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
6. Grant Institutional Proposal #30090237: VentNet: A Real-Time Multimodal Data Integration Model for Prediction of Respiratory Failure in Patients with COVID-19
7. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–44.
8. Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755.
9. LeCun Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D. (1989). *Neural Computation*, 1, 541-551.
10. WHO.int. The top 10 causes of death, 2020 [Internet]. Geneva: World Health Organization; [accessed 2021 June 28]. WHO Global Health Estimates. Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
11. Assessing national capacity for the prevention and control of noncommunicable diseases: report of the 2019 global survey. Geneva: World Health Organization; 2020. License: CC BY-NC-SA 3.0 IGO
12. CDC.gov. CDC National Health Report Highlights, 2012 [Internet]. Atlanta, GA: U.S. Department of Health and Human Services Centers for Disease Control and Prevention; [accessed 2021 June 14]. CDC Health Report. Available from: https://www.cdc.gov/healthreport/publications/compendium.pdf
13. Hasenstab, K., Yuan, N., Retson, T., Conrad, D., Kligerman, S., Lynch, D., Hsiao, A. Automated CT Staging of COPD Severity for Predicting Disease Progression and Mortality with a Deep Learning Convolutional Neural Network. *Radiology: Cardiothoracic Imaging*.
14. Leung JM, Niikura M, Yang CWT, et al. COVID-19 and COPD. *Eur Respir J* 2020; 56:2002108 [https://doi.org/10.1183/13993003.02108-2020].
15. Vestbo J, Hurd SS, Agusti AG, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med*. Feb 15 2013;187(4):347-65. doi:10.1164/rccm.201204-0596PP
16. Lanclus M, Clukers J, Van Holsbeke C, Vos W, Leemans G, Holbrechts B, Barboza K, De Backer W, De Backer J. Machine Learning Algorithms Utilizing Functional

Respiratory Imaging May Predict COPD Exacerbations. *Academic radiology*. 2019;26(9):1191–1199.

17. Esteban C, Moraza J, Esteban C, Sancho F, Aburto M, Aramburu A, Goiria B, Garcia-Loizaga A, Capelastegui A. The prevalence of asthma–chronic obstructive pulmonary disease overlap syndrome (ACOS) in a Northern Italy general population sample. *European Respiratory Journal*. 2015;46: OA3282. doi: 10.1183/13993003.

18. Zwanenburg, A. (2019). Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *Eur J Nucl Med Mol Imaging*, 46, 2638–2655. https://doi.org/10.1007/s00259-019-04391-8

19. Erikainen S and Chan S. (2019). Contested futures: envisioning "Personalized," "Stratified," and "Precision" medicine. *New genetics and society*, *38*(3), 308–330. https://doi.org/10.1080/14636778.2019.1637720

20. Inke R. König, Oliver Fuchs, Gesine Hansen, Erika von Mutius, Matthias V. Kopp. (2017). What is precision medicine?. *European Respiratory Journal*, 50(4) 1700391. https://doi.org/10.1183/13993003.00391-2017

21. Venkataramana K. Sidhaye, Kristine Nishida, Fernando J. Martinez. (2018). Precision medicine in COPD: where are we and where do we need to go? *European Respiratory Review*, 27(149) 180022. https://doi.org/10.1183/16000617.0022-2018

22. Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy (Basel, Switzerland)*, *23*(1), 18. https://doi.org/10.3390/e23010018

23. Shah, P., Kendall, F., Khozin, S. *et al.* (2019). Artificial intelligence and machine learning in clinical development: a translational perspective. *npj Digit. Med.*2, 69 https://doi.org/10.1038/s41746-019-0148-3

24. Holzinger, A., Haibe-Kains, B., & Jurisica, I. (2019). Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. *European journal of nuclear medicine and molecular imaging*, *46*(13), 2722–2730. https://doi.org/10.1007/s00259-019-04382-9

25. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. May 2015;521(7553):436–44.

26. Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*. Jun 10 2016;6(1):27755.

27. Ohno-Machado L. Data science and artificial intelligence to improve clinical practice and research. *J Am Med Inform Assoc*. Jan 2018;25(10):1273.

28. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep*. Jan 2014;16(1):441.

29. Johnson AEW, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE*. Feb 2016;104(2):444–66.

30. Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*. Jan 2017;21(1):4–21.
31. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med*. Jan 2019;25(1):24–9.
32. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1(1):18.
33. Butte AJ. Big data opens a window onto wellness. *Nature Biotechnology*. Aug 2017;35(8):720–1.
34. Fu Y, Ling Z, Arabnia H, Deng Y. Current trend and development in bioinformatics research. *BMC Bioinformatics*. Dec 3 2020;21(Suppl 9):538.
35. Prescott E, Vestbo J. Socioeconomic status and chronic obstructive pulmonary disease. *Thorax* 1999;54:737-741.
36. Eisner MD, Blanc PD, Omachi TA, et al. Socioeconomic status, race and COPD health outcomes. *J Epidemiol Community Health*. 2011;65(1):26-34.
37. Lowe KE, Regan EA, Anzueto A, Austin E, Austin JHM, Beaty TH, Benos P V, Benway CJ, Bhatt SP, Bleecker ER, Bodduluri S, Bon J, Boriek AM, Boueiz AR, Bowler RP, Budoff M, Casaburi R, Castaldi PJ, Charbonnier J-P, et al. COPDGene® 2019: Redefining the Diagnosis of Chronic Obstructive Pulmonary Disease. Chronic Obstr Pulm Dis J COPD Found. 2019;6(5):384–99. doi: 10.15326/jcopdf.6.5.2019.0149
38. Huang, SC., Pareek, A., Seyyedi, S. *et al.* Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit. Med.* 3, 136 (2020). https://doi.org/10.1038/s41746-020-00341-z
39. Huang, SC., Pareek, A., Zamanian, R. *et al.* Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Sci Rep* 10, 22147 (2020). https://doi.org/10.1038/s41598-020-78888-w
40. (2018). *RSNA Pneumonia Challenge (2018)*. https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/RSNA-Pneumonia-Detection-Challenge-2018.
41. Chowdhury, M.E.H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al-Emadi, N., Reaz, M.B.I., & Islam, M. T. (2020). "Can AI help in screening Viral and COVID-19 pneumonia?" *IEEE Access*, 8, 132665 – 132676.
42. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Maadeed, S.A., Zughaier, S.M., Khan, M.S. and Chowdhury, M.E., 2020. Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-ray Images

CHAPTER 2: DESIGN AND IMPLEMENT A DATA-DRIVEN APPROACH TO IDENTIFY PRIMARY DRIVERS OF DISEASE PROGRESSION, EXACERBATION, AND MORTALITY IN COPD PATIENTS

When we meet someone in person, attraction can happen in a magical instant. Something about the twinkle of their eyes, their crescent smile, their unique fragrance draws us closer to them. We hear their familiar voice for the first time, feel the touch of their hand on ours. All these sensory inputs map to pools of preconceived representations encoded within the labyrinth of over 100 trillion synaptic connections that comprise our mind palace.[1] Specifically, neuroimaging studies have found that falling in love activates cortical areas including the cortex, the medial insula, anterior cingulate, and hippocampus, as well as subcortical areas including parts of the striatum and the ventral tegmental area, namely, the core regions of our brain's reward system.[1] Additionally, the frontal cortex, including regions involved in decision making, and the amygdala, involved in regulating the fear response, become deactivated.[2]

This intricate chemical dance that takes place in our minds when we fall in love has inspired innumerable works in art, literature, cinema, and perhaps more indirectly, in every other field known to man. The ability to feel love is arguably the most prominent trait developed in the course of our human evolution that continues to drive forward our evolution. A curious question would be whether we can evolve AI to feel love. Can we evolve AI to feel any emotions at all? Practically speaking, the answer remains uncertain, as the field of AI is still at its infancy in terms of growth potential. So far, we have successfully trained AI models to detect and identify facial expressions, vocal tones, and other manifestations of human emotions.[3] However, even the most advanced AI today is still a machine, bereft of the physiological and biological substrates needed to feel emotions.

Perhaps the more ethically meaningful question to ask is: *should* we evolve AI to feel emotions? What purpose would that serve? One could be to provide a realistic proxy for human companionship. Many people suffer from loneliness, especially amongst the elderly living in populous metropolitan areas. Imagine having a service where any lonely person can commission an AI robot to live with them, empathize with them, and emotionally respond to their needs. How would the AI robot feel about providing this service? At what point should we care about how this AI robot feels? At what point does AI embody enough human-like qualities and traits such that we must grapple with the reality that it is no longer a *what* but a *who*? Accepting the profound meaning behind this shift in pronouns would mark a huge stepping stone in our journey of contending with and perhaps ultimately redefining our preconceived notions of what it means to be human. At the point when we do recognize an AI machine as an AI *being*, we are confronted with yet another set of ethical questions related to whether these AI *beings* should be granted the same rights, protections, and privileges as human beings have in varying degrees.

A broader, more thorough discussion on the extent of human-driven AI evolution and its ethical implications is beyond the scope of this thesis. Suffice it to say, it was to my relief, given my limited computational skills when I started my graduate studies, that my first first-author project involved exploring a much more fundamental process in human and machine cognition: pattern recognition. Here I present the first chapter of this thesis, which applies a well-established, widely used dimensionality reduction method of pattern recognition on the richly curated COPDGene data to elucidate disease heterogeneity on the patient population.

2.1 Introduction

Chronic obstructive pulmonary disease (COPD) is a heterogeneous syndrome, with phenotypic manifestations that tend to be distributed along a continuum. Unsupervised machine

learning based on broad selection of imaging and clinical phenotypes may be used to identify primary variables that define disease axes and stratify patients with COPD. The objective of this aim is to identify primary variables driving COPD heterogeneity using principal component analysis (PCA), and to define disease axes and assess the prognostic value of these axes across three outcomes: progression, exacerbation, and mortality.

Several investigators have begun to explore unsupervised machine learning algorithms to understand the factors underlying the heterogeneous COPD patient population. This includes the use of clustering algorithms in an effort to define disease subtypes that aggregate patients who share common phenotypes.[4,17,18,19,20,21,22,23,24,25,26,27,28] While the concept of disease subtypes may have intuitive appeal, clustering approaches have neither achieved strong separation for proposed subgroups nor shown high reproducibility across different studies.[4]

In contrast, studies suggest that COPD phenotypes tend to be distributed along a continuum, which may be more accurately represented by disease axes.[9,30] We define a disease axis as a continuous representation of phenotypic variability present across the patient population. Principal component analysis (PCA) is a well-established dimensionality reduction technique that produces linear combinations of the features that maximally explain the observed population variance and thus can be applied to define disease axes. To this end, a few groups have begun to explore PCA to create disease axes, but have been limited by narrow variable selection without broad use of quantitative CT imaging and spirometric phenotypes.[23,31] Further, it is unclear from these prior studies whether disease axes ultimately provide prognostic value for patients with COPD.

To address these challenges, we sought to apply PCA to the wide range of clinical and imaging phenotypic data from patients from the COPDGene study. Leveraging the COPDGene cohort, we hypothesized that applying PCA to the wide range of clinical and imaging characteristics of these patients would be able to stratify the heterogenous COPD population and ultimately inform their long-term outcomes.

2.2 Design and implementation of the unsupervised machine learning pipeline

2.2.1 Unsupervised learning using principal component analysis

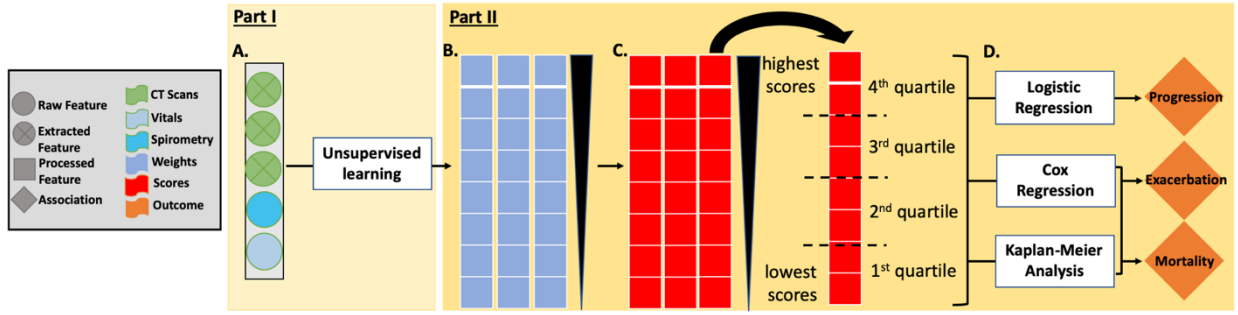

Figure 2.1: Unsupervised learning framework. Part I: (A) Extracted CT imaging and non-imaging clinical features from the COPDGene Phase 1 data are fed into an unsupervised learning method such as PCA. Part II: (B) A matrix of feature weights. (C) A matrix of scores computed using feature weights. Each score distribution can be used to stratify patients into quartiles. (D) Quartiles of stratified patients are fed into logistic regression, Cox regression, and Kaplan-Meier analyses to assess their association with progression, exacerbation, and mortality.

To understand the spectrum of phenotypes that comprise COPD, we applied the PCA method to tabulated clinical, spirometric, and imaging data collected in Phase 1 of the COPDGene study (Figure 2.1A). PCA aligns highly correlated, continuous variables together into principal components, such that most of the information within the initial set of variables is compressed into the first few principal components.

Out of a total of 916 phenotypes, 147 continuous features were available for 7331 patients included in the PCA (Table 2.1). The features were standardized using mean centering and standard deviation. Clinical features included systolic and diastolic blood pressure, height,

weight, age, heart rate, body mass index (BMI), and oxygen saturation. Spirometric features included forced expiratory volume in the first second (FEV_1), forced vital capacity (FVC), and the FEV_1/FVC ratio. Computed tomographic (CT) features included emphysema, air trapping, and airway wall thickness measures. Outcome variables were excluded from the PCA, as they were used at a later stage for assessment.

Table 2.1: List of CT, spirometric, and demographic features included in the PCA.

CT Measures							Spirometric Measures	Demographics, Vitals, Smoking History
Air Trapping	Emphysema	Airway Wall Thickness, Hyperinflation	Lung Volume, Lung Mass, Lung Density	Other: Perc15	Other: Attenuation	Other: CT textures	FEV1_post	age_baseline
Exp_Below856_Vida	Insp_Below950_Vida	Pi10_SRWA_Vida	TLC_Thirona	Perc15_Insp_Vida	AttenMean_Insp_Vida	Slicer_Insp_Kurtosis	FEV6_post	Height_CM
Exp_LAA856_total_Thirona	pctEmph_UpperLobes	Pi15_SRWA_Vida	TLC_Slicer	Perc15_Exp_Vida	AttenStdDev_Insp_Vida	Slicer_Insp_Skewness	FVC_post	Weight_KG
Exp_LAA856_LUL_Thirona	pctEmph_LowerLobes	WallAreaPct_seg_Vida	TLC_Vida	perc15_density	AttenMean_Exp_Vida	Slicer_Insp_Histo_Mode	FEV1pp_post	sysBP
Exp_LAA856_LLL_Thirona	pctEmph_UpperLobes_Thirona	Pi10_Thirona	TLC_CT	Perc15_Insp_Slicer	AttenStdDev_Exp_Vida	Slicer_Exp_Kurtosis	FVCpp_post	diasBP
Exp_LAA856_RUL_Thirona	pctEmph_LowerLobes_Thirona	AWT_seg_Thirona	TLC_pp_plethy	Perc15_Exp_Slicer	ExpInspMeanAtten_ratio_Vida	Slicer_Exp_Skewness	FEV1_FVC_post	HR
Exp_LAA856_RML_Thirona	Insp_LAA950_total_Thirona	WallAreaPct_seg_Thirona	TLV_pp_mesa	Perc15_Insp_total_Thirona	MeanAtten_Insp_total_Thirona	Slicer_Exp_Histo_Mode	PEF_post	Resting_SaO2
Exp_LAA856_RLL_Thirona	Insp_LAA950_LUL_Thirona	FRC_Thirona	UL_pctTotalAirVol	Perc15_Insp_LUL_Thirona	MeanAtten_Insp_LUL_Thirona		FEF2575_post	BMI
Exp_Below856_Slicer	Insp_LAA950_LLL_Thirona	FRC_Vida	Insp_totalvolume_total_Thirona	Perc15_Insp_LLL_Thirona	MeanAtten_Insp_LLL_Thirona		FEV1_pre	O2_suppl_6MW
PRM_pct_airtrapping_Thirona	Insp_LAA950_RUL_Thirona	FRC_Slicer	Insp_totalvolume_LUL_Thirona	Perc15_Insp_RUL_Thirona	MeanAtten_Insp_RUL_Thirona		FEV6_pre	distwalked
PRM_pct_normal_Thirona	Insp_LAA950_RML_Thirona	FRC_TLC_ratio_Thirona	Insp_totalvolume_LLL_Thirona	Perc15_Insp_RML_Thirona	MeanAtten_Insp_RML_Thirona		FVC_pre	SmokStartAge
PRM_pct_other_Thirona	Insp_LAA950_RLL_Thirona	FRC_TLC_total_Thirona	Insp_totalvolume_RUL_Thirona	Perc15_Insp_RLL_Thirona	MeanAtten_Insp_RLL_Thirona		FEV1_FVC_pre	CigPerDaySmokAvg
	PRM_pct_emphysema_Thirona	FRC_TLC_LUL_Thirona	Insp_totalvolume_RML_Thirona	Perc15_Exp_total_Thirona	MeanAtten_Exp_total_Thirona		PEF_pre	OthSmokChildYrs
	pctEmph_UpperThird_Slicer	FRC_TLC_LLL_Thirona	Insp_totalvolume_RLL_Thirona	Perc15_Exp_LUL_Thirona	MeanAtten_Exp_LUL_Thirona		FEF2575_pre	OthSmokYrs
	pctEmph_LowerThird_Slicer	FRC_TLC_RUL_Thirona	Exp_totalvolume_total_Thirona	Perc15_Exp_LLL_Thirona	MeanAtten_Exp_LLL_Thirona		deltaFEV1	ExpSmokAtWorkYr
	Insp_Below950_Slicer	FRC_TLC_RML_Thirona	Exp_totalvolume_LUL_Thirona	Perc15_Exp_RUL_Thirona	MeanAtten_Exp_RUL_Thirona		deltaFVC	ATS_PackYears
		FRC_TLC_RLL_Thirona	Exp_totalvolume_LLL_Thirona	Perc15_Exp_RML_Thirona	MeanAtten_Exp_RML_Thirona		BDR_pct_FEV1	Duration_Smoking
		FRC_CT	Exp_totalvolume_RUL_Thirona	Perc15_Exp_RLL_Thirona	MeanAtten_Exp_RLL_Thirona		BDR_pct_FVC	
		FRC_TLC_ratio	Exp_totalvolume_RML_Thirona		Slicer_ExpInspMeanAtten_ratio			
			Exp_totalvolume_RLL_Thirona		Slicer_IntensityMean_In			
			Slicer_Insp_Lung_Mass		Slicer_IntensityStdDev_In			
			Slicer_Exp_Lung_Mass		Slicer_IntensityMean_Ex			
			adj_density_plethy		Slicer_IntensityStdDev_Ex			
			adj_density_mesa					

2.2.2 Assessment of prognostic value of principal components

To assess the value of the principal components generated by PCA, we computed principal component scores (PCS) for each component for every patient. To compute the PCS, we multiplied the weights of the features in each component by the original features and performed a summation, [which is effectively a dot product] (Figure 2.1B). We partitioned PCS distributions by quartiles and then evaluated how patients stratified along these axes differed in their 5-year and 10-year outcomes (Figure 2.1C). Specifically, we evaluated the association of each PCS quartile with progression, exacerbation, and all-cause mortality (Figure 2.1D).

Progression was defined as a decline in FEV_1 by greater than 350 mL over 5 years or 10 years.[33] Longitudinal spirometric data for calculation of FEV_1 change were available for 4159 and 1487 patients for 5-year and 10-year analyses, respectively. An exacerbation was defined as a patient-reported adverse event that had occurred within the previous 12 months at the 5-year or 10-year follow-up (an adverse event may include any of the following: ER visit for lung problems, hospitalization for lung problems, antibiotic treatment for a chest illness, and/or steroid treatment for a chest illness). All-cause mortality refers to a death event due to any cause at the 5-year or 10-year follow-up. Longitudinal follow-up survey data for exacerbation and death events were available for 6532 and 7331 patients, respectively.

2.2.3 Statistical analyses

Statistical analyses were performed in Python-v3.6.12 and R-v3.6.0 software. Patient demographics, smoking history, and spirometry metrics are summarized in Table 2.2.

Table 2.2: Demographics, smoking history, and spirometry of individuals included in the sample cohort used in the PCA and regression analyses.

	Sample Cohort (n = 7331)
Characteristics	Number, Mean (SD), or %
Demographics	
Age (years)	60 (9)
Sex (F)	45.8 %
Ethnicity (Black)	40.3 %
BMI	28.8 (6.0)
Height (cm)	171.3 (9.5)
Weight (kg)	83.8 (19.2)
Smoking history	
Current smokers	49.1 %
Pack-years of smoking	44.6 (24.8)
Spirometry	
FEV_1 % predicted	76.7 (25.2)
FEV_1/FVC	0.66 (0.16)
GOLD stage	
0	3148
1	595
2	1506
3	865
4	393
PRISM	824

PCA was performed using the sci-kit learn machine learning library for Python. Multivariable logistic regression analyses were applied to estimate odds of progression, adjusting for baseline age, sex, race, pack years, current smoking status, and FEV_1 percent predicted FEV_1% predicted is equivalent to the FEV_1% of the patient divided by the average FEV_1% in the population for any person of similar age, sex, and body composition.[34] Cox regression and Kaplan-Meier analyses were performed to estimate risk of exacerbation and all-cause mortality, using right-censoring [to limit the estimate to 10 years]. Cox regression analyses were adjusted for baseline age, sex, race, pack years, and current smoking status. Confidence intervals for odds and hazard ratios were analytically calculated [per standard formulas]. Regression *P*-values were adjusted using Bonferroni correction. Statistical significance was determined using a 5% Type I error threshold.

2.3 Results

2.3.1 Principal axes

We included [in the PCA] 7331 patients between 39 and 85 years, of which 40.3% are black and 45.8% are female smokers with a mean of 44.6 pack years from the COPDGene Phase 1 cohort (2008-2011). Additional demographic information may be found in the study by Regan *et al*.[35] A correlation heatmap of weights for input features across the top 5 principal components is shown in Figure 2.2. For visual clarity, a representative set of imaging, spirometric, and demographic features was chosen.

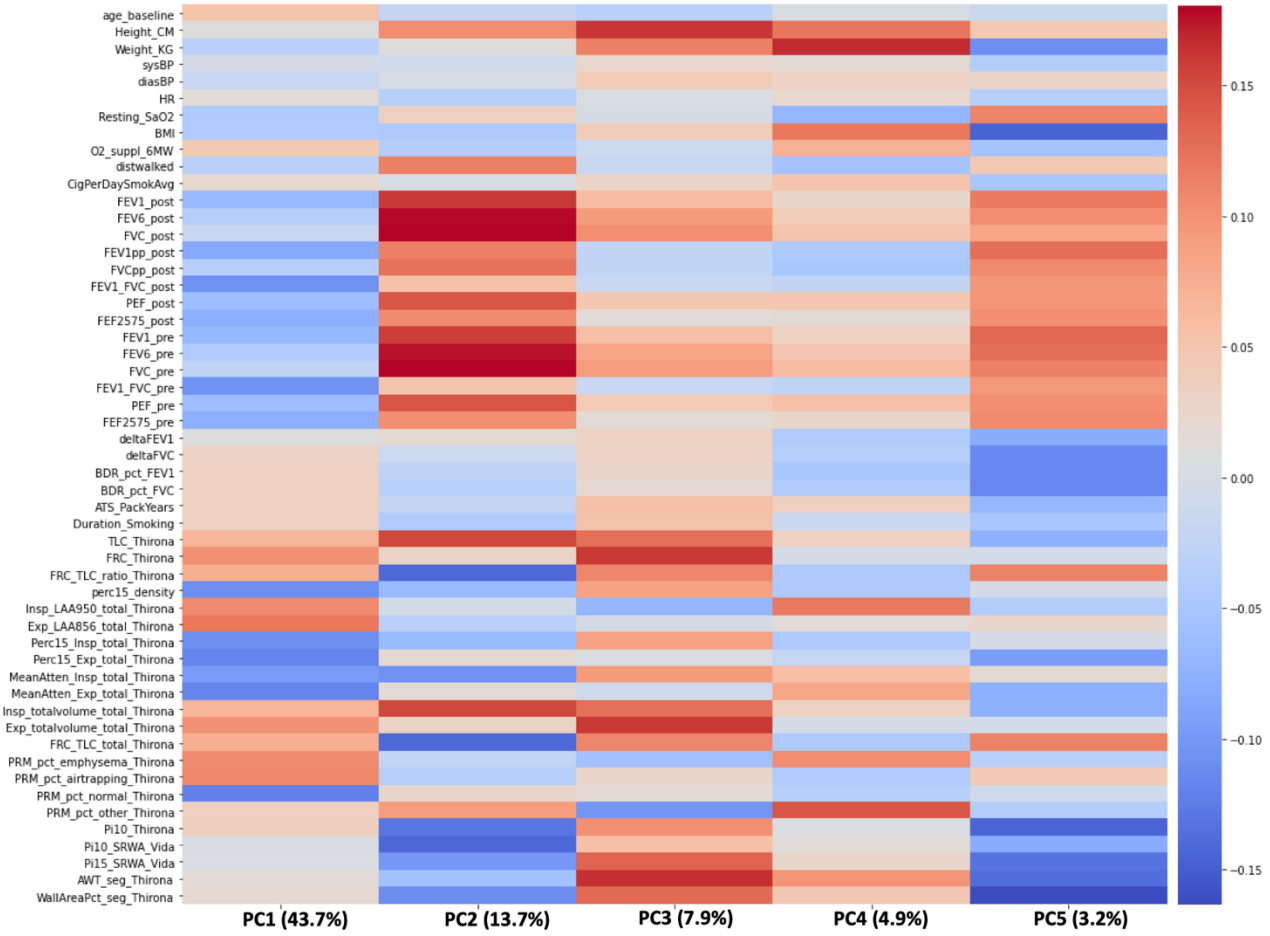

Figure 2.2: Correlation heatmap of loadings for input features across the top 5 principal components, PC1-5 (percentage of variance explained). For visual clarity, a representative set of features was chosen.

A unique set of the top 25 principal component features and their weights are listed in Table 2.3. The first component (Component 1), accounting for 43.7% of variance, is defined by CT measurements of air trapping and emphysema. The second component (Component 2), accounting for 13.7% of variance, is defined by CT and spirometric measurements of vital capacity and lung volume. The third component (Component 3), accounting for 7.9% of the variance, is defined by lung mass, airway thickening, and body habitus. Out of a total of 147 continuous features included in the PCA (Table 2.1), these features are the primary drivers of the outcomes we later assessed.

Table 2.3: Principal component analysis reveals highest weighted features that define each of first three principal components. Definition of abbreviations: CT = computed tomography; FEV1 = forced expiratory volume in 1 second; FEV6 = forced expiratory volume in 6 seconds; FRC = functional residual capacity; FVC = forced vital capacity; HU = Hounsfield units; pct = percent; Pi15 = internal perimeter 15 mm; PRM = parametric response mapping; SRWA = square root wall area; TLC = total lung capacity. CT and spirometric features appear prominently across all three components.

Component 1 (Variance Explained: 43.7%)		Component 2 (Variance Explained: 13.7%)		Component 3 (Variance Explained: 7.9%)	
Feature	**Feature Weight**	**Feature**	**Feature Weight**	**Feature**	**Feature Weight**
CT measures of air trapping		Spirometric measures of lung volume and vital capacity		CT measures of lung mass	
Less than −856 HU on expiratory, total	0.120	FVC, prebronchodilator, liters	0.180	Lung mass on inspiratory	0.237
Less than −856 HU on expiratory, left upper lobe	0.117	FVC, post-bronchodilator, liters	0.180	Lung mass on expiratory	0.219
Less than −856 HU on expiratory, right upper lobe	0.116	FEV_6, post-bronchodilator, liters	0.178	CT measures of airway wall thickening and hyperinflation	
Less than −856 HU on expiratory, right lower lobe	0.111	FEV_6, prebronchodilator, liters	0.176	Airway wall thickness, segmental	0.165
Less than −856 HU on expiratory, right middle lobe	0.111	FEV_1, post-bronchodilator, liters	0.161	FRC	0.161
Less than −856 HU on expiratory, left lower lobe	0.110	FEV_1, prebronchodilator, liters	0.158	Pi15 SRWA	0.135
PRM_pct_airtrapping	0.109	CT measures of lung volume		FRC/TLC ratio	0.110
CT measures of emphysema		TLC	0.153	Body habitus	
Less than −950 HU on inspiratory, total	0.107	Total volume on inspiratory	0.153	Height (cm)	0.162
PRM_pct_emphysema	0.105	Right lower lobe volume on inspiratory	0.152	Weight (kg)	0.114

Table 2.4: Prognostic ability of the first three principal components for progression, exacerbation, and all-cause mortality. Patients in the 2nd - 4th quartiles show significantly greater odds of progression and risk of exacerbation and mortality compared to patients in the 1st quartile.

	5-Year FEV1 Progression		10-Year FEV1 Progression		5-Year Exacerbation		10-Year Exacerbation		5-Year All-Cause Mortality		10-Year All-Cause Mortality	
No. of subjects	4,159		1,487		6,532		6,532		7,331		7,331	
Principal Component Analysis (PCA)	Odds Ratio (95% CI)	Adj. P-Value	Odds Ratio (95% CI)	Adj. P-Value	Hazard Ratio (95% CI)	Adj. P-Value	Hazard Ratio (95% CI)	Adj. P-Value	Hazard Ratio (95% CI)	Adj. P-Value	Hazard Ratio (95% CI)	Adj. P-Value
Component 1												
1st Quartile	1		1		1		1		1		1	
2nd Quartile	1.50 (0.97, 1.44)	0.896	1.33 (0.98, 1.81)	0.597	1.50 (1.29, 1.75)	<0.001	1.43 (1.27, 1.62)	<0.001	1.37 (1.02, 1.85)	0.306	1.32 (1.07, 1.62)	0.068
3rd Quartile	1.40 (1.13, 1.73)	0.017	1.03 (0.75, 1.42)	1	2.00 (1.72, 2.33)	<0.001	1.79 (1.59, 2.03)	<0.001	1.89 (1.41, 2.51)	<0.001	2.08 (1.71, 2.52)	<0.001
4th Quartile	1.74 (1.37, 2.23)	<0.001	1.29 (0.90, 1.86)	1	3.78 (3.26, 4.39)	<0.001	3.13 (2.77, 3.53)	<0.001	5.14 (3.94, 6.70)	<0.001	5.00 (4.16, 6.00)	<0.001
Component 2												
1st Quartile	1		1		1		1		1		1	
2nd Quartile	1.36 (1.11, 1.67)	0.027	1.27 (0.94, 1.74)	1	1.69 (1.44, 1.99)	<0.001	1.48 (1.30, 1.67)	<0.001	1.15 (0.84, 1.56)	1	1.15 (0.93, 1.41)	1
3rd Quartile	1.81 (1.45, 2.26)	<0.001	1.45 (1.04, 2.02)	0.255	2.40 (2.05, 2.81)	<0.001	2.02 (1.79, 2.29)	<0.001	1.58 (1.17, 2.12)	<0.001	1.66 (1.36, 2.02)	<0.001
4th Quartile	3.20 (2.41, 4.26)	<0.001	1.97 (1.29, 3.00)	0.015	5.36 (4.59, 6.26)	<0.001	4.24 (3.74, 4.80)	<0.001	5.01 (3.82, 6.57)	<0.001	4.81 (4.00, 5.80)	<0.001
Component 3												
1st Quartile	1		1		1		1		1		1	
2nd Quartile	1.37 (1.11, 1.69)	0.034	1.27 (0.94, 1.72)	1	1.05 (0.92, 1.19)	1	1.08 (0.97, 1.20)	1	1.18 (0.96, 1.45)	1	1.20 (1.03, 1.39)	0.124
3rd Quartile	1.36 (1.10, 1.69)	0.043	1.66 (1.21, 2.26)	0.013	1.08 (0.94, 1.23)	1	1.06 (0.95, 1.19)	1	1.36 (1.10, 1.70)	0.042	1.40 (1.20, 1.63)	<0.001
4th Quartile	1.53 (1.23, 1.90)	0.001	1.46 (1.05, 2.04)	0.231	1.05 (0.91, 1.20)	1	1.02 (0.91, 1.15)	1	1.37 (1.07, 1.74)	0.098	1.38 (1.16, 1.63)	<0.001

2.3.2 Evaluating association of components with disease progression

For 5-year progression, stratifying patients along the first three component score distributions all had significant association with 5-year progression. Specifically, stratification of patients by the first PCS showed 1.7-fold greater odds of progression between the 4th quartile and the 1st quartile, while the second PCS showed 3.2-fold greater odds, and the third PCS had 1.5-fold greater odds ($p<0.001$) (Figure 2.3A-C, Table 2.4).

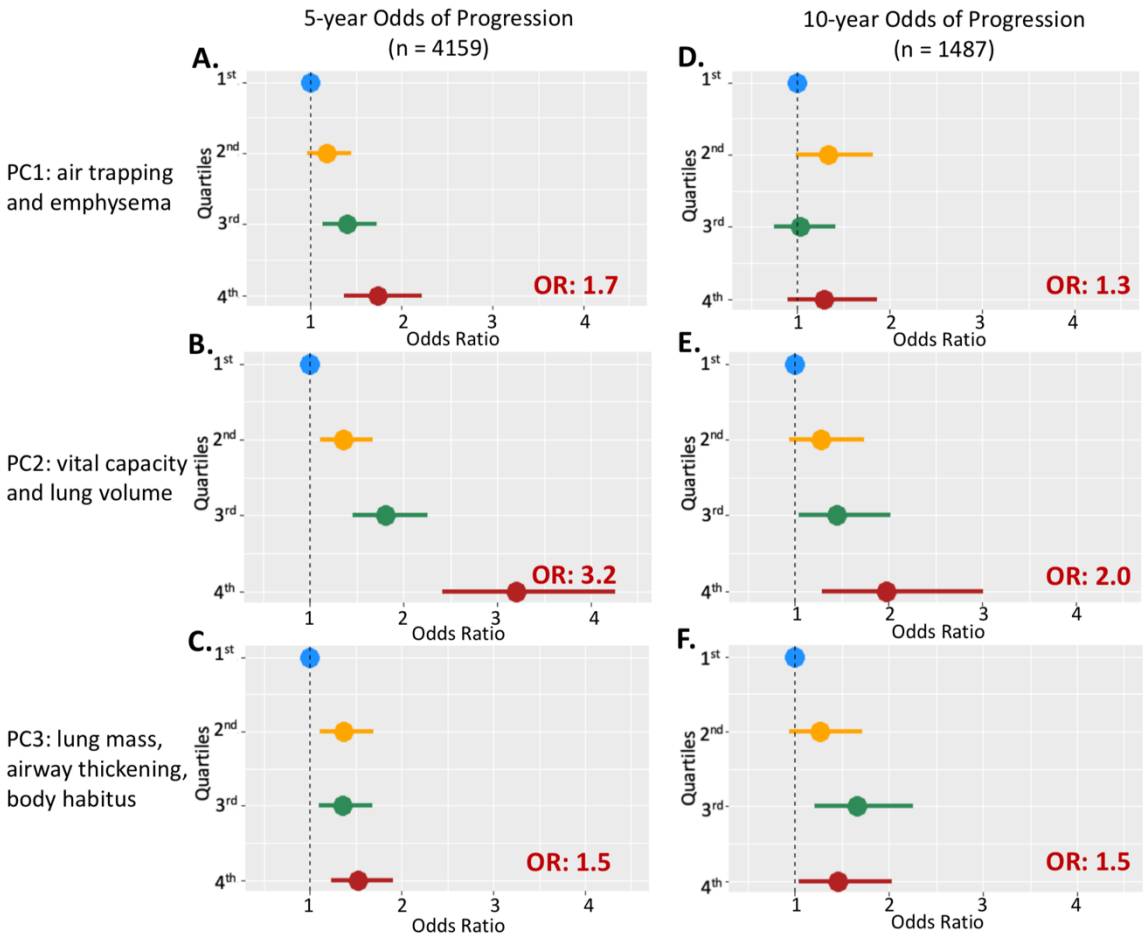

Figure 2.3: Evaluating 5-year (A-C) and 10-year (D-F) odds of progression. Forest plots reveal that stratifying patients along the PC1-3 score distributions has significant associations with progression. The red represents the odds ratio between the highest and lowest quartiles. Vertical dotted line represents no effect. 95% confidence intervals that lie to the right of the dotted line represents significant association with outcome. Abbreviations: OR, odds ratio between the 1st quartile and 4th quartile.

For 10-year progression, stratification of patients by the first PCS showed 1.3-fold greater odds of progression between the 4th quartile and the 1st quartile, while the second PCS showed 2.0 greater odds. The third PCS showed no significant degree of differential odds of progression (Figure 2.3D-F). Notably, for the first and second PCS, the magnitude of the odds ratios from the 10-year analyses is less than that from the 5-year analyses, an observation which is discussed more at length in the Discussion and Limitations section below.

2.3.3 Evaluating association of components with exacerbation

Table 2.5: Exacerbation and mortality by year of study. A total of 2049/6532 participants experienced at least one exacerbation in the last 12 months within 5 years, and a total of 2899/6532 experienced an exacerbation within 10 years. A total of 691/7331 participants died within 5 years, and a total of 1448/7331 died within 10 years.

	Exacerbation	Mortality
Year of Study	Number of Subjects who Experienced at Least One Exacerbation in the Past Year	Number of Deaths
1	512	63
2	580	130
3	426	163
4	295	171
5	236	164
5-year Total	2049 / 6532 subjects	691 / 7331 subjects
6	208	150
7	206	172
8	174	160
9	136	144
10	126	131
10-year Total	2899 / 6532 subjects	1448 / 7331 deaths

Cox regression analyses revealed that stratifying patients along the first and second PCS had significant association with 5-year risk of exacerbation. Specifically, stratification of patients by the first PCS revealed 3.8-fold greater risk of exacerbation between the 4th quartile and the 1st quartile, while the second PCS showed 5.4-fold greater risk ($p<0.001$) (Figure 2.4A-C, Table 2.4). For the 10-year analysis, stratifying patients across the first PCS revealed up to 3.1-fold greater risk of exacerbation, while the second PCS showed 4.2-fold greater risk (Figure 2.5A-C, Table 2.4). Stratifying patients along the third PCS had no significant association with exacerbation in both 5-year and 10-year analyses.

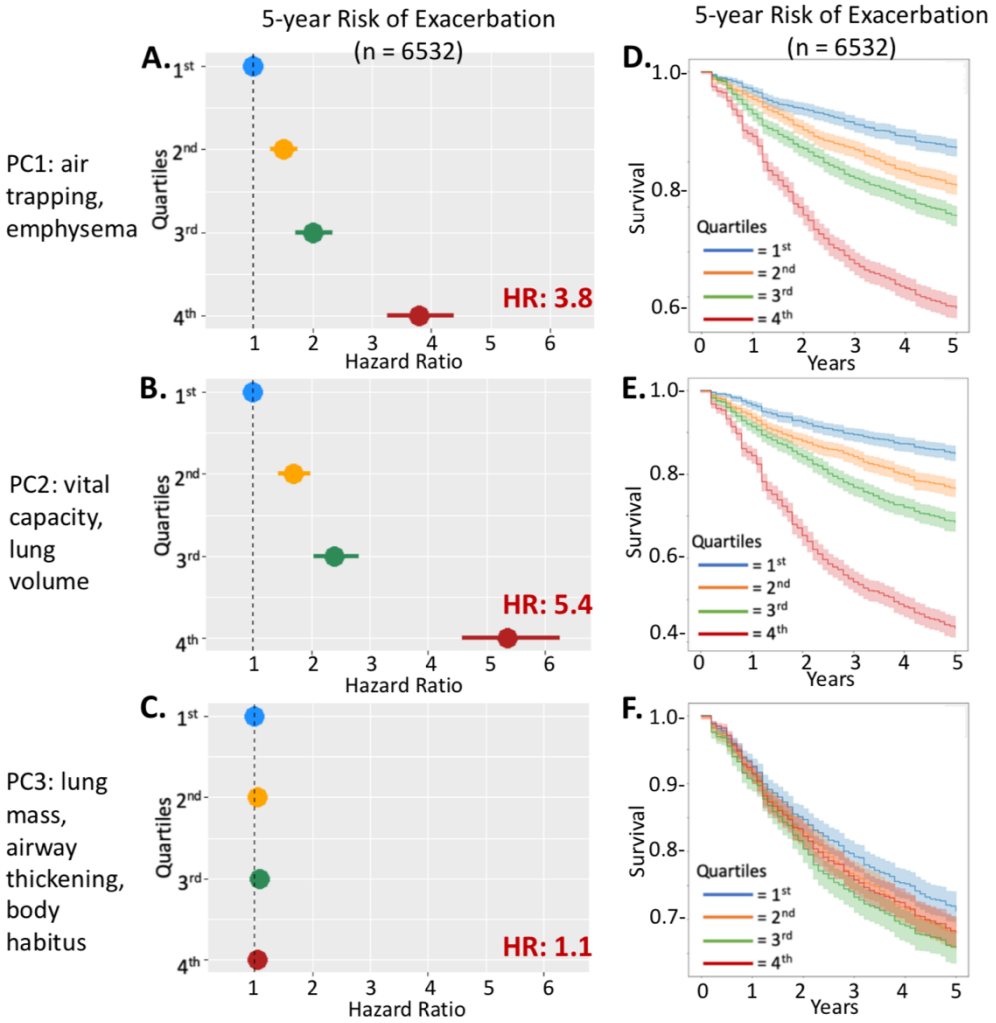

Figure 2.4: Evaluating 5-year risk of exacerbation. (A)-(C) Forest plots reveal that stratifying patients along the PC1 and PC2 score distributions has significant associations with exacerbation. Vertical dotted line represents no effect. 95% confidence intervals that lie to the right of the dotted line represents significant association with outcome. Abbreviations: HR, hazard ratio between the 1st quartile and 4th quartile. (D)-(F) Kaplan-Meier survival curves estimate the fraction of 6532 patients who experience at least one exacerbation in 5 years, stratified into quartiles by the first three principal component score distributions (PC1-3).

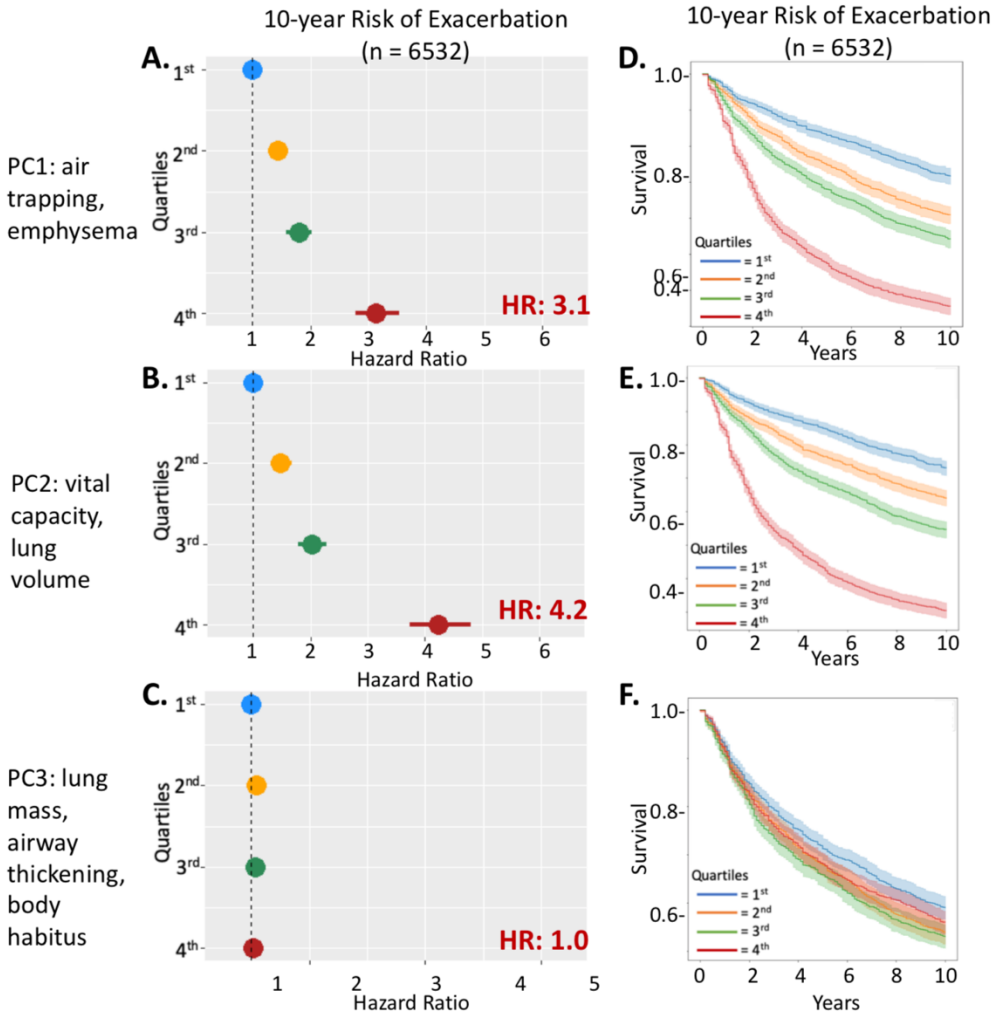

Figure 2.5: Evaluating 10-year risk of exacerbation. (A)-(C) Forest plots reveal that stratifying patients along the PC1 and PC2 score distributions has significant associations with exacerbation. Vertical dotted line represents no effect. 95% confidence intervals that lie to the right of the dotted line represents significant association with outcome. Abbreviations: HR, hazard ratio between the 1st quartile and 4th quartile. (D)-(F) Kaplan-Meier survival curves estimate the fraction of 6532 patients who experience at least one exacerbation in 10 years, stratified into quartiles by the first three principal component score distributions (PC1-3).

As shown in the Kaplan-Meier survival curves for 5-year and 10-year risk of exacerbation, stratification of patients by the first and second PCS revealed that a significantly larger fraction of patients in the 4th quartile experienced an exacerbation compared to those in the 1st quartile ($p<0.005$). Stratification of patients by the third PCS showed no significant degree of differential risk of exacerbation (Figure 2.4D-F, 2.5D-F). The number of subjects who experienced at least one exacerbation in the previous year, per year of the study, is shown in Table 2.5.

2.3.4 Evaluating association of components with mortality

Cox regression analyses revealed that stratifying patients along the first, second, and third component score distributions all had significant association with 5-year risk of mortality. Specifically, stratification of patients by the first PCS revealed 5.1-fold greater risk of mortality between the 4th quartile and the 1st quartile, while the second PCS showed 5.0-fold greater risk ($p<0.001$) (Figure 2.6A-C, Table 2.4). Stratifying patients along the third PCS had no significant association with mortality for 5-year analysis. For the 10-year analysis, stratifying patients across the first PCS revealed 5.0-fold greater risk of mortality, while the second PCS had 4.8-fold greater risk, and the third PCS had 1.4-fold greater risk (Figure 2.7A-C, Table 2.4).

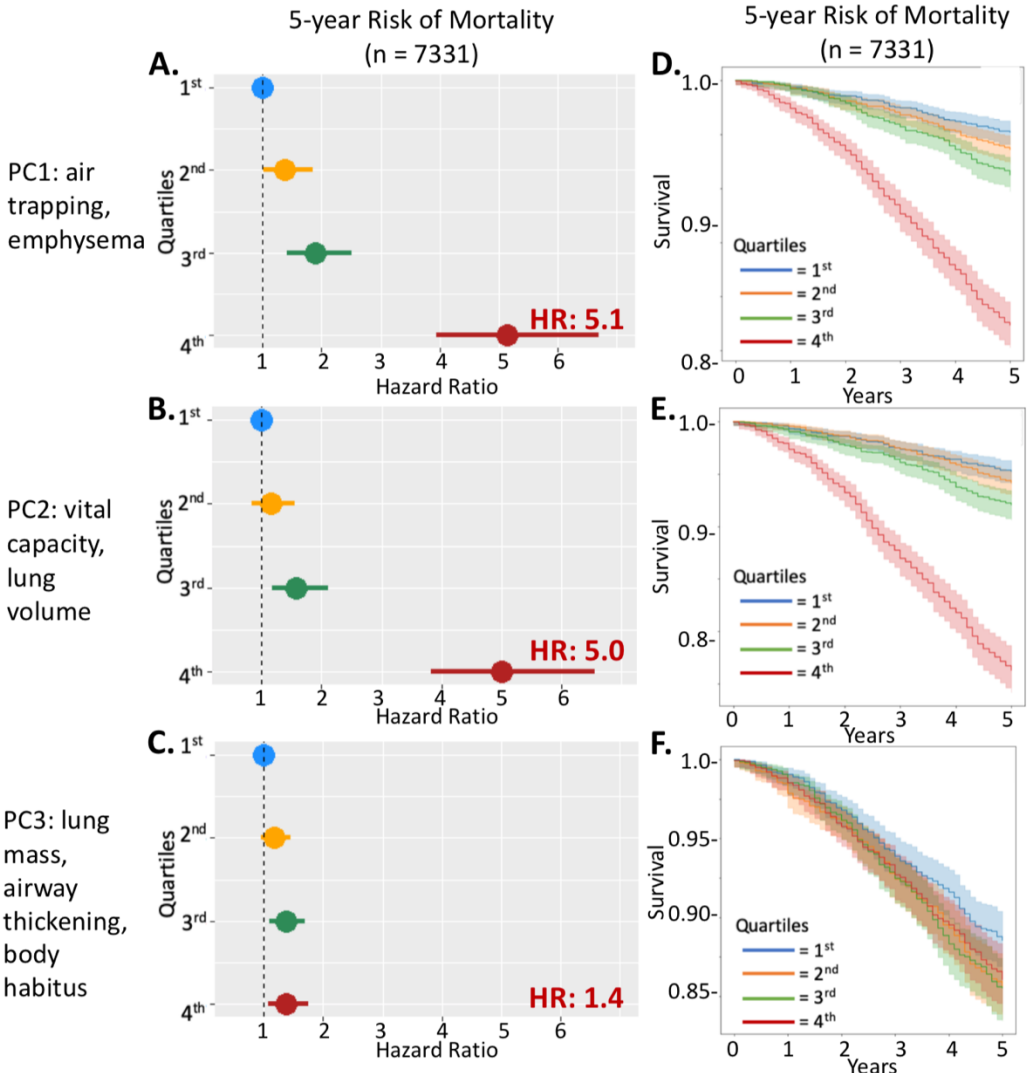

Figure 2.6: Evaluating 5-year risk of all-cause mortality. (A)-(C) Forest plots reveal that stratifying patients along the PC1-3 score distributions has significant associations with mortality. Vertical dotted line represents no effect. 95% confidence intervals that lie to the right of the dotted line represents significant association with outcome. Abbreviations: HR, hazard ratio between the 1st quartile and 4th quartile. (D)-(F) Kaplan-Meier curves estimate the fraction of 7331 patients who die in 10 years, stratified into quartiles by the first three principal score distributions (PC1-3).

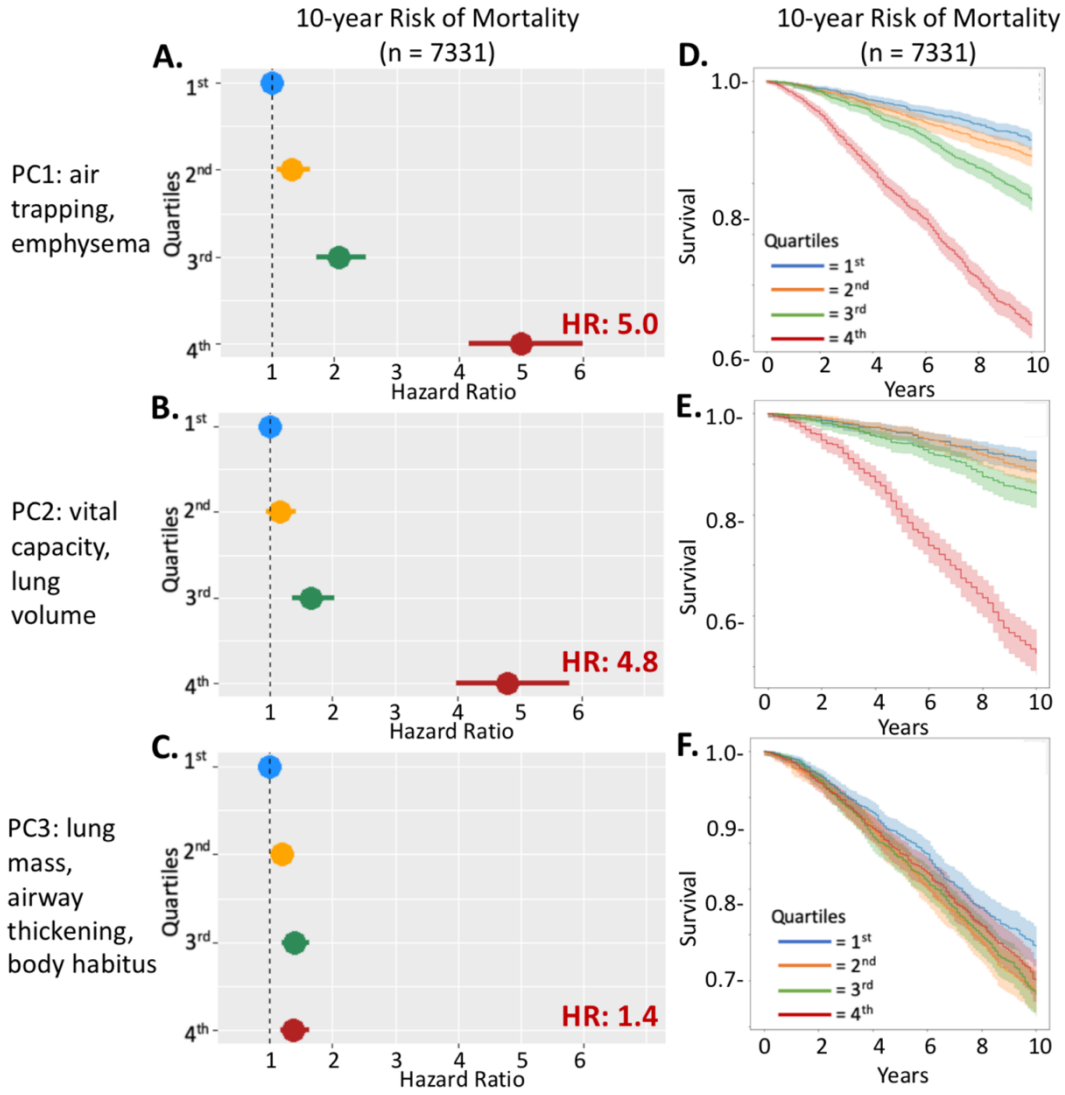

Figure 2.7: Evaluating 10-year risk of all-cause mortality. (A)-(C) Forest plots reveal that stratifying patients along the PC1-3 score distributions has significant associations with mortality. Vertical dotted line represents no effect. 95% confidence intervals that lie to the right of the dotted line represents significant association with outcome. Abbreviations: HR, hazard ratio between the 1st quartile and 4th quartile. (D)-(F) Kaplan-Meier curves estimate the fraction of 7331 patients who die in 10 years, stratified into quartiles by the first three principal score distributions (PC1-3).

Additionally, in the Kaplan-Meier survival analyses for 5-year and 10-year risk of all-cause mortality, stratification of patients by the first, second, and third PCS revealed that a significantly larger fraction of patients in the 4th quartile died compared to those in the 1st quartile ($p<0.005$) (Figure 2.6D-F, 2.7D-F). The number of subjects who died per year of the study is shown in Table 1.5.

2.4 Discussion and limitations

For over fifty years, spirometry has been a mainstay of pulmonary function testing (34). The Global Obstructive Lung Disease (GOLD) system relies on spirometric evidence of airway obstruction to stage patients with COPD.[35] However, Lynch et al. have shown that patients classified within the same GOLD stage can have vastly different degrees of emphysema and air trapping, which can be quantified by CT.[36] Hasenstab et al have shown that quantitative measurements from CT confer significant prognostic value.[5] Furthermore, COPDGene investigators have included CT measurements of emphysema and air trapping in the recently updated diagnostic criteria for COPD.[33] We expand on these concepts, applying unsupervised learning using the broadest selection of variables to discover that CT measurements prominently define principal axes of disease in patients with COPD. This finding provides additional evidence for the importance of incorporating CT as part of COPD patient management.

In our assessment of 5-year and 10-year odds of progression, we noted that for the first and second PCS, the magnitude of the odds ratios from the 10-year analyses is less than that from the 5-year analyses. We hypothesized that this result was related to how we defined progression as a loss of at least 350 mL of lung volume in 5 or 10 years. There is a physiological limitation to how much lung volume a patient may lose overtime, which may explain why the odds of further progression in the 10-year assessment are lower than that in the 5-year assessment. Our assessment of 5-year and 10-year exacerbation risk revealed that PC3 did not yield significant hazards ratios. Similarly, our assessment of 5-year and 10-year mortality risk revealed that PC3 did not yield significant hazard ratios. The reason may be that PC3 is composed primarily of lung mass and body habitus features, which do not correlate strongly with exacerbation risk or mortality risk.

Previous attempts to create disease axes applied similar unsupervised learning methods to other COPD cohorts but were limited by their narrow variable selection.[9,23,31] Burgel et al selected 8 COPD-related variables to perform PCA on 322 COPD patients from a French multicenter cohort.[23] Dyspnea score, smoking history, and BMI appeared prominently in their top three components, explaining 61% of the variance.[23] Roy et al selected 9 pulmonary function variables and inflammatory biomarkers to perform PCA on 127 COPD patients recruited from primary care in the UK.[31] Sputum neutrophil cell count, eosinophil percentage, and bronchodilator reversibility appeared prominently in their top three components, explaining 53.5% of the variance.[31] In our study, we included a total of 171 continuous variables from 7331 patients in the PCA, using the broadest selection of imaging, pulmonary function, and demographic features among related studies. CT measurements appeared saliently in the top three components, explaining 65.3% of the variance, greater than what has been found in prior studies.

Several investigators have begun applying strategies to improve the interpretability of disease axes. Chen et al relied on predefined subtypes to create and interpret their disease axes.[30] In contrast, we implemented a purely data-driven approach, using PCA to create and interpret our disease axes. Our mortality association results are consistent with those found by Kinney et al (2018), who applied factor analysis to discover that CT measures of emphysema and air trapping appeared prominently in two separate disease axes.[9] In spite of using a different data reduction approach, our findings from PCA echo the importance of imaging phenotypes in explaining a significant proportion of variance in the data. We show further that our air trapping/emphysema disease axis can be predictive of future FEV_1 decline and exacerbation, in addition to mortality.

Our study focused on the prognostic value of each disease axis individually. Here, our primary aim was to evaluate the association of each of the disease axes with three clinically relevant outcomes. Future work may interrogate whether using different combinations of the disease axes would confer additional prognostic value.

A limitation in this study may have been that the exacerbations events were self-reported which are less specific than physician-adjudicated or ICD-coded events. Future work in this area may benefit from characterizing reliability of self-reported exacerbation events.

Another limitation is that we did not train a model to predict the outcomes directly, as our primary goal was to define interpretable disease axes to better understand COPD heterogeneity. However, if the sole aim was prognostication, other approaches using supervised learning may build upon this work and yield increased performance. For example, González et al. applied a deep learning approach to predict mortality with fair discrimination using only four selected slices from inspiratory CT.[37] Tavakoli et al. applied several machine learning algorithms to predict severe exacerbations using administrative health data.[14] Future work may investigate the application of supervised algorithms for COPD prognostication using a combination of imaging, spirometric, and demographic variables.

2.5 Conclusions

Leveraging the broadest range of imaging and clinical characteristics available in the COPDGene cohort, we observed that the top principal components that emerged from our unsupervised analysis were predominantly related to CT imaging and spirometric phenotypes. Specifically, the top three components (1) air trapping/emphysema, (2) vital capacity/lung volume, and (3) lung mass/airway thickening/body habitus each explain a large fraction of the variability of the population. We used these components to create disease axes. Each of these

axes also individually demonstrate prognostic potential, predictive of future FEV_1 decline, exacerbation, and mortality. In summary, this work shows that imaging phenotypes explain a significant proportion of COPD heterogeneity. Including CT measurements in COPD patient management may assist in identifying those most in need of therapeutic interventions early in their disease trajectory.

2.6 Acknowledgments

Chapter 2, in full, is adapted from a paper that appeared in Annals of the American Thoracic Society, Volume 19, Pages 1993-2002, Dec 2022. The dissertation author was the primary author of this paper, which was coauthored by Kyle Hasenstab, Tara Retson, Douglas Conrad, David Lynch, and Albert Hsiao.

The dissertation author thanks Drs. Hasenstab, Retson, Conrad, and Hsiao for providing meaningful feedback and discussion.

2.7 References

1. Zeki, S. (2007) The neurobiology of love. *FEBS Letters*. https://www.sciencedirect.com/science/article/pii/S0014579307004875.

2. Seshadri K. G. (2016). The neuroendocrinology of love. *Indian journal of endocrinology and metabolism*, *20*(4), 558–563. https://doi.org/10.4103/2230-8210.183479

3. Loghmani, M. R., Rovetta, S., and Venture, G. 2017. "Emotional intelligence in robots: Recognizing human emotions from daily-life gestures," *IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, pp. 1677-1684, doi: 10.1109/ICRA.2017.7989198.

4. WHO.int. The top 10 causes of death, 2020 [Internet]. Geneva: World Health Organization; [accessed 2021 June 28]. WHO Global Health Estimates. Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

5. Assessing national capacity for the prevention and control of noncommunicable diseases: report of the 2019 global survey. Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IGO

6. CDC.gov. CDC National Health Report Highlights, 2012 [Internet]. Atlanta, GA: U.S. Department of Health and Human Services Centers for Disease Control and Prevention; [accessed 2021 June 14]. CDC Health Report. Available from: https://www.cdc.gov/healthreport/publications/compendium.pdf

7. Castaldi PJ, Benet M, Petersen H, Rafaels N, Finigan J, Paoletti M, *et al*. Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. *Thorax* 2017;72:998-1006.

8. Hasenstab K, Yuan N, Retson T, Conrad D, Kligerman S, Lynch D, Hsiao A, *et al*. Automated CT Staging of COPD Severity for Predicting Disease Progression and Mortality with a Deep Learning Convolutional Neural Network. *Radiology: Cardiothoracic Imaging* 2021;3.

9. Lynch DA, Moore CM, Wilson C, Nevrekar D, Jennermann T, Humphries SM, Austin JHM, *et al*. CT-based Visual Classification of Emphysema: Association with Mortality in the COPDGene Study. *Radiology* 2018;288:859–66.

10. Young KA, Regan EA, Han MK, Lutz SM, Ragland M, Castaldi PJ, *et al*. Subtypes of COPD Have Unique Distributions and Differential Risk of Mortality. *Chronic Obstr Pulm Dis* 2019;6:400–413.

11. Burgel PR, Paillasseur JL, Peene B, Dusser D, Roche N, Coolen J, *et al*. Two distinct chronic obstructive pulmonary disease (COPD) phenotypes are associated with high risk of mortality. *PloS One* 2012;7.

12. Kinney GL, Santorico SA, Young KA, Cho MH, Castaldi PJ, San José Estépar R, *et al*. Identification of Chronic Obstructive Pulmonary Disease Axes That Predict All-Cause Mortality: The COPDGene Study. *Am J Epidemiol*2018;187:2109–2116.

13. Moll M, Qiao D, Regan EA, Hunninghake GM, Make BJ, Tal-Singer R, *et al*. Machine Learning and Prediction of All-Cause Mortality in COPD. *Chest* 2020;158:952-964.

14. Strand M, Austin E, Moll M, Pratte KA, Regan KA, Hayden LP, *et al*. A risk prediction model for mortality among smokers in the COPCGene® study. *Chronic Obstr Pulm Dis* 2020; 7:346-361.

15. Han M, Kazerooni E, Lynch D, Liu LX, Murray S, Curtis JL, *et al*. Chronic Obstructive Pulmonary Disease Exacerbations in the COPDGene Study: Associated Radiologic Phenotypes. *Radiology* 2011;261:274-282.

16. Müllerova H, Maselli DJ, Locantore N, Vestbo J, Hurst JR, Wedzicha JA, *et al*. Hospitalized exacerbations of COPD: risk factors and outcomes in the ECLIPSE cohort. *Chest* 2015;147:999-1007.

17. Tavakoli H, Chen W, Sin DD, FitzGerald JM, Sadatsafavi M. Predicting Severe Chronic Obstructive Pulmonary Disease Exacerbations: Developing a Population Surveillance Approach with Administrative Data. *Ann Am Thorac Soc* 2020;17:1069-1076.

18. Bhatt SP, Soler X, Wang X, Murray S, Anzueto AR, Beaty TH, *et al*. Association between functional small airway disease and FEV1 decline in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2016;194:178–84.

19. Ross JC, Castaldi PJ, Cho MH, Hersh CP, Rahaghi FN, Sánchez-Ferrero GV, *et al*. Longitudinal Modeling of Lung Function Trajectories in Smokers with and without Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 2018;198:1033-1042.

20. Paoletti M, Camiciottoli G, Meoni E, Bigazzi F, Cestelli L, Pistolesi M, *et al*. Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of Chronic Obstructive Pulmonary Disease (COPD) phenotypes. *J Biomed Inform* 2009;42:1013-1021.

21. Cho MH, Washko GR, Hoffmann TJ, Criner GJ, Hoffman EA, Martinez FJ, *et al*. Cluster analysis in severe emphysema subjects using phenotype and genotype data: an exploratory investigation. *Respir Res* 2010;11:30.

22. Vanfleteren LEGW, Spruit MA, Groenen M, Gaffron S, van Empel VPM, Bruijnzeel PLB, *et al*. Clusters of Comorbidities Based on Validated Objective Measurements and Systemic Inflammation in Patients with Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 2013;187:728-735.

23. Fens N, van Rossum A, Zanen P, van Ginneken B, van Klaveren R, Zwinderman A, *et al*. Subphenotypes of Mild-to-Moderate COPD by Factor and Cluster Analysis of Pulmonary Function, CT Imaging and Breathomics in a Population-Based Survey. *COPD* 2013;10:277-285.

24. Castaldi PJ, Dy J, Ross J, Chang Y, Washko GR, Curran-Everett D, *et al*. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax* 2014;69:415–422.

25. Rennard S, Locantore N, Delafont B, Tal-Singer R, Silverman EK, Vestbo J, *et al*. Identification of Five Chronic Obstructive Pulmonary Disease Subgroups with Different Prognoses in the ECLIPSE Cohort Using Cluster Analysis. *Ann Am Thorac Soc* 2015;12:303-312.

26. Burgel PR, Paillasseur JL, Caillaud D, Tillie-Leblond I, Chanez P, Escamilla R, *et al*. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur Respir J* 2010;36:531-539.

27. Castaldi PJ, Boueiz A, Yun J, Estepar RSJ, Ross JC, Washko G, *et al*. Machine Learning Characterization of COPD Subtypes: Insights From the COPDGene Study. *Chest* 2020;157:1147-1157.

28. Pistolesi M, Camiciottoli G, Paoletti M, Marmai C, Lavorini F, Meoni E, *et al*. Identification of a predominant COPD phenotype in clinical practice. *Respir Med* 2008;102:367-376.

29. Garcia-Aymerich J, Gómez FP, Benet M, Farrero E, Basagaña X, Gayete À, *et al*. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax* 2011;66:430-437.

30. Boueiz A, Chang Y, Cho MH, Washko GR, San José Estépar R, Bowler RP, Crapo JD, *et al*. Lobar Emphysema Distribution Is Associated With 5-Year Radiological Disease Progression. *Chest* 2018;153:65-76.

31. Burgel P, Paillasseur J, Janssens W, Piquet J, Riet GT, Garcia-Aymerich J, *et al*. A simple algorithm for the identification of clinical COPD phenotypes. *Eur Respir J* 2017;50.

32. Friedlander AL, Lynch D, Dyar LA, Bowler RP. Phenotypes of chronic obstructive pulmonary disease. *COPD*2007;4:355-384.

33. Chen J, Cho M, Silverman EK, Hokanson JE, Kinney GL, Crapo JD, *et al*. Turning subtypes into disease axes to improve prediction of COPD progression. *Thorax* 2019;74:906-909.

34. Roy K, Smith J, Kolsum U, Borrill Z, Vestbo J, Singh D. COPD phenotype description using principal components analysis. *Respir Res* 2009;10.

35. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, *et al*. Genetic epidemiology of COPD (COPDGene) study design. *COPD* 2010;7:32-43.

36. Lowe KE, Regan EA, Anzueto A, Austin E, Austin JHM, Beaty TH, *et al*. COPDGene® 2019: Redefining the Diagnosis of Chronic Obstructive Pulmonary Disease. *Chronic Obstr Pulm Dis J COPD Found* 2019;6:384–99.

37. Cho, O., Oh, Y. T., Chun, M., Noh, O. K., & Heo, J. S. (2018). Prognostic implication of FEV1/FVC ratio for limited-stage small cell lung cancer. *Journal of thoracic disease*, *10*(3), 1797–1805. https://doi.org/10.21037/jtd.2018.02.14

38. Johns DP, Walters JAE, Walters EH. Diagnosis and early detection of COPD using spirometry. *J Thorac Dis* 2014;6:1557-1569.

39. Vestbo J, Hurd S, Agustí A, Jones P, Vogelmeier C, Anzueto A, *et al*. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med*2013;184:347–65.

40. Lynch D, Austin JHM, Hogg JC, Grenier PA, Kauczor HU, Bankier AA, *et al*. CT-Definable Subtypes of Chronic Obstructive Pulmonary Disease: A Statement of the Fleischner Society. *Radiology* 2015;277:192-205.

41. González G, Ash SY, Vegas-Sánchez-Ferrero G, Onieva JO, Rahaghi FN, Ross JC, *et al*. Disease Staging and Prognosis in Smokers Using Deep Learning in Chest Computed Tomography. *Am J Respir Crit Care Med* 2018;197:193-203.

CHAPTER 3: DEVELOP AUTOMATED COMPUTATIONAL METHODS TO DEFINE CLINICALLY MEANINGFUL STAGES OF DISEASE USING COMPUTED TOMOGRAPHY (CT) FOR PATIENTS WITH COPD

Online dating platforms such as Tinder have significantly facilitated romantic attraction based on visual appeal. A swipe left means that the user is not interested in the person the algorithm presented them through the app, while a swipe right means that the user is interested. A recent sociological study conducted by researchers from Michigan State University and the University of Maryland has shown that it takes less than a second for people on dating apps to decide whether or not they would be interested in getting to know someone further, and their decision is almost solely based on the other person's appearance.[1] We owe the rapidity of this virtual mate selection process to the ability of our visual system to capture, process, and discriminate among all the physical features presented in the picture before we decide whether we want to invest additional time on getting to know this person. Indeed, all the efforts that many males make to advertise their mating potential through these pictures – from pumping iron in the gym to cuddling suspiciously docile tiger cubs – would be for naught had we not evolved a set of eyes to appreciate these forms of human peacocking.

The human eye is so complex that creationists cite it as a product of intelligent design rather than natural selection.[2] Indeed, even Charles Darwin himself seemed to marvel at the intricate nature of the eye, expressing in his 1859 magnus opus, *On the Origin of Species*, that it would be "absurd" to think that the eye evolved through natural selection, though he reaffirmed that the eye did evolve this way, despite not having direct evidence of intermediate forms in his time.[2] It has been challenging to study eye evolution, given how difficult and rare it is for soft tissue structures to be preserved over time.

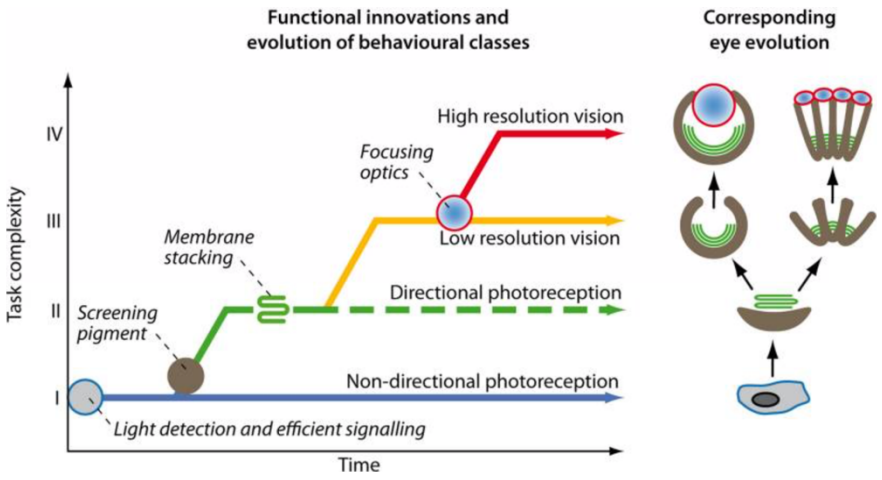

Figure 3.1[1]: Nilsson's conceptual diagram of eye evolution showing the emergence of key functional innovations on a vertical scale of task complexity. The green dashed line for directional photoreception indicates that class II task may become superfluous by the evolution of class III tasks. Other classes remain relevant after the evolution of a higher class. For example, eye structures capable of class IV tasks can still perform class III tasks. The receptor/eye morphologies corresponding to each task are shown on the right.

Biologists have nonetheless made significant progress by studying how the eye forms in developing embryos and through comparative genomics and comparative anatomy studies.[2] They have concluded that our human eye, and more generally speaking, the eye of jawed vertebrates, evolved in less than 100 million years, from a simple light sensor for circadian and seasonal rhythms around 600 million years ago to a complex camera-like organ by 500 million years ago.[2,3] Nilsson characterized four key innovations in eye evolution, namely, light sensitivity, screening structures, membrane stacking, and focusing optics.[1] Figure 3.1 summarizes the development of sensory tasks for different classes of photoreceptive behaviors that emerged as corresponding types of sensory structures evolved.[1]

The ability of this optically and neurologically complex organ to perform these sensory tasks has inspired researchers in computer vision to develop algorithms that mimic this ability *in silico*. Our efforts to model the visual cortex harken back to the mid-twentieth century, when Hubel and Wiesel discovered two major cell types in the feline primary visual cortex: simple

cells and complex cells.[4] Simple cells respond to bars of light or dark when placed at specific special orientations.[4] Each simple cell has an orientation of the bar at which it fires the most, which a response curve corresponding to how far the angle of the bar changes from its preferred orientation.[4] In comparison, complex cells still have preferred orientations but have more spatially invariant responses.[4] Hubel and Wiesel postulated that these complex cells are receiving input from several simple cells, all having the same preferred orientation but at slightly different locations.[4] Figure 3.2 shows the spatial relationship between simple cells and complex cells, which has inspired the foundational operations of a convolutional neural network.[4]

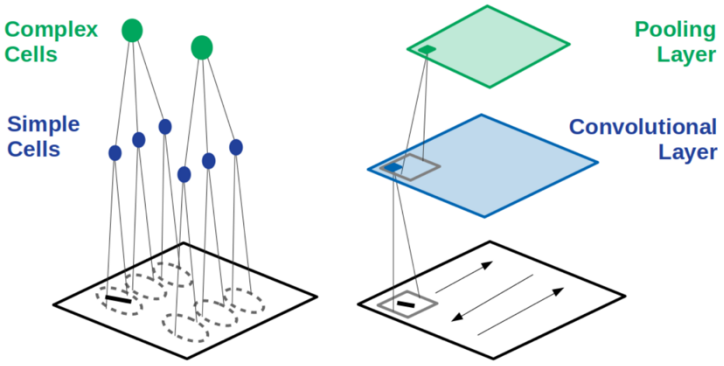

Figure 3.2[4]: Lindsay's conceptual diagram showing the relationship between components of the visual system and the fundamental framework of a convolutional neural network. The left summarizes Hubel and Wiesel's findings on the dynamics between simple cells and complex cells. The right shows these dynamics in the context of a convolutional neural network. Specifically, the application of a small filter (gray box) to every location in the image creates a feature map.[4] A convolution operation is then applied to the image, creating a convolutional layer that has as many feature maps as the number of filters applied. A max-pooling operation is then applied to the convolutional layer, wherein the maximal activation is taken in a small section of each feature box, which downsamples the image and creates a pooling layer comprising downsampled feature maps.

In 1980, Fukushima integrated Hubel and Wiesel's findings into a computational model of the visual system, called the Neocognitron, which is the precursor to modern convolutional neural networks.[4,5] The Neocognitron has two main cell types, S-cells and C-cells, named and functionally modeled after simple cells and complex cells, respectively.[5] Specifically, applying a 2D grid of weights at each location in the input image creates a "plane" of S-cell responses with

all cells sharing the same preferred visual features.[5] C-cells create a response that is defined as a nonlinear function of several S-cells coming from the same plane but at different locations.[5] Figure 3.3 shows the correspondence between the biologically inspired hierarchy model by Hubel and Wiesel and the neural network architecture of the Neocognitron. Figure 3.4 shows how each layer in the Neocognitron is interconnected.[5] Figure 3.5 provides an example of how the Neocognitron "reads" the letter "A".[5]

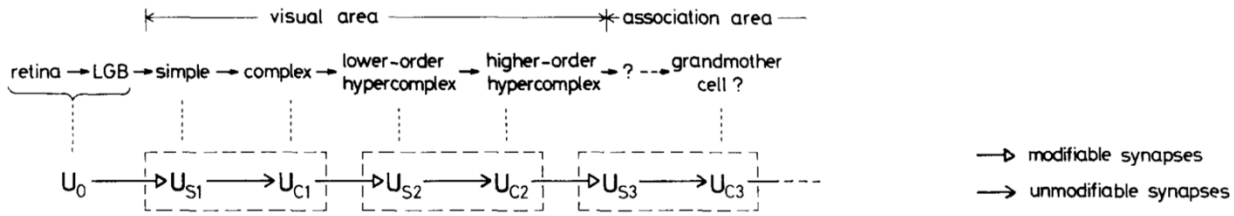

Figure 3.3[5]: Fukushima's schematic diagram showing how the S-cells and C-cells of the Neocognitron correspond to simple and complex cells, respectively. The Neocognitron contains a "cascade connection" of an input layer U_0, followed by several modules, each of which is composed of two layers of cells connected in a "cascade".[5] The S-layer (with S-cells) in the *l*-th module is labeled as U_{Sl}, while the C-layer (with C-cells) in the *l*-th module is labeled as U_{Cl}.[5]

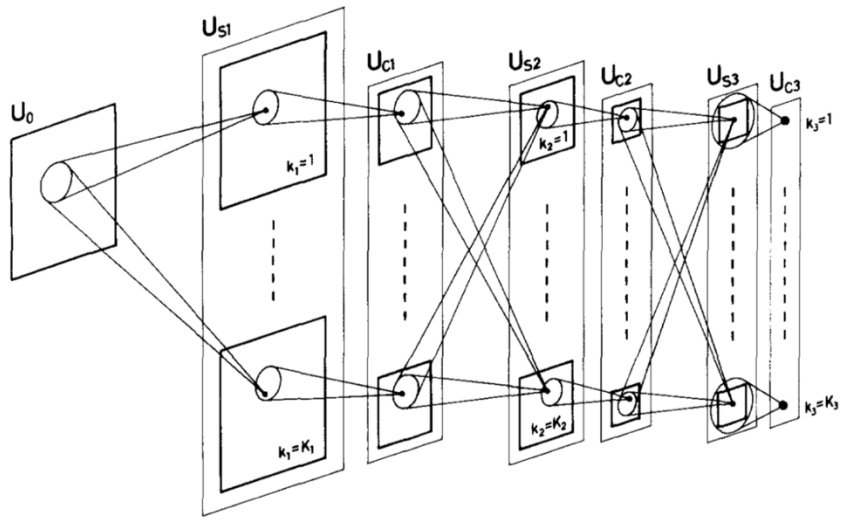

Figure 3.4[5]: Fukushima's schematic diagram showing the interconnected layers in the Neocognitron architecture.

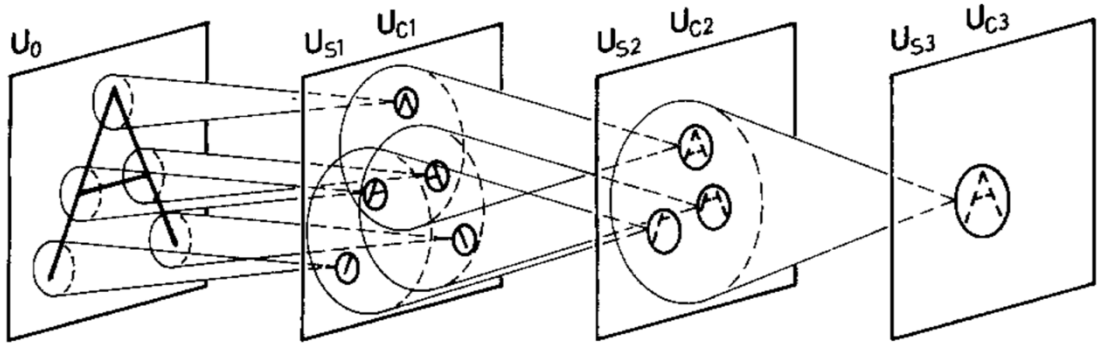

Figure 3.5[5]: Fukushima's schematic diagram showing an example of how the Neocognitron processes the input, the letter "A", through its interconnected layers.

Building on the work of Hubel, Wiesel, and Fukushima, LeCun and his group demonstrated the significant predictive power of the convolutional neural network in their 1989 canonical paper, which describes the successful training of a small convolutional neural network using backpropagation to perform handwritten digit classification.[6] The neural network architecture is shown in Figure 3.6.

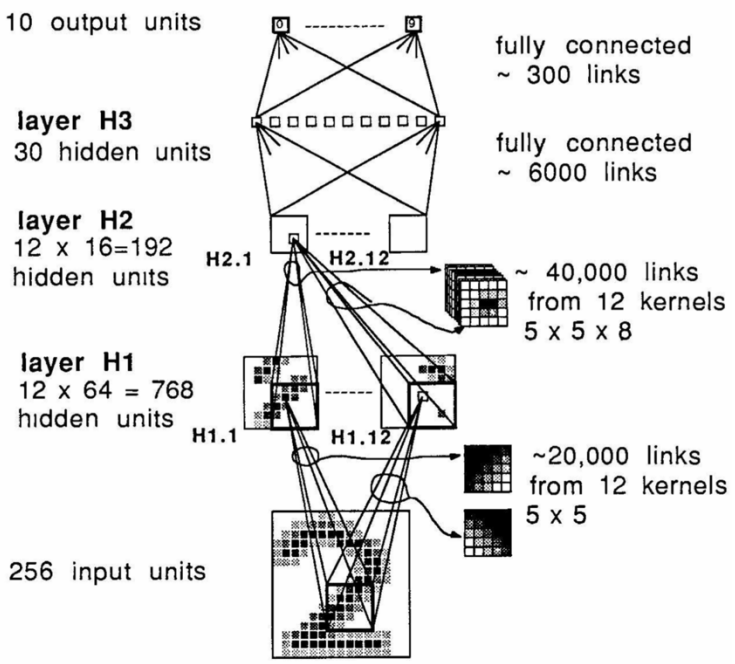

Figure 3.6[6]: Architecture of LeCun *et al.*'s convolutional neural network trained using backpropagation.

Since this demonstration there have been many different model architectures, varying in parameters such as network depth, placement and number of pooling layers, number of feature maps per layer, training procedures, and the number of residual connections that skip layers.[4] Figure 3.7 shows a timeline highlighting some of the widely used model architectures that have emerged from the collective efforts of researchers over the past three decades.[7]

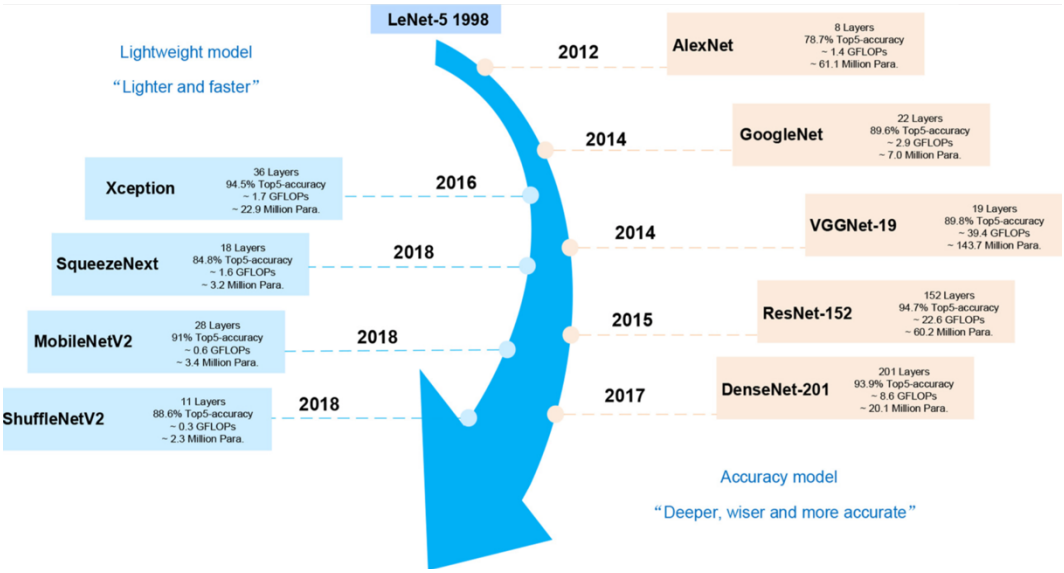

Figure 3.7[7]: Conceptual timeline outlined by Peng Xiyuan *et al.* showing the concomitant evolution of increasingly lighter architectures and more accurate architectures.

Not represented on this timeline is the U-Net convolutional neural network architecture, developed by Ronneberger *et al.* in 2015 in part to address the drawbacks inherent in Ciresan *et al.*'s sliding-window training strategy.[8] Specifically, the sliding-window setup is less computationally efficient because the network has to run separately for each patch, with much redundancy due to overlapping patches.[8] Additionally, researchers see a significant trade-off between localization accuracy and the use of context: Larger patches need more max-pooling layers that reduce localization accuracy, while smaller patches allow the network to see only limited context.[8]

The U-Net architecture provides mechanisms to mitigate this trade-off, making it possible to have high localization accuracy and good use of context (Figure 3.8).[8] The U-Net achieves this by adding successive convolution layers to the usual contracting network, thereby increasing the resolution of the output.[8] High localization accuracy is achieved when high resolution features from the contracting arm of the network are combined with the upsampled output, which then allows a successive convolution layer to learn to generate a more precise output.[8] Notably, there is a large number of feature channels in the upsampling part of the expansive arm of the network, which allow the network to propagate "context information" upward to higher resolution layers.[8] The two arms of the network are quite symmetrical, hence forming a u-shaped architecture.

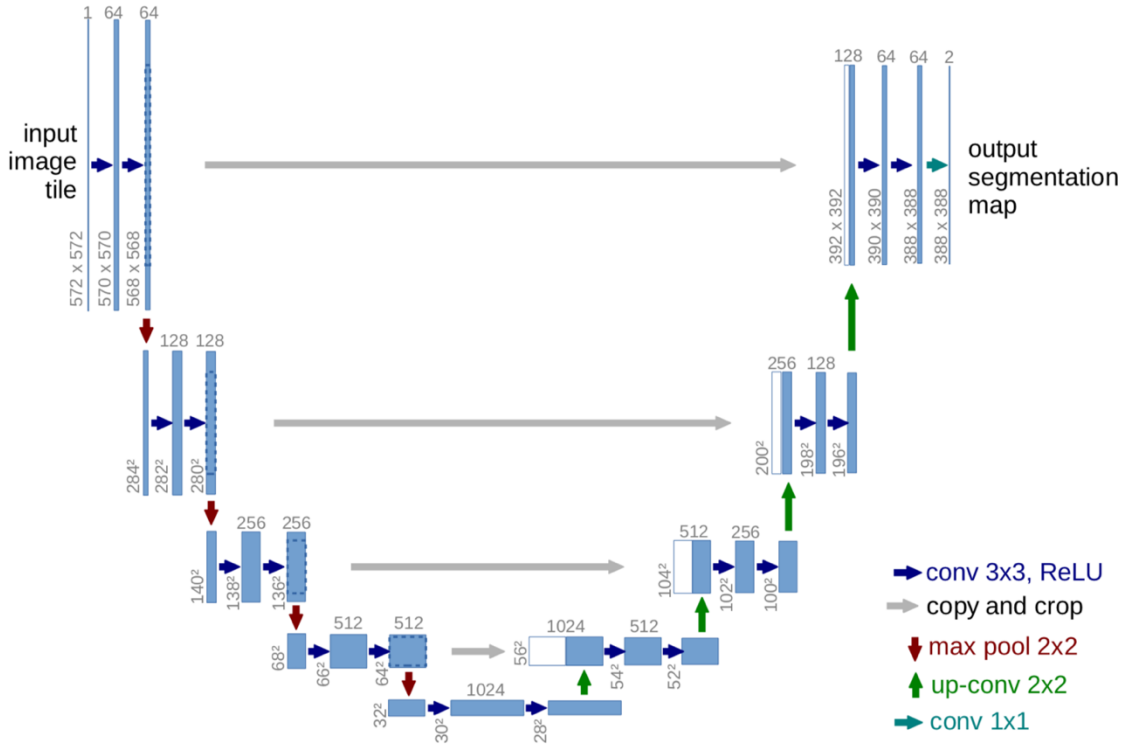

Figure 3.8[8]: U-Net architecture developed by Ronneberger *et al*. Each blue box represents a multi-channel feature map, with the number of channels found on top of the box. White boxes correspond to copied features maps. Arrows show the different operations.

Since 2015 the U-Net has been widely used for fast and precise segmentation of biomedical images.[8] Here I present the second chapter of this thesis, in which we developed an automated lung segmentation algorithm using a U-Net inspired 3D convolutional neural network to quantify pathophysiological imaging features and define stages of disease on CT for patients in the COPDGene cohort.

3.1 Introduction

We can infer from our unsupervised analysis that CT imaging measures of emphysema and air trapping, as well as pulmonary function measures of vital capacity, are important for predicting COPD patient outcomes. We know that for the past 60 years, pulmonary function measures such as FEV_1pp and FVC have been used to define the Global Initiative for Obstructive Lung Disease (GOLD) stages of COPD severity. Currently, the ratio of FEV_1 to FVC is used in the diagnosis of COPD. According to GOLD criteria, if the FEV_1 to FVC ratio falls below 0.7, the patient has a confirmed COPD diagnosis. The GOLD criteria further classify patients into mild, moderate, severe, and very severe groups based on their predicted normal FEV_1 percentage (FEV_1pp) (Table 3.1).

Table 3.1. GOLD defines a FEV_1/FVC threshold of less than 0.7 as a confirmed COPD diagnosis.

GOLD Stages		
In patients with FEV1/FVC < 0.70 (based on post-bronchodilator FEV1)		
GOLD 1	Mild	FEV_1 > 80% predicted
GOLD 2	Moderate	50% ≤ FEV_1 < 80% predicted
GOLD 3	Severe	30% ≤ FEV_1 < 50% predicted
GOLD 4	Very Severe	FEV_1 < 30% predicted

Lower FEV_1 values indicate more severe disease. The main shortcoming of using these two spirometric measurements to diagnose and stage COPD is that many times they fail to reflect the extent of the underlying lung damage that can otherwise be captured with CT imaging.

COPD diagnosis and staging typically relies on history of tobacco use or second-hand exposure, symptomatology, and pulmonary function testing. However, it is becoming apparent that quantitative measurements from CT have potential diagnostic and prognostic value as these measurements correlate with spirometry-based pulmonary function testing.[9,10,11] As such, CT measurements have been included in the recently updated diagnostic criteria for COPD from investigators of the National Institutes of Health-supported COPDGene study.[12]

Despite the emerging use of quantitative CT, stratification of patients with suspected COPD is currently based on spirometric pulmonary function testing according to Global Initiative for Chronic Obstructive Lung Disease (GOLD) stages.[13] Analogous staging criteria by CT have not been defined to date. As diagnostic criteria begin to incorporate CT, an open question becomes how to use quantitative CT measurements to establish disease severity in the clinic. Current clinical determinations rely heavily on qualitative visual assessment.[14,15] Air trapping, in particular, is an imaging feature associated with small airway disease and is a

clinically significant component of COPD.[15,16] However, it is often overlooked as it typically cannot be directly visualized like emphysema, bronchial wall thickening, or endobronchial mucus plugging from a single inspiratory phase. Even so, deformable registration algorithms have been used to both quantify and visualize the distribution of air trapping and recent innovations in deep convolutional neural networks (CNN) have shown promise to automate and supplement a variety of tasks in medical imaging, including automated lung and lung lobe segmentation.[17,18,19,20,21,22,23,24,25,26,27,28] This study is motivated by our recognition that quantitative CT measurements, though explored in the research domain, have not yet permeated the clinical arena. Our goal in this was to provide the computational infrastructure to define a set of clinically pragmatic CT stages of COPD severity.

We build on prior work by others to assess the feasibility of defining CT-based stages of COPD severity based on a combination of air trapping and emphysema obtained from inspiratory and expiratory phase CT. With the hypothesis that these two CT metrics alone may provide clinical prognostic value, we leveraged deep learning and deformable registration techniques to automate quantification of emphysema and air trapping across patients in the COPDGene study. We used these imaging features to propose CT-based stages for disease severity. We then assessed the ability of our CT-based criteria to prognosticate disease progression and mortality. To further illustrate the clinical applicability of our lung quantification algorithm, we tested our model on a small cohort of patients from UCSD Health who may have COPD, cystic fibrosis, and other life-threatening conditions.

3.2 Design and implementation of DeepLungQuant CNN model

3.2.1 Study cohorts

This retrospective Health Insurance Portability and Protection Act-compliant study was approved by the institutional review boards of the participating institutions with waived requirement for written informed consent. There were two distinct sets of patients included in this study: one for development of a lung segmentation CNN, and a second utilizing this CNN and deformable registration to define CT-based stages for COPD severity.

3.2.2 Dataset for DataLungQuant model development

We retrospectively collected a convenience sample of 1037 volumetric CT series from 888 current or former smokers undergoing lung cancer screening using low radiation exposure protocols to develop our own deep learning-based lung segmentation algorithm. Studies included 621 non-contrast CT examinations obtained as part of the National Lung Screening Trial (NLST) and 416 non-contrast CT examinations sampled from our institution between August 2008 and October 2018.[29] We used a semi-automated annotation software called Coreline to perform segmentations of each lung using volumetric CT series as the input, as shown in Figure 3.9.

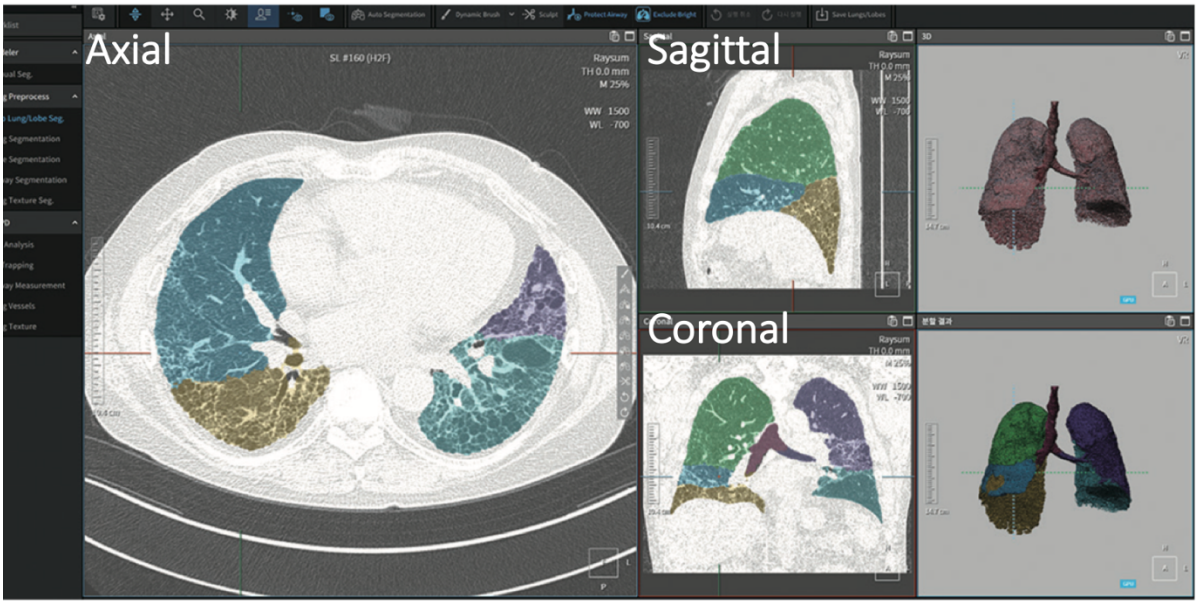

Figure 3.9. **Coreline is a semi-automated annotation software that performs lobar segmentations of each lung and outputs these segmentations as masks that were then used to train the DeepLungQuant model**.

Coreline outputs these segmentations as ground-truth masks for each lung, lung lobe, and trachea, which were then visually inspected and rectified using ITK-SNAP, another annotation software, on an as needed basis. Imaging data was anonymized prior to CNN training, therefore demographic information is not available.

3.2.3 Dataset for CT-based staging of COPD severity

To develop a CT-based strategy for staging COPD severity, we used baseline non-contrast CT and spirometric data from all 9,652 patients with CT examinations from the COPDGene study.[30] The COPDGene project has resulted in almost 400 publications from other research groups.[31] The analysis presented here, which defines CT-based criteria for staging COPD, had not been previously performed on the COPDGene dataset. Exclusion criteria were missing inspiratory or expiratory CT (n = 536) and missing all-cause mortality status (n = 165),

resulting in 8951 patients that were included in this study. Patients from this dataset were used for testing of the CNN and derivation of stages of disease severity.

Table 3.2. Demographic, spirometric, and CT imaging characteristics of the study cohort.

Characteristic	Value
Demographics (n = 8951)	
Men	53.35% (4775/8951)
Race (% NHW)	68.04% (6090/8951)
Age	60 ± 9
Body mass index	28.8 ± 6.2
Pack years	44.2 ± 24.9
Current smokers	52.54% (4703/8951)
Spirometry (n = 8895)*	
FEV1pp	77.3 ± 27.0
FVCpp	89.2 ± 18.1
FEV1/FVC	0.65 ± 0.17
GOLD 0	3855
GOLD 1	695
GOLD 2	1723
GOLD 3	1022
GOLD 4	529
PRISm	1071
Imaging (n = 8951)	
%EM	6.2 ± 9.8
%AT	41.5 ± 20.4
%TLI	47.9 ± 23.6
Normal	3278
Mild	1405
Moderate	1456
Severe	1175
Very Severe	1637

Spirometric measurements included forced expiratory flow in one second (FEV_1), percent predicted FEV_1 (FEV_1pp), and forced vital capacity (FVC) following administration of 180 g of albuterol at baseline and 5-year follow-up.[20] Baseline CT images and spirometric measurements were collected between 2007-2011; 5-year follow-up measurements were collected between 2013-2017. Data also included 10-year all-cause mortality, determined through longitudinal monitoring of patients every 6 months by the COPDGene investigators since 2009. Patients with

FEV_1/FVC of less than 0.7 were classified using the GOLD staging system for COPD severity; patients with FEV_1/FVC of greater than 0.7 and FEV_1pp of greater than 80% were used as the normal spirometric reference group (GOLD 0).[21] Other patients had airflow obstruction (FEV_1pp < 80%) and reduced FVC on spirometry, leading to a preserved FEV_1/FVC ratio (> 0.7). These patients, comprising 12% of the COPDGene cohort, were classified as having preserved ratio impaired spirometry (PRISm), and were not included in the analysis of COPD status for this study. Demographics are provided in Table 3.2.

3.2.4 CT image acquisition

CT scans for the patients within the COPDGene cohort were acquired using multi-detector General Electric, Philips, or Siemens CT scanners, each with at least 16 detector channels. Images comprised inspiratory (200 mAs) and expiratory (50 mAs) acquisitions of the entire thorax without contrast. Images were reconstructed using a standard soft tissue kernel, utilized sub-millimeter slice thickness (0.625-0.9mm) and intervals (0.45-0.625mm), with smooth and edge-enhancing algorithms. Further details about the COPDGene study design and imaging protocols are described in Regan *et al.*, keeping in mind that phenotyping depends on data richness and integrity.[31]

3.2.5 Quantitative analysis of emphysema and air trapping

Dr. Kyle Hasenstab designed and implemented our deep learning lung segmentation algorithm, DeepLungQuant, shown in Figure 3.10, which is a U-Net-inspired 3D CNN trained to segment the lungs, five lung lobes, and trachea on volumetric CT series. The input to the CNN is a 192x192x192x1 array representing the preprocessed CT series. The output to the CNN is a

192x192x192x10 array whose channels comprise binary masks of the left and right lungs, left and right lung lobes, left and right lung fissures, and trachea.

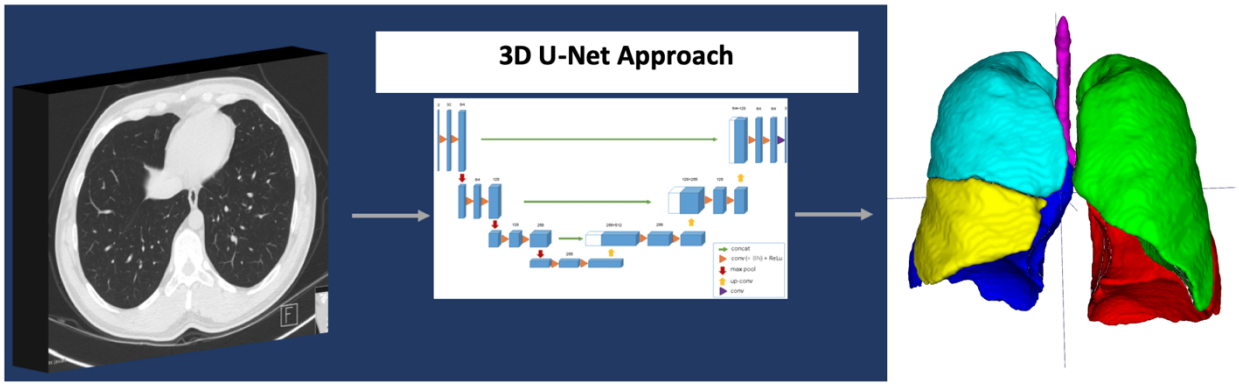

Figure 3.10. **DeepLungQuant is a whole lung and lobar segmentation model whose architecture is a U-net-inspired 3D convolutional neural network (CNN) trained to segment the lungs, five lung lobes, and trachea on volumetric CT series.**

The analysis pipeline, shown in Figure 3.11, combines our deep learning lung segmentation algorithm with deformable registration to automate quantification of emphysema and air trapping.[32] Deformable registration, generally speaking, is a process by which two or more 3D images are aligned into a common coordinate system.[33] In our case, the expiratory series on CT were deformably registered to the inspiratory series on CT for each patient.

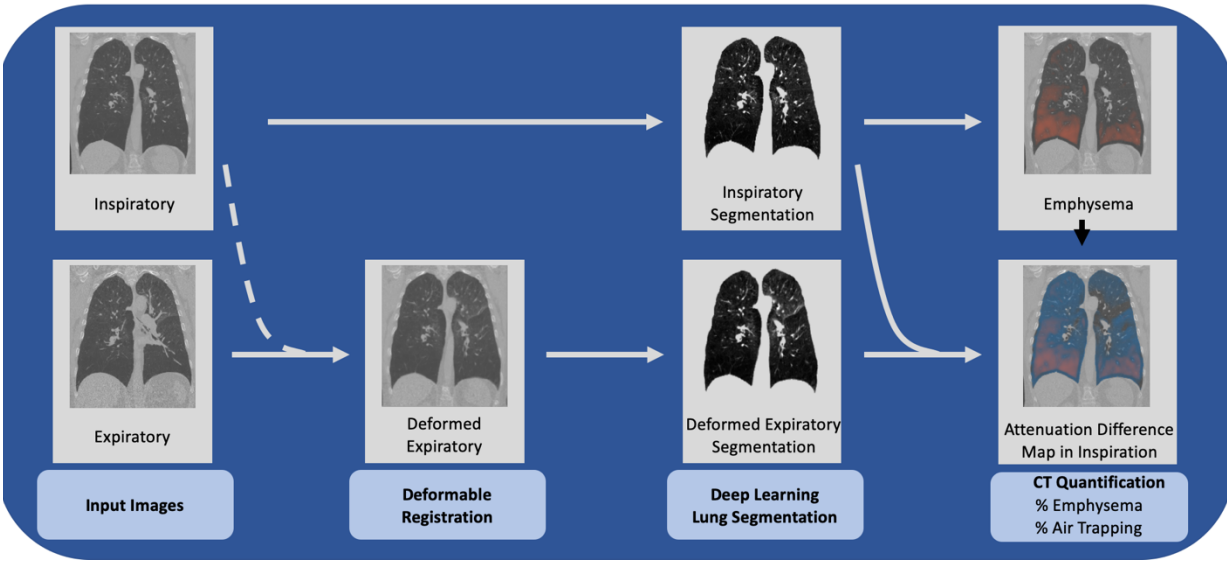

Figure 3.11. DeepLungQuant is a deep learning algorithm that directly processes CT images and quantifies imaging features such as emphysema and air trapping.

Specifically, expiratory series were deformably registered to inspiratory series using a symmetric diffeomorphic registration algorithm.[34,35] Lungs masks were then generated from inspiratory images using our lung segmentation CNN. Pixel-wise subtractions of co-registered inspiratory and expiratory attenuation were then used to compute attenuation difference maps, henceforth referred to as "lung disease maps". These lung disease maps are overlaid spatially to visualize total lung involvement (TLI) of the disease.

Our goal was to separate distinct regions of the lungs affected by emphysema from those affected by air trapping using inspiratory images and lung disease maps. As shown in Figure 3.12, we may use these lung maps to visualize how each region of the lungs are affected by disease.[32]

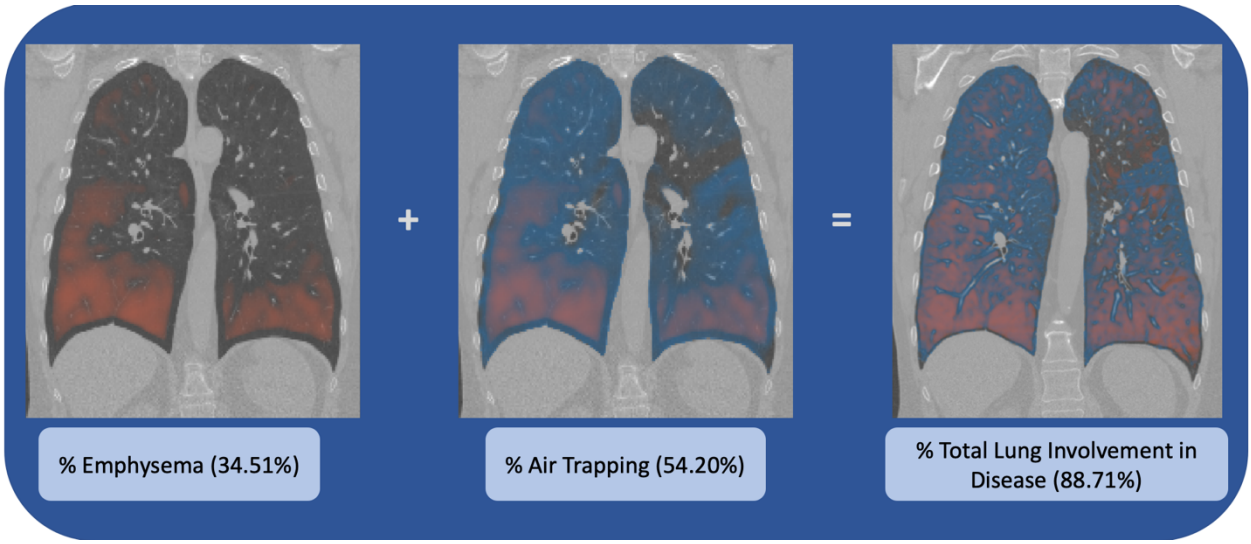

Figure 3.12. Intermediate captures from the DeepLungQuant can be overlaid to assess the pathophysiological changes due to COPD overtime.

Our approach leveraged standard imaging parameters to define regions of emphysema (EM) and air trapping (AT). Specifically, the percent emphysema (%EM) was computed as the percentage of lung on the inspiratory images with attenuation less than or equal to -950 HU; percent air trapping (%AT) was computed as the percentage of voxels with attenuation differences of less than or equal to 100 HU and inspiratory attenuation of greater than -950 HU.[28,29] We then defined the percentage of total lung involvement (%TLI) as the total lung affected by either emphysema or air trapping (%TLI = %EM + %AT). Note our approach for quantification of air trapping is similar to that of parametric response maps.[18,19]

3.2.6 Selection of thresholds to define CT-based stages of COPD

Applying this deep learning lung segmentation algorithm, DeepLungQuant, we had the emphysema and air trapping measures automatically quantified. Using these measures, we defined CT-based stages of COPD severity for patients with normal, mild, moderate, severe, and very severe disease on CT, as shown in Table 3.3.

Table 3.3. Five pragmatic CT-based stages of COPD: normal, mild, moderate, severe, and very severe.

CT Stage	% Emphysema		% Air Trapping/Total Lung Involvement
Normal	< 1%	or	<10%
Mild	> 1%	and	10-33%
Moderate	1-1.5%	or	33-50%
Severe	1.5-3%	or	50-66%
Very Severe	>3%	and	>66%

To derive these CT-based stages, GOLD stages of COPD severity were converted into four categories: *(a)* GOLD 1-4 versus not, *(b)* GOLD 2-4 versus not, *(c)* GOLD 3-4 versus not, and *(d)* GOLD 4 versus not. We then performed a search of thresholds on %EM and %TLI imaging features that jointly maximized the Youden index for predicting these GOLD categories. The Youden index (also known as Youden's J statistic) integrates sensitivity and specificity information with a value that falls between 0 and 1, and is defined as:

$$J = sensitivity + specificity - 1$$

Thresholds were then used to guide the proposal of five CT-based stages of COPD: normal, mild, moderate, severe, and very severe. Note that we intentionally defined CT-based

stages in this way to parallel the widely used functional GOLD stage classes. Further discussion regarding how these CT-based stages were derived can be found in the supplement.[32]

3.2.7 Prognostic evaluation of proposed CT-based stages of COPD

To evaluate the prognostic potential of the CT-based stages of COPD, we assessed the ability of these proposed CT-based stages alone and in conjunction with the GOLD stages to prognosticate two clinically relevant outcomes, abnormal spirometric progression of FEV_1 (FEV_1 loss > 350 mL between baseline and 5-year follow-up) and 10-year all-cause mortality. Additional details on the definition of FEV_1 progression and all-cause mortality can be found in Lowe *et al*.[12]

3.2.8 Statistical analyses

Deformable registrations and lung segmentations were performed in Python-v3.6.2 using the DIPY-v0.14.0 and Keras-v2.2.0 libraries.[34,39] Statistical analysis was performed using R-v3.4.0 software. Patient demographics, imaging features, and spirometry metrics were summarized using descriptive statistics. We used univariable and multivariable logistic regressions to predict GOLD stage using %EM and %AT as inputs and evaluated performance using area under the receiver operating characteristic curve (AUC). Multivariable logistic regression was used to estimate odds of FEV_1 loss across GOLD and CT-based staging, adjusting for patient age at study entry, sex, race, current smoking, pack-years, and baseline FEV_1 post-bronchodilator. We used Cox regression with right censoring to estimate hazard ratios for all-cause mortality across GOLD and CT-based staging, adjusting for sex, race, pack-years and current smoking; age was used as the underlying timescale. Sensitivity and specificity, and confidence intervals for odds and hazard ratios were analytically calculated as appropriate.

Regression *P*-values were adjusted using Bonferroni correction; statistical significance was determined using a 5% Type I error threshold.

3.3 Results

3.3.1 Patient demographics, spirometry, GOLD stage, and CT characteristics

Spirometric measurements and GOLD stage are summarized in Table 3.2. This is the dataset that was used in our logistic regression and Cox proportional hazards modeling. A total of 56 patients had missing spirometric measurements and were omitted from analyses involving GOLD stages. GOLD 0 contained the largest number of patients (43%; 3855 of 8895); however, mean FEV_1pp (77.3% ± 27.0%) and FEV_1/FVC (0.65 ± 0.17) were less than the thresholds (FEV_1pp = 80% and FEV_1/FVC = 0.70) separating GOLD 0 from other GOLD stages, reflecting the distributional left skewness of the spirometric measurements. GOLD 4 contained the least number of patients (6%; 529 of 8895). Quantitative measures across the entire group of 8951 patients showed a mean %EM of 6.2 ± 9.8, %AT of 41.5 ± 20.4, and %TLI of 47.9 ± 23.6 (Table 3.2).

3.3.2 Evaluation of prognostic potential of CT-based stages

Using logistic and Cox regressions, we assessed the ability of these proposed CT-based stages alone and in conjunction with the GOLD stages to prognosticate two clinically relevant outcomes, abnormal spirometric progression of FEV_1 (FEV_1 loss > 350 mL between baseline and 5-year follow-up) and 10-year all-cause mortality. Odds ratios and hazard ratios for FEV_1 progression and all-cause mortality are summarized in Table 3.4.

Table 3.4. Predictive ability of imaging and spirometry for COPD progression and all-cause mortality including patients with PRISm (Continued). Patients with PRISm showed significantly greater odds of all-cause mortality compared to GOLD 0 patients.

	FEV1 Progression			All-Cause Mortality		
Characteristic	N	Odds Ratio [95% CI]	Adj. P-Value	N	Hazard Ratio [95% CI]	Adj. P-Value
Imaging only						
None	1742	1		3278	1	
Mild	853	1.07 [0.88, 1.31]	1.000	1405	0.83 [0.67, 1.04]	0.807
Moderate	825	**1.50 [1.22, 1.84]**	**0.001**	1456	0.82 [0.68, 0.99]	0.294
Severe	595	**2.88 [2.29, 3.63]**	**<0.001**	1175	1.11 [0.93, 1.32]	1.000
Very Severe	550	**2.67 [2.02, 3.53]**	**<0.001**	1637	**2.23 [1.93, 2.58]**	**<0.001**
Spirometry only						
GOLD0	2238	1		3855	1	
GOLD1	405	**1.55 [1.22, 1.97]**	**0.004**	695	0.83 [0.64, 1.07]	1.000
GOLD2	877	**2.06 [1.60, 2.65]**	**<0.001**	1723	**1.52 [1.29, 1.79]**	**<0.001**
GOLD3	408	1.53 [1.01, 2.30]	0.468	1022	**2.80 [2.39, 3.30]**	**<0.001**
GOLD4	85	0.10 [0.01, 0.75]	0.274	529	**7.15 [6.04, 8.46]**	**<0.001**
PRISm	552	0.77 [0.58, 1.02]	0.746	1071	**1.80 [1.46, 2.20]**	**<0.001**
Imaging and Spirometry						
GOLD0						
None	1124	1		2014	1	
Mild	626	1.02 [0.82, 1.30]	1.000	1004	0.88 [0.65, 1.19]	1.000
Moderate	352	1.35 [1.03, 1.79]	0.923	570	0.69 [0.50, 0.95]	0.692
Severe	102	1.62 [1.04, 2.52]	0.928	209	0.73 [0.44, 1.20]	1.000
Very Severe				58	0.62 [0.23, 1.69]	1.000

Table 3.4. Predictive ability of imaging and spirometry for COPD progression and all-cause mortality including patients with PRISm (Continued). Patients with PRISm showed significantly greater odds of all-cause mortality compared to GOLD 0 patients

GOLD1						
None	81	1.25 [0.73, 2.14]	1.000	151	0.59 [0.33, 1.05]	1.000
Mild	77	1.82 [1.11, 2.99]	0.498	137	0.91 [0.52, 1.61]	1.000
Moderate	131	1.71 [1.15, 2.57]	0.250	224	0.69 [0.46, 1.06]	1.000
Severe	**97**	**2.76 [1.74, 4.37]**	**<0.001**	153	0.73 [0.46, 1.14]	1.000
GOLD2						
None	164	1.06 [0.67, 1.66]	1.000	318	1.29 [0.91, 1.83]	1.000
Mild	63	1.44 [0.79, 2.63]	1.000	104	1.18 [0.65, 2.12]	1.000
Moderate	226	1.71 [1.20, 2.56]	0.096	427	1.34 [1.00, 1.78]	1.000
Severe	265	**4.55 [3.24, 6.38]**	**<0.001**	517	1.36 [1.06, 1.75]	0.499
Very Severe	159	**4.80 [3.20, 7.22]**	**<0.001**	357	1.21 [0.92, 1.59]	1.000
GOLD3						
None				76	**2.64 [1.65, 4.23]**	**0.002**
Moderate				51	1.98 [1.10, 3.58]	0.729
Severe	98	1.98 [1.12, 3.51]	0.549	215	**2.20 [1.63, 2.97]**	**<0.001**
Very Severe	252	**2.23 [1.50, 3.63]**	**0.005**	677	**2.43 [1.97, 3.01]**	**<0.001**
GOLD4						
Very Severe	78	0.16 [0.02, 1.19]	1.000	496	**6.05 [4.92, 7.44]**	**<0.001**
PRISm						
None	341	0.84 [0.59, 1.19]	1.000	688	**1.59 [1.21, 2.10]**	**0.026**
Mild	85	0.68 [0.36, 1.30]	1.000	146	**2.26 [1.43, 3.59]**	**0.016**
Moderate	91	1.17 [0.67, 2.06]	1.000	168	1.39 [0.91, 2.11]	1.000
Severe				55	1.12 [0.57, 2.19]	1.000

In the imaging-only reference category in Figure 3.13, odds of disease progression for patients with moderate (OR, 1.50 [95% CI: 1.22, 1.84]), severe (OR, 2.88 [95% CI: 2.29, 3.63]) and very severe (OR, 2.67 [95% CI: 2.01, 3.53]) were greater than the normal imaging reference category ($P < .001$ for all). Patients with severe to very severe disease on CT had greater odds of progression than patients with mild to moderate disease, suggesting likelihood of progression increases with CT-based severity.

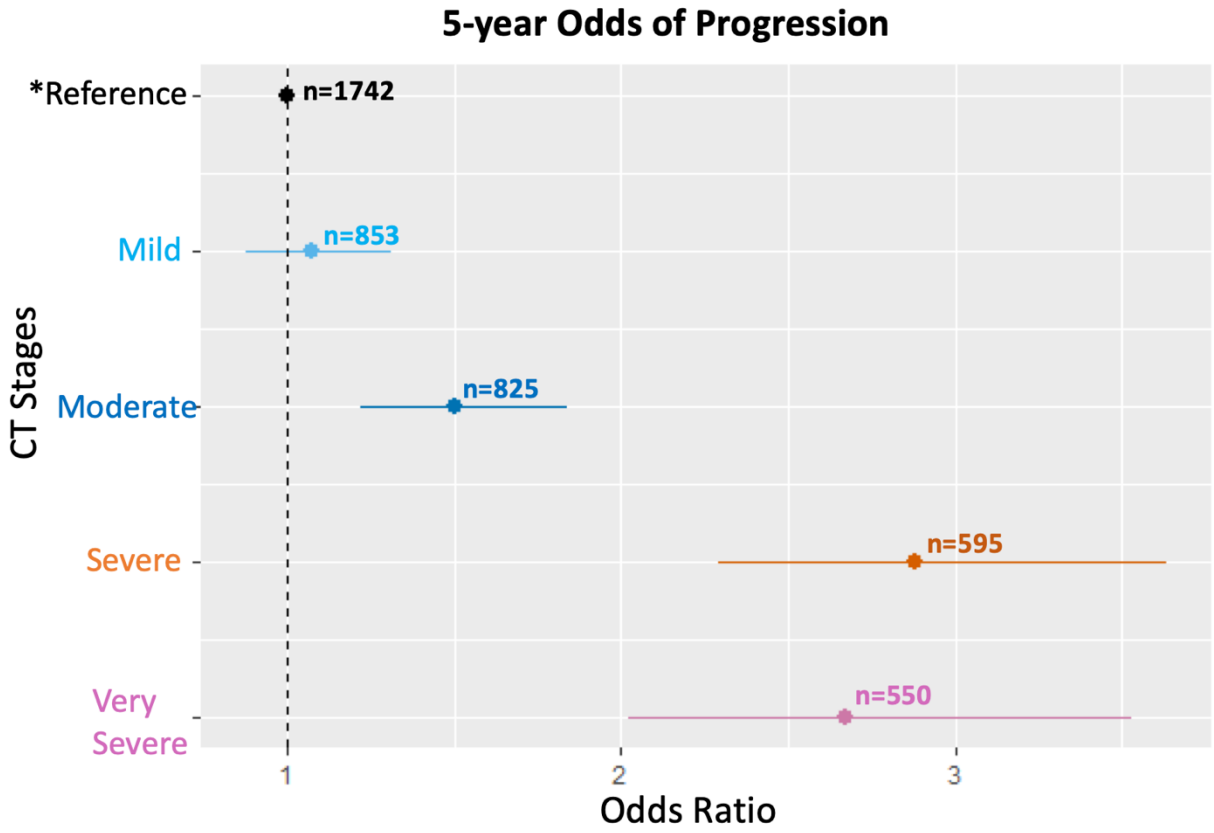

Figure 3.13. Patients with moderate to very severe disease on CT have significantly greater odds of progression in 5 years.

Hazard ratio of mortality for patients with very severe (HR, 2.23 [95% CI: 1.93, 2.58]) disease on CT (Figure 3.14**)** was approximately twice the hazard ratio of patients with normal to severe disease (severe HR, 1.11 [95% CI: 0.93, 1.32]). For the normal imaging reference category, hazard ratios for mild (HR, 0.83 [95% CI: 0.67, 1.04]) to severe (HR, 1.11 [95% CI: 0.93, 1.32]) disease on CT were not different.

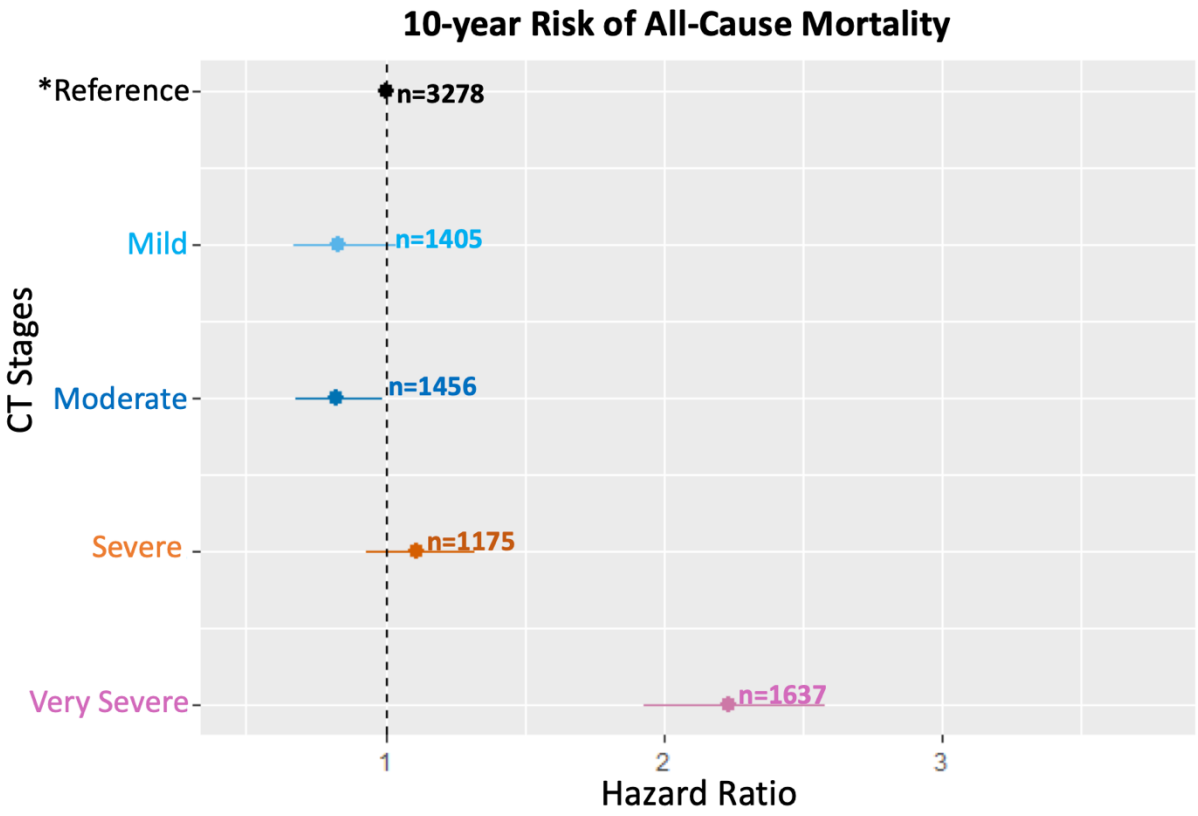

Figure 3.14. Patients who have very severe disease on CT also have significantly greater risk of mortality in 10 years.

For spirometry-based staging, as shown in Table 3.4, only GOLD stages 1-2 showed significantly greater odds (OR, 1.53 [95% CI: 1.20, 1.95] and OR, 2.01 [95% CI: 1.55, 2.61], respectively) of FEV_1 progression compared to GOLD 0. In contrast, hazard ratios for all-cause mortality showed a significant increasing trend with GOLD stage severity (GOLD 3 HR, 2.80 [95% CI: 2.38, 3.30] and GOLD 4 HR, 7.15 [95% CI: 6.03, 8.46]), with exception to GOLD 1 (HR 0.83 [95% CI: 0.64, 1.06]), which was not different from the GOLD 0 reference.

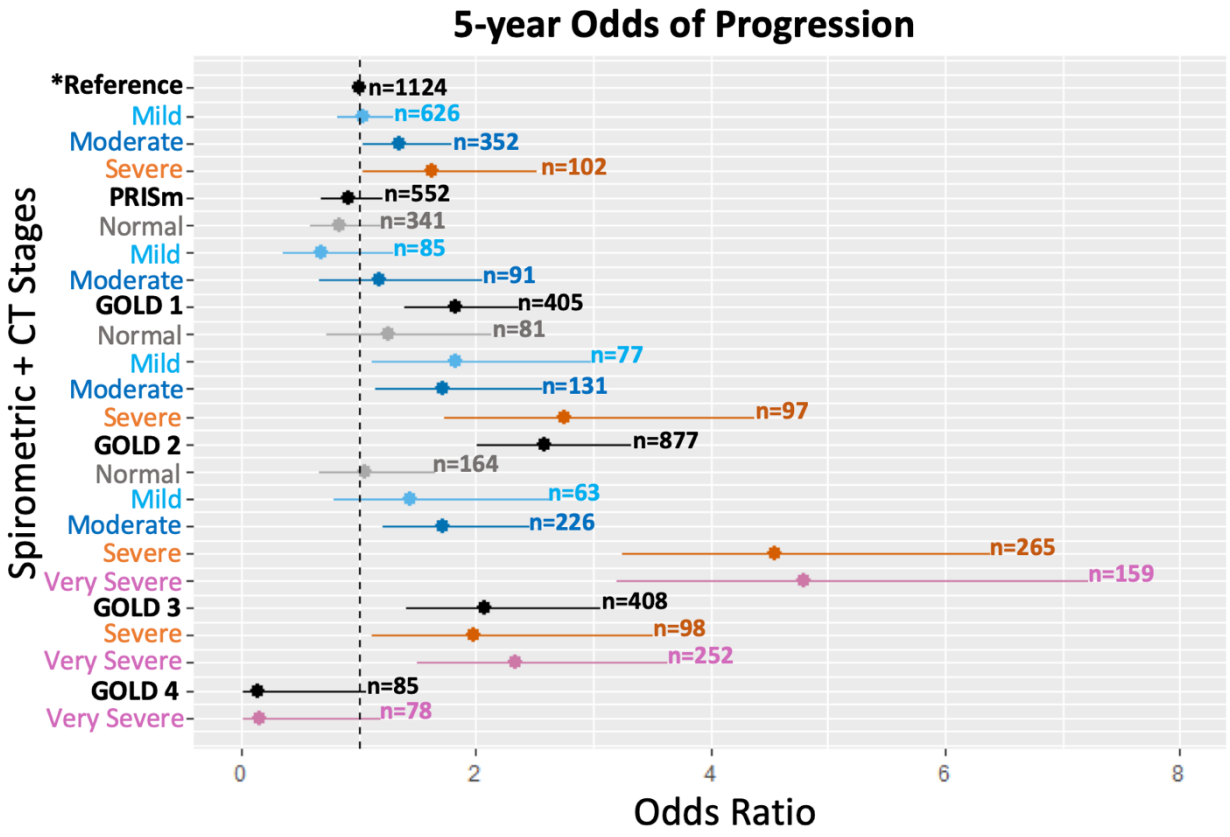

Figure 3.15. Analysis of FEV_1 progression using both imaging and spirometry as predictors showed significant interactions. Amongst patients with GOLD 2 disease, severe to very severe disease on CT showed higher likelihood of progression.

Analysis of FEV_1 progression using both imaging and spirometry as predictors, as shown in Figure 3.15, showed significant interactions. Amongst patients with GOLD 2 disease, severe to very severe disease on CT showed higher likelihood of progression (4.48 [95% CI: 3.18, 6.31] and 4.72 [95% CI: 3.13, 7.13]).

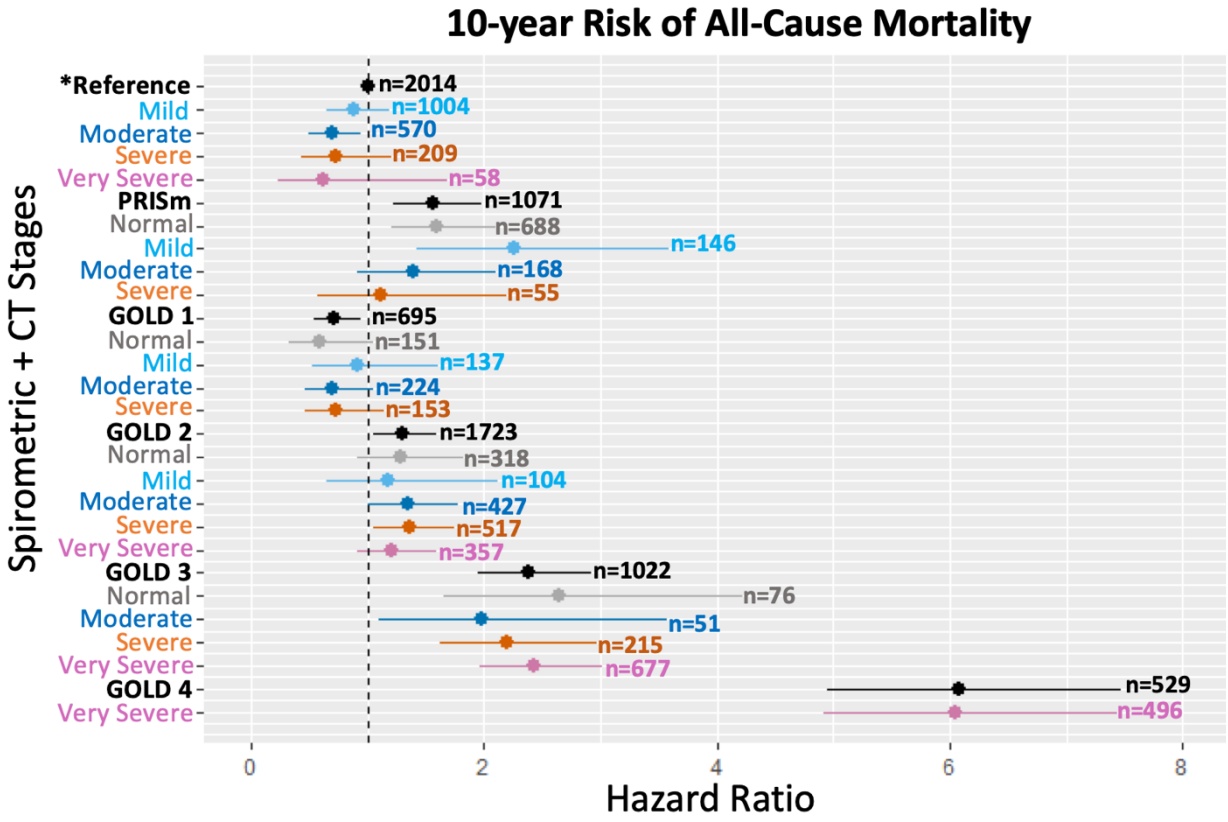

Figure 3.16. Mortality hazard ratios for CT-based stages within their respective GOLD stage categories were not significantly different from each other.

Mortality hazard ratios for CT-based stages within their respective GOLD stage categories were not significantly different from each other (Figure 3.16**)**. Using logistic and Cox regression analyses, we have clearly demonstrated the clinical utility of our proposed CT-based stages of COPD.

3.4 Discussion and limitations

We used a purely data-driven, machine learning approach to define CT-based stages of COPD that demonstrate significant prognostic potential for predicting disease progression and all-cause mortality.

Our results show that open-source deep learning and deformable image registration tools can be used to automate CT-based COPD stages, which can be made readily available in a clinical picture archiving and communication system for routine use.

Our study leverages several years of recent work that highlights the potential prognostic value of CT for COPD. Qualitative assessments and quantitative measurements have each been proposed by others for characterization of the lung parenchyma and airways, with inter- and intra-software reproducibility of these measurements having been thoroughly assessed.[10,11,14,15,17,40,41] The COPDGene investigators recently proposed a new set of criteria to assist in the diagnosis of COPD, combining smoking exposure, symptoms, spirometry, and imaging to provide categories of diagnostic certainty – a departure from historical definitions that did not incorporate CT imaging.[12] The new criteria were driven, in part, by the recognition that patients with COPD may not all exhibit abnormal pulmonary function tests.

In our study, quantification of air trapping is most similar to parametric response maps, which place thresholds on densities of deformed images. However, a fundamental difference is the removal of an upper threshold on inspiratory to define air trapping, considering voxels with attenuation difference of less than or equal to 100 HU and inspiratory attenuation of greater than -950 HU. Further investigation may be required to detail the advantages and disadvantages of each method.

Similar approaches using deformable registration have also been proposed, requiring aggregation of multiple commercial software tools. Boes et al and Pompe et al categorically defined the proportion of lung affected by emphysema and air trapping by placing predetermined attenuation thresholds on deformably registered inspiratory and expiratory images to create

parametric response maps.[18,19,20,21] Ostridge et al and Kirby et al computed disease probability maps, describing the probability each colocalized lung voxel contains emphysema or air trapping.[21,22]

Other investigators have also begun to explore the value of deep learning to automate analysis of CT. Gonzalez et al developed a CNN to predict GOLD stage using CT images and Humphries et al used deep learning to automate the Fleischner visual stages of emphysema.[25,41] These classification-based applications of deep learning illustrate the potential for deep learning to perform prognostication, but may be challenging to integrate into the clinical workflow or to quality control. Specifically, they are challenged by "explainability," as it can be difficult to distinguish between successful execution of the algorithm and 'edge' cases at the boundaries of the capabilities of the algorithm. Algorithms that leverage deep learning to automate segmentation tasks, like those implemented in the strategy we use here, may be more readily assessed visually to ensure quality for clinical use.

Our study has several limitations. Inspiratory and expiratory phase imaging requires adequate patient cooperation. Suboptimal expiratory phase images may result in apparent 'air trapping' leading to mischaracterization of patients. When these algorithms are deployed in the clinic, it would important to assess the adequacy of expiratory phase acquisitions, which may be assessed by evaluating concavity of the posterior, membranous portion of the trachea. Future studies may investigate the use of deep learning approaches to automatically evaluate the quality of expiratory series. Another limitation pertains to the newly recognized COPD patient subpopulation, known as the PRISm group. These PRISm patients have preserved (normal) FEV_1/FVC ratio but impaired FEV_1pp of less than 80%. Studies have found PRISm is associated with increased respiratory symptoms, inflammation, and mortality, and may contain patients with

overlap of fibrosis, larger body habitus, and COPD.[33] Future studies may be helpful to evaluate the benefit of quantitative CT in the PRISm population. Finally, our proposed CT-based staging system incorporates measurements of emphysema and air trapping, but does not incorporate measurements of airway caliber or wall thickening. It is possible that airway measurements may add additional prognostic value. Automated methods to standardize and accurately quantify airway abnormalities are reserved as a future direction.

3.5 Conclusions

In this study, we leveraged deep learning and deformable registration techniques to automate quantification of emphysema and air trapping. Using these imaging features, we defined a pragmatic CT-based system for staging COPD, which parallels spirometry-defined GOLD stages. Our CT-based stages are prognostically valuable, whether used independently or in concert with pulmonary function tests, predicting both future FEV1 decline and mortality. Most notably, we found patients with moderate spirometric impairment and moderate-severe disease on CT had the highest likelihood of progression. These results point to the synergistic value of CT and functional testing to identify patients at the highest risk of progression, and potentially those who might have the greatest benefit from medical or behavioral intervention to address ongoing tobacco or other particulate exposures.

We designed, developed, and clinically validated this deep learning pipeline that automates quantification of emphysema and air trapping for patients with COPD and other lung conditions. In conclusion, we show that quantitative CT-based stages of COPD severity can prognosticate progression and mortality when implemented independently or in conjunction with the spirometric GOLD criteria. Fully-automated deep learning algorithms can facilitate

translation of the proposed stages into clinical practice, potentially allowing us to better manage our patients with COPD.

3.6 Acknowledgments

Chapter 3, in full, is adapted from a paper that appeared in Radiology: Cardiothoracic Imaging, Volume 3, Issue 2, e200477, April 8, 2021. The dissertation author was one of the primary authors of this paper, which was coauthored by Kyle Hasenstab, Tara Retson, David Lynch, Douglas Conrad, Seth Kligerman, and Albert Hsiao. The dissertation author thanks Kyle Hasenstab, Tara Retson, David Lynch, Douglas Conrad, Seth Kligerman, and Albert Hsiao for providing meaningful feedback and discussion.

3.7 References

1. Nilsson DE (2013). Eye evolution and its functional basis. *Visual neuroscience*, *30*(1-2), 5–20. https://doi.org/10.1017/S0952523813000035

2. Lamb TD (2011, July 1). *Evolution of the Eye*. Scientific American. https://www.scientificamerican.com/article/evolution-of-the-eye/
3. Lamb, T., Collin, S. & Pugh, E. (2007). Evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup. *Nat Rev Neurosci* 8, 960–976. https://doi.org/10.1038/nrn2283

4. Lindsay G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

5. Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybernetics* 36, 193–202. https://doi.org/10.1007/BF00344251

6. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4), 541-551, doi: 10.1162/neco

7. Xiyuan, P, Jinxiang, Y, Bowen, Y, Liansheng, L, Yu, P. (2021). A Review of FPGA-Based Custom Computing Architecture for Convolutional Neural Network Inference. *Chinese Journal of Electronics*. https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/cje.2020.11.002

8. Ronneberger, O, Fischer, P, Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Arxiv*. https://arxiv.org/pdf/1505.04597.pdf

9. Bhatt SP, Soler X, Wang X, Murray S, Anzueto AR, Beaty TH, Boriek AM, Casaburi R, Criner GJ, Diaz AA, Dransfield MT, Curran-Everett D, Galban CJ, Hoffman EA, Hogg JC, Kazerooni EA, Kim V, Kinney GL, Lagstein A, et al. Association between functional small airway disease and FEV1 decline in chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2016;194(2):178–84. doi: 10.1164/rccm.201511-2219OC

10. Lynch DA, Moore CM, Wilson C, Nevrekar D, Jennermann T, Humphries SM, Austin JHM, Grenier PA, Kauczor H-U, Han MK, Regan EA, Make BJ, Bowler RP, Beaty TH, Curran-Everett D, Hokanson JE, Curtis JL, Silverman EK, Crapo JD. CT-based Visual

Classification of Emphysema: Association with Mortality in the COPDGene Study. Radiology. 2018;288(3):859–66. doi: 10.1148/radiol.2018172294

11. Schroeder JD, McKenzie AS, Zach JA, Wilson CG, Curran-Everett D, Stinson DS, Newell JD, Lynch DA. Relationships between airflow obstruction and quantitative CT measurements of emphysema, air trapping, and airways in subjects with and without chronic obstructive pulmonary disease. Am J Roentgenol. 2013;201(3):460–70. doi: 10.2214/AJR.12.10102

12. Lowe KE, Regan EA, Anzueto A, Austin E, Austin JHM, Beaty TH, Benos P V, Benway CJ, Bhatt SP, Bleecker ER, Bodduluri S, Bon J, Boriek AM, Boueiz AR, Bowler RP, Budoff M, Casaburi R, Castaldi PJ, Charbonnier J-P, et al. COPDGene® 2019: Redefining the Diagnosis of Chronic Obstructive Pulmonary Disease. Chronic Obstr Pulm Dis J COPD Found. 2019;6(5):384–99. doi: 10.15326/jcopdf.6.5.2019.0149

13. Vestbo J, Hurd S, Agustí A, Jones P, Vogelmeier C, Anzueto A, Barnes P, Fabbri L, Martinez F, Nishimura M, Stockley R, Sin D, Rodriguez-Roisin R. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. Am J Respir Crit Care Med. 2013;184(4):347–65. doi: 10.1164/rccm.201204-0596PP

14. Lynch DA, Austin JHM, Hogg JC, Grenier PA, Kauczor H-U, Bankier AA, Barr RG, Colby T V, Galvin JR, Gevenois PA, Coxson HO, Hoffman EA, Newell JD, Pistolesi M, Silverman EK, Crapo JD, Crapo JD. CT-Definable Subtypes of Chronic Obstructive Pulmonary Disease: A Statement of the Fleischner Society. Radiology. 2015;277(1):192–205. doi: 10.1148/radiol.2015141579

15. Ostridge K, Wilkinson TMA. Present and future utility of computed tomography scanning in the assessment and management of COPD. European Respiratory Journal. European Respiratory Society; 2016;48(1):216-28. doi: 10.1183/13993003.00041-2016

16. Matsuoka S, Kurihara Y, Yagihashi K, Hoshino M, Watanabe N, Nakajima Y, Matsuoka S. Quantitative Assessment of Air Trapping in Chronic Obstructive Pulmonary Disease Using Inspiratory and Expiratory Volumetric MDCT. AJR. 2008;190(3):762-9. doi: 10.2214/AJR.07.2820

17. Lynch DA, Al-Qaisi MA. Quantitative computed tomography in chronic obstructive pulmonary disease. Journal of Thoracic Imaging. 2013;28(5):284-90. doi: 10.1097/RTI.0b013e318298733c

18. Boes JL, Hoff BA, Bule M, Johnson TD, Rehemtulla A, Chamberlain R, Hoffman EA, Kazerooni EA, Martinez FJ, Han MK, Ross BD, Galbán CJ. Parametric Response Mapping Monitors Temporal Changes on Lung CT Scans in the Subpopulations and Intermediate Outcome Measures in COPD Study (SPIROMICS). Acad Radiol. 2015;22(2):186–94. doi: 10.1016/j.acra.2014.08.015

19. Pompe E, Galban CJ, Ross BD, Koenderman L, Hacken NHT, Postma DS, Berge MVD, Jong PAD, Lammers JWJ, Hoesein FAM. Parametric response mapping on chest computed tomography associates with clinical and functional parameters in chronic obstructive pulmonary disease. Respir Med. 2017;123:48–55. doi: 10.1016/j.rmed.2016.11.021

20. Kirby M, Yin Y, Tschirren J, Tan WC, Leipsic J, Hague CJ, Bourbeau J, Sin DD, Hogg JC, Coxson HO, Network CCRG and the CRR. A Novel Method of Estimating Small Airway Disease Using Inspiratory-to-Expiratory Computed Tomography. Respiration. 2017;94(4):336–45. doi: https://doi.org/10.1159/000478865

21. Ostridge K, Gove K, Paas KH, Burke H, Freeman A, Harden S, Kirby M, Peterson S, Sieren J, McCrae C, Vaarala O, Staples KJ, Wilkinson TM. Using Novel Computed Tomography Analysis to Describe the Contribution and Distribution of Emphysema and Small Airways Disease in Chronic Obstructive Pulmonary Disease. Ann Am Thorac Soc. 2019;16(8):990–7. doi: 10.1513/AnnalsATS.201810-669OC

22. Wang K, Mamidipalli A, Retson T, Bahrami N, Hasenstab K, Blansit K, Bass E, Delgado T, Cunha G, Middleton MS, Loomba R, Neuschwander-Tetri BA, Sirlin CB, Hsiao A, Network on behalf of the members of the NCR. Automated CT and MRI Liver Segmentation and Biometry Using a Generalized Convolutional Neural Network. Radiol Artif Intell. 2019;1(2):180022. doi: 10.1148/ryai.2019180022

23. Blansit K, Retson T, Masutani E, Bahrami N, Hsiao A. Deep Learning–based Prescription of Cardiac MRI Planes. Radiol Artif Intell. 2019;1(6):e180069. doi: 10.1148/ryai.2019180069

24. Bahrami N, Retson T, Blansit K, Wang K, Hsiao A. Automated selection of myocardial inversion time with a convolutional neural network: Spatial temporal ensemble myocardium inversion network (STEMI-NET). Magn Reson Med. 2019;81(5):3283–91. doi: 10.1002/mrm.27680

25. Humphries SM, Notary AM, Centeno JP, Strand MJ, Crapo JD, Silverman EK, Lynch DA, Genetic Epidemiology of COPD (COPDGene) Investigators. Deep Learning Enables Automatic Classification of Emphysema Pattern at CT. Radiology. 2020;294(2):434-444. doi: 10.1148/radiol.2019191022

26. Gerard SE, Patton TJ, Christensen GE, Bayouth JE, Reinhardt JM. FissureNet: A Deep Learning Approach For Pulmonary Fissure Detection in CT Images. IEEE Trans Med Imaging. 2019;38(1):156–66. doi: 10.1109/TMI.2018.2858202

27. Park B, Park H, Lee SM, Seo JB, Kim N. Lung Segmentation on HRCT and Volumetric CT for Diffuse Interstitial Lung Disease Using Deep Convolutional Neural Networks. J Digit Imaging. 2019;32(6):1019–26. doi: 10.1007/s10278-019-00254-8

28. Gerard SE, Herrmann J, Kaczka DW, Musch G, Fernandez-Bustamante A, Reinhardt JM. Multi-resolution convolutional neural networks for fully automated segmentation of acutely injured lungs in multiple species. Med Image Anal. 2020;60(101592). doi: https://doi.org/10.1016/j.media.2019.101592

29. Park J, Yun J, Kim N, Park B, Cho Y, Park HJ, Song M, Lee M, Seo JB. Fully Automated Lung Lobe Segmentation in Volumetric Chest CT with 3D U-Net: Validation with Intra- and Extra-Datasets. J Digit Imaging. 2020;33(1):221–30. doi: 10.1007/s10278-019-00223-1

30. Team NLSTR, Aberle D, Berg C, Black W, Church T, Fagerstrom R, Galen B, Gareen I, Gatsonis C, Goldin J, Gohagan J, Hillman B, Jaffe C, Kramer B, Lynch D, Marcus P, Schnall M, Sullivan D, Sullivan D, et al. The National Lung Screening Trial: overview and study design. Radiology. 2011;258(1):243–53. doi: 10.1148/radiol.10091808

31. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD. Genetic epidemiology of COPD (COPDGene) study design. COPD. 2010;7(1):32–43. doi: 10.3109/15412550903499522
32. Hasenstab K, Yuan N, Retson T, Conrad D, Kligerman S, Lynch D, Hsiao A, *et al*. Automated CT Staging of COPD Severity for Predicting Disease Progression and Mortality with a Deep Learning Convolutional Neural Network. *Radiology: Cardiothoracic Imaging* 2021;3.

33. Shackleford, J., Kandasamy, N., Sharp, G. (2013). *High Performance Deformable Image Registration Algorithms for Manycore Processors*. https://doi.org/10.1016/B978-0-12-407741-6.00001-3.

34. Garyfallidis E, Brett M, Amirbekian B, Rokem A, van der Walt S, Descoteaux M, Nimmo-Smith I, Contributors D. DIPY, a library for the analysis of diffusion MRI data. 2014.

35. Avants B, Epstein C, Grossman M, Gee J. Symmetric Diffeomorphic Image Registration with CrossCorrelation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain. Med Image Anal. 2008;12(1):26–41. doi: 10.1016/j.media.2007.06.004

36. Heussel CP, Herth FJF, Kappes J, Hantusch R, Hartlieb S, Eberhardt R, Weinheimer O, Kauczor HU. Fully automatic quantitative assessment of emphysema in computed tomography: Comparison with pulmonary function testing and normal values. Eur Radiol. 2009;19(10):2391–402. doi: 10.1007/s00330-009-1437-z

37. Madani A, Zanen J, de Maertelaer V, Gevenois PA. Pulmonary emphysema: objective quantification at multi-detector row CT--comparison with macroscopic and microscopic morphometry. Radiology. 2006;238(3):1036–43. doi: 10.1148/radiol.2382042196

38. Kotikalapudi R, Contributors. keras-vis. 2017. https://github.com/raghakot/keras-vis.

39. Kim S, Seo J, Lee H, Nevrekar D, Forssen A, Crapo J, Schroeder J, Lynch D. Chronic obstructive pulmonary disease: lobe-based visual assessment of volumetric CT by Using standard images--comparison with quantitative CT and pulmonary function test in the COPDGene study. Radiology. 2013;266(2):626–35. doi: 10.1148/radiol.12120385

40. Kirby M, Hatt C, Obuchowski N, Humphries SM, Sieren J, Lynch DA, Fain SB, Committee QLD. Inter- And Intra-Software Reproducibility of Computed Tomography Lung Density Measurements. Med Phys. 2020; Online ahead of print. doi: 10.1002/mp.14130.

41. González G, Ash SY, Vegas Sanchez-Ferrero G, Onieva Onieva J, Rahaghi FN, Ross JC, Díaz A, San José Estépar R, Washko GR, COPDGene and ECLIPSE investigators. Disease Staging and Prognosis in Smokers Using Deep Learning in Chest Computed Tomography. Am J Respir Crit Care Med. 2018;197(2):193-203. doi: 10.1164/rccm.201705-0860OC

CHAPTER 4: DEVELOP AND EVALUATE A MACHINE LEARNING APPROACH TO FACILITATE DIAGNOSIS FOR PATIENTS WITH COVID-19

Large scale cohort analysis has shown that the mortality of hospitalized patients with COVID-19 was associated with health-care burden, reflected by the simultaneous number of hospitalized patients with severe illness.[1] Chest X-rays are often part of the initial diagnostic workup of pneumonia and are used to monitor progression or resolution.[2]

The ongoing COVID-19 pandemic has increased the sheer volume of chest X-rays performed annually, highlighting a critical need to develop automated methods to assist in detection and diagnostic interpretation.[3,4,5] Improved automated methods to triage, diagnose, and predict prognosis for hospitalized COVID-19 patients, especially those with chronic co-morbidities such as COPD, has the potential to assist and guide trained personnel to provide timely and proactive personalized care.

Here I present the third chapter of this thesis, in which we developed a novel automated method for COVID-19 pneumonia detection, training the algorithm on a subset of the 2018 RSNA pneumonia challenge chest radiographs and validating it on a subset of the COVID-19 radiography dataset, COV-RAD.

4.1 Introduction

Pneumonia is the eighth leading cause of death and first among infectious causes of death, with a prevalence of 10-65% among hospitalized patients (In the United States alone, pneumonia is responsible for over 250,000 emergency room visits and nearly 50,000 deaths per year.[7] Chest X-rays are often part of the initial diagnostic workup of pneumonia and are used to monitor progression or resolution.[7] Given the ongoing COVID-19 pandemic, the sheer volume of

chest X-rays performed annually highlights a critical need to develop automated methods to assist in detection and diagnostic interpretation.[10,11,12]

Medical literature has seen 37 machine learning algorithms developed in 2020 to detect COVID-19 pneumonia in either chest X rays or CT images (Figure 4.1).[3,8,9,10]

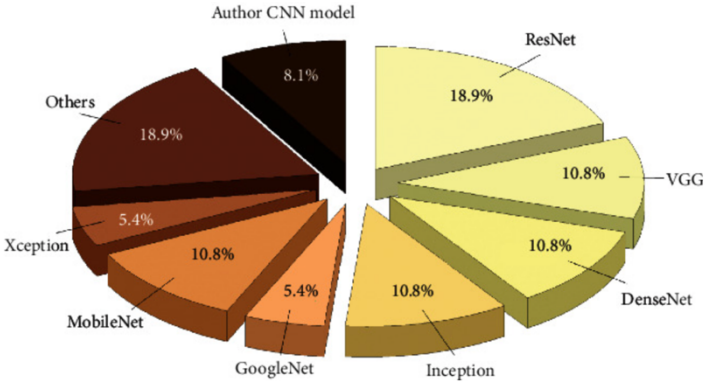

Figure 4.1[3]. Percentage breakdown of CNN architectures that have been developed to automate the analysis of COVID-19 chest radiographs.

According to the findings, deep learning-based models have great capacity to offer an accurate and efficient system for the detection and diagnosis of COVID-19 in the field of COVID-19 radiologic image processing[8]. Specifically, convolutional neural networks (CNNs) show significant capacity in implicitly learning structural features of an image or volume, making them highly effective in performing a variety of tasks, including image classification, object detection, and segmentation[7].

In spite of their advantages in computer vision tasks, CNNs often have complex network structures which may include hundreds of convolutional layers and pooling layers[5]. It takes a significant amount of computing and memory resources to train these deep CNN models. While this may not pose a challenge in resource-rich environments such as well-funded labs in academia, many medically underserved areas in the United States are not equipped with the

hardware to supply the computing power needed to train and develop these models in real time[6].

Several groups have successfully developed shallow CNNs for various image detection and classification tasks[5]. Inspired by these efforts to overcome the computational limitations imposed by training complex networks, we developed a shallow 3-layer CNN classifier, SimplePNUnet, to detect the presence of pneumonia using a subset of 2000 frontal chest X-rays from the publicly accessible dataset released as part of the 2018 RSNA Pneumonia Detection Challenge.

4.2 Design and implementation of shallow CNN for Real-Time COVID-19 Pneumonia Detection

4.2.1 Study cohort

We used a publicly available dataset of frontal chest X-rays with bounding boxes representing pneumonia annotated by radiologists, released as part of the 2018 RSNA pneumonia challenge. The RSNA pneumonia challenge included 25,684 chest radiographs spanning the range of pathologies from this data set, distributed as DICOM images. The chest X-rays in this dataset are a subset of a larger NIH database comprising 112,120 frontal chest radiographs[7], each of which was labeled with findings and diagnoses extracted from radiology text reports using natural language processing. These pathologies included atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia[7].

Additionally, we used the COVID-19 Radiography Dataset (henceforth referred to as COV-RAD) that was compiled by a team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh along with their collaborators from Pakistan and Malaysia.

Collaborating with physicians at these sites, they have created a database of chest X-ray images for COVID-19 positive cases along with Normal and Viral Pneumonia images. Over the past two years, they have increased the database to 3616 COVID-19 positive cases along with 10,192 Normal, 6012 Lung Opacity (Non-COVID lung infection), and 1345 Viral Pneumonia images and corresponding lung masks[8,9].

4.2.2 Image processing

We randomly sampled 2000 chest X-rays from the RSNA pneumonia detection challenge. Each image has a spatial resolution of 1024x1024 pixels and an 8-bit pixel depth. We randomly divided the chest radiographs into 1000 for training, 500 for testing, and 500 for validation. Each set of radiographs was normalized separately using two different normalization techniques: Technique 1 normalizes each pixel between 0 and 1 by dividing each pixel by the maximum value 255. Technique 2 uses the mean and standard deviation computed pixel-wise from the training set to normalize images from the validation and test sets. Additionally, we tested five interpolation techniques (i.e. nearest neighbor, bilinear, bicubic, biquartic, and biquintic) to down-sample the resolution of each image to 256x256 to facilitate computation.

4.2.3 Model development

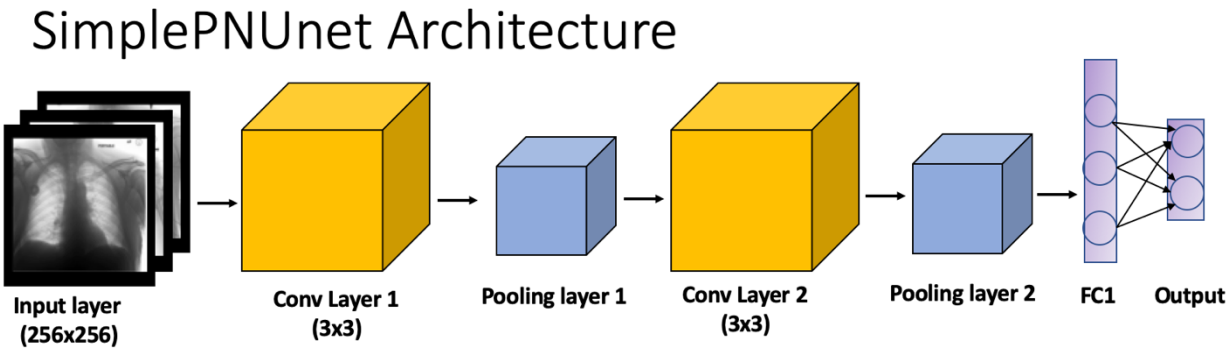

Figure 4.2. SimplePNUnet architecture includes a 3-layer CNN.

We built a shallow 3-layer convolutional neural network (CNN), SimplePNUnet, composed of two convolutional layers and one dense layer. The architecture of the model is shown in Figure 4.2. Using 50/50 class balance during training, we monitored the validation performance to stop training early to avoid overtraining the model. The optimal weights for each model were saved based on the epoch that achieved the highest validation performance. Additionally, we tested the stability of model training across different batch sizes.

We applied Bayesian optimization using 5-fold cross validation to fine-tune the following set of hyperparameters: learning rate, L1 and L2 regularization, kernel size, number of filters, pool size, number of strides, number of dense nodes, dropout rate, and type of activation function. Layer-wise lambda optimization was performed by using higher L1 and L2 values for layers closer to the input layer.

We implemented and trained SimplePNUnet in Python (version 3.6.6; Python Software Foundation, Wilmington, DE), using Keras 2.4 and Tensorflow 2.2.0 on a GPU workstation running Windows 10, equipped with a Titan Xp graphics card (NVIDIA, Mountain View, CA).

4.2.4 Model evaluation

We externally evaluated our model using 1500 COVID-19 positive and 1500 normal chest radiographs from COV-RAD. The 1500 COVID-19 positive radiographs were curated from various sources, as shown in Table 4.1.

Table 4.1. **Curation of COVID-19 positive chest radiographs from COV-RAD dataset**.

Source of Image	Sample Size
BIMCV	391
Eurorad	258
COVID-CXNet	400
COVID-ChestXRay	154
Institute for Diagnostic and Interventional Radiology (Hannover Medical School)	183
SIRM	114
Total	1500

Using the same model architecture and hyperparameters, we retrained SimplePNUnet across five iterations, each time increasing the proportion of COV-RAD data in the training set up to 50% out of a total of 1000 samples. The training methodology mirrors that used in the development of SimplePNUnet. The COV-RAD samples used in retraining SimplePNUnet were excluded from the external validation set of 3000 COV-RAD samples.

4.3 Results

4.3.1 Patient demographics

Among the patients from the RSNA Pneumonia Challenge dataset, 43% were females aged 1 to 93 years. Thirty-six percent of radiographs were assigned as positive for pneumonia, while 64% were either normal or diagnosed as abnormal but did not have pneumonia. No additional exclusion criteria were applied for the current study[7]. Demographic information was not available for the COV-RAD cohort.

4.3.2 Image preprocessing and downsampling

Table 4.2. Impact of different normalization and downsampling techniques on model performance. The combination of techniques that yields the highest AUC is highlighted.

Normalization and Downsampling techniques

Normalization \ Downsampling	Nearest neighbor	Bilinear	Bicubic	Biquartic	Biquintic
pixel/255	AUC = 81.64%	AUC = 81.78%	AUC = 82.99%	AUC = 82.17%	AUC = 82.34%
mean, standard deviation	AUC = 79.91%	AUC = 81.14%	AUC = 84.37%	AUC = 80.41%	AUC = 84.48%

We found that normalizing each set separately using the mean and standard deviation computed from the training set (Technique 2) and downsampling using biquintic interpolation yielded the highest performance (Table 4.2).

4.3.3 Model evaluation

We define the best model as the one with the highest AUC with minimal overfit. As shown in Table 4.3 the performance of the best model across the training, validation, and test sets was comparable (AUC, 0.870 vs 0.812 vs 0.834, respectively) using a batch size of 96 and layer-wise lambda optimization.

Table 4.3. Impact of different minibatch size and lambda optimization techniques on model performance. The model with the highest AUC with minimal overfit on training set is highlighted.

Performance AUC	Minibatch size: N = 32	N = 64	N = 96	N = 128
Same lambda per layer (search: 1e-4 to 1e-1)				
Train AUC (n = 1000)	0.8829	0.8882	0.9113	0.9216
Val AUC (n = 500)	0.8097	0.8152	0.8136	0.8146
Test AUC (n = 500)	0.8432	0.8362	0.8437	0.8519
Layer-wise lambda optimization (higher for lower layers)				
Train AUC (n = 1000)	0.8884	0.9055	0.8704	0.9144
Val AUC (n = 500)	0.7970	0.8162	0.8119	0.8078
Test AUC (n = 500)	0.8264	0.8434	0.8335	0.8348

We evaluated our best model using 3000 samples from COV- RAD as our external validation set. Results reveal that the original version of SimplePNUnet (i.e. not retrained with COV-RAD samples incorporated into training set) had the highest AUC out of all the iterations we tested (Table 4.4, Figure 4.3). The model with the highest sensitivity was retrained using a training set comprising 30% of COV-RAD data.

Table 4.4. Performance of iterations of the SimplePNUnet model retrained using a training set comprising varying proportions of the COV-RAD data.

Proportion of COV-RAD Data Incorporated into Training Set	Performance Metric				
	AUC	Sensitivity	Specificity	PPV	NPV
None	0.712	0.725	0.561	0.623	0.671
10%	0.678	0.604	0.669	0.646	0.628
20%	0.708	0.707	0.583	0.629	0.666
30%	0.705	0.755	0.508	0.606	0.675
40%	0.693	0.729	0.528	0.607	0.661
50%	0.710	0.739	0.537	0.615	0.673

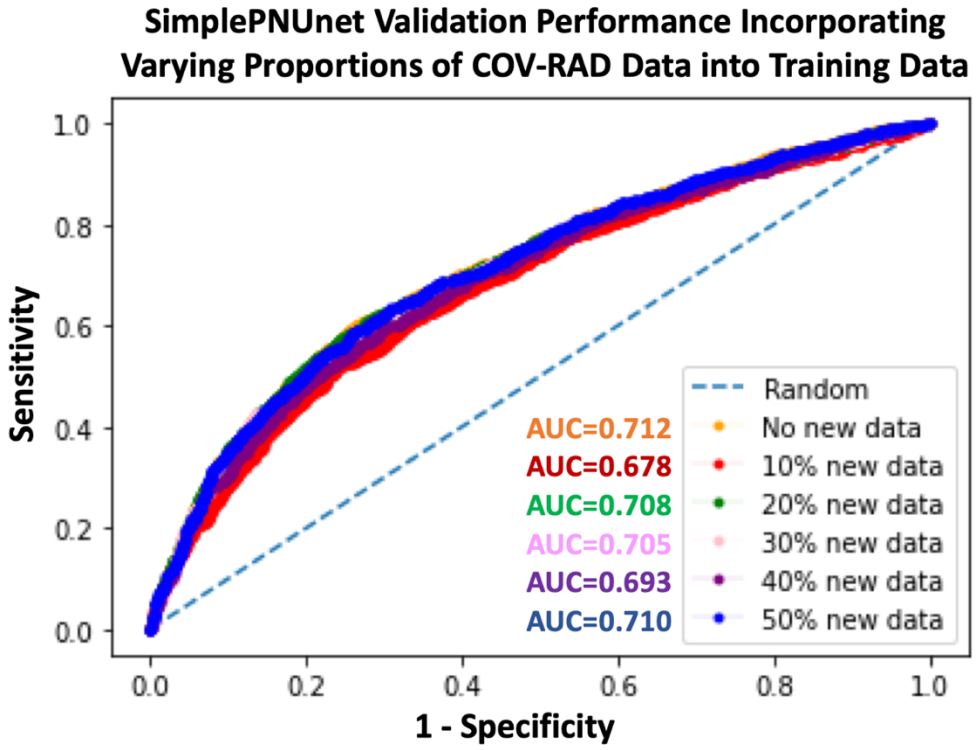

Figure 4.3. AUROC curves comparing the performance of SimplePNUnet retrained with varying proportions of COV-RAD data incorporated into the training set and validated on a separate subset of COV-RAD data.

4.4 Discussion and future work

This study aimed to investigate the feasibility of using Bayesian hyperparameter optimization to tune a shallow CNN, SimplePNUnet, to achieve performance comparable to that of much deeper, more computationally expensive CNNs. Our experiments show that our model consistently performed above an AUC of 0.8. The results are promising: SimplePNUnet provides a light-weight architecture for high performance pneumonia detection. Additionally, SimplePNUnet performed well in detecting COVID-19 pneumonia when evaluated on a subset of the external dataset, COV-RAD.

One limitation of this study is that it remains unclear that all the COVID-19 positive radiographs from COV-RAD were screened and confirmed by radiologists. Chowdhury *et al*. stated that while a subset of the COVID-19 positive radiographs were collected from the Italian Society of Medical and Interventional Radiology (SIRM) COVID-19 Database, Hannover Medical School, and the Valencia Region Image Bank (BIMCV), other images were gathered from articles and online resources. Future work may investigate using other image datasets that are confirmed and validated by radiologists.

Future work may explore the application of transfer learning to adapt SimplePNUnet into to detect COVID-19 pneumonia with greater performance. The objective of transfer learning is to take learned features from SimplePNUnet (i.e. source model) that may be useful when training the target model to detect COVID-19 pneumonia. Specifically, we may try keeping the weights of the initial layer of SimplePNUnet fixed, and retrain the deeper layers using the same approach described in the Methods section.

We hypothesize that transferring learned features from SimplePNUnet will expedite training time, save computational resources, and improve model performance. We assume that

SimplePNUnet will learn transferable features from chest X-rays positive for pneumonia. However, one potential pitfall is negative transfer, a phenomenon that may occur when the learned features from SimplePNUnet do not improve model performance in detecting COVID-19 pneumonia. Further investigation into image quality, with potential integration of additional datasets and model optimization may be necessary to overcome this challenge.

4.5 Conclusions

Recent developments in AI have shown that deep learning algorithms such as convolutional neural networks (CNN) are effective in automating various tasks in medical imaging such as pneumonia detection. The challenge remains in implementing these algorithms efficiently in the clinic where there may be insufficient hardware to support CNN with hundreds of hidden layers and millions of parameters. To address this challenge, we successfully developed a novel, automated shallow 3-layer CNN classifier, SimplePNUnet, to detect the presence of pneumonia using a subset of 2000 frontal chest X-rays from the publicly accessible dataset released as part of the 2018 RSNA Pneumonia Detection Challenge. Additionally, we assessed the robustness of SimplePNUnet across different downsampling techniques and batch sizes. Performance of the algorithm was evaluated using the area under receiver operating characteristic curve (AUROC). The performance of the best model across the training (n=1000), validation (n=500), and test (n=500) sets was comparable (AUC, 0.870 vs 0.812 vs 0.834, respectively). SimplePNUnet achieved good performance on an external validation set comprising COVID-19 chest radiographs (AUC, 0.712).

4.6 Acknowledgments

Chapter 4 presents unpublished material, and is coauthored with Theresa Gaasterland. The dissertation author was the primary author of this chapter. The dissertation author thanks Theresa Gaasterland, Albert Hsiao, and Shamim Nemati for providing meaningful feedback and discussion on the computational experiments performed in this chapter.

4.7 References

1. Kiranyaz S, Avci O, and Abdeljaber O, Ince T, Gabbouj M, and Inman D. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*.
2. Hasenstab K, Yuan N, Retson T, Conrad D, Kligerman S, Lynch D, and Hsiao A. (2021). Automated CT Staging of COPD Severity for Predicting Disease Progression and Mortality with a Deep Learning Convolutional Neural Network. *Radiology: Cardiothoracic Imaging*.
3. Ghaderzadeh, M., & Asadi, F. (2021). Deep Learning in the Detection and Diagnosis of COVID-19 Using Radiology Modalities: A Systematic Review. *Journal of healthcare engineering*, *2021*, 6677314. https://doi.org/10.1155/2021/6677314
4. Regunath SH and Oba Y. (2021). Community-Acquired Pneumonia. *Treasure Island (FL): StatPearls Publishing*.
5. Lei F, Liu X, Dai Q, and Wing-Kuen, BL. (2019). *Shallow convolutional neural network for image classification*. *SN Appl. Sci.*
6. Yu S, Wu S, and Wang L. (2017). A shallow convolutional neural network for blind image sharpness assessment. *Plos Journal*.
7. Madhavan S, Borker RD, Fernandes AW, Amonkar MM, and Rosunbluth SA. (2004). Assessing Predictors of Influenza and Pneumonia Vaccination in Rural Senior Adults. *Journal of Health & Social Policy*.
8. Hurt B, Yen A, Kligerman S, and Hsiao A. (2020). Augmenting Interpretation of Chest Radiographs With Deep Learning Probability Maps. *J Thorac Imaging*.
9. Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, Islam KR, Khan MS, Iqbal A, Al-Emadi N, Reaz MBI, and Islam MT. (2020). "Can AI help in screening Viral and COVID-19 pneumonia?"*IEEE Access*.
10. Rahman T, Khandakar A, Qiblawey Y, Tahir A, Kiranyaz S, Kashem SBA, Islam MT, Maadeed, SA, Zughaier SM, Khan MS and Chowdhury ME. (2020). Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-ray Images.
11. Kun-Hsing Y, Beam AL, and Kohane IS. (2018). Artificial intelligence in Healthcare. *Nature Biomedical Engineering*.
12. Peláez E, Serrano R, Murillo G, and Cárdenas W. (2021). A Comparison of Deep Learning Models for Detecting COVID-19 in Chest X-ray Images. *IFAC-PapersOnLine*.
13. Kriza C, Amenta V, Zenié A, Panidis D, Chassaigne H, Urbán P, Holzwarth U, Sauer AV, Reina V, and Griesinger CB. (2021). Artificial intelligence for imaging-based COVID-19 detection: Systematic review comparing added value of AI versus human reader. *European Journal of Radiology*.

CHAPTER 5: DEVELOP AND EVALUATE AN INTERPRETABLE MULTIMODAL MACHINE LEARNING FRAMEWORK FOR PREDICTING SEVERE EXACERBATIONS IN COPD

To advance the goals of the Tobacco-Related Disease Research Program (TRDRP) in Chapter 5, we developed a novel, interpretable multimodal machine learning approach that combines pulmonary function, quantitative CT imaging, and socioeconomic features to predict one of the clinically relevant patient outcomes, severe exacerbations.

COPD exacerbations are characterized by sudden worsening of symptoms that often result in increased future exacerbation and mortality risk.[1,2,3] Although there is no consensus on how to assess exacerbation severity, classification systems have primarily relied on the level and type of care needed to treat symptoms of exacerbation.[4] Severe exacerbations often require hospitalization or admission into the ER, are responsible for more than 90% of the total COPD-related healthcare costs, and cause significant decline in pulmonary function and worsening of health status.[4,5,6,7] Nearly 50% of hospitalizations due to severe exacerbations are considered preventable with proper and timely outpatient care.[3,8,9] Identifying clinical and sociodemographic characteristics that are strongly predictive of severe exacerbations may help screen for patients at high risk for experiencing an exacerbation that may significantly increase their burden of healthcare costs.[8,9]

5.1 Introduction

Machine learning algorithms have been developed to predict severe exacerbation in COPD cohorts and to identify factors that may put a patient at high risk for experiencing an exacerbation.[6,10,11,12] The prevailing standard for assessing the risk of future exacerbations uses previous history of exacerbations with a modest discriminatory performance.[6,10,11,12] Advancements in machine learning approaches (e.g., gradient boosting) using clinical and

administrative data in EHR have outperformed previous approaches using linear models with higher discriminatory power.[11,12] Although these models achieved good performance [AUC 0.80-0.85], they used a limited selection of data types as their input features and had limited interpretability.[6,10,11,12] While studies have identified clinical factors associated with severe exacerbations, these too, were limited by input feature selection, and the models showed moderate performance [AUC 0.65-0.71].[10]

Multimodal machine learning models that integrate multiple data sources have shown to be effective in predicting Alzheimer's disease and presence of pulmonary embolisms.[13] Specifically, social determinants of health (SDH) such as lower educational attainment and household income are known risk factors for various adverse COPD health outcomes, including exacerbations, that have been underutilized in COPD outcome risk predictions.[13] Integrating clinical, imaging, and sociodemographic data into a multimodal offers valuable information and has been shown to improve model predictive power.[13]

The development of model-agnostic post-hoc explainers has significantly improved model interpretability.[15] Techniques such as SHapley Additive exPlanations (SHAP) compute the additive marginal contribution of each feature toward the model's prediction.[15] SHAP is called model-agnostic because it can explain how a model arrives at its predictions for any model architecture used. Building such interpretability modules into a machine learning model makes the preconceived "black box" nature model transparent to the end-user, which would pave the way to facilitating integration of decision support models into the clinical workflow.

Integrating a broad selection of patient-level clinical, imaging, and SDH data from the feature-rich COPDGene longitudinal cohort, we built a novel interpretable, multimodal machine learning framework for predicting severe exacerbations in patients with COPD.[4] We further

identified the top contributing factors associated with severe exacerbations, which may be used to screen for patients at high risk for developing exacerbation-related complications.

5.2 Design and implementation of the interpretable, multimodal machine learning framework

5.2.1 Interpretable, multimodal machine learning framework for predicting COPD severe exacerbations

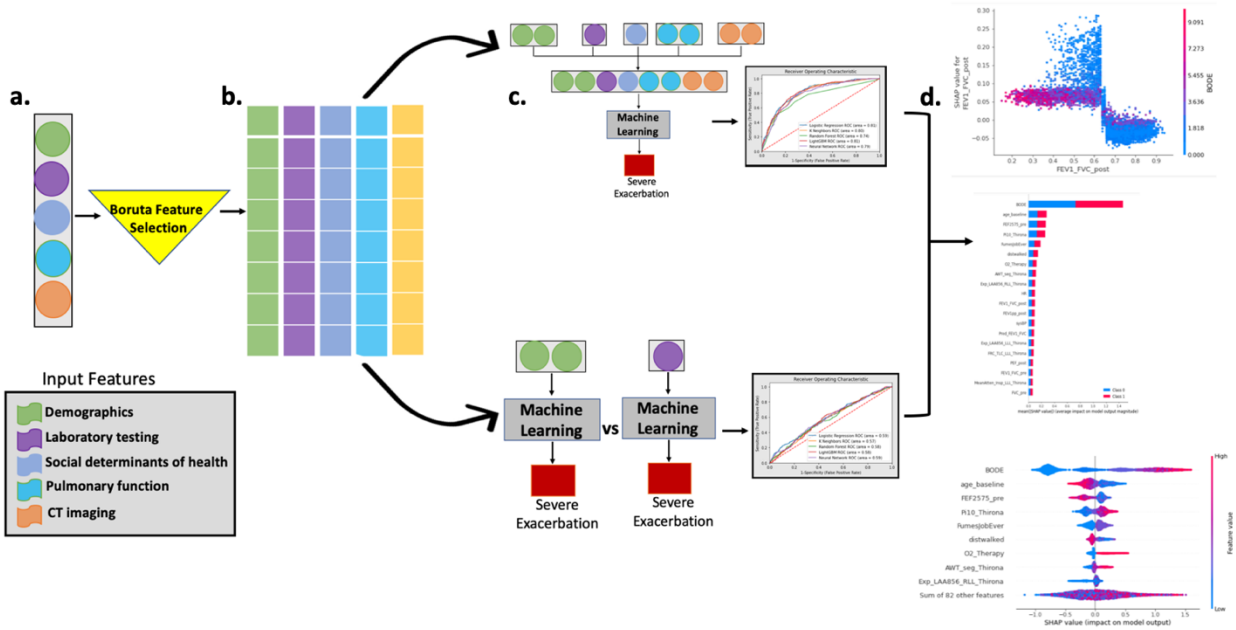

Figure 5.1: Interpretable, multimodal machine learning framework. (a) Data curation (b) Boruta feature selection (c) Model development and hyperparameter tuning (d) Model evaluation and interpretability

5.2.2 Boruta Feature Selection

To curate the data, a total of 118 demographic, socioeconomic, laboratory testing, pulmonary function testing, and CT imaging features available for 7781 patients were curated from the Phase 1 COPDGene data (Figure 5.1a). Next, for each data type, we applied the Boruta feature selection method using a random forest classifier to identify only the features that may be relevant for predicting severe exacerbation (Figure 5.1b).

The Boruta algorithm determines feature relevance by comparing the relevance of real features to that of randomly shuffled "shadow" features.[16] Both the real and "shadow" features are then used as input to a random forest classifier, which is a constructed set of decision trees, each of which contains a set of internal nodes and leaves.[16,17] Inside the internal nodes, the selected feature is used to decide how to divide the data into two separate sets by a criterion called Gini impurity or information gain for our classification task.[16,17] Specifically, Gini impurity measures how on average each feature decreases the impurity of the split, selecting for the feature with the greatest decrease inside each node.[16,17] Feature importances are computed as the mean and standard deviation of the accumulation of impurity decrease within each tree.[18]

In the Boruta algorithm, the importance of each original feature is then compared against a threshold, which is defined as the highest feature importance computed among the "shadow" features. After performing 100 iterations of the algorithm, we had a selected set of features for each type of data. In each iteration, a feature becomes a "hit" if its importance surpasses the threshold. For each data type, we included all the features that yielded at least 1 "hit" across all the iterations. These selected features were included in the feature set for training our models.

5.2.3 Model Development and Hyperparameter Tuning

After performing Boruta feature selection, we trained a total of five unimodal models using data from only a single data type: demographic data (i.e. demo-only), laboratory testing data (i.e. lab-only), socioeconomic data (i.e. SDH-only), pulmonary function data (i.e. PFT-only), and imaging data (i.e. CT-only). We further trained four multimodal models using different combinations of features: demographic and laboratory testing data (i.e. demo-lab); demographic, laboratory testing, and socioeconomic data (i.e. demo-lab-SDH); demographic, laboratory testing, socioeconomic, and pulmonary function data (i.e. demo-lab-SDH-PFT);

demographic, laboratory testing, socioeconomic, pulmonary function, and imaging data (i.e. all-features).

For each model, the data was split 70% for training and 30% for testing, using a total of four different model architectures to predict severe exacerbation: logistic regression, K-nearest neighbors, random forest classifier, and lightGBM. The hyperparameters of each model were optimized using Bayesian optimization (Figure 5.1c).

5.2.4 Model Evaluation and Interpretability

Each model was evaluated using the area under the receiver operating characteristic curve (AUROC), among other performance metrics. Model interpretability was achieved using SHAP global and local interpretation plots that summarize the average marginal contribution of each feature to severe exacerbation (Figure 5.1d).

5.3 Results

5.3.1 Patient characteristics

Our patient cohort composed of 53.2% men, 28.9% black, aged 39-85 years with an average of 44.1 pack year smoking history (Table 5.1). Also shown in Table 5.1 are pulmonary function, imaging, lab results, and social determinants of health data, all of which are relevant to developing our models.

Table 5.1. Demographic, spirometric, imaging, clinical, and socioeconomic characteristics of 7781 patients curated from Phase 1 COPDGene data

Characteristic	Value	Characteristic	Value
Demographics		Laboratory Testing	
Men	53.2% (4141/7781)	Body mass index	28.8 ± 6.1
Race (% Black)	28.9% (2247/7781)	Heart rate	75 ± 13
Age	60 ± 9	O2 therapy	0.11 ± 31
Pack years	44.1 ± 25.2	O2 supple for 6MW	0.21 ± 0.89
Current smokers	50.5% (3933/7781)	Resting SaO2	96.1 ± 2.76
Spirometry		Weight (kg)	83.6 ± 19.3
FEV1pp	76.8 ± 25.4	Systolic BP	129.1 ± 16.7
FVCpp	87.7 ± 17.8	Diastolic BP	76.9 ± 10.8
FEV1/FVC	0.66 ± 0.16	Social Determinants of Health	
GOLD 0	42.2% (3284/7781)	Dusty job ever	47.0% (3655/7781)
GOLD 1	8.0% (621/7781)	Fumes job ever	47.4% (3687/7781)
GOLD 2	20.2% (1574/7781)	Education level (college)	24.8% (1927/7781)
GOLD 3	11.8% (922/7781)	Employment status	33.0% (2565/7781)
GOLD 4	5.5% (426/7781)		
PRISm	11.1% (864/7781)		
Imaging			
%Emphysema	6.4 ± 9.8		
%Air trapping	21.2 ± 19.6		
Pi10	2.3 ± 0.6		
Hyperinflation	0.58 ± 0.11		

5.3.2 Unimodal model performance

Unimodal model performance on validation set using only demographics data was relatively poor, while model performance using only lab testing data was relatively strong (Figure 5.2a,b). Model performance on validation set using only social determinants of health data was also relatively poor (Figure 5.2c). Model performance on validation set using only pulmonary function test data was relatively strong, with lightGBM and logistic regression algorithms outperforming the other models (Figure 5.2d). Model performance on validation set

using only pulmonary function test data was relatively strong, with lightGBM slightly outperforming the other models (Figure 5.2e).

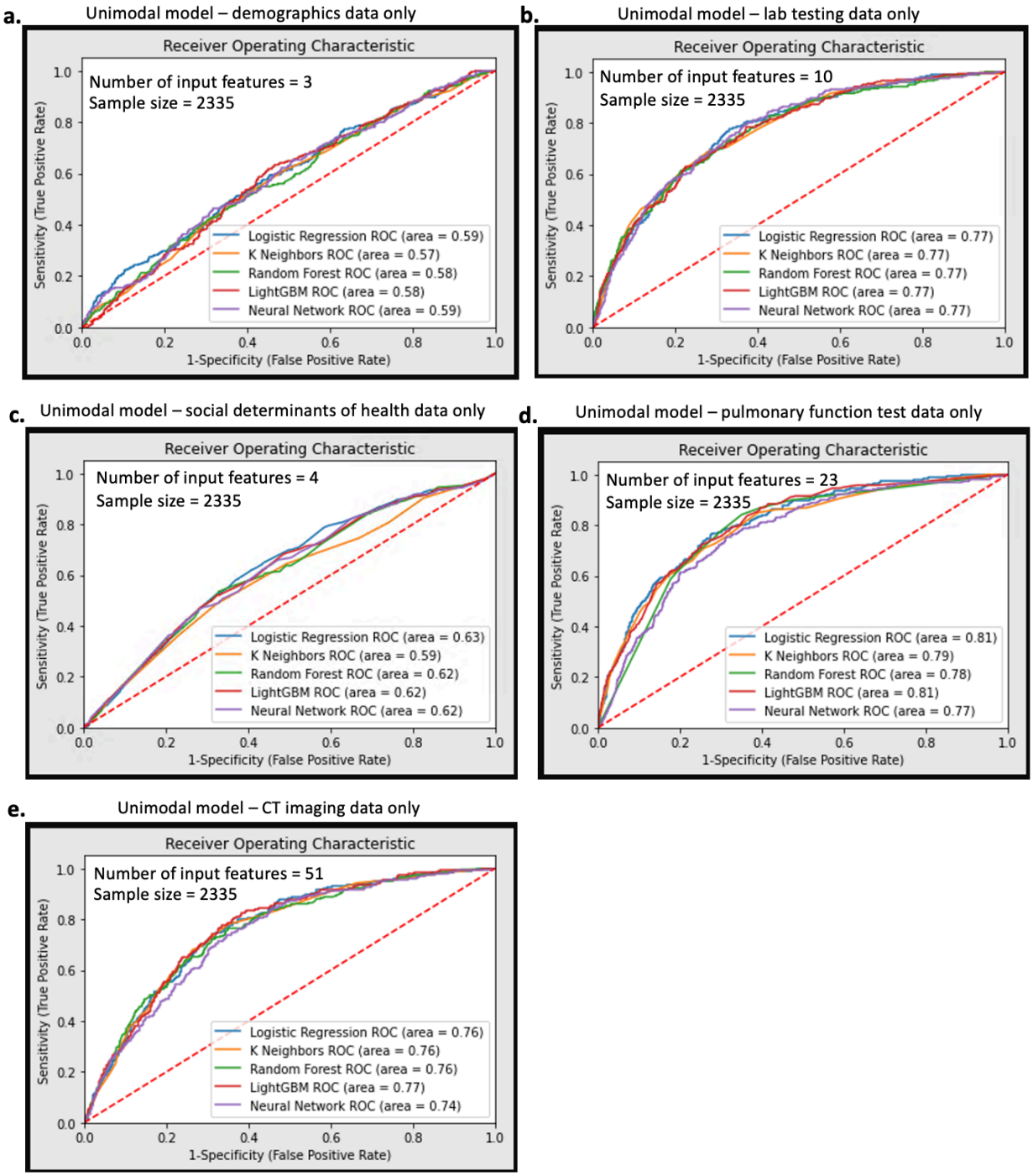

Figure 5.2: Performance (AUROC curves) of unimodal models trained using the following data: (a) demographics only (b) labs only (c) social determinants of health only (d) pulmonary function test only (e) CT imaging only

5.3.3 Multimodal model performance

Multimodal model performance on validation set using a combination of demographics and lab test data was relatively strong, with lightGBM and logistic regression again outperforming other models (Figure 5.3a). Model performance using a combination of demographics, lab test, and social determinants of health data was relatively strong, with the neural network and logistic regression slightly outperforming other models (Figure 5.3b). Model performance using a combination of demographics, lab test, social determinants of health, and pulmonary function test data was relatively strong, with the lightGBM and logistic regression slightly outperforming other models (Figure 5.3c). Model performance after adding CT imaging data was strong, with the lightGBM and logistic regression slightly outperforming other models (Figure 5.3d). Overall, the lightGBM and logistic regression models performed the best out of all the models we evaluated (AUC, 0.81).

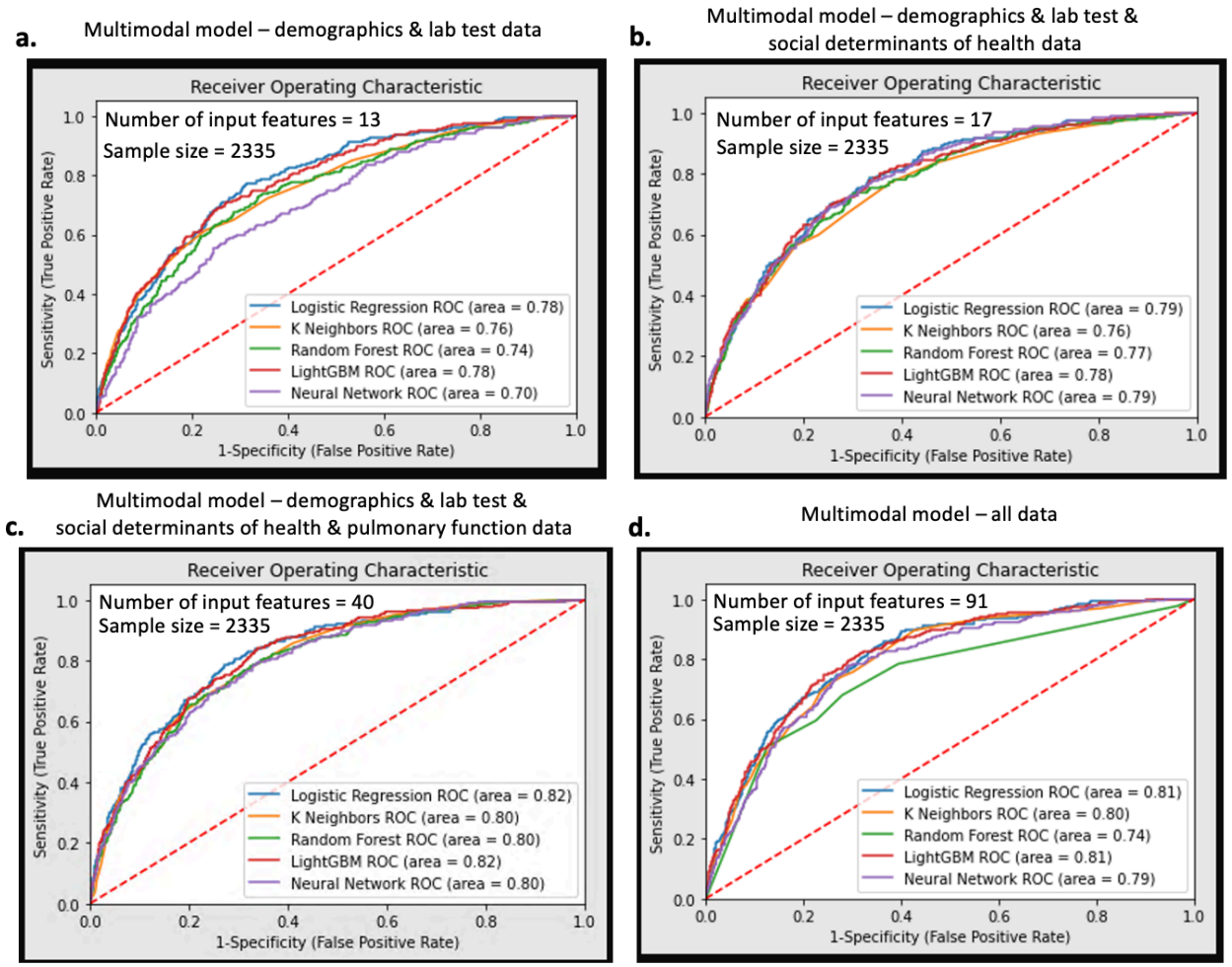

Figure 5.3: Performance (AUROC curves) of multimodal models trained using the following data: (a) demographics and labs (b) demographics, labs, and social determinants of health (c) demographics, labs, social determinants of health, and pulmonary function (d) demographics, labs, social determinants of health, pulmonary function, and CT imaging (all data)

5.3.4 Model interpretability using SHAP

To demonstrate model interpretability, we generated local and global SHAP interpretations for the observations used in evaluating the lightGBM multimodal model that was trained using all the features. SHAP force plots identify which features had the most influence on the model's prediction for a single observation. Figure 5.4 is an illustrative example of a force plot showing a local interpretation generated for a single observation with its own set of SHAP values. This example of a local interpretation reveals that BODE, $FEF_{25-75\%}$ (measured prior to

bronchodilator administration), occupational exposure to fumes, and resting oxygen saturation rank among the top contributing factors to this prediction.

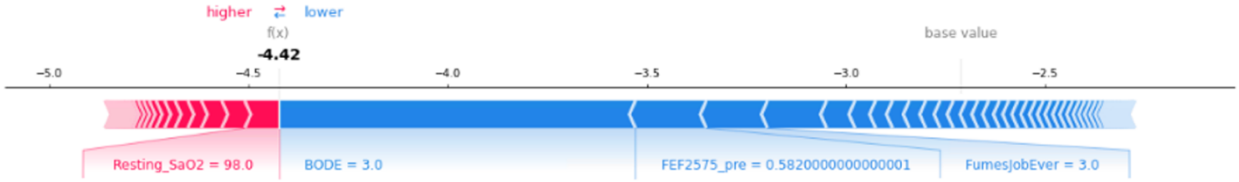

Figure 5.4. SHAP force plot generated using TreeSHAP on a single observation. SHAP force plots show the features from left to right, with the pink color representing positive contributions on the left, influencing the prediction value closer to 1, and the blue color representing negative contributions on the right, influencing the prediction value closer to 0. The distance between each feature corresponds to the magnitude of its influence. Raw values of the top contributing features to this prediction are shown on the bottom. The bolded value represents the output of the SHAP analysis for this single observation and does not represent the outcome label used in training and testing the multimodal model.

SHAP feature importance plots provide a global interpretation of the top contributing features for both positive and negative predictions. Figure 5.5 shows that BODE, age, and $FEF_{25-75\%}$ rank among the most important factors for the model's prediction of severe exacerbation.

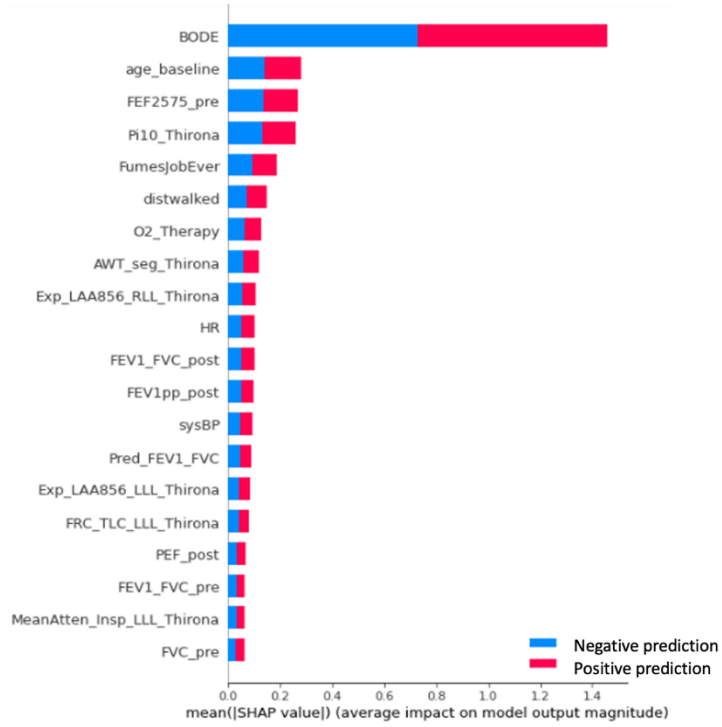

Figure 5.5. SHAP feature importance plot showing the global importance of the features. Feature importances are computed by taking the average of the absolute Shapley values per feature across the data and sorting them by decreasing importance.

The SHAP beeswarm plot is designed to display an information-dense summary of how the top features in a dataset influence the model's output. Figure 5.6 reveals that higher BODE indices correspond to higher SHAP values, which implies that the model tends to make positive predictions and vice versa. Greater age at baseline corresponds to lower SHAP values, suggesting that the model tends to make negative predictions and vice versa. Lower $FEF_{25-75\%}$ values correspond to higher SHAP values.

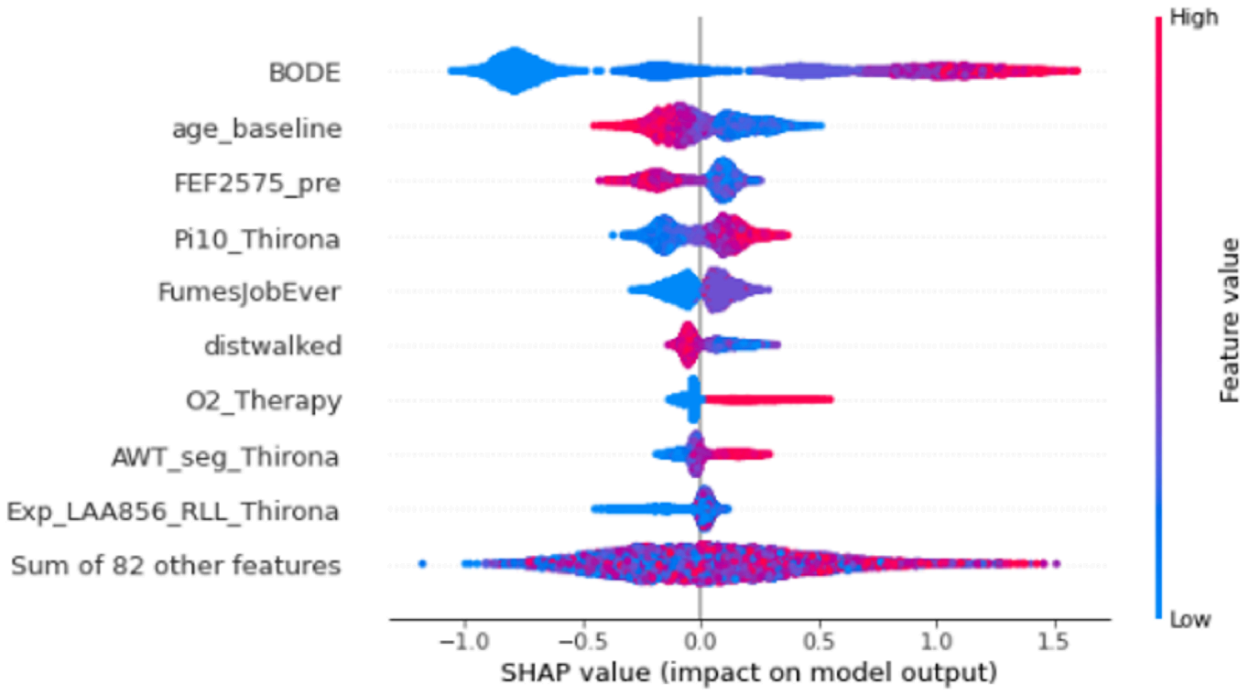

Figure 5.6. SHAP beeswarm summary plot reveals the top contributing features that influence the model's output.

SHAP dependence plots show the effect of a single feature across the entire dataset, and highlight possible interactions with another selected feature. The dependence plot in Figure 5.7 reveals that higher BODE values associate with lower FEV_1/FVC ratios, with the shift from higher to lower BODE values appearing notably close to the 0.7 FEV_1/FVC ratio diagnostic threshold for COPD. Lower BODE values and lower ratios have higher SHAP values, which trend the model toward positive prediction of severe exacerbation.

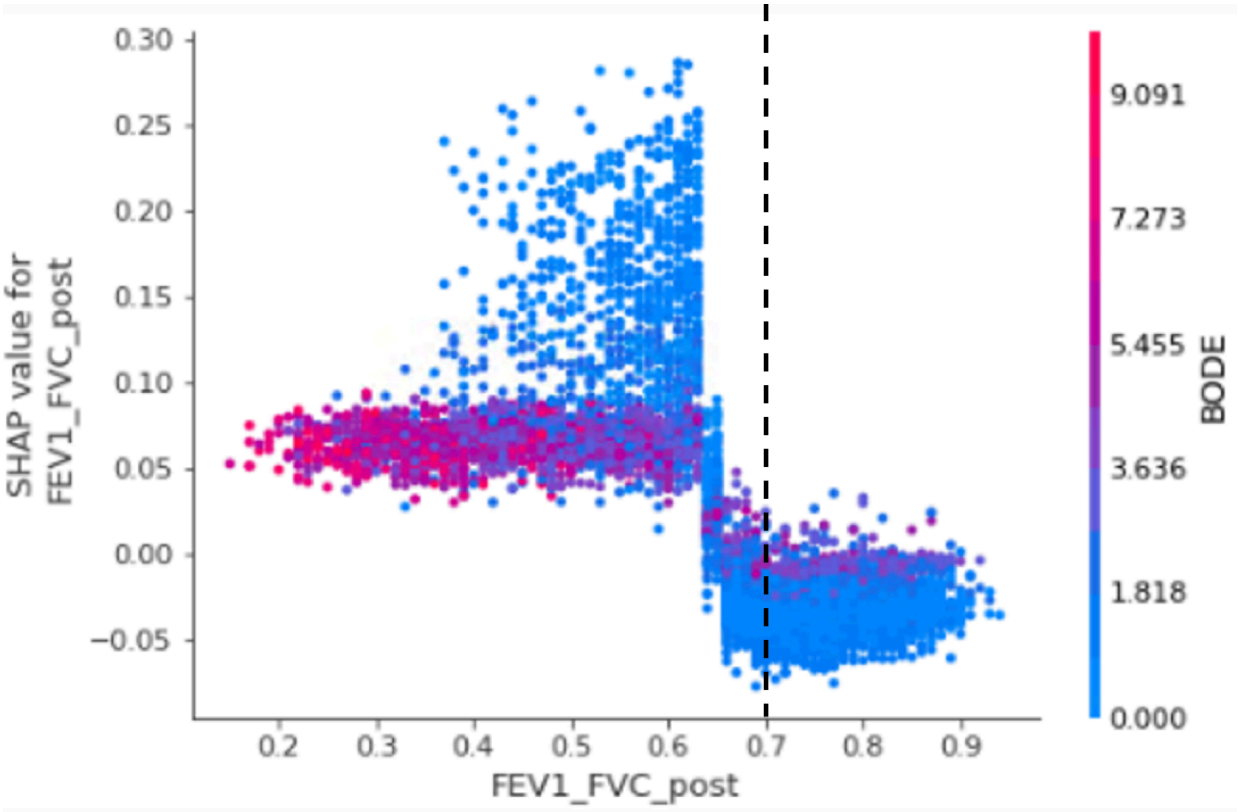

Figure 5.7. SHAP dependence plot reveals the impact of FEV_1/FVC across the whole dataset and interaction effects with BODE. The vertical dispersion of SHAP values at a single FEV_1/FVC value is driven by interaction effects, and BODE is chosen for coloring to highlight possible interactions. The 0.7 FEV_1/FVC ratio diagnostic threshold for COPD is marked by the dashed line.

5.4 Conclusions

In conclusion, we built a novel, interpretable multimodal machine learning framework for predicting severe exacerbations in patients with COPD (AUC, 0.81) using a broad selection of clinical, imaging, and socioeconomic data from the COPDGene cohort. We found that multimodal models significantly outperform a few unimodal models in predicting severe

exacerbations due to the incorporation of different data types. We identified BODE, baseline age, and $FEF_{25-75\%}$ among the top contributing features contribution to a positive prediction.

5.5 Discussion

We built a novel, high-performing (AUC, 0.81), interpretable multimodal machine learning framework for predicting severe exacerbations in patients with COPD using a broad selection of clinical, imaging, and socioeconomic data from the COPDGene cohort. We found that multimodal models significantly outperform unimodal models in predicting severe exacerbations. We further identified BODE, age, and $FEF_{25-75\%}$ to be among the top contributing features associated with severe exacerbations. Future work may investigate how these factors may be used to screen for patients at high risk for developing exacerbation-related complications.

Bayesian model-based optimization methods have been shown to be a powerful method for optimizing objective functions that are computationally expensive to evaluate.[19,20] For hyperparameter tuning, Bayesian optimization builds a probabilistic model of the objective function to propose a more optimal set of hyperparameters to evaluate that may improve model performance.[19,20] The goal is to find the most optimal set of hyperparameters that maximizes model performance. This iterative process is faster and more efficient than other common methods of hyperparameter optimization, such as grid search, which evaluates every position in a grid of pre-defined hyperparameter values, and random search, which randomly evaluates points in a bounded domain of hyperparameter values.[19,20]

Ongoing work includes applying the Delong's test to assess whether the differences in AUC between the multimodal and unimodal models are statistically significant. Additionally, several additional evaluation metrics will be calculated to assess the model performance across

the validation set, including: accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Accuracy refers to the fraction of outcome predictions that the model predicts correctly. Sensitivity is the ability of the model to correctly classify a patient as *having* the outcome.[21,22,23] Specificity refers to the ability of the model to correctly classify a patient as *not having* the outcome.[21,22,23] PPV is the percentage of patients classified as *having* the outcome who actually have the outcome.[21,22,23] NPV is the percentage of patients classified as *not having* who actually do not have the outcome.[21,22,23] PPV and NPV can be used to assess the clinical relevance of the model by accounting for the prevalence of severe exacerbations in the COPDGene population. A high PPV indicates that a positive outcome prediction is likely correct, while a high NPV indicates that a negative outcome prediction is likely to be correct.

Derived from Lloyd Shapley's studies in cooperative game theory, the Shapley values are calculated by averaging the marginal contribution of each feature across all potential permutations of the feature set.[15,26] The Shapley value of one feature is estimated as the difference between the actual prediction of the model when the feature is included and the expected prediction when the feature is excluded.[15,26,29] SHAP offers several advantages over other commonly used model interpretability methods such as Local Interpretable Model-agnostic Explanations (LIME) and Permutation Feature Importance (PFI).[30,31] While SHAP can be used to explain both global and local behaviors of the model, LIME only attempts to explain the model's behavior in the locality (i.e. neighborhood) of a specific data point.[30,32] While SHAP has been shown to uncover non-linear dependencies amongst features, LIME makes the tenuous assumption that the data has linear relationships locally, which fails to account for more complex, non-linear relationships that may exist in the underlying data structure.[30,32] Unlike SHAP, PFI only produces insights on the global behavior of the model and becomes more

susceptible to biased interpretations when features are correlated.[34] While SHAP measures the marginal contributions of the features toward a prediction, which lends a more intuitive understanding of model behavior, PFI quantifies the increase in prediction error of the model after permuting the feature's values and does not quantify the contribution of each feature toward the prediction.[34]

For this study, SHAP identifies BODE, age, and $FEF_{25-75\%}$ as among the top contributing features to predicting severe exacerbation. BODE is an index that integrates body mass index, airflow limitation (FEV_1), dyspnea and 6-min walk distance.[35] It has been shown to be predictive of long-term adverse health outcomes for COPD patients. The score of the BODE index is divided into four quartiles, where the highest quartile indicates the most serious disease.[35] Increasing BODE quartiles were associated with greater mortality risk. Initially it seemed counterintuitive that greater age associated with a negative prediction of severe exacerbation, as several other papers found age as one of the main predictors of lung function decline.[35] However, a study by Stone *et al.* reveals that older patients (especially those above 80) with severe COPD tend to be more socially isolated, have a reduced likelihood of self-transfer to hospital, and reduced awareness of symptoms related to exacerbation.[36] This suggests that severe exacerbation events experienced by older patients may be underreported in the COPDGene cohort, which may lead to a seemingly negative association between increasing age and severe exacerbation. The forced mid-expiratory flow ($FEF_{25-75\%}$) value is a potentially sensitive marker of obstructive peripheral airflow.[37] Previous study has found that lower $FEF_{25-75\%}$ can be an early predictor of COPD.[37] While SHAP provides these unique insights into model behavior, it is important to recognize that model interpretability does not mean causality.

One limitation of this study is that we did not validate on external dataset. However, future work may use the 5-year and 10-year, Phase 2 and Phase 3 data, respectively, to validate our models. Since severe exacerbations lead to further lung function decline which increases mortality risk, future work may apply our current machine learning framework to predict disease progression and mortality.

5.6 Acknowledgments

Chapter 5 presents unpublished material, and is coauthored with Theresa Gaasterland. The dissertation author was the primary author of this chapter. The dissertation author thanks Theresa Gaasterland for providing meaningful feedback and discussion on the computational experiments performed in this chapter.

5.7 References

1. Capriotti, T., Tomy, R., & Morales, M. (2023, May 9). *COPD Updates: 2023 GOLD Report for Primary Care Providers*. Clinical Advisor. https://www.clinicaladvisor.com/home/topics/copd-information-center/copd-updates-2023-gold-report-primary-care/#:~:text=Chronic%20lower%20respiratory%20disease%2C%20primarily,death%20in%20the%20United%20States.&text=Almost%2015.7%20million%20Americans%20(6.4%25)%20have%20been%20diagnosed%20with%20COPD
2. National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. Basics about COPD. Centers for Disease Control and Prevention: June 9, 2021. Accessed April 28, 2023. https://www.cdc.gov/copd/basics-about.html
3. Wedzicha, J. A., & Seemungal, T. A. (2007). COPD exacerbations: defining their cause and prevention. *Lancet (London, England)*, *370*(9589), 786–796. https://doi.org/10.1016/S0140-6736(07)61382-8
4. Reis, A. J., Alves, C., Furtado, S., Ferreira, J., Drummond, M., Robalo-Cordeiro, C., & GI DPOC-Grupo de Interesse na Doença Pulmonar Obstrutiva Crónica (2018). COPD exacerbations: management and hospital discharge. *Pulmonology*, *24*(6), 345–350. https://doi.org/10.1016/j.pulmoe.2018.06.006
5. Press, V. G., Konetzka, R. T., & White, S. R. (2018). Insights about the economic impact of chronic obstructive pulmonary disease readmissions post implementation of the hospital readmission reduction program. *Current opinion in pulmonary medicine*, *24*(2), 138–146. https://doi.org/10.1097/MCP.0000000000000454
6. Zeng, S., Arjomandi, M., Tong, Y., Liao, Z. C., & Luo, G. (2022). Developing a Machine Learning Model to Predict Severe Chronic Obstructive Pulmonary Disease Exacerbations: Retrospective Cohort Study. *Journal of medical Internet research*, *24*(1), e28953. https://doi.org/10.2196/28953
7. Blanchette CM, Dalal AA, Mapel D. (2012). Changes in COPD demographics and costs over 20 years. *J Med Econ*,*15*(6):1176-1182
8. 2023 Gold reports. Global Initiative for Chronic Obstructive Lung Disease - GOLD. 2023. https://goldcopd.org/gold-reports [accessed 2023-10-20]
9. Johnston J, Longman J, Ewald D, King J, Das S, Passey M. Study of potentially preventable hospitalisations (PPH) for chronic conditions: what proportion are preventable and what factors are associated with preventable PPH? BMJ Open 2020 Nov 09;10(11):e038415
10. Singh, D., Hurst, J. R., Martinez, F. J., Rabe, K. F., Bafadhel, M., Jenkins, M., Salazar, D., Dorinsky, P., & Darken, P. (2022). Predictive modeling of COPD exacerbation rates using baseline risk factors. *Therapeutic advances in respiratory disease*, *16*, 17534666221107314. https://doi.org/10.1177/17534666221107314
11. Tavakoli, H., Chen, W., Sin, D. D., FitzGerald, J. M., & Sadatsafavi, M. (2020). Predicting Severe Chronic Obstructive Pulmonary Disease Exacerbations. Developing a Population Surveillance Approach with Administrative Data. *Annals of the American Thoracic Society*, *17*(9), 1069–1076. https://doi.org/10.1513/AnnalsATS.202001-070OC
12. Adibi, A., Sin, D. D., Safari, A., Johnson, K. M., Aaron, S. D., FitzGerald, J. M., & Sadatsafavi, M. (2020). The Acute COPD Exacerbation Prediction Tool (ACCEPT): a

modelling study. *The Lancet. Respiratory medicine*, *8*(10), 1013–1021. https://doi.org/10.1016/S2213-2600(19)30397-2
13. Huang, S. C., Pareek, A., Zamanian, R., Banerjee, I., & Lungren, M. P. (2020). Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports*, *10*(1), 22147. https://doi.org/10.1038/s41598-020-78888-w
14. Eisner, M. D., Blanc, P. D., Omachi, T. A., Yelin, E. H., Sidney, S., Katz, P. P., Ackerson, L. M., Sanchez, G., Tolstykh, I., & Iribarren, C. (2011). Socioeconomic status, race and COPD health outcomes. *Journal of epidemiology and community health*, *65*(1), 26–34. https://doi.org/10.1136/jech.2009.089722
15. Lundberg, S.M. & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *NeurIPS Proceedings*, 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
16. Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11
17. Płoński, P. (2020, June 29). *Random Forest Feature Importance Computed in 3 Ways with Python*. mljar. https://mljar.com/blog/feature-importance-in-random-forest/
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html
19. Snoek, J., Larochelle, H., & Adams, R.P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *NeurIPS Proceedings*, 25. https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf
20. Dewancker, I., McCourt, M., & Clark, S. (2015). Bayesian optimization primer. *SIGOPT*. https://static.sigopt.com/b/20a144d208ef255d3b981ce419667ec25d8412e2/static/pdf/SigOpt_Bayesian_Optimization_Primer.pdf
21. Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, *56*(1), 45–50. https://doi.org/10.4103/0301-4738.37595
22. Safari, S., Baratloo, A., Elfil, M., & Negida, A. (2015). Evidence Based Emergency Medicine Part 2: Positive and negative predictive values of diagnostic tests. *Emergency (Tehran, Iran)*, *3*(3), 87–88.
23. Tenny, S., Hoffman, M. (2020). Prevalence. *StatPearls [Internet]*. https://www.ncbi.nlm.nih.gov/books/NBK430867/
24. Aboze, B.J. (2023, May 16). *A Comprehensive Guide into SHAP (SHapley Additive exPlanations) Values*. deepchecks. https://deepchecks.com/a-comprehensive-guide-into-shap-shapley-additive-explanations-values/
25. Kuo, C. (2019, Sept 13). *Explain Your Model with the SHAP Values*. Medium. https://medium.com/dataman-in-ai/explain-your-model-with-the-shap-values-bc36aac4de3d

26. Molnar, C. (2022). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/shap.html
27. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (1135–1144)
28. Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.
29. Fadel, S. (2022, Feb 28). *Explainable Machine Learning, Game Theory, and Shapley Values: A technical review*. Statistics Canada. https://www.statcan.gc.ca/en/data-science/network/explainable-learning#:~:text=When%20using%20SHAP%20%2C%20the%20aim,values%20from%20coalitional%20game%20theory.
30. Strumbelj, E. & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems, 41*(3), 647-665.
31. Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure, *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134
32. Fryer, D. V., Strumke, I., & Nguyen, H. (2021). Model independent feature attributions: Shapley values that uncover non-linear dependencies. *PeerJ. Computer science*, *7*, e582. https://doi.org/10.7717/peerj-cs.582
33. Ruggero, A. (2020, Jul). *Explaining Learning to Rank Models with Tree Shap*. sease. https://sease.io/2020/07/explaining-learning-to-rank-models-with-tree-shap.html
34. Bagheri, R. (2022, Mar 31). *Introduction to SHAP Values and their Application in Machine Learning*. Medium. https://towardsdatascience.com/introduction-to-shap-values-and-their-application-in-machine-learning-8003718e6827
35. Li, C. L., Lin, M. H., Chen, P. S., Tsai, Y. C., Shen, L. S., Kuo, H. C., & Liu, S. F. (2020). Using the BODE Index and Comorbidities to Predict Health Utilization Resources in Chronic Obstructive Pulmonary Disease. *International journal of chronic obstructive pulmonary disease*, *15*, 389–395. https://doi.org/10.2147/COPD.S234363
36. Stone, R. A., Lowe, D., Potter, J. M., Buckingham, R. J., Roberts, C. M., & Pursey, N. J. (2012). Managing patients with COPD exacerbation: does age matter?. *Age and ageing*, 41(4), 461–468. https://doi.org/10.1093/ageing/afs039
37. Kwon, D. S., Choi, Y. J., Kim, T. H., Byun, M. K., Cho, J. H., Kim, H. J., & Park, H. J. (2020). $FEF_{25-75\%}$Values in Patients with Normal Lung Function Can Predict the Development of Chronic Obstructive Pulmonary Disease. *International journal of chronic obstructive pulmonary disease*, *15*, 2913–2921. https://doi.org/10.2147/COPD.S261732

CHAPTER 6: CONCLUSIONS

In our first aim (CHAPTER 2), we leveraged the broadest range of imaging and clinical characteristics available in the COPDGene cohort. We observed that the top principal components that emerged from our unsupervised analysis were predominantly related to CT imaging and spirometric phenotypes. Specifically, the top three components (1) air trapping/emphysema, (2) vital capacity/lung volume, and (3) lung mass/airway thickening/body habitus each explain a large fraction of the variability of the population. We used these components to create disease axes. Each of these axes also individually demonstrate prognostic potential, predictive of future FEV_1 decline, exacerbation, and mortality. In summary, this work shows that imaging phenotypes explain a significant proportion of COPD heterogeneity. Including CT measurements in COPD patient management may assist in identifying those most in need of therapeutic interventions early in their disease trajectory.

In our second aim (CHAPTER 3), we leveraged deep learning and deformable registration techniques to automate quantification of emphysema and air trapping. Using these imaging features, we defined a pragmatic CT-based system for staging COPD, which parallels spirometry-defined GOLD stages. Our CT-based stages are prognostically valuable, whether used independently or in concert with pulmonary function tests, predicting both future FEV_1 decline and mortality. Most notably, we found patients with moderate spirometric impairment and moderate-severe disease on CT had the highest likelihood of progression. These results point to the synergistic value of CT and functional testing to identify patients at the highest risk of progression, and potentially those who might have the greatest benefit from medical or behavioral intervention to address ongoing tobacco or other particulate exposures. We demonstrated that quantitative CT-based stages of COPD severity can prognosticate progression

and mortality when implemented independently or in conjunction with the spirometric GOLD criteria. Fully-automated deep learning algorithms can facilitate translation of the proposed stages into clinical practice, potentially allowing us to better manage our patients with COPD.

In our third aim (CHAPTER 4), we successfully developed a novel, automated shallow 3-layer CNN classifier, SimplePNUnet, to detect the presence of pneumonia using a subset of 2000 frontal chest X-rays from the publicly accessible dataset released as part of the 2018 RSNA Pneumonia Detection Challenge. Additionally, we assessed the robustness of SimplePNUnet across different downsampling techniques and batch sizes. Performance of the algorithm was evaluated using the area under receiver operating characteristic curve (AUROC). The performance of the best model across the training (n=1000), validation (n=500), and test (n=500) sets was comparable (AUC, 0.870 vs 0.812 vs 0.834, respectively). SimplePNUnet achieved good performance on an external validation set comprising COVID-19 chest radiographs (AUC, 0.712).

Lastly in our fourth aim (CHAPTER 5), we built a novel, interpretable multimodal machine learning framework for predicting severe exacerbations in patients with COPD (AUC, 0.81) using a broad selection of clinical, imaging, and socioeconomic data from the COPDGene cohort. We found that multimodal models significantly outperform a few unimodal models in predicting severe exacerbations due to the incorporation of different data types. We identified BODE, baseline age, and $FEF_{25-75\%}$ among the top contributing features contribution to a positive prediction.