Effort-Independent Asthma Severity Classification

by

James C. Lynch III

B.S., University of Rhode Island (2015)

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Electrical Engineering and Computer Science
May 20, 2021

Certified by .....
Professor Thomas Heldt
Associate Professor of Electrical and Biomedical Engineering
Thesis Supervisor

Accepted by .....
Professor Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Effort-Independent Asthma Severity Classification

by

James C. Lynch III

Submitted to the
Department of Electrical Engineering and Computer Science
on May 20, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Asthma is an obstructive pulmonary disorder. It impacts the lives of over 24 million individuals in the United States alone, a large segment of which are children. We propose to investigate capnography as a viable diagnostic modality to guide the treatment of asthma as an alternative to the gold standard, spirometry. Capnography shows promise in the detection of similar pulmonary disorders, and would serve as a noninvasive and effort-independent tool, providing critical information to clinicians when patients are unable or unwilling to comply with spirometry testing. In this work, we demonstrate the viability of using features extracted from time-based capnography to determine underlying patient symptom severity, using logistic regression classification models. Applications in both controlled, pulmonary function laboratories and emergency department triage conditions are explored. We show that for an adult population undergoing methacholine challenge pulmonary function testing, capnography recordings from subjects with asthmatic exacerbation may be distinguished from their normal/baseline recordings with an AUROC of 0.92 (0.84 – 1.00). Additionally, using data from an acute pediatric setting we show that recordings from subjects with severe asthmatic exacerbation may be distinguished from subjects with mild or moderate asthma symptoms with an AUROC of 0.86 (0.72 – 1.00).

Thesis Supervisor: Professor Thomas Heldt
Title: Associate Professor of Electrical and Biomedical Engineering

Acknowledgements

Without question, the individual to whom I owe the greatest professional debt is Professor Thomas Heldt. I am profoundly grateful for his mentorship, guidance, patience, and of course, his keen insight. I have observed, and myself experienced Professor Heldt effortlessly draw out intellectual curiosity and enthusiastic scholarship in his students, both advised and taught. It has been, and continues to be an honor to learn from him, and work with him and those he brings into the lab.

Dr. Baruch Krauss was instrumental both in facilitating data acquisition for these studies, as well as providing crucial clinical insight. This work would not have been possible without his guidance and expertise.

My parents, sister, and grandparents are enduring supporters, without whom I simply could not be who I am or where I am today. Their selflessness and love ground me in ways they cannot imagine.

Equal parts friends, colleagues, and peer mentors I thank Rohan, Imad, Jonathan, Varesh, Jeff, Freddie, and Stéphane for countless hours of thoughtful conversation, and their support and insight over the years.

This work was made possible in part by the Frederick R. (1953) and Barbara Cronin Fellowship from the Department of Electrical Engineering and Computer Science and by Philips Healthcare through funding to the MIT Medical Electronic Device Realization Center.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Asthma is an obstructive disorder of the upper respiratory system. It manifests as wheezing, difficulty breathing, shortness of breath, and coughing – all symptoms which may be exacerbated by increased physical activity [1]. As it is a chronic condition, asthma may persist throughout the lifetime of an affected individual and as a result can reduce the quality of life. Although rare, asthma can even result in death [2]. In the United States, approximately 24.7 million individuals have asthma as of 2018 [1]. Asthma often manifests early in childhood, and affects some 5.5 million children in the US [1].

A diagnosis of asthma is usually confirmed by spirometry, a pulmonary function test that measures whether a patient's lungs are partially obstructed [3]. This test is labor-intensive on the part of the patient, and requires that they exhale forcibly into a measurement device several times. In situations where a patient is unable or unwilling to comply with the test, it can be difficult for clinicians to adequately confirm a diagnosis and determine optimal treatment. Due to asthma's prevalence in children, pediatric patients are often asked to perform spirometry. When the patient is in distress, or is simply too small or young to adequately perform the test, getting a reliable reading can be nearly impossible [3]. This particular problem motivates the development of diagnostic tools that provide an accurate, alternative method of measuring lung performance that does not impose such a burden on the patient. A passive, effort-independent method of determining pulmonary function in these cases

would substantially increase the amount of information available to clinicians in these circumstances.

1.1 Thesis Contributions

The goal of this work is to investigate the viability of using time-based capnography as a diagnostic aid in certain clinical settings, as applied to asthma. We investigate two clinical settings: a pulmonary function laboratory in which adult subjects are undergoing methacholine challenge testing to determine whether they have asthma, and a pediatric emergency department setting in which children are presenting with asthma symptoms seeking emergent treatment. In both settings, we attempt to identify features of the time series of exhaled partial pressure of carbon dioxide, termed the capnogram, taken at various points in testing and/or treatment that correlate with the documented concurrent patient state.

In Chapters 4 and 5, we show that it is indeed possible to use the information contained in these time-based capnography recordings to successfully differentiate between subjects with more severe asthmatic exacerbation from those with no or minimal symptoms. These investigations serve as confirmation of the viability of using capnogram features in the development of effort-independent diagnostic tools to augment the triage and diagnosis of asthma.

1.2 Thesis Structure

This work is divided into six Chapters. Chapter 2 is dedicated to the description of capnography and its current use in clinical settings relevant to asthma treatment and diagnosis. Chapter 3 describes the human subjects protocols and associated data collection for this work, as well as the data preprocessing and annotation necessary to perform the analysis. Chapter 4 explores three asthma severity classification tasks using the methacholine challenge pulmonary function testing dataset. Chapter 5 explores classifying asthma symptom severity in a pediatric emergency department

triage setting. Finally, Chapter 6 summarizes the results of the classification tasks, and discusses the challenges encountered over the course of this research. Further, we speculate as to the next steps and future contributions that may follow this body of work.

Chapter 2

Background

2.1 Asthma

Asthma can present as a multitude of symptoms including difficulty breathing, chest tightness, and coughing caused by airflow obstruction and inflammation [3, 4]. Around the world, it affects approximately 339 million individuals globally, and more than 24.7 million individuals in the United States [1, 5]. It is estimated that the total cost of asthma-related healthcare in the United States exceeds $62 billion every year [4]. The disease often presents for the first time in childhood, affecting as many as 1 in 12 children in the United States [1, 3, 4]. For many children diagnosed with asthma, symptoms cease by early adulthood [6]. Industrialization is implicated in the increase in prevalence of asthma around the world [6].

Environmental factors such as allergens, active respiratory infection, seasonal allergies, exercise, and airborne contaminants such as tobacco smoke can trigger or exacerbate symptoms [3]. The typical disease process is, at root, caused by inflammation. In response to environmental stimulus, immune cells enter into the epithelium of the upper respiratory tract. This infiltration causes inflammation and may cause excessive mucus secretion, damage to the epithelium, and damage to smooth muscles that line the airway [3]. These changes to the airway partially obstruct the flow of air to and from the lungs, which then manifest as dyspnea, cough, wheezing, and tightness in the chest [3, 6]. Over time, these processes can cause permanent remodeling

or structural alteration of the airway, further worsening symptoms, and reducing the effectiveness of treatment. [3, 6].

While the symptoms of chronic asthma are often mild, acute asthma can present as severe bronchospasm that must be treated with bronchodilators [3]. During bronchospasm, the bronchi suddenly contract and significantly reduce airflow to the lungs. Short-acting β_2 agonists, such as albuterol, act as bronchodilators and are delivered by inhaler, by nebulizer, orally, or intravenously to relieve the symptoms [3]. The reversibility of airway obstruction/airway hyper-responsiveness with bronchodilators is considered a distinct feature of asthma that distinguishes it from the symptoms of other pulmonary disorders such as chronic obstructive pulmonary disease (COPD) or congestive heart failure (CHF), which present with similar symptom complexes in adults [3, 7]. As these and other respiratory diseases share symptoms with asthma, pulmonary performance testing, such as with spirometry, may be used to confirm an asthma diagnosis.

2.2 Spirometry

In the clinic, the extent to which respiratory performance is diminished is determined through the use of pulmonary function tests. The test most commonly used, particularly in the context of asthma diagnosis, is spirometry [3, 7, 8]. Spirometry is administered using a hand-held flow meter that quantifies the volume of air forcefully exhaled over a period of time [8]. To appropriately administer spirometry, the patient must force air out as hard and rapidly as possible, starting from the point of maximum inhalation [8]. This can be strenuous, and must be repeated multiple times before the measured value is considered reliable [8]. When a patient is suffering from a major asthma attack, the test can be all but impossible to perform. Children under five years of age in most cases cannot perform the test reliably [9, 10]. Spirometry is also inappropriate or unusable for individuals that are unable to understand or follow instructions [11]. However, the information that spirometry provides in these circumstances, namely the ability to compare lung performance before and af-

ter bronchodilator treatment, aids the clinician in determining whether the treatment was successful and whether asthma was indeed the cause of the patient's symptoms and is considered the gold standard for diagnosing asthma [8, 12].

Spirometry measures the volume of air expired over time and a number of diagnostically useful parameters can be extracted from this time series. The specific parameters that clinicians typically use are the forced vital capacity (FVC), the forced expiratory volume expired in one second (FEV_1), and the forced expiratory volume expired in six seconds (FEV_6) [8]. FVC measures the volume of air that is expired over the course of a full exhalation, starting from a point of effort-induced maximal inhalation and ending at a maximally forced exhalation [8]. For the purposes of spirometry, this represents the total volume of air that a patient can exhale. The FEV_1 and FEV_6 measure the volume of air that may be exhaled, as rapidly as possible, by the patient in one or six seconds, respectively, as counted from the beginning of exhalation [8]. These individual measures and the derived measure FEV_1 / FVC are determined for the patient's pulmonary state before and after bronchodilator treatment and compared. These measures are also referenced against known normal ranges for a patient's age. Again, as asthma is defined by the recovery of pulmonary performance post-bronchodilator treatment, a significant improvement in these measures implicates asthma as the likely underlying disease process.

2.3 Methacholine Challenge

When a clinician would like to definitively determine in a controlled setting whether a patient has asthma, they may order a bronchial provocation test such as a methacholine challenge. This test is performed when the patient is at their baseline, with no acute symptoms of asthma. During the methacholine challenge, the bronchoconstrictor methacholine chloride is used to stimulate a bronchial response and a tightening of the airways [13].

To complete the methacholine challenge, spirometry is first performed prior to the administration of the bronchoconstrictor in order to establish a baseline measurement

of pulmonary performance [13]. Successively more concentrated doses of the bronchoconstrictor are administered with a diluent (such as saline), up to a maximum concentration of 16 mg/mL [13]. At each dosage step, the patient repeats the spirometry test. If at any dosage, up to the maximum dose, the patient's FEV_1 volume drops by more than 20% of their baseline, the patient is considered positive for asthma and a bronchodilator may be administered to reverse the effects of the methacholine [13]. If the patient's FEV_1 is not diminished by at least 20% at the maximal dosage, they are considered negative for airway hyper-responsiveness, and thus asthma is ruled out from the differential diagnosis [13].

2.4 Capnography

Capnography is the measurement of carbon dioxide (CO_2) content in the exhaled breath of a subject [14]. It is expressed as a trace of the partial pressure of CO_2 (PCO_2), specifically the partial pressure of CO_2 in exhaled air ($PeCO_2$), and is conventionally measured in millimeters of mercury (mmHg) [14]. Capnography is useful for the passive monitoring of patient respiration, such as during surgery or when a patient is intubated [14]. Modern capnographs use infrared light to measure the amount of CO_2 present by utilizing CO_2's partial absorption of infrared light centered about the wavelength 4,256 μm [14]. This particular absorption peak is chosen as water vapor's absorption of infrared light interferes with that of CO_2 at other wavelengths [14].

Exhaled air is sampled using one of two methods, namely mainstream and sidestream capnography [15]. Mainstream capnography measures the CO_2 content of air directly as it enters or exits a subject's mouth or nose [15, 16]. By placing a sensor directly inline with the airflow, it is also possible to measure the volume flow rate of air, and thus the total volume of air inhaled and exhaled over a breathing cycle. A mainstream capnograph's sensor may be used connected to, or as a part of, a ventilator [15]. In contrast with mainstream, sidestream capnography instead samples air from the patient's breathing line or a nose/mouth cannula using a small pump with a constant

flow rate [15]. This allows the capnograph's infrared sensor to sit away from the patient, connected to the cannula or breathing line by a long, flexible tube.

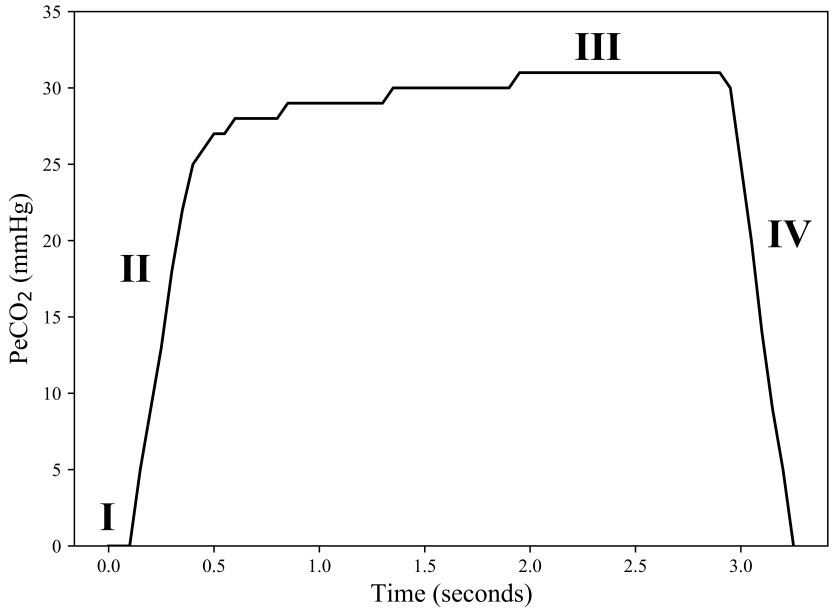

Figure 2-1: The capnogram of a single exhalation, with Phases I through IV annotated.

Over the course of a single exhalation, the gas content of the exhaled air changes as air from different parts of the respiratory system reach the point of measurement, as indicated in Figure 2-1. From a normal physiological perspective, the ambient air contains essentially no CO_2. Hence, air from the oropharynx and conducting airways where gas exchange with the blood does not occur has essentially no CO_2 content [17]. This air corresponds to that which is measured by the sensor during Phase I in the Figure. These anatomical structures are often called the anatomical deadspace of the respiratory system [17]. $PeCO_2$ rapidly rises as air from deeper within the lungs mixes with air from the deadspace and passes through the sensor, corresponding to Phase II [17]. The partial pressure of CO_2 eventually plateaus at the highest level of CO_2 at Phase III, which provides a measure of the CO_2 content of gas from alveoli deep within the lungs [17]. The final (often also the maximum) value of the plateau is called end-tidal CO_2 ($ETCO_2$) and is a closely-monitored vital sign in critical care and surgical settings, as it provides information about ventilation performance [16].

As inspiration begins, the sensor registers a rapid falloff in PCO_2 as it is flushed with inspired, atmospheric air that lacks physiologically significant levels of CO_2, shown during Phase IV in Figure 2-1 [17].

$PeCO_2$ as measured by a capnograph is plotted as a function of time or as a function of exhaled volume [16]. These traces, and the systems that generate them, are therefore considered either time-capnography and volume-capnography. The term capnogram usually refers to the time-based trace of $PeCO_2$ [16]. Mainstream capnography is capable of generating plots with respect to both time and volume, when a flow sensor is present. As sidestream capnographs lack flow sensors, they can only generate time-based capnograms [16].

2.5 Interpretations of Capnography

Early analysis of the capnogram generally considered the gross features of the shape, such as the slope of the exhalation onset and the plateau, as well as the curvature between these segments [17]. Most of the interpretation of the capnogram shape is qualitative, often looking for intra-patient variations over the course of treatment, or simply confirming the presence of normal respiration such as during endotracheal tube placement [14]. Quantitative analysis beyond determining the respiratory rate and $ETCO_2$ was not commonly used in clinical practice [17]. More recently, work has been undertaken to develop automated quantitative methods of capnogram interpretation in the context of specific diseases [17]. These approaches can be divided into feature-based and model-based techniques.

2.5.1 Feature-Based Approaches

Feature-based capnogram analysis enumerates several features of the capnogram shape and investigates the correlation and classification of particular ranges of these indices with a patient's underlying disease or condition. The capnogram is typically divided into three segments: the rising edge of the expiration's onset when CO_2 enriched air reaches the airway opening, the plateau leading to the end of the expiration and

subsequent initiation of inspiration, and the curved middle segment connecting these two slopes [16, 17]. Each of these segments has an associated slope and duration. Additional measures, such as area ratio of the middle capnogram segment, and the ratio of the slopes of the first and third segment may also be considered [17]. You, *et al.* established correlations between these indices and the results of spirometry in asthmatic patients, suggesting that evidence of asthma is present in the capnogram shape [17].

Mieloszyk, *et al.* developed classifiers that are capable of differentiating between the capnograms of patients with COPD and patients with CHF, as well as between patients with COPD and healthy adults [18]. In this study, the authors trained a series of classifiers based on quadratic discriminant analysis (QDA) for each task. The features chosen were exhalation duration, end-exhalation slope, $ETCO_2$, and the duration the $PeCO_2$ remains at the $ETCO_2$ value of each exhalation. Multiple classifier "voters" were trained on different, overlapping cuts of the training set. Based on the results of these voting classifiers, decision thresholds were selected that provided approximately equal sensitivity and specificity on the test data. This approach resulted in good performance of the COPD/CHF and COPD/normal classifiers, quantified using the area under the receiver operating characteristic (ROC) curve. [18]. These methods expanded significantly on earlier techniques by successfully implementing automated preprocessing, artifact rejection, and feature extraction in conjunction with powerful classification methods.

2.5.2 Model-Based Approaches

Model-based approaches to capnogram analysis apply mechanistic models to the physiology of the respiratory system, and parameterize the underlying physiological state [19, 20]. The general shape of the time-based capnogram may be derived from such a model [19]. Abid *et al.* implemented a simple model of the lungs and upper airways that accounts for the mixing of exhaled CO_2 with air in the deadspace. This model may be solved analytically for an expression for the partial pressure of exhaled CO_2, p_D, as a function of time:

$$p_D(t) = p_A \left(1 - e^{-\alpha} e^{\alpha e^{-t/\tau}}\right)$$

where p_A, τ, and α represent the constant (over an exhalation) CO_2 gas concentration in the alveoli, the pulmonary time constant associated with the expiration, and a constant quantity that represents the ratio of tidal volume to deadspace volume in the respiratory system, respectively. This function may then be fit to the capnograms of individual exhalations to derive these parameters. The authors implemented a classifier that was able to successfully discriminate between capnograms from patients with COPD and healthy adults with good accuracy [19].

Chapter 3

Clinical Data Collection and Preprocessing

To pursue the development of an effort-independent, capnography-based diagnostic aid for asthma, it is necessary to acquire representative data across patient demographics and asthma severity. In principle, such an analysis requires recorded capnograms describing subjects' respiration over a period of time, and corresponding asthma symptom severity labels. Our clinical collaborators undertook two studies that attempt to capture these data elements: a methacholine challenge study, which takes capnography measurements throughout the standard pulmonary function test, and a pediatric asthma study that captures capnograms from pediatric patients presenting to the emergency room with acute asthma symptom exacerbation.

For both of these studies, the Oridion Capnostream 20 was used as the capnograph. It is a sidestream capnograph that samples a small amount of air at a constant rate (approximately 50 mL/min) from a patient's nasal cannula and measures the exhaled CO_2 as a function of time. The device is relatively small and portable, making it ideal for use in diverse clinical settings. It is capable of sampling CO_2 at 20 Hz with a reported $PeCO_2$ precision of 1 mmHg. For these studies, the waveform measurements were recorded to directly attached USB flash storage directly from the Capnostream. Measurement accuracy falls within ± 2 mmHg between $0 - 38$ mmHg $PeCO_2$ and may decrease slightly above from 38 mmHg.

3.1 Methacholine Challenge Study

Studying patients undergoing a methacholine challenge offers the opportunity to collect data on potential asthmatics under precise, controlled conditions. As described previously, the methacholine challenge is performed when trying to confirm whether a patient has asthma when they are not actively experiencing acute symptoms. From 2011 to 2012, an IRB-approved study was undertaken at the Beth Israel Deaconess Medical Center with the aim of assessing how a patient's capnogram changes under different levels of airway obstruction. In this study, subjects were recruited from adult patients already scheduled to undergo the methacholine challenge pulmonary function test. After consenting to participate in the study, the methacholine challenge was conducted as normal with the addition of capnography recordings taken after each spirometry test. The Oridion Capnostream was used to record the time-based capnogram in this study.

The test proceeded as follows: first, baseline measurements of the subject's FVC and FEV_1 were made with spirometry. After these usual measurements were taken, the subject was connected to the capnograph via a nasal cannula and their capnogram was recorded for approximately three minutes. Then, the subject was administered pure diluent after which both the spirometry and capnography measurements were repeated. This pattern of drug administration followed by measurements was repeated multiple times with increasingly larger doses of methacholine: 0.0625 mg/mL, 0.25 mg/mL, 1.0 mg/mL, 4.0 mg/mL, and 16 mg/mL. If after the administration of any concentration of methacholine (or the pure diluent), the subject's FEV_1 were measured to have fallen below 80% of the baseline value, the subject would be considered positive for airway hyper-responsiveness and asthma. After a positive measurement, the subject would be administered a bronchodilator to ease their symptoms. Spirometry and capnography measurements would be taken a final time some ten minutes after the administration of the bronchodilator in order to confirm the recovery of pulmonary performance. In the event that the subject's FEV_1 did not drop below 80% of their baseline through the administration of the 16 mg/mL dose, the test would

indicate a negative result for airway hyper-responsiveness and asthma.

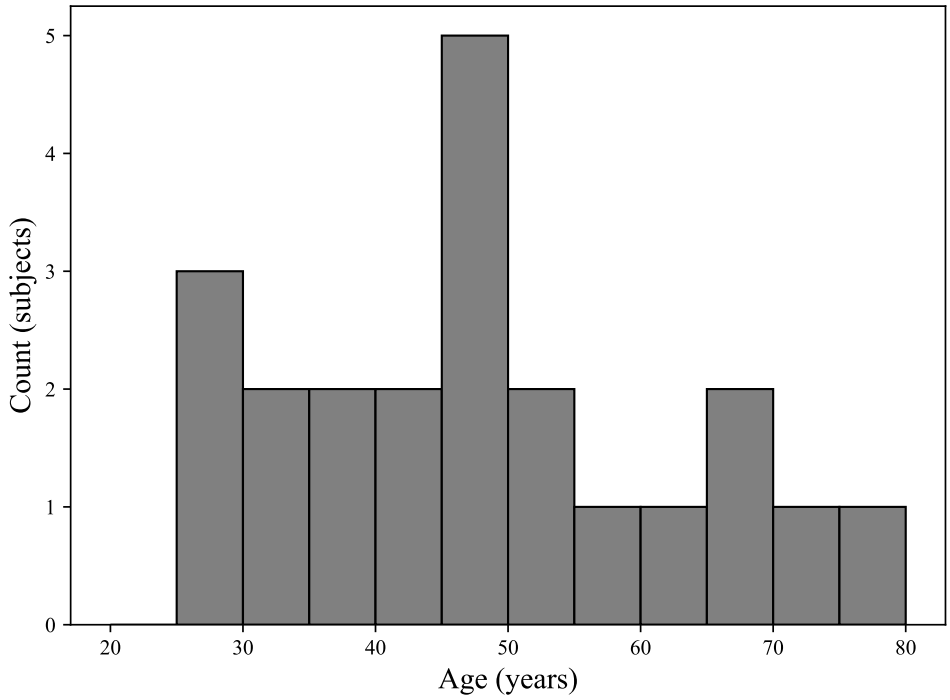

Figure 3-1: The distribution of age of subjects considered in the methacholine study. Four subjects are missing demographics information, including their age.

Data on 26 subjects ranging in age from 25 to 79 were collected. Four of the 26 subjects are missing demographic data, including their age and sex. For the 22 patients with demographics information, a distribution of their ages are presented in Figure 3-1. Sex and airway hyper-responsiveness/positivity to the methacholine challenge test are reported in Table 3.1 for all subjects. To optimize data collection, often the capnography recordings were skipped for intermediate concentrations of methacholine. In all subjects, the baseline and largest doses of methacholine had corresponding capnography recordings, and most subjects that received bronchodilator treatment also have a capnography recording 10 minutes post-bronchodilator administration. Specifically, in the case a subject were negative for airway hyper-responsiveness and received the 16 mg/mL methacholine dose, the subject's capnogram would be retained at baseline and after the 16 mg/mL dose. If a subject tested positive for airway hyper-responsiveness, the subject's capnogram would be recorded

Table 3.1: Subject characteristics in methacholine challenge study.

	n (%)
Sex	
Male	10 (39%)
Female	12 (46%)
Unknown	4 (15%)
Methacholine Challenge Result*	
Negative	10 (38%)
Positive	16 (62%)

* Subject exhibits airway hyper-responsiveness to methacholine. See Chapter 2.3.

at baseline, after the methacholine dose that induced a drop in FEV_1 below 80% of their baseline FEV_1, and additionally 10 minutes after the administration of a bronchodilator. These historical de-identified data, including the capnography recordings and spirometry test results were made available for analysis.

3.2 Pediatric Asthma Study

In contrast to the well-controlled experimental conditions afforded by the methacholine challenge study, it is both necessary and useful to study patients who present to an emergency department (ED) in asthmatic distress. Studying acute asthma in the ED setting makes it possible to capture capnography recordings from a wide variety of patients from different demographics and asthma symptom severity. Further, the development of an effort-independent diagnostic aid for determining asthma severity stands to benefit these patients the most, particularly when a patient is very young, in distress, unable to perform spirometry, or otherwise unable to follow instructions.

In this study, pediatric and young-adult patients who were both being treated for, and had a diagnosis of, asthma were asked to participate on presentation the ED prior to treatment. After consenting to participate in the study (or in the case of pediatric patients, after the healthcare proxy such as a parent consents to the patient's participation), the standard-of-care continued as usual with the addition of capnography recording both before and after treatment. The study proceeded as

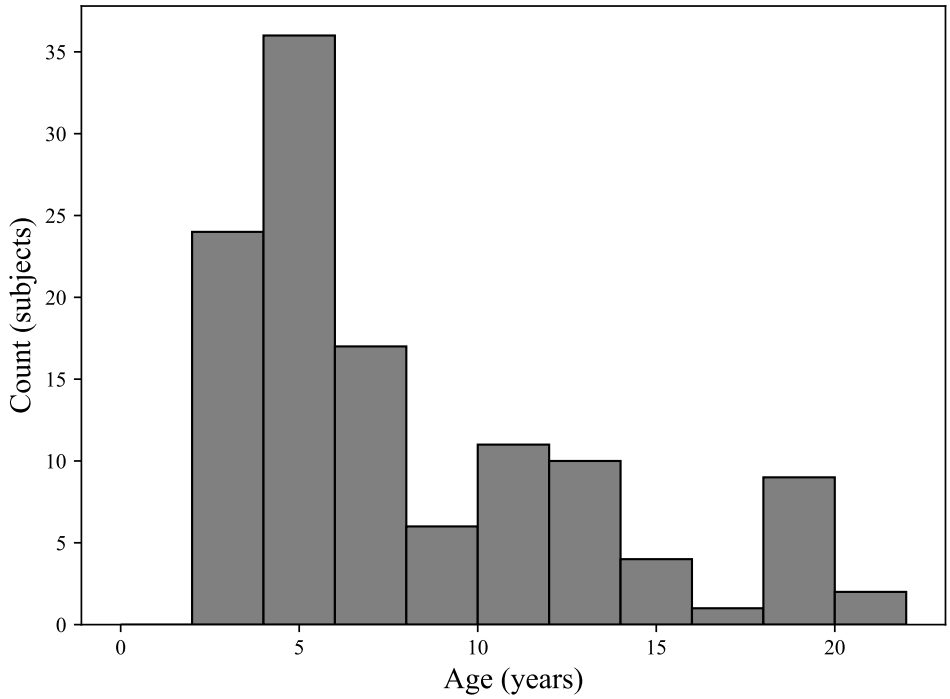

Figure 3-2: The distribution of ages for all subjects with complete datasets in the pediatric asthma study.

follows: subjects were first evaluated in accordance with standard facility procedure. Demographic information, as well as a pre-treatment Hospital Asthma Severity Score (HASS) were recorded in the study database. The subject then gave an approximately three-minute capnography recording using a nasal cannula with mouth scoop to sample exhaled air from both the nose and mouth. Treatment with fast-acting nebulized bronchodilator followed. After treatment, usually 1 to 2 hours later, another capnography recording was taken and a clinician then assessed the subject's post-treatment HASS. De-identified capnography recordings and coded patient metadata from this study were made available for analysis. We received these de-identified data from Dr. Baruch Krauss of Boston Children's Hospital, who collected these data with IRB approval and informed consent by the patient or legally authorized representative. Data collection took place in three separate phases; Phase 1 took place from December 2015 to May 2016, Phase 2 from October 2017 to February 2018, and Phase 3 from August 2019 to March 2020.

Given the dynamic and unpredictable environment presented by the ED, only approximately 120 subjects of 213 initially-enrolled patients had complete datasets. In this study cohort, the ages of subjects are distributed as shown in Figure 3-2, ranging from 2 to 21 years. 63 of the 120 subjects were male and 57 were female.

HASS: HOSPITAL ASTHMA SEVERITY SCORE TOOL			
Parameter	**1 point**	**2 points**	**3 points**
Pulse Oximetry	> 94% on room air	90-94% on room air	< 90% on room air, or requires supplemental O_2 to maintain O_2 saturation > 94%*
Auscultation	Clear or end expiratory wheezing	Expiratory wheeze	Inspiratory, expiratory wheezing, and/or diminished or no breath sounds
Retractions (Muscle groups include: intercostal, substernal, supraclavicular)	None or 1 muscle group	2 muscle groups	3 muscle groups
Dyspnea	Full sentences	Partial sentences	Single words or grunts
Respiratory Rate	2-5y: <30/min 6-12y: <25/min >12y: <20/min	2-5y: 30-40/min 6-12y: 25-30/min >12y: 20-25/min	2-5y: >40/min 6-12y: >30/min >12y: >25/min
TOTAL SCORE	**Mild (<7)**	**Moderate (7-9)**	**Severe (10-13)**
*assign all patients on continuous nebs 3 points when scoring pulse oximetry			

Figure 3-3: The HASS as implemented in the BCH ED and described in [21].

3.3 The Hospital Asthma Severity Score

The HASS, or Hospital Asthma Severity Score, is a metric used at BCH to assess the overall intensity of asthma symptoms that a patient is experiencing. Five components comprise the score: the oxygen saturation of the blood, auscultation of the airway, the extent to which the patient's muscles are visibly retracting to maintain adequate respiration, apparent dyspnea, and the patient's respiratory rate (3-3). Each component receives a score of 1 (normal or mild severity) to 3 (most severe). These components are summed, and the resulting total of 5 – 15 is used as the value of the HASS, where 5 – 6 is considered "Mild," 7 – 9 is considered "Moderate," and 10 – 15 is considered "Severe." At BCH, Abecassis, *et al.* completed a study that suggests that the HASS correlates with spirometry test results [21].

For pediatric asthma study, the pre-treatment and post-treatment HASS distributions are shown in Figure 3-4.

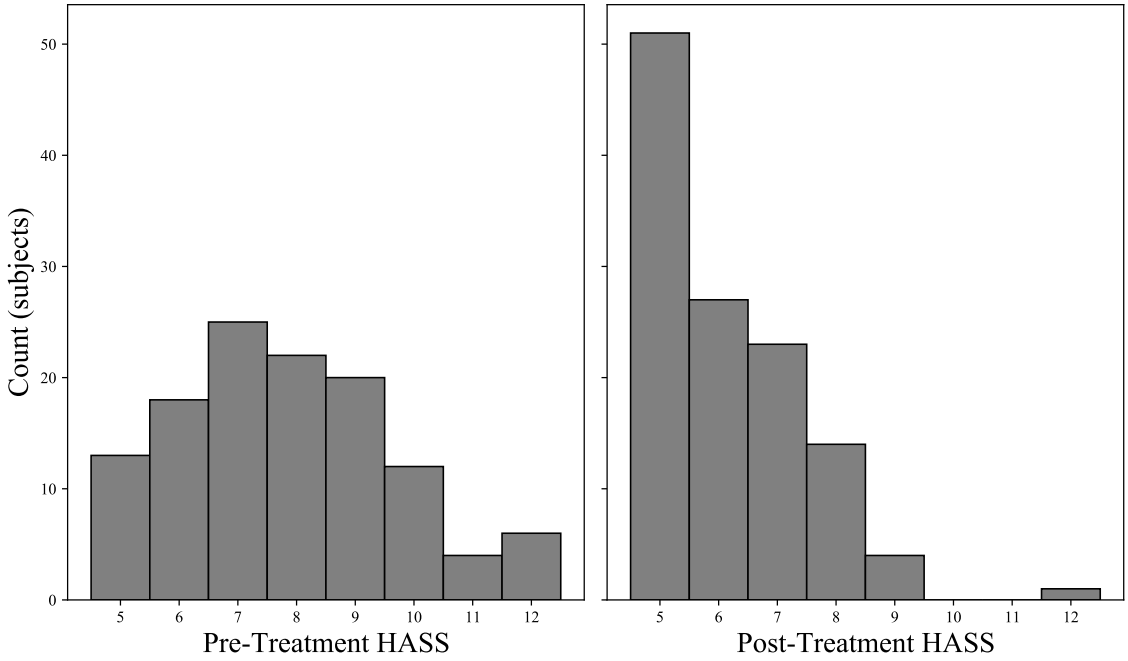

Figure 3-4: The distribution of pre-treatment and post-treatment HASS for all 120 subjects with complete datasets in the pediatric asthma study.

3.4 Data Preprocessing and Record Annotation

As described in Chapter 2.5, capnography recording interpretation involves segmenting the capnogram into individual exhalations, and then quantifying the morphology of those individual exhalations in terms of a choice of features, or by fitting a parameterized expiration model. For both the methacholine challenge study and the pediatric asthma study, data were preprocessed to ensure only high quality exhalation capnograms were considered for algorithm development. For all analysis based on these studies, only subjects for whom we had complete data were considered. This requirement often led to the exclusion of records in the pediatric asthma study, as a complete dataset for a particular subject had four data elements that needed to be collected with patient cooperation and clinician feedback at different times in their

ED stay. In particular, these elements include the pre-treatment and post-treatment HASS and the pre-treatment and post-treatment capnography recordings. Often, the brief time between initial patient triage and the bronchodilator treatment left little time for the pre-treatment recording for some potential subjects. On the other hand, the methacholine challenge study had consistent, good data quality, due in part to the non-emergent study conditions and a more cooperative adult subject population.

Prior work describes the development of automated expiration segmentation from the capnography recordings of the Oridion Capnostream [18, 19]. Due to the extreme variation in record quality, numerous artifacts, and limited number of "good" expirations in the pediatric asthma data, we decided to annotate the records by hand rather than using automated means. To implement a consistent preprocessing procedure, we applied the same annotation process to the methacholine study data and the pediatric study data. For the pediatric study data, we prioritized annotating records that covered the widest range of HASS values, particularly in the extremes. Before deciding to include a study for analysis, and therefore annotation, we visually inspected the all the recordings from the study, and elected to prefer records with longer runs of clean exhalations. Only complete exhalations that were free of noise were annotated.

The annotation process itself was carried out using a purpose-built graphical utility, shown in Figure 3-5. The utility was written in Python 3, and utilizes the Matplotlib plotting library and Qt5 graphical toolkit. For both the methacholine challenge study and the pediatric asthma study, the annotation process proceeded as follows: first, recordings were reviewed and selected, and all recordings for the corresponding subjects were loaded into the annotator. Each of two reviewers ran a local copy of the annotation software that tracked their individual annotation progress. All recordings were presented with the same fixed scale. Each annotator used the keyboard to advance the recording, and used the mouse to add flags directly onto the recording to indicate the beginning of Phase II and end of Phase III of the capnogram. For all subjects included in analysis, two independent reviewers annotated the recordings. For a small subset of recordings, each reviewer annotated each record twice. After

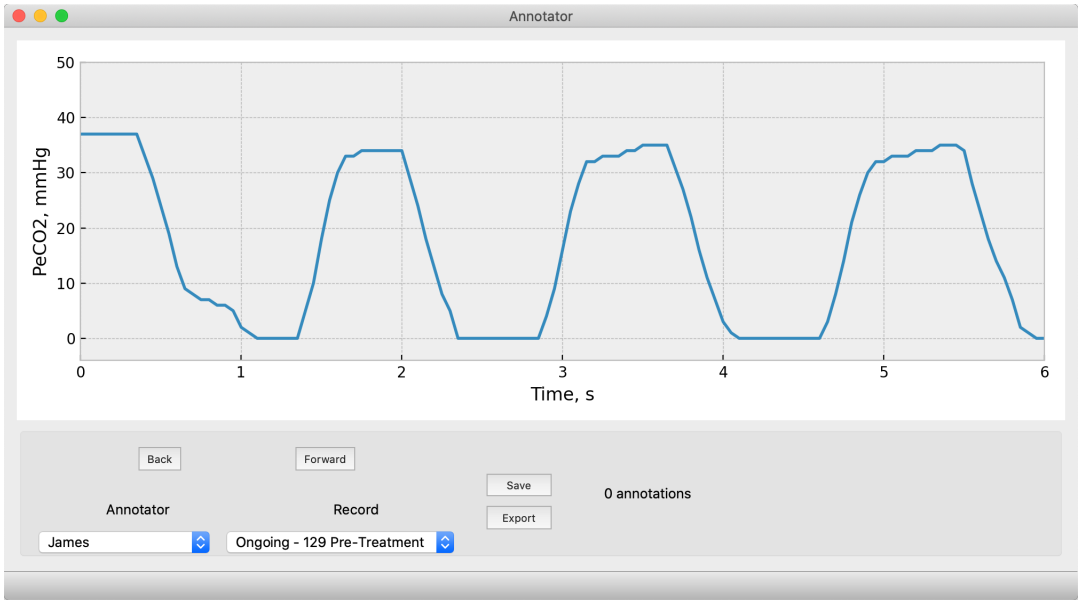

Figure 3-5: The annotator, showing the beginning of a recording from the pediatric asthma study, without annotations entered.

annotation, the exhalation Phase II start and Phase III end labels were compared for inter-rater and intra-rater reliability assessment, and downstream analysis. A record with annotations visible is shown in Figure 3-6, where the beginning of Phase II and the end of Phase III of each valid exhalation are marked with red triangles. The resulting exhalation trace between the endpoints is automatically highlighted in red.

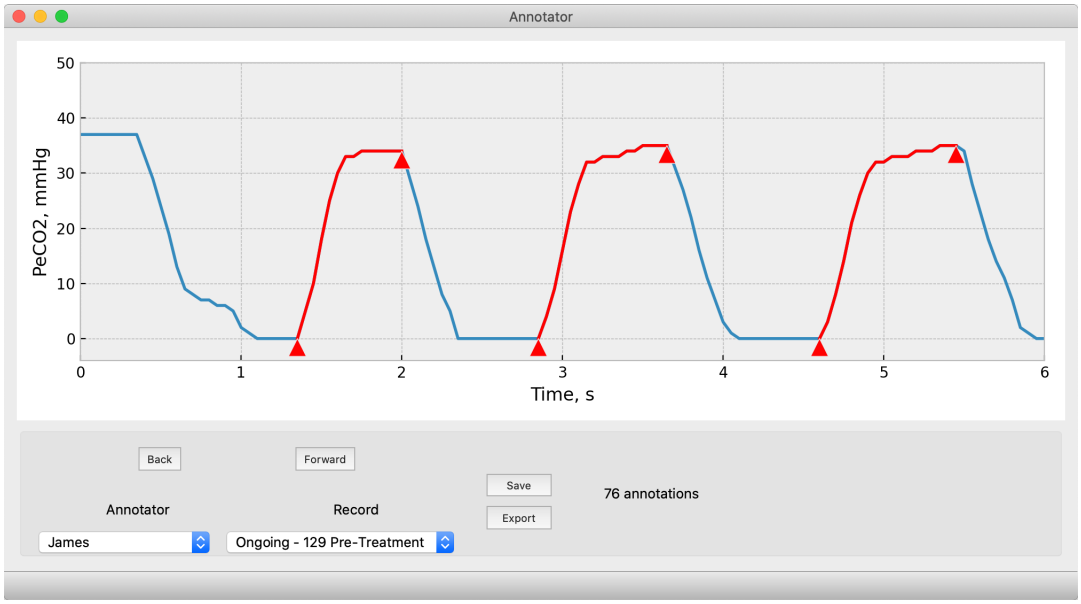

Figure 3-6: The annotator, showing the same record and time range as in Figure 3-5 after annotations have been entered by a reviewer.

Chapter 4

Methacholine Challenge-Induced Asthma Symptom Severity Classification

4.1 Classification Objective

The main objective of this work is to develop a means of assessing the severity of a subject's asthma symptoms based solely on their capnogram. Prior contributions describe successfully distinguishing between different disease processes using capnogram features [18] and provide a framework by which we may investigate more subtle differences between classes of severity of the same disease. Using records from the methacholine challenge dataset, we seek to develop a proof-of-concept for the construction of a classifier that can distinguish between a normal and exacerbated asthma symptom state.

The methacholine study data provide high quality capnogram recordings from a controlled environment in which the subject is precisely administered an agent that induces airway hyper-responsiveness, if the subject has asthma. As described in Section 3.1, in subjects that exhibited no airway hyper-responsiveness to the methacholine challenge (the negative test result), the post-test capnography recordings are

all taken at the maximum concentration of 16 mg/mL. For subjects that did exhibit airway hyper-responsiveness, defined by a 20% or greater drop in FEV_1 (the positive test result), the post-test capnography recording is taken at whatever concentration produced this effect. In most cases for these positive subjects, there is an additional capnography recording taken several minutes after the administration of a bronchodilator that is intended to recover pulmonary function and that is typically a successful asthma treatment.

4.2 Annotations and Record Preprocessing

For all 26 subjects in this dataset, there are 67 recordings – one baseline and one test recording for each of the 26 subjects, plus one post-bronchodilator administration recording for 15 out of 16 positive subjects (where one positive subject has a missing post-bronchodilator recording). In total, these 236 minutes of capnography recordings yield 3200 exhalations that were annotated as artifact-free and complete, with all capnogram phases present as shown in Figure 2-1. Each of these 67 recordings was annotated once by each annotator, and a selection of ten recordings were annotated twice by each annotator as described in Section 3.4. For inclusion into the analysis, an exhalation must have been annotated as valid by both annotators. It is not necessary for both annotators to choose precisely the same Phase II start and Phase III end for a particular exhalation, as the primary purpose of the annotation process was to identify exhalations for inclusion in downstream analysis. If both annotators have marked overlapping annotations in a particular region of a recording, the greatest extent (earliest start time, latest end time) of the interval is taken as the exhalation to capture Phases II and III entirely.

The inter-annotator contingency table across all methacholine study records, where the binary classes are whether to include or exclude an individual exhalation, is shown in Table 4.1. The annotators doubly-annotated ten recordings across the methacholine dataset to provide a measure of the consistency of their inclusion/exclusion of exhalations across multiple trials. The intra-annotator contingency table for each annotator

Table 4.1: Inter-annotator contingency table for selecting valid exhalations across all records from the methacholine challenge study.

	Annotator 2 Included	Annotator 2 Excluded
Annotator 1 Included	3200	72
Annotator 1 Excluded	92	N/A

Table 4.2: Intra-annotator contingency table for the same set of ten methacholine challenge study recordings.

Annotator 1	Round 2 Included	Round 2 Excluded
Round 1 Included	379	31
Round 1 Excluded	2	N/A

Annotator 2	Round 2 Included	Round 2 Excluded
Round 1 Included	395	7
Round 1 Excluded	23	N/A

is presented in Table 4.2. Due to unclear exhalation waveform boundaries caused by artifacts, we do not quantify the total number of potential exhalations. Therefore, the count of potential exhalations that correspond to the contingency in which both annotators, or each annotator across both rounds, did not annotate a region of the recording as an exhalation is not reported, and is indicated by 'N/A' in these tables.

Across all recordings the annotators perform comparably, both including an additional 2%–3% exhalations that the other did not consider valid. For intra-rater performance, the ten recordings were selected from among the more noise-laden records, and this is reflected by the 7%–8% difference in number of exhalations annotated between rounds for both annotators. As expressed in Table 4.2, Annotator 1 included more exhalations in their first round, whereas Annotator 2 included more annotations in their second.

4.3 Features

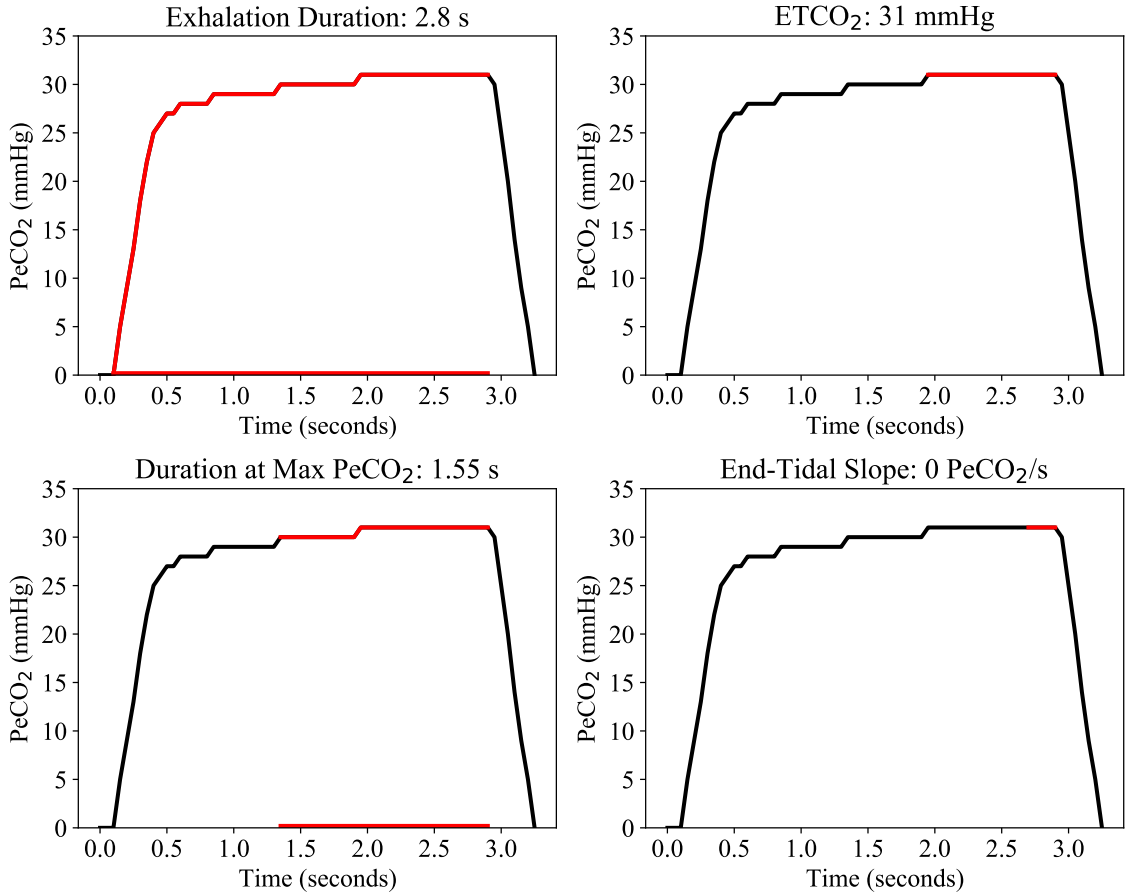

Figure 4-1: Visual representation of the four figures used in this work are shown on a single, prototypical exhalation. The values of the features are given in the titles of the individual subplots, and for time-based features both the extent of the exhalation itself and the corresponding projection onto the horizontal axis is highlighted for clarity.

As a natural starting point, we adopt the features used in [18]: exhalation duration, end-exhalation slope, $ETCO_2$, and the duration of $PeCO_2$ at its maximum value. Exhalation duration is defined as the duration bounded between the beginning of the capnogram's Phase II and the end of its Phase III. The end-exhalation slope is determined by taking the last five points in an exhalation and computing the slope with respect to time using a linear least squares fit. $ETCO_2$ is calculated as the maximum value of the exhalation, which typically occurs at the end of Phase III. Duration of $PeCO_2$ at its maximum value is the duration that the exhalation's $PeCO_2$

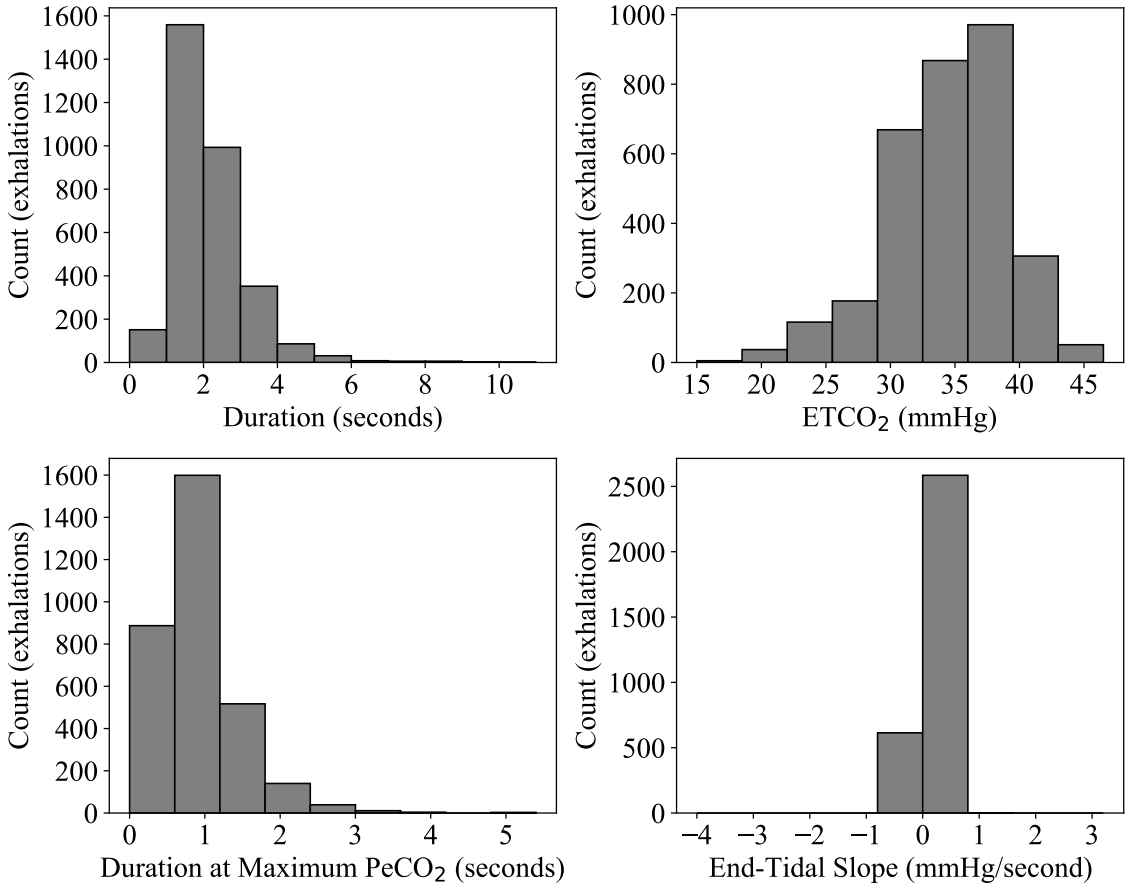

Figure 4-2: The distributions of capnogram feature values across all exhalations in the methacholine dataset.

is at it's greatest value (equivalent to $ETCO_2$ in the absence of artifacts), and ideally represents the duration of the plateau. We tolerate 1 mmHg $PeCO_2$ lower than the maximum value to account for quantization artifacts. These features are shown on a prototypical exhalation in Figure 4-1. The distributions of these features across all recordings in the methacholine study are shown in Figure 4-2.

4.4 Classification Tasks

The methacholine data contain recordings from a variety of different physiological states that support multiple classification tasks. The capnography recordings may be divided into five categories: baseline recordings from subjects that eventually test positive, baseline recordings from subjects that eventually test negative, test recordings (at the patient-dependent maximum concentration of methacholine) from subjects that test positive, test recordings from subjects that test negative, and finally post-bronchodilator administration recordings from subjects that tested positive for airway hyper-responsiveness.

From these data, we implemented three different classification tasks: differentiation of exhalations from the baseline recordings and test recordings in positive subjects, the differentiation of exhalations from all subjects at baseline versus exhalations from positive subjects' test recordings, and finally the differentiation of positive subjects' exhalations at test and those after the administration of a bronchodilator. The tasks are enumerated below in terms of the negative class versus the positive class.

Task 1	Negative:	Baseline (eventually) positive subject exhalations
	Positive:	Positive test exhalations
Task 2	Negative:	All baseline recording exhalations
	Positive:	Positive test exhalations
Task 3	Negative:	Positive subjects' post-bronchodilator exhalations
	Positive:	Positive test exhalations

4.5 Classification Model

To complete these three classification tasks, we implemented a logistic regression. A logistic regression is a statistical model that models binary outcome variables with respect to any number of continuous or discrete input variables. The PDF of the logistic regression assumes a characteristic sigmoidal shape, representing probabilistic regimes near each binary class (close to zero or one), and the boundary between them. The output of a trained logistic regression model is a value representing the probability that a particular vector of input variables belong to the positive class. An arbitrary threshold may then be applied to these probabilities to reduce the continuous probability value to a classification prediction. Different threshold levels trade between higher false positive or false negative rates: a higher threshold will generally result in false positives but will cause more false negatives, and a lower threshold will produce more false positives but fewer false negatives. Therefore, the determination of the appropriate threshold value is dependent on the nature of the classification task itself. For example, if for a particular health condition it is more desirable to catch as many true positive cases as possible to prevent untreated illness, it is better to use a lower testing threshold. In this case, a higher number of false positives may be acceptable if the burden of additional, more accurate testing is low in those patients that receive false positives as a result of this threshold choice.

All of the logistic regressions used in this work were implemented in Python using the Scikit-learn software package [22]. We used the default L-BFGS solver, and included an L_2 regularization penalty for all fits of the model. We selected six subjects (three positive, three negative) to serve as hold-out records to validate the final implementation of each classifier. For each classification task, the remaining recordings from the relevant categories as described above were randomly divided into four folds, as indicated in Tables 4.3 and 4.4. The input features fed to the logistic regression are not normalized in these implementations. Each individual annotated exhalation was assigned a label of 0 or 1, depending on whether its source recording was a member of the positive or negative class for a particular classification task. Four logistic regres-

sions were then trained per classification task, each using a different fold as the test set and the remaining three folds as the training data. Finally, each logistic regression model was evaluated using the test fold data, invoking receiver operating characteristic analysis as the basis for comparison. The results report the classification results of individual exhalations from the training and test folds.

Table 4.3: Record and exhalation counts for each classification task's folds, as well as the hold-out data.

Task 1	Training		Test	
Fold	N Records	N Exhalations	N Records	N Exhalations
1	19	904	7	252
2	19	880	7	276
3	20	835	6	321
4	20	849	6	307
Hold-Out	N/A	N/A	6	364

Task 2	Training		Test	
Fold	N Records	N Exhalations	N Records	N Exhalations
1	24	1152	9	392
2	25	1198	8	346
3	25	1092	8	452
4	25	1190	8	354
Hold-Out	N/A	N/A	9	542

Task 3	Training		Test	
Fold	N Records	N Exhalations	N Records	N Exhalations
1	18	888	7	236
2	19	855	6	269
3	19	803	6	321
4	19	826	6	298
Hold-Out	N/A	N/A	6	294

Table 4.4: Positive/negative class balance for each task fold.

Task 1	Training		Test	
Fold	Negative	Positive	Negative	Positive
1	485	419	111	141
2	424	456	172	104
3	466	369	130	191
4	413	436	183	124
Hold-Out	N/A	N/A	147	217

Task 2	Training		Test	
Fold	Negative	Positive	Negative	Positive
1	733	419	251	141
2	742	456	242	104
3	723	369	261	191
4	754	436	230	124
Hold-Out	N/A	N/A	325	217

Task 3	Training		Test	
Fold	Negative	Positive	Negative	Positive
1	469	419	95	141
2	399	456	165	104
3	434	369	130	191
4	390	436	174	124
Hold-Out	N/A	N/A	77	217

4.6 By-Exhalation Performance

These by-exhalation classification results report the aggregate training and test performance of multiple logistic regression models: four univariate implementations with a single feature each per classification task, and a multivariate implementation using all four features per classification task. For each task, the same fold selection and exhalation class balance were used between univariate and multivariate implementations. The training/test balance of records and exhalations across each classification task and fold is given in Table 4.3, and the positive/negative class balance across each task and fold is given in Table 4.4.

The area under the receiver operating characteristic (AUROC) curve is the principal metric used to describe the performance of the logistic regression classifier throughout this work. The receiver operating characteristic (ROC) curve is graphical representation of the performance of any classifier that may operate with arbitrary decision thresholds. The curve itself indicates the classifier's sensitivity against $1 -$ specificity, that is, the trend of the true positive rate as the false positive rate increases (and, as the decision threshold is lowered) [23]. The AUROC is calculated as the integral of the ROC, and takes a value between 0 and 1, where an area of 1 represents an ideal classifier capable of perfect sensitivity with zero false positives. A perfectly random classifier has an AUROC of 0.5, and the corresponding ROC curve would be rendered as a straight line from (0.0, 0.0) to (1.0, 1.0) in ROC space [23]. In the following results, the average AUROC calculated across all folds is given with the corresponding 95% error bounds for the implementation of each logistic regression model.

4.6.1 Univariate Regression

To determine the predictive value of each individual feature, we first trained the logistic regression classifier using a single feature at a time. The resulting performance of each implementation is given in terms of the AUROC in Table 4.5 along with estimates of the confidence bounds computed through four-fold cross validation. This

table shows the calculated AUROC measures for twelve univariate logistic regression models implemented in the methacholine dataset, one for each feature and task combination. Both the training and test results are given in terms of the average area (95% confidence interval) as calculated across the four training/test folds.

As one would expect, the training performance is marginally higher on average for each task and feature given that the model is directly trained on these data. Overall, the univariate test set performance is rather low across all classification tasks and features with the exception of $ETCO_2$, particularly in Task 3 for which the test AUROC is 0.77 (0.68 – 0.86). This might suggest that across otherwise healthy adults, there is a notable change in $ETCO_2$ during the methacholine challenge.

Table 4.5: Mean AUROC and 95% confidence intervals for each of the univariate classification tasks, implemented by-exhalation and calculated using four fold cross validation.

	Task 1 AUROC (95% CI)	
Feature	Training	Test
Exhalation Duration	0.55 (0.46 – 0.64)	0.25 (0.14 – 0.36)
$ETCO_2$	0.74 (0.69 – 0.79)	0.69 (0.56 – 0.82)
Duration at Max $PeCO_2$	0.72 (0.65 – 0.79)	0.72 (0.56 – 0.88)
End-Tidal Slope	0.61 (0.57 – 0.65)	0.60 (0.52 – 0.68)

	Task 2 AUROC (95% CI)	
Feature	Training	Test
Exhalation Duration	0.55 (0.48 – 0.62)	0.41 (0.21 – 0.61)
$ETCO_2$	0.77 (0.71 – 0.83)	0.76 (0.51 – 1.00)
Duration at Max $PeCO_2$	0.67 (0.61 – 0.73)	0.66 (0.51 – 0.81)
End-Tidal Slope	0.59 (0.56 – 0.61)	0.58 (0.52 – 0.64)

	Task 3 AUROC (95% CI)	
Feature	Training	Test
Exhalation Duration	0.58 (0.50 – 0.66)	0.28 (0.13 – 0.43)
$ETCO_2$	0.80 (0.77 – 0.83)	0.77 (0.68 – 0.86)
Duration at Max $PeCO_2$	0.63 (0.56 – 0.70)	0.42 (0.21 – 0.63)
End-Tidal Slope	0.56 (0.52 – 0.60)	0.56 (0.48 – 0.64)

4.6.2 Multivariate Regression

The multivariate logistic regression model uses all four features to implement a classifier for the three classification tasks. The same training method that was used for the univariate models was used to train the multivariate models; for each task, recordings from the appropriate positive and negative recording categories were divided into four approximately equally-sized folds, the same folds that were used in the univariate regression (Table 4.3). Each fold was used once as a test set, and the remaining three folds were used to train the multivariate logistic regression model. The resulting trained model's performance was evaluated in terms of ROC curve generated by varying the decision threshold. These results report the classification performance of individual exhalations from the training and test folds of each classification task.

Fitting the logistic regression to the training data using all four features, the performance improves to an average AUROC greater than 0.80 across all tasks, as shown in Table 4.6. There is considerable variance in the performance of the second task across the test AUROC of the models. The second task represents a slightly more difficult classification task than Task 1 as the Task's negative class additionally includes exhalations from baseline recordings of subjects that (eventually) tested negative for airway hyper-responsiveness. That said, the mean performance of Tasks 1 and 2 are nearly identical by the AUROC metric.

The ROC curves for Tasks 1, 2, and 3 are rendered in Figures 4-3, 4-4, and 4-5, respectively. These plots are generated by varying the classification cutoff threshold from 0.0 to 1.0. 95% error bars are shown at select points along these curves. As the threshold decreases (from the top right of ROC space), fewer exhalations are included in the positive class, and the number of true positive exhalations and false positive exhalations drop. The red dot in each plot indicates the point along the ROC curve where the sensitivity (the true positive rate) equals the specificity ($1 -$ the false positive rate). This point represents an even balance between the number of true positives, among all exhalations from each classification task's positive class, and the number of true negatives, among all exhalations from each classification task's

negative class. The probability threshold value for this labeled operating point is given in the legend of each ROC curve plot.

Classification Tasks 1 and 3 aim to serve as relatively straightforward tests of the discriminatory capability of a logistic regression-based model to determine whether a subject with asthma is undergoing asthmatic exacerbation. Task 1 represents the forward process, where a subject starts with a baseline "negative" respiration, and experiences a specific drop in pulmonary performance to a "positive" respiration. Task 3 represents the reverse process, where a subject starts with the same "positive" respiration and is brought back towards their baseline (but may not completely recovery their baseline pulmonary performance by the time the post-bronchodilator capnography recording is taken). These are the specific transitions in severity that are the subject of our investigation, in which it is known or highly suspected that the patient has asthma, and the aim of testing is to determine whether the patient is experiencing extreme distress or mild respiratory discomfort in order for clinicians to triage them appropriately.

Table 4.6: Mean AUROC and 95% confidence intervals for each of the multivariate classification tasks, implemented by-exhalation and calculated using four fold cross validation.

	AUROC (95% CI)	
Classification Task	Training	Test
Task 1	0.87 (0.83 – 0.91)	0.82 (0.71 – 0.93)
Task 2	0.88 (0.83 – 0.93)	0.82 (0.59 – 1.00)
Task 3	0.88 (0.85 – 0.91)	0.84 (0.72 – 0.96)

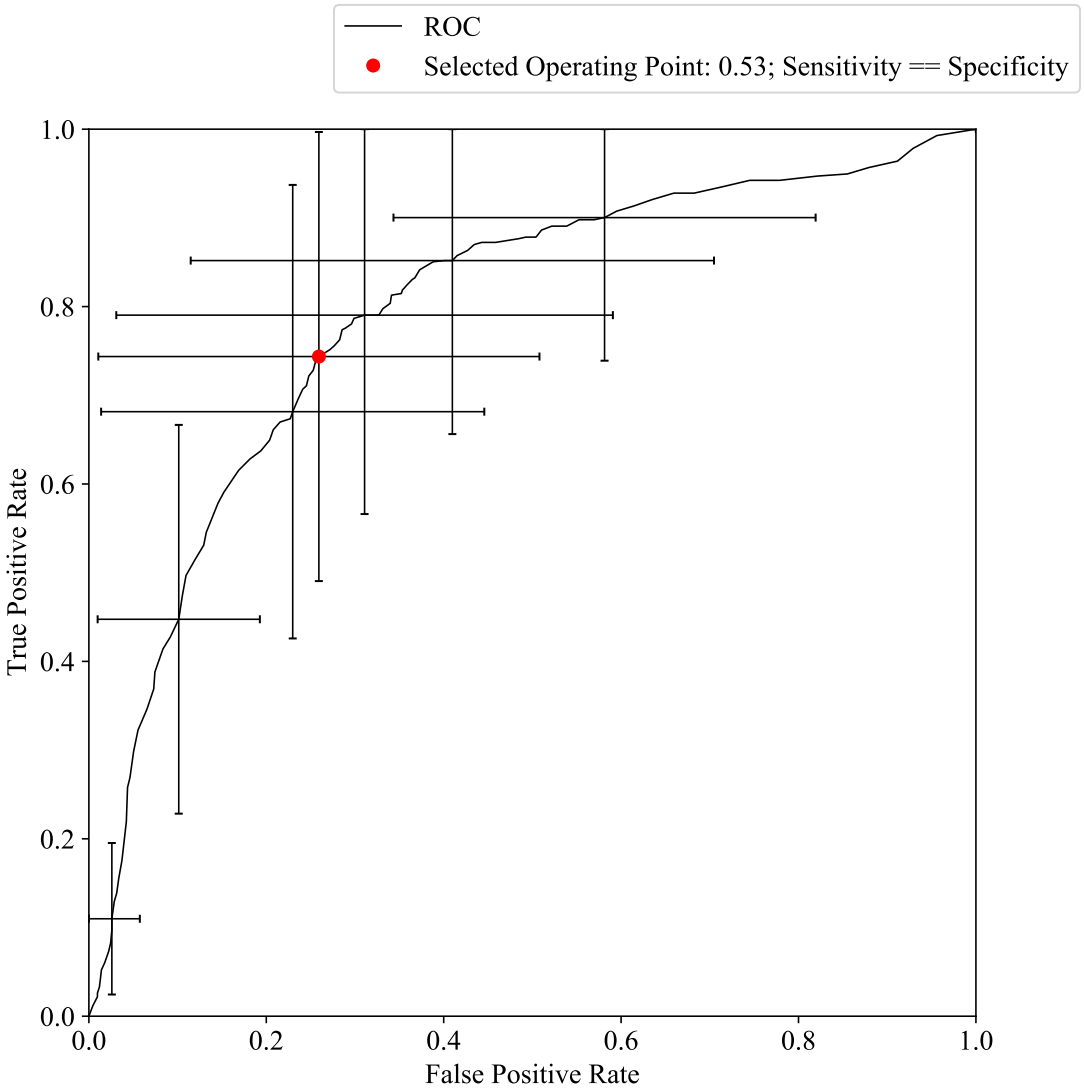

Figure 4-3: Mean ROC curves and selected 95% confidence intervals for the Task 1 multivariate logistic regression classification results. The selected operating point, at which sensitivity equals specificity, is marked in red. Sensitivity = specificity = 0.74.

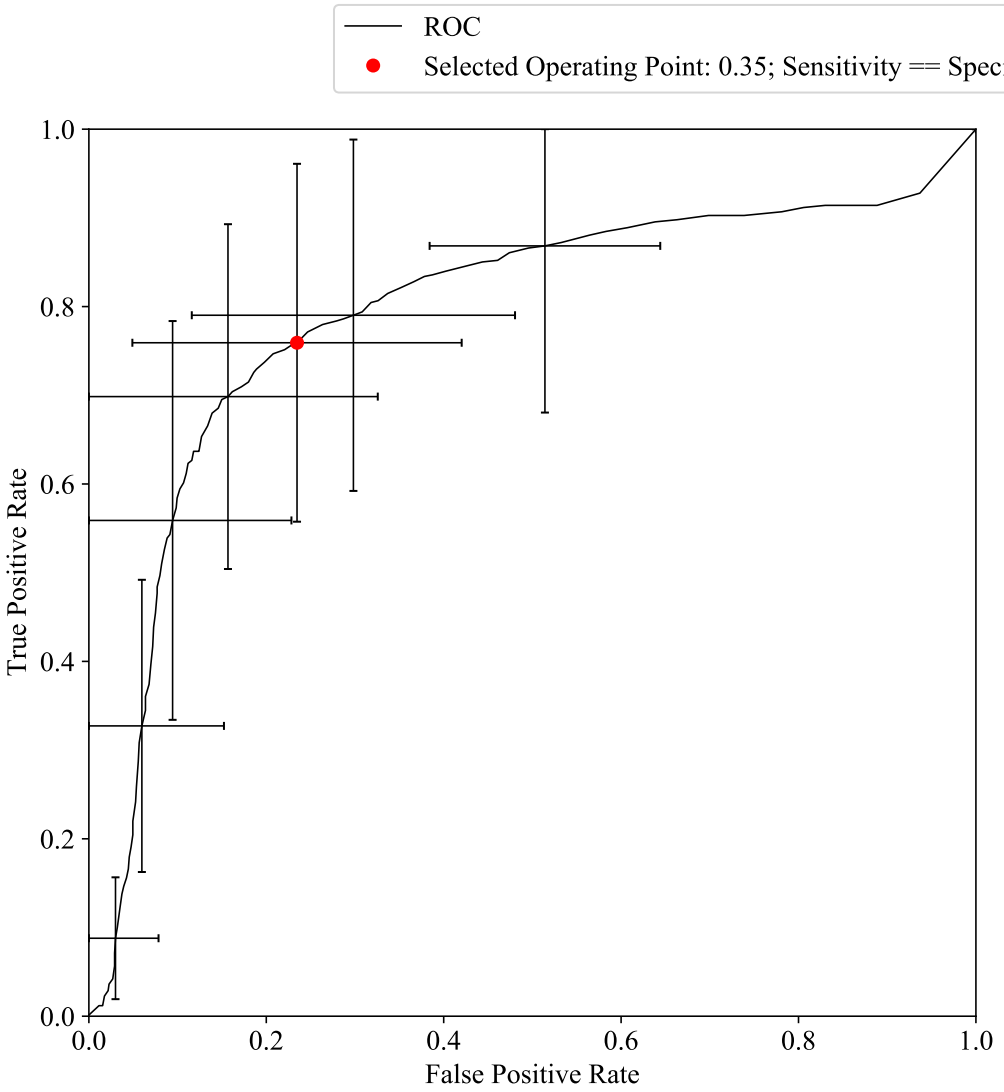

Figure 4-4: Mean ROC curves and selected 95% confidence intervals for the Task 2 multivariate logistic regression classification results. The selected operating point, at which sensitivity equals specificity, is marked in red. Sensitivity = specificity = 0.76.

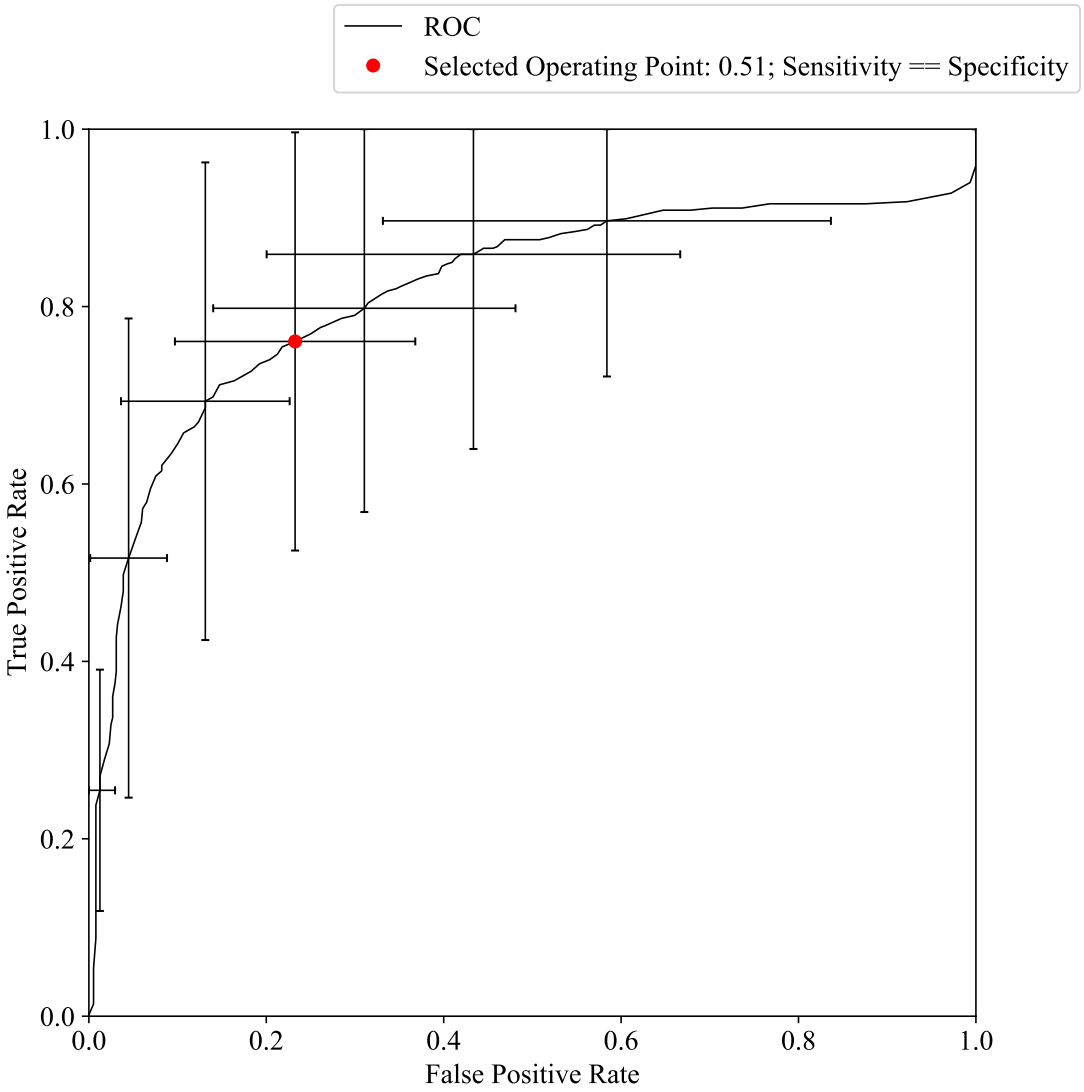

Figure 4-5: Mean ROC curves and selected 95% confidence intervals for the Task 3 multivariate logistic regression classification results. The selected operating point, at which sensitivity equals specificity, is marked in red. Sensitivity = specificity = 0.76.

4.6.3 By-Record Performance

Beyond classifying individual exhalations, a meaningful diagnostic tool would need to make a determination on whether the record of a subject should be considered positive. The exhalations from each Task's test folds were first classified individually using the previously trained four-feature, multivariate logistic regression models described in Section 4.6.2. The optimal probability thresholds, determined by finding the threshold at which the classifier sensitivity is equivalent to its specificity, were then applied to the predicted probabilities of the individual exhalations from the training folds for each Task. The training and test performance of the by-record classification are given in terms of the AUROC for each classification task in Table 4.7, as calculated across the same four folds used previously. The threshold varied along these ROC curves is the fraction of the total number of exhalations in a record that need to be individually classified as positive before an entire record is considered positive.

Table 4.7: Mean AUROC and 95% confidence intervals for each of the multivariate classification tasks, implemented by-record using the operating points described in Section 4.6.2, and calculated using four fold cross validation.

	AUROC (95% CI)	
Classification Task	Training	Test
Task 1	0.91 (0.88 – 0.94)	0.92 (0.84 – 1.00)
Task 2	0.91 (0.87 – 0.95)	0.84 (0.59 – 1.00)
Task 3	0.89 (0.85 – 0.93)	0.93 (0.86 – 1.00)

By-record ROC curves are given for the three classification tasks in Figures 4-6, 4-7, and 4-8. Overall, the performance of logistic regression classifier works well when applied across whole records. Task 2 has marginally weaker performance with a wider variance in terms of classification error across the four folds.

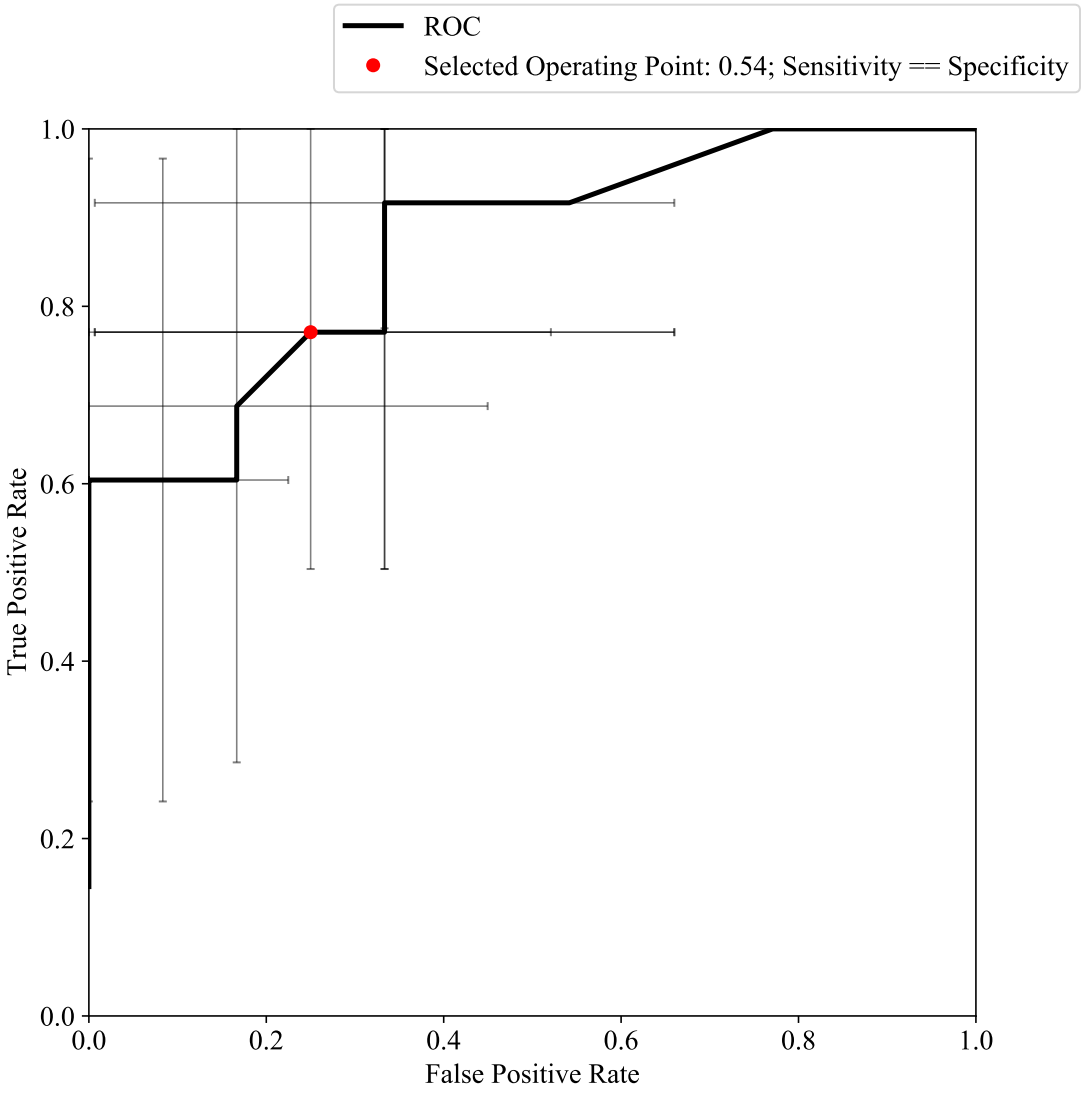

Figure 4-6: Mean ROC curves and selected 95% confidence intervals for the Task 1 by-record multivariate logistic regression classification results. The selected operating point, at which sensitivity equals specificity, is marked in red. Sensitivity = specificity = 0.76.

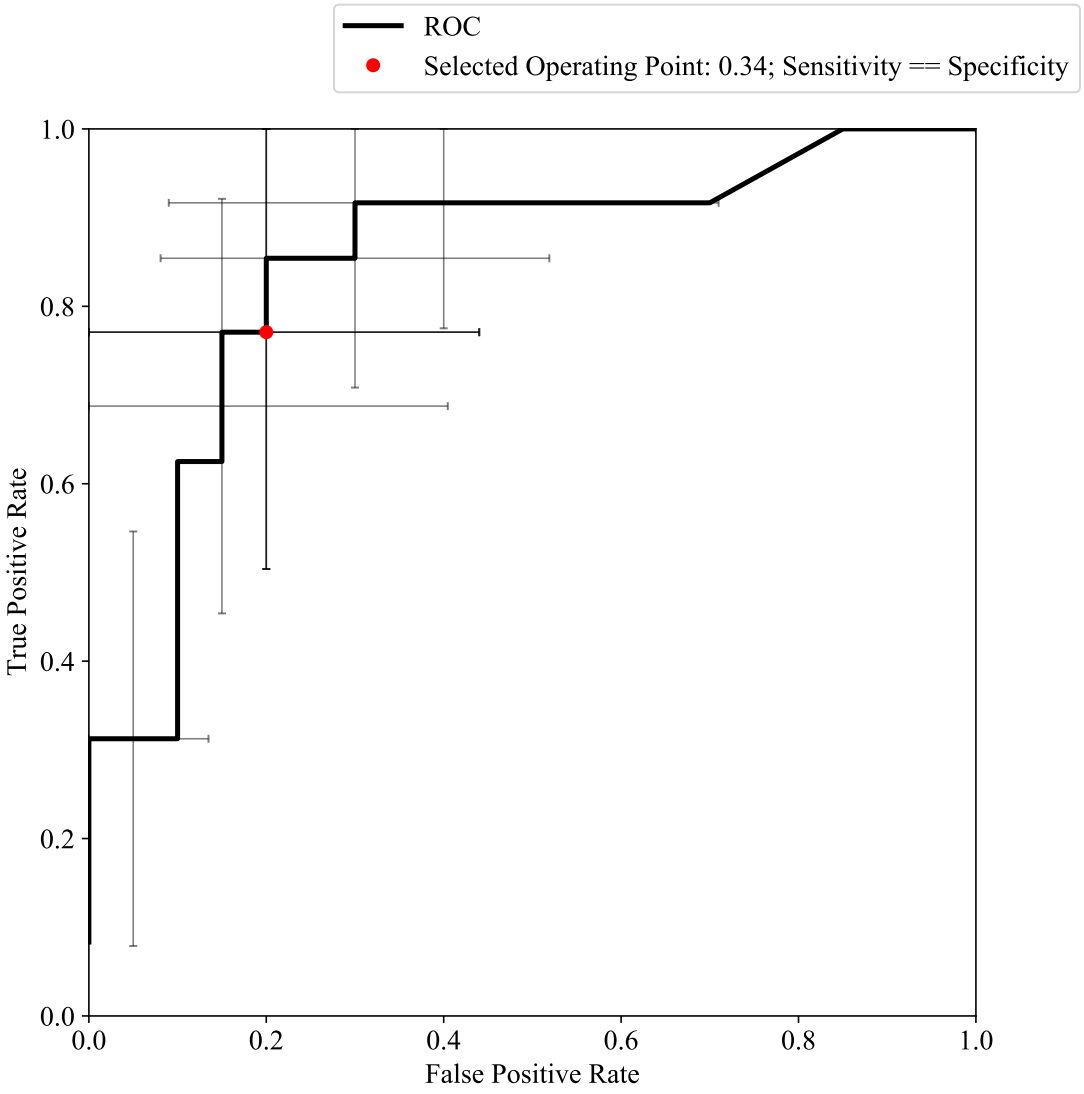

Figure 4-7: Mean ROC curves and selected 95% confidence intervals for the Task 2 by-record multivariate logistic regression classification results. The selected operating point, at which sensitivity equals specificity, is marked in red. Sensitivity = specificity = 0.78.

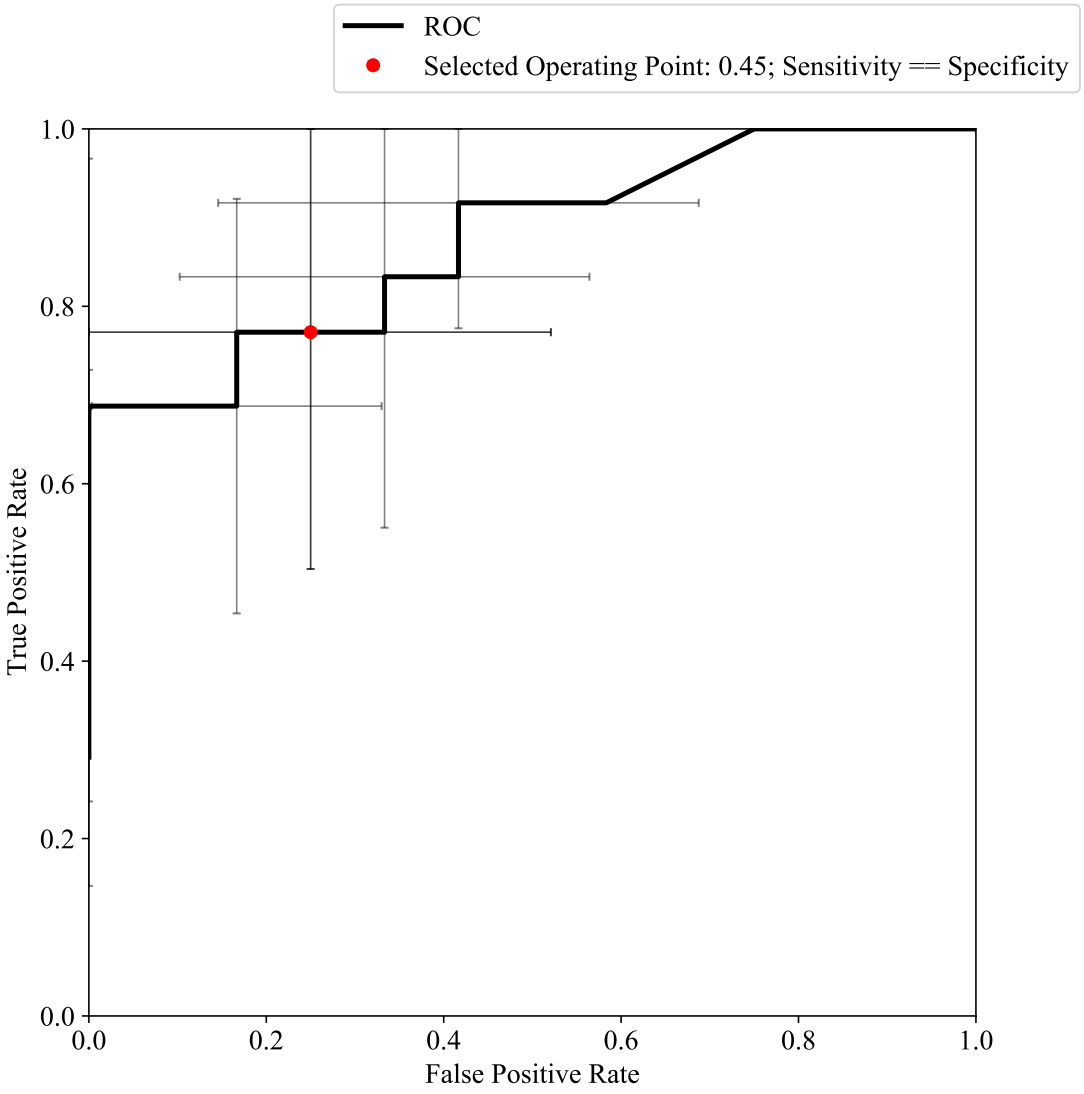

Figure 4-8: Mean ROC curves and selected 95% confidence intervals for the Task 3 by-record multivariate logistic regression classification results. The selected operating point, at which sensitivity equals specificity, is marked in red. Sensitivity = specificity = 0.76.

4.7 Hold-Out Validation

4.7.1 Hold-Out Test Performance

To validate the multivariate results, we use the hold-out dataset as a test set and incorporate all non hold-out records (Folds 1 through 4 from Table 4.3 and Table 4.4) into the training set for the logistic regression model. The hold-out set performs extremely well in Tasks 1 and 2 when the classifier is trained on the remaining exhalation. This might suggest there are records among those used for the training dataset for which the exhalations are misclassified more than those from other records. For a subset of all of the methacholine subjects, there are detailed spirometry reports that show the reduction in pulmonary performance at each methacholine concentration during the challenge. Some subjects approach the 80% reduction in FEV_1 threshold very closely, but do not quite meet the requirements to be considered "positive" for the test. As such, the use of a particular threshold used by the methacholine challenge may cause a greater number of exhalation misclassifications for subjects whose final reduction in FEV_1 falls near the methacholine challenge decision threshold value.

The ROC curves capture the high performance of Task 1 and Task 2, in Figures 4-9 and 4-10, respectively. The relatively poorer performance of Task 3 may be explained by the greater variability in underlying physiological state associated with the post-bronchodilator capnography recordings. Subjects recover to different relative FEV_1 levels after the administration of the bronchodilator, some falling short near around 90% of their baseline value, but all positive subjects have approximately the same value of FEV_1 relative to their baseline when the test capnography recording is taken. Controlling for this variability may improve the performance of the logistic regression model for this particular classification task as-framed.

Table 4.8: AUROC for each of the by-exhalation multivariate classification tasks using the hold-out dataset as the test set, and all remaining records as the training set.

	AUROC	
Classification Task	Training	Test (Hold-Out Set)
Task 1	0.87	0.94
Task 2	0.88	0.95
Task 3	0.88	0.70

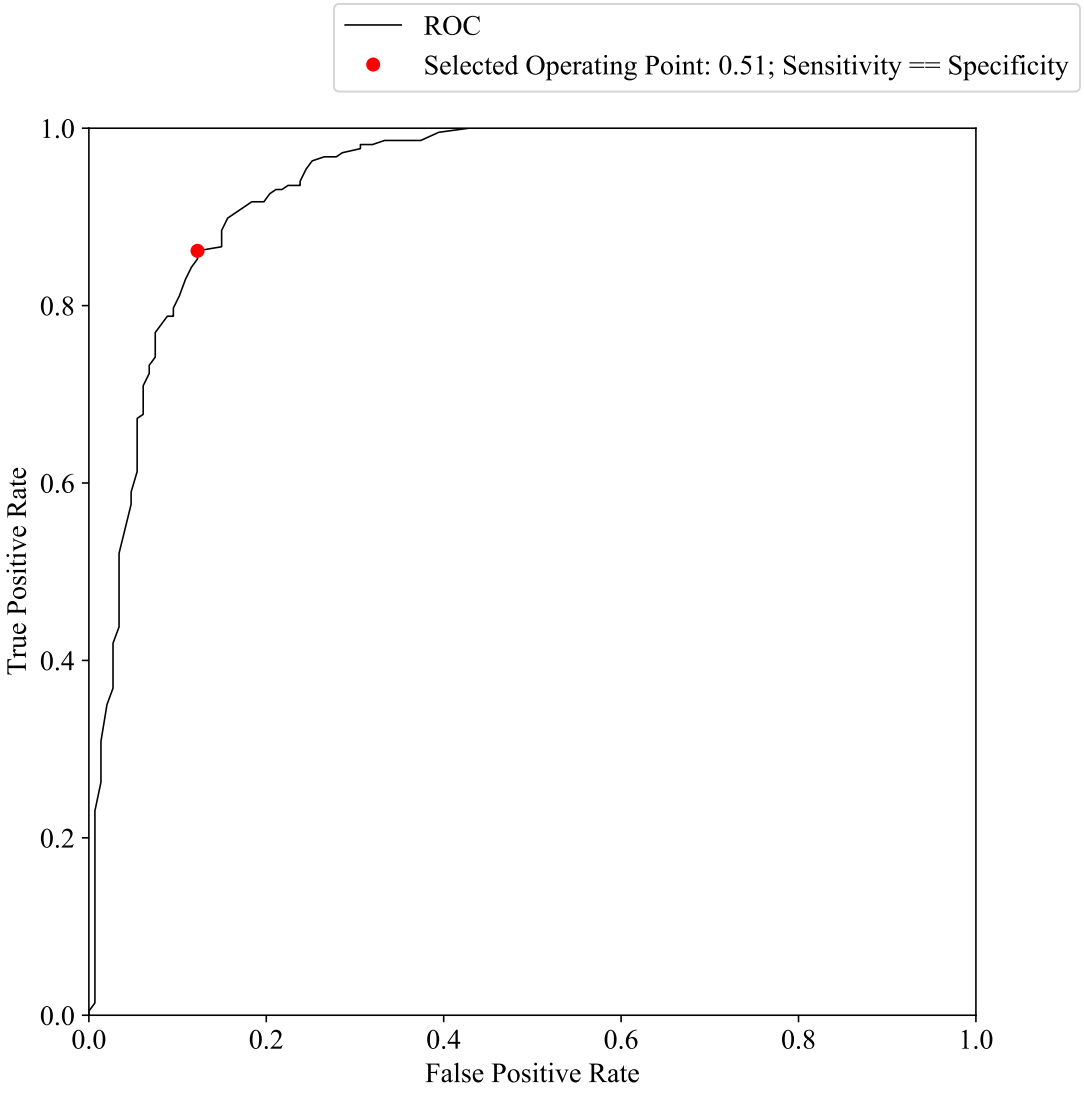

Figure 4-9: ROC curve for the Task 1 hold-out dataset. The selected operating point, at which sensitivity equals specificity, is marked in red. Sensitivity = specificity = 0.87.

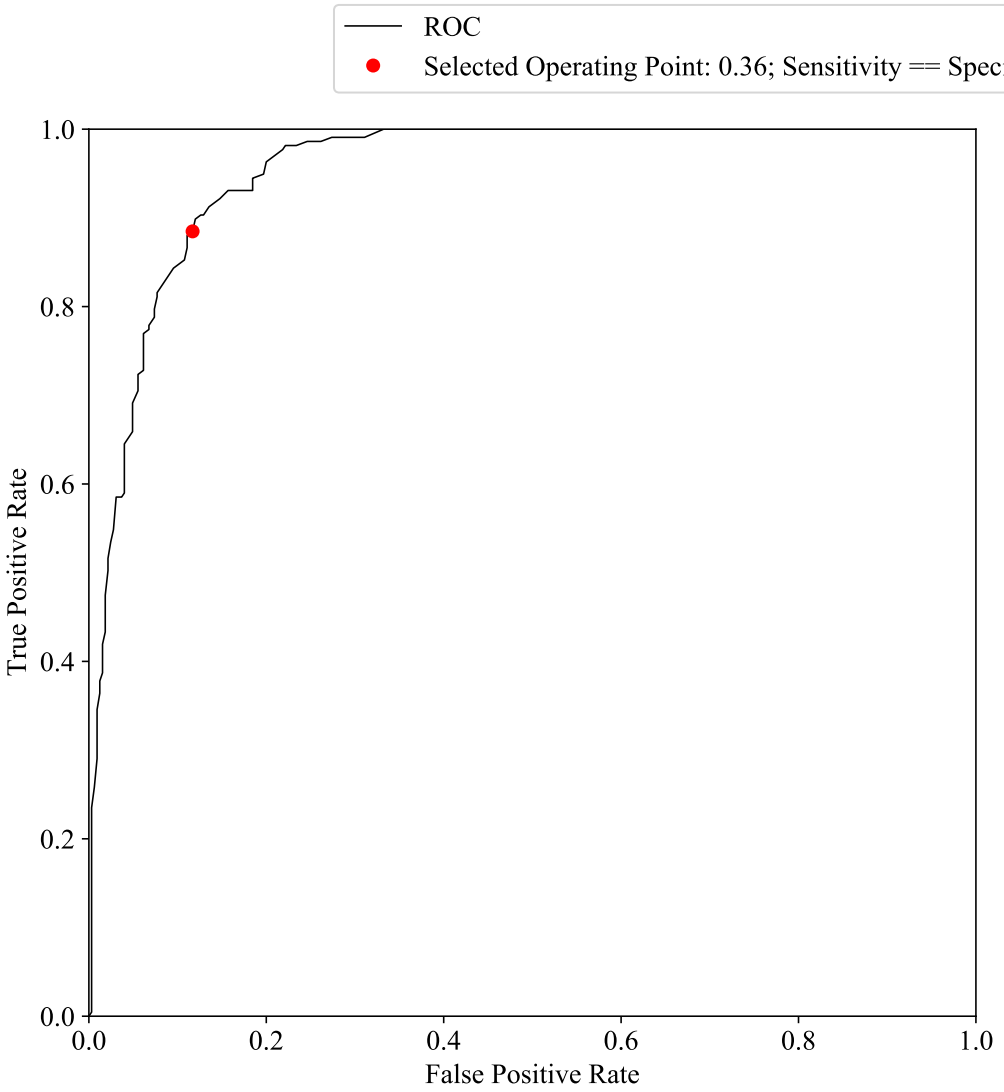

Figure 4-10: ROC curve for the Task 2 hold-out dataset. The selected operating point, at which sensitivity equals specificity, is marked in red. Sensitivity = specificity = 0.88.

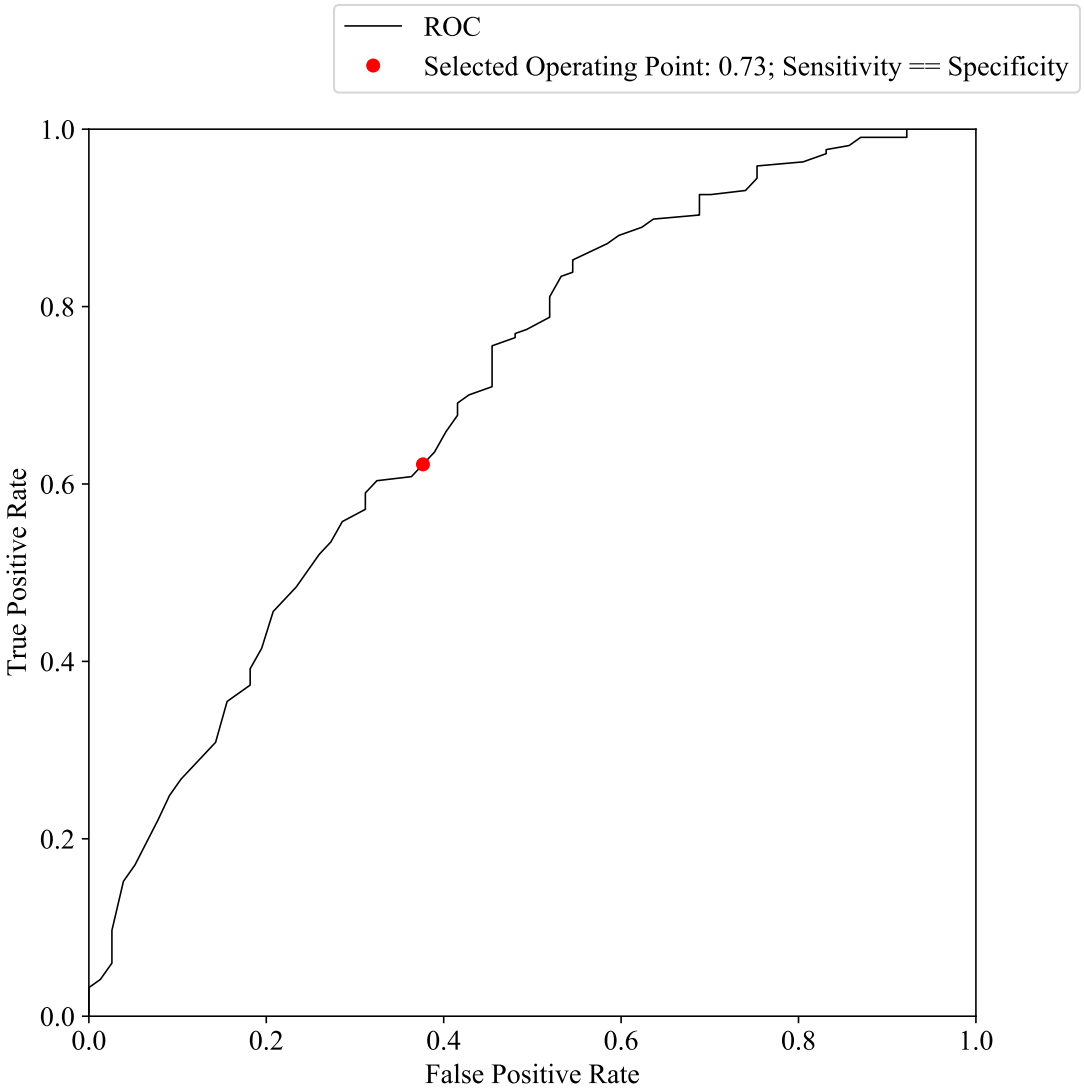

Figure 4-11: ROC curve for the Task 3 hold-out dataset. The selected operating point, at which sensitivity equals specificity, is marked in red. Sensitivity = specificity = 0.62.

4.7.2 Hold-Out By-Record Performance

We also evaluated the by-record performance of the hold-out set. The exhalations from the hold-out records were first classified individually using the trained logistic regression models from Section 4.7.1, for which all non hold-out records were considered as training data. The optimal probability thresholds, shown in the ROC plots above, were then applied to the predicted probabilities of the individual exhalations from the hold-out records. We used the optimal fraction of positive exhalation thresholds as calculated in Section 4.6.3 to determine the predicted class of each of the hold-out records for each Task: 0.51, 0.36, and 0.73, respectively.

For Task 1, all three (eventually) positive baseline recordings and all three positive test recordings are classified correctly. For Task 2, all three test positive recordings are properly classified, but only five of six (eventually) positive or (eventually) negative baseline recordings are correctly classified (that is, there is one false positive). For Task 3, only two out of three positive test recordings are correctly classified (there is one false negative), but all three post-bronchodilator administration recordings are properly classified.

4.8 Summary

In this Chapter, we investigated three classification tasks based on the methacholine pulmonary function dataset. Starting with a by-exhalation logisitic regression classifier model, we build up to classifying entire subject's capnography recordings across these three classification tasks, incorporating decision thresholds determined using ROC analysis. With these clean, largely artifact-free recordings taken in ideal laboratory conditions, the performance of the logistic regression classification models perform well in both the by-exhalation and by-record implementations.

Task 3 did suffer a drop in by-exhalation classification performance, suggesting there may be subtle differences in how the reversal of asthma symptoms manifest in the exhalation features versus the forward, methacholine-induced exacerbation of asthma symptoms. For Tasks 1 and 2, however, the test performance in both the

four-fold implmentation and hold-out validation consistently realize greater than an AUROC of 0.80, suggesting good classifier viability.

Chapter 5

Acute ED Asthma Symptom Severity Classification

Following the asthma severity classification implemented on the methacholine challenge study data, we sought to apply the same technique to the pediatric emergency department dataset. As described in Section 3.2, this dataset includes patients that present to the BCH ED due to suspected symptoms of asthmatic exacerbation. Once diagnosis is confirmed, they are typically administered bronchodilator treatment. The triage and diagnosis process involves scoring patients with a HASS to determine overall severity of the presenting symptoms. An automated, patient effort-independent diagnostic aid to provide additional information to clinicians triaging these young patients would help improve this process.

5.1 Classification Objective

The objective of this work is to investigate the viability of identifying capnography recordings, taken before any treatment (the pre-treatment recordings), from subjects that present with a low HASS and from those that present with a high HASS (as determined by a clinician) using a feature-based logistic regression model.

5.2 Subject Population

The overall subject population is described in Section 3.2. Of all 120 subjects with complete datasets, 39 were selected for exhalation annotation by two reviewers for downstream analysis. Subjects were largely selected at random, but we ensured the smaller number of subjects with a high pre-treatment HASS were included (omitting those subjects with recordings that had a large number of artifacts) in order to cover a suitable pre-treatment HASS range. The distribution of these selected subjects' pre-treatment HASS is given in Figure 5-1. The distribution of these subjects' ages is shown in Figure 5-2. All of these subjects underwent treatment for asthma with nebulized bronchodilator after the pre-treatment recordings were taken.

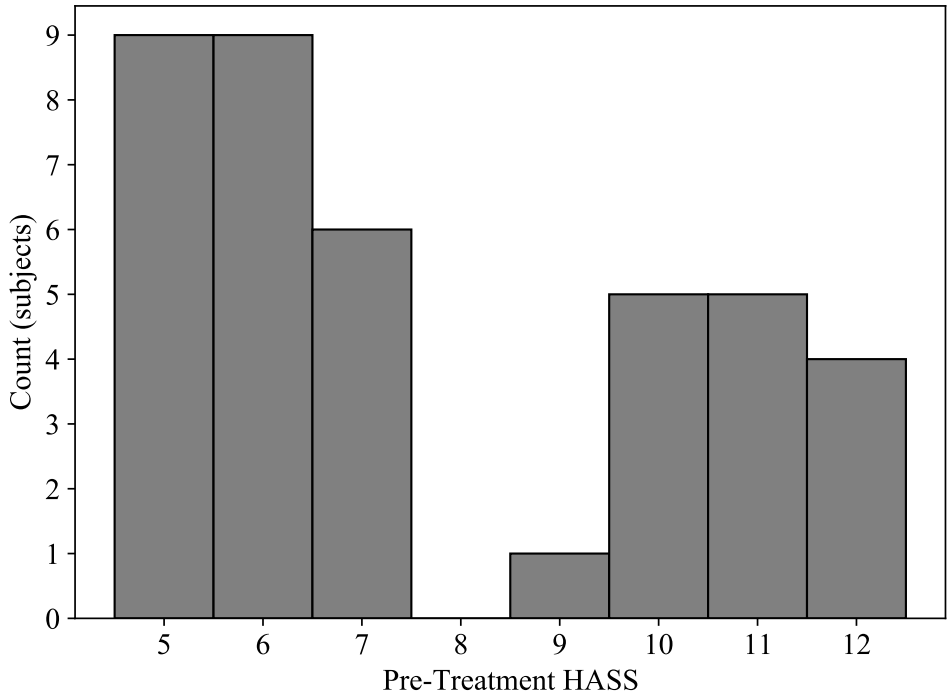

Figure 5-1: The distribution of pre-treatment HASS for subjects in the pediatric asthma dataset that were included in analysis.

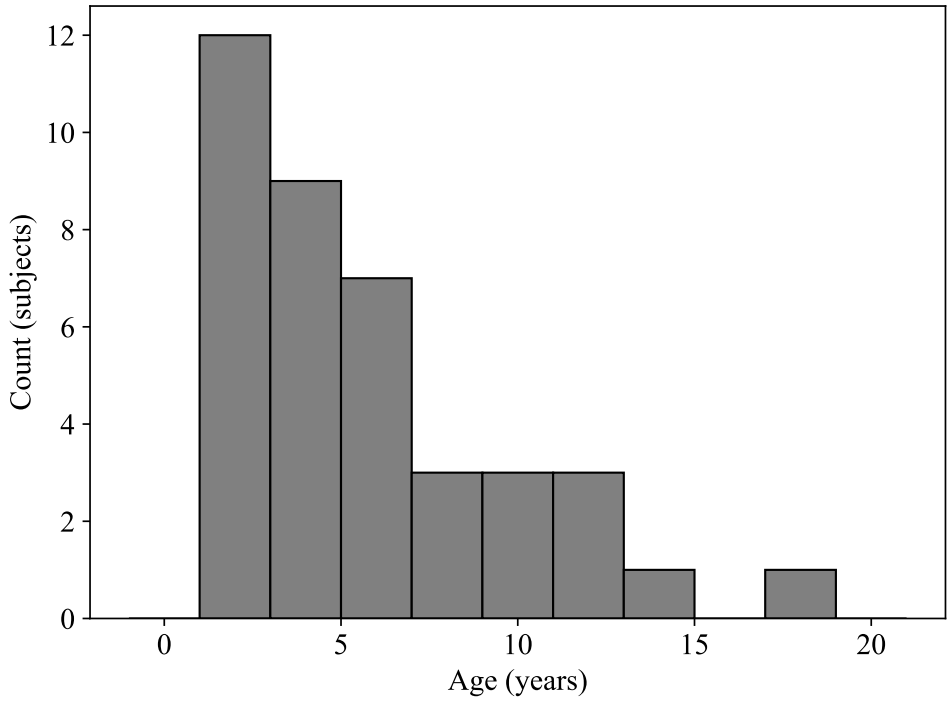

Figure 5-2: The distribution of ages for subjects in the pediatric asthma dataset that were included in analysis.

5.3 Exhalation Annotation

The pre-treatment recordings from the selected 39 subjects were annotated following the procedure described in in a manner identical to the process described in Sections 3.4 and 4.2. Annotating a particular exhalation indicates that it should be included in analysis: that the included segment of the recording includes Phases II and III of the exhalation and that it is uninterrupted by any artifacts.

In addition to annotating all 39 pre-treatment recordings once, the two annotators doubly-annotated five subjects' pre-treatment recordings in order to provide a measure of intra-rater consistency. The inter-rater and intra-rater contingency tables for the exhalation annotations for the two annotators are given in Table 5.1. In total, the two annotators agreed on 2219 out of 2535 total annotated exhalations across 99 minutes of recordings. In order for a particular exhalation to be included in downstream feature extraction and classification analysis, the exhalation needed to be annotated by both annotators. Minor disagreements as to the endpoints of a par-

ticular exhalation on the order of a few sample points did not disqualify the inclusion of a particular exhalation. The raters' agreement on whether to include a particular exhalation was approximately consistent with the methacholine results. Annotator 1 was slightly more consistent between annotation rounds.

Table 5.1: Inter- and intra-annotator contingency tables for the same set of five pre-treatment recordings, corresponding to five subjects from the pediatric asthma study.

	Annotator 2 Included	Annotator 2 Excluded
Annotator 1 Included	2219	74
Annotator 1 Excluded	242	N/A

Annotator 1	Round 2 Included	Round 2 Excluded
Round 1 Included	291	8
Round 1 Excluded	7	N/A

Annotator 2	Round 2 Included	Round 2 Excluded
Round 1 Included	255	4
Round 1 Excluded	28	N/A

The four exhalation features used in classification, described in Section 4.3, were extracted from the annotated exhalations. Distributions of these features across all 39 pre-treatment recordings are given in Figure 5-3. Versus the feature distributions in the methacholine dataset, the two duration features in the pediatric dataset, namely Exhalation Duration (top left) and Duration at Maximum $PeCO_2$ (bottom left), have considerably lower means. This is supported by the demographics of each population; children, particularly those under 12 have higher respiratory rates than adults. Age differences in normal and abnormal respiratory rate ranges are referenced in the HASS rubric, Figure 3-3.

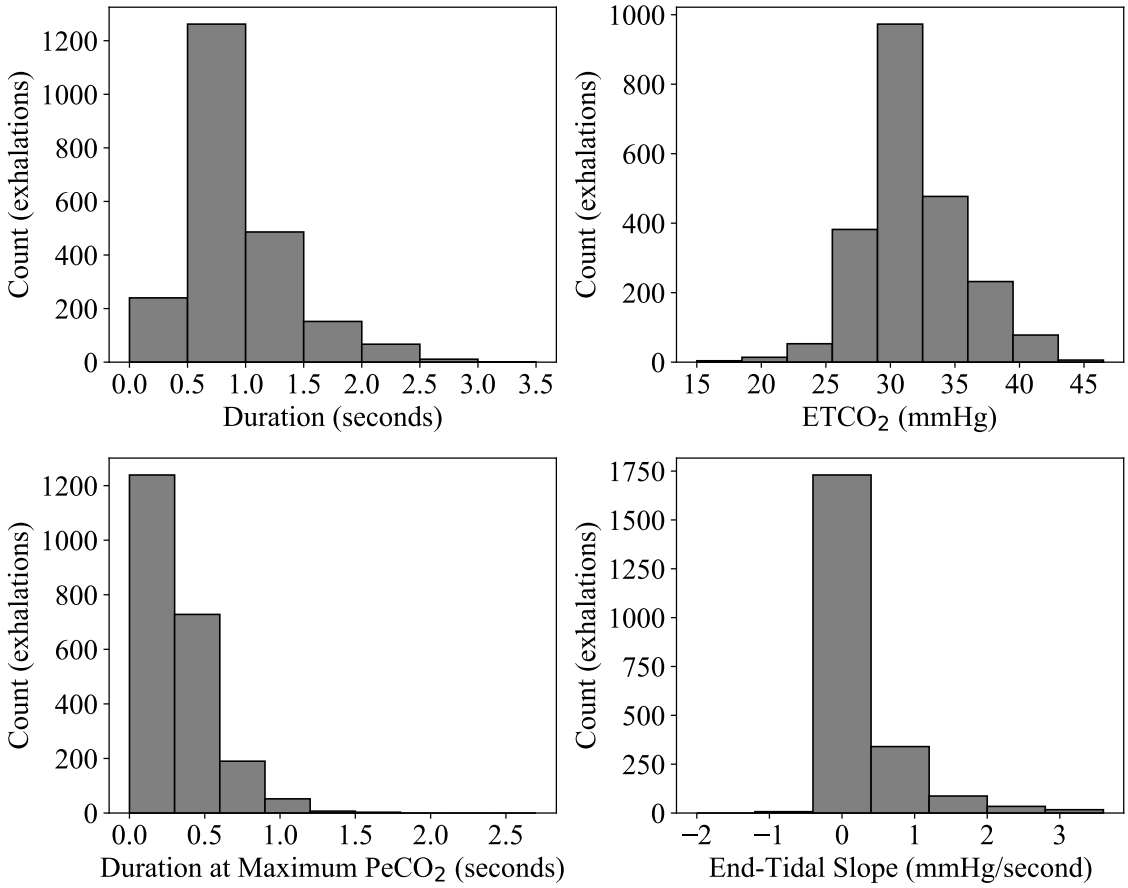

Figure 5-3: The distributions of capnogram feature values across all annotated exhalations in the pre-treatment recordings from selected subjects in the pediatric dataset.

5.4 Classification

A simple classification implementation that identifies the subjects most severely suffering from asthmatic exacerbation is a suitable starting point for the development of a diagnostic aid. We selected a pre-treatment HASS cutoff of 9 above which we consider a subject to have severe asthma symptoms. This is consistent with the Severe HASS category, which constitutes the "positive" class in our classification task. The remaining subjects, falling into the Mild or Moderate HASS category form the "negative" class in our classification task. Under this scheme, 25 subjects fall into the negative class, and 14 fall into the positive, Severe class.

To implement a classifier, we followed an approach similar to that explored in Section 4.5. We first identified four records, two "positive" and two "negative," to hold out of the classifier training and testing (the hold-out set), then divided the remaining 35 remaining pre-treatment recordings (23 negative, 12 positive) into three folds for training and testing. Subjects were randomly divided into these three folds, as given in Tables 5.2 and 5.3. Each subject/pre-treatment recording, and corresponding annotated exhalations, appears in the test set of one fold, and the training set in the remaining two folds. For each individual feature and for the multivariate four-feature implementation we trained three logistic regression models, each using a particular fold as the source of test exhalations and using the remaining two folds as sources of the training exhalations.

Table 5.2: The training and test counts for the subjects/records and corresponding annotated exhalations are given below for the three folds. The folds 1, 2, and 3 refer to the unique test records and test exhalations in a particular row, which appear in the training sets in other two folds. Record and exhalation counts are also given for the hold-out set.

	Training		Test	
Fold	N Records	N Exhalations	N Records	N Exhalations
1	23	1382	12	636
2	23	1271	12	747
3	24	1383	11	635
Hold-Out	N/A	N/A	4	201

Table 5.3: Positive/negative class exhalation balance for each classification fold, and the four hold-out records.

	Training		Test	
Fold	Negative	Positive	Negative	Positive
1	749	633	335	301
2	661	610	423	324
3	758	625	326	309
Hold-Out	N/A	N/A	99	102

5.5 By-Exhalation Performance

After training the logistic regression models, two implementations emerged as having reasonable by-exhalation performance in terms of the AUROC of the resulting classifier. The by-exhalation performance of all four univariate and one multivariate logistic regression models across the three folds is given in Table 5.4. This table reports the average AUROC across all three folds, as well as the 95% confidence interval about the mean value.

The univariate model that uses the Duration at Maximum $PeCO_2$ feature performs comparably to the multivariate model, both performing rather well with mean AUROC greater than 0.80. Plots of the corresponding by-exhalation ROC curves for these selected implementations, the univariate Duration at Maximum $PeCO_2$ logistic regression model, and the multivariate four-feature model are given in Figure 5-4 and Figure 5-5, respectively. These by-exhalation ROC curves were generated by varying the classification probability threshold between 0 and 1, and calculating the mean TPR and FPR across the corresponding three logistic regressions. The chosen by-exhalation operating point to be used for the by-record classification in the following section is indicated in each plot by a red dot. At this probability threshold, the specificity of the model is equivalent to the sensitivity (the true positive rate is equal to 1 − false positive rate).

Table 5.4: Mean AUROC and 95% confidence intervals for the four, single-feature univariate implementations, and the multivariate implementation of the pre-treatment severity classifier, implemented by-exhalation and calculated using three fold cross validation.

	AUROC (95% CI)	
	Training	Test
Exhalation Duration	0.78 (0.71 – 0.85)	0.78 (0.65 – 0.91)
$ETCO_2$	0.66 (0.65 – 0.67)	0.67 (0.65 – 0.69)
Duration at Max $PeCO_2$	0.81 (0.76 – 0.86)	0.82 (0.72 – 0.92)
End-Tidal Slope	0.74 (0.70 – 0.78)	0.74 (0.66 – 0.82)
Multivariate	0.82 (0.76 – 0.88)	0.81 (0.69 – 0.93)

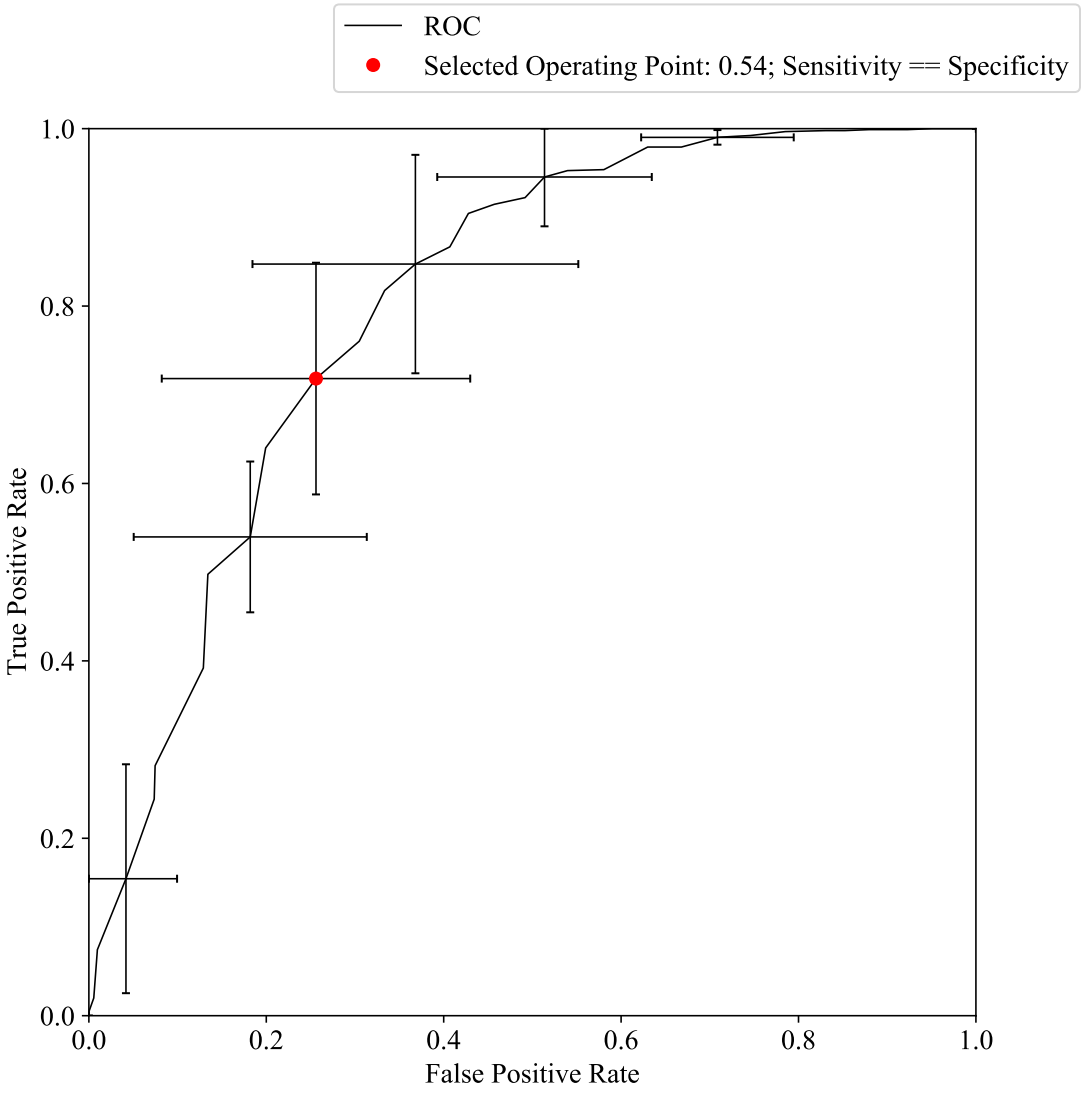

Figure 5-4: The ROC curve and with 95% CI bounds at select thresholds for the Duration at Maximum $PeCO_2$ univariate logistic regression model implementation. The red dot indicates the probability threshold at which the sensitivity equals the specificity for the model, and the value of this threshold is given in the legend. Sensitivity = specificity = 0.74.

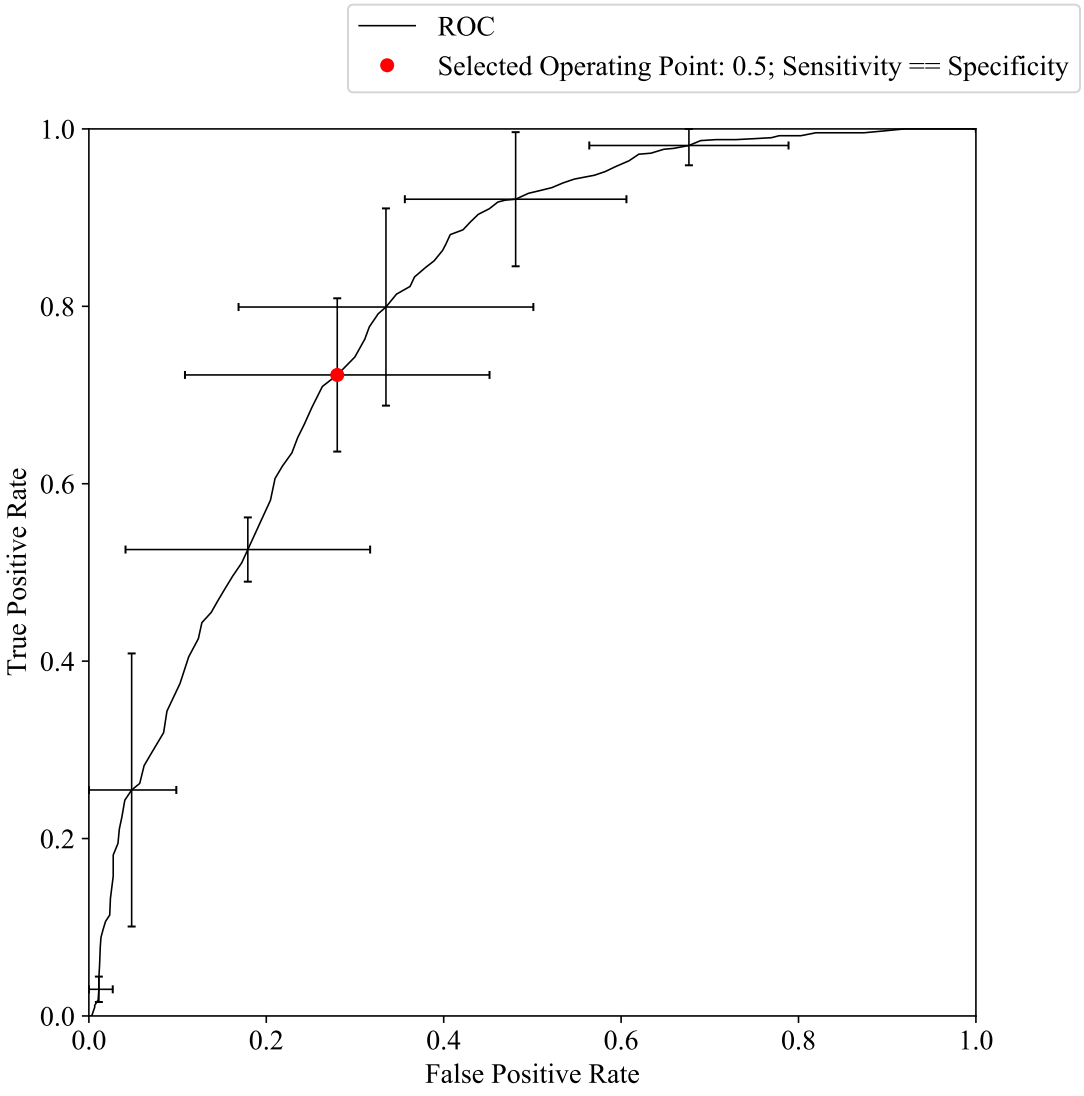

Figure 5-5: The ROC curve and with 95% CI bounds at select thresholds for the four-feature multivariate logistic regression model implementation. The red dot indicates the probability threshold at which the sensitivity equals the specificity for the model, and the value of this threshold is given in the legend. Sensitivity = specificity = 0.72.

5.6 By-Record Performance

Taking the Duration at Maximum $PeCO_2$ univariate model and the four-feature logistic regression model implementations, the two best-performing model implementations in the by-exhalation classification, we built by-record classifiers from the trained logistic regressions. Each test fold's exhalations were classified using the logistic regression corresponding to that fold, and the selected operating point was applied as cutoff. The by-exhalation classification results were then aggregated by source recording, and the fraction of exhalations classified as positive were calculated. Interpreting this by-record fraction as a probability, we calculated the AUROC and 95% CI across all three folds for both the Duration at Maximum $PeCO_2$ univariate model and the four-feature logistic regression model. These measures are reported in Table 5.5. Again, both the multivariate model and the univariate Duration at Maximum $PeCO_2$ model perform comparably. At the operating point, this yields approximately a by-record classification accuracy of 80% for the Duration at Maximum $PeCO_2$ univariate model, and 77% for the four-feature multivariate model.

The ROC curves with 95% CI bars corresponding to the univariate Duration at Max $PeCO_2$ and the multivariate logistic regression classifier are given in Figures 5-6 and 5-7, respectively. The indicated operating point represents the by-recording positive exhalation ratio threshold at which specificity equals sensitivity for by-record classification.

Table 5.5: By-record AUROC and 95% CI of the Duration at Max $PeCO_2$ univariate logistic regression model implementation, and the four-feature multivariate implementation across all three folds, using the specified classification operating point thresholds.

	AUROC (95% CI)	
	Training	Test
Univariate, Duration at Max $PeCO_2$	0.86 (0.79 – 0.93)	0.86 (0.72 – 1.00)
Multivariate, all 4 features	0.87 (0.80 – 0.94)	0.85 (0.73 – 0.97)

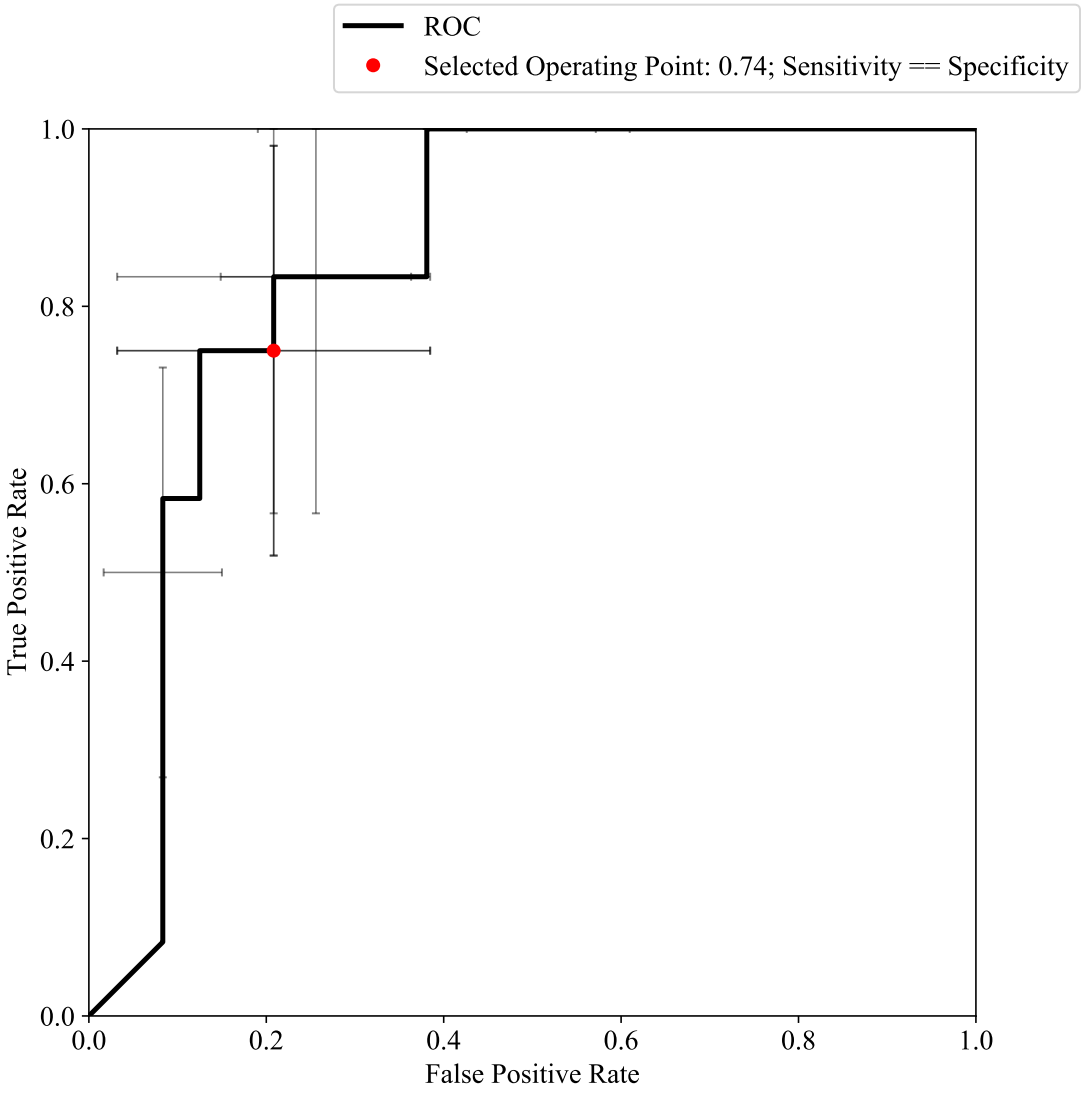

Figure 5-6: By-record performance of the univariate Duration at Maximum $PeCO_2$ logistic regression model. The operating point at which specificity equals sensitivity for the by-record classification task is indicated by the red dot. Sensitivity = specificity = 0.77.

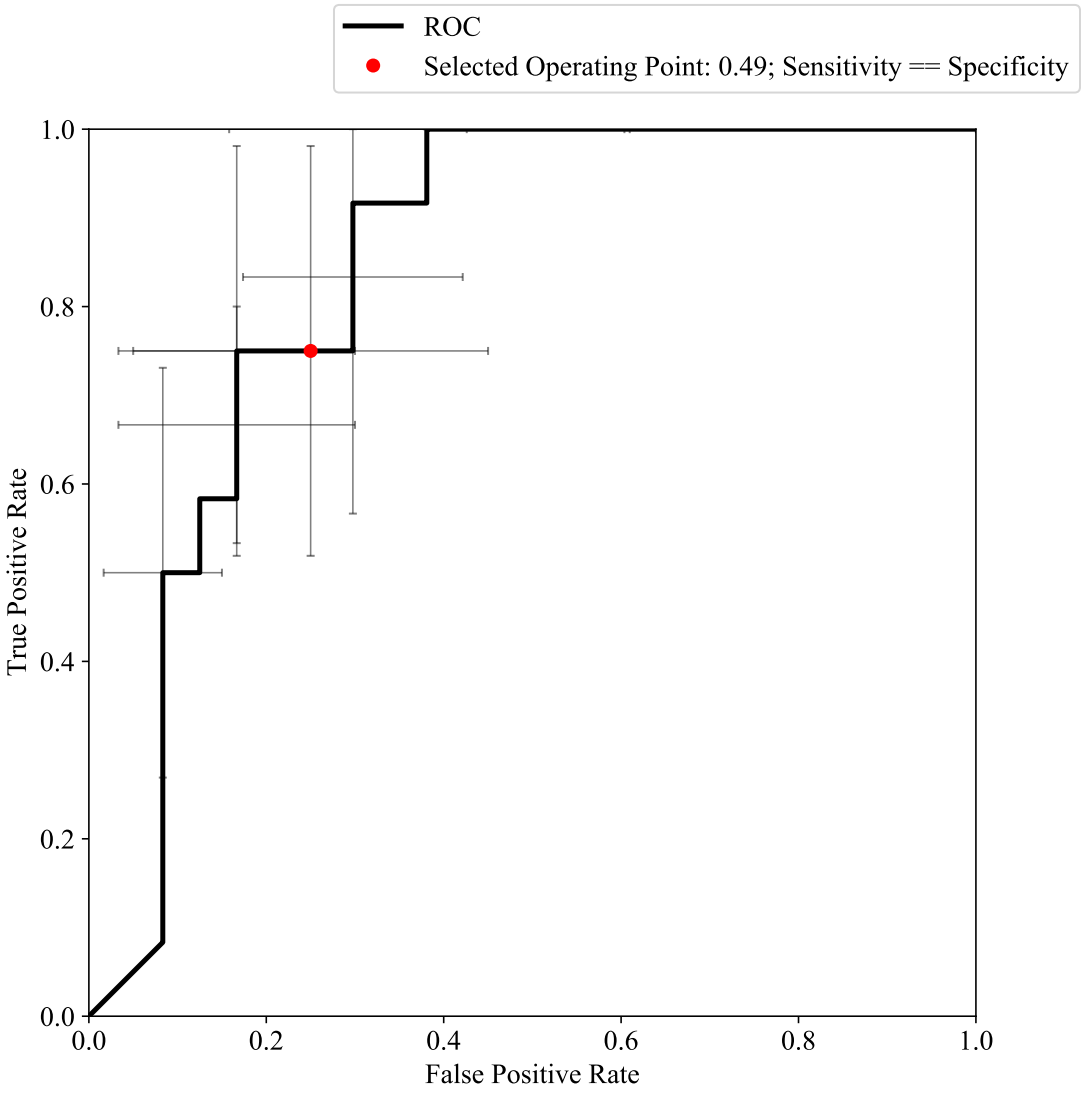

Figure 5-7: By-record performance of the four-feature multivariate logistic regression model. The operating point at which specificity equals sensitivity for the by-record classification task is indicated by the red dot. Sensitivity = specificity = 0.76.

5.7 Hold-Out Validation

5.7.1 Hold-Out Test Performance

To validate the previous results, we trained the two implementations of the logistic regression classifier on all 35 non-hold-out records, and applied the resulting models to the four hold-out records. The by-exhalation results are provided in Table 5.6 for the two models. By exhalation, the two models again performed comparably, both having AUROCs of approximately 0.88. The by-exhalation ROC curves are shown in Figures 5-8 and 5-9. Using the thresholds determined in the previous section, we can apply these models to the hold-out data to evaluate their by-record performance.

Table 5.6: By-exhalation AUROCs of the two logistic regression models, trained on all non hold-out data and applied to the four hold-out recordings.

	AUROC	
	Training	Test (Hold-Out Set)
Univariate, Duration at Max $PeCO_2$	0.81	0.88
Multivariate, all 4 features	0.82	0.88

For the Duration at Maximum $PeCO_2$ univariate logistic regression model, using a positive exhalation fraction of 0.74, and a by-exhalation threshold of 0.54, the model correctly predicts the class of three out of four records (one false negative). In comparison, the four-feature multivariate model using a positive exhalation fraction of 0.49 and a by-exhalation threshold of 0.5, classifies all four hold-out pre-treatment recordings correctly.

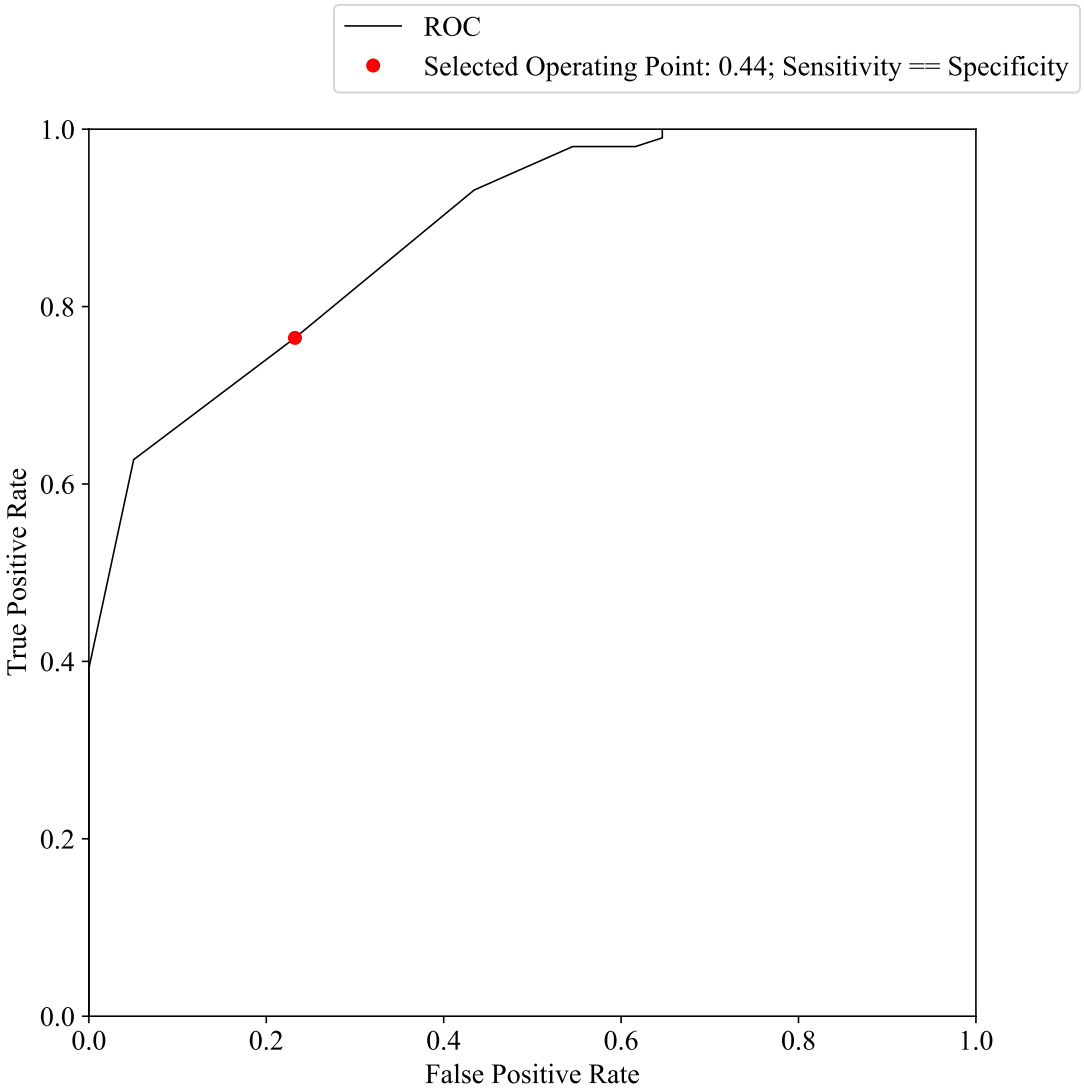

Figure 5-8: By-exhalation ROC curve of the Duration at Maximum $PeCO_2$ univariate model, as trained on all non hold-out records and applied to the four hold-out pre-treatment recordings. Sensitivity = specificity = 0.76.

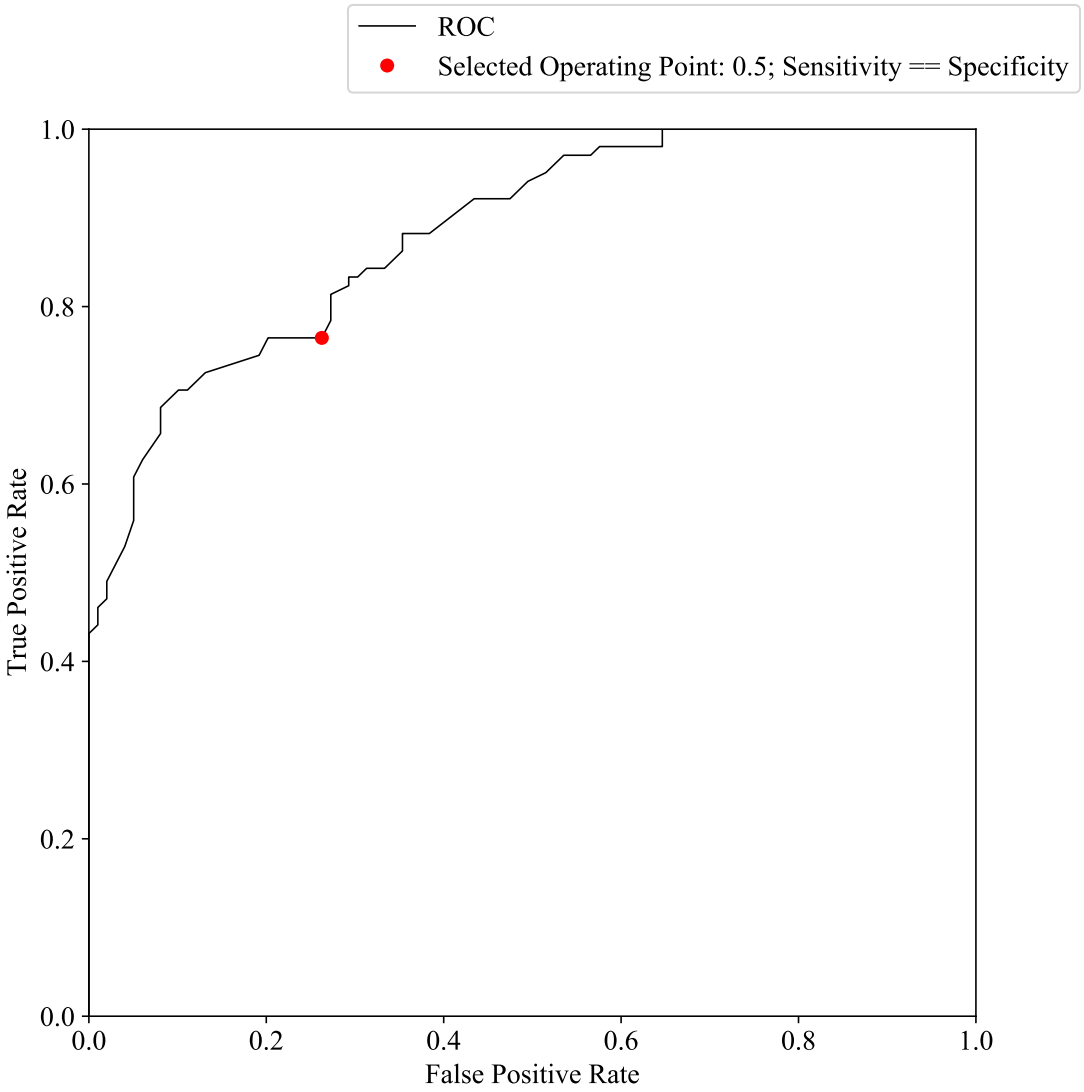

Figure 5-9: By-exhalation ROC curve of the four-feature multivariate model, as trained on all non hold-out records and applied to the four hold-out pre-treatment recordings. Sensitivity = specificity = 0.75.

5.8 Summary

In this Chapter we show the viability of implementing two variations of a feature-based asthma severity classifier. It is capable of distinguishing between the individual exhalations of pediatric subjects with severe asthmatic exacerbation and those with only mild or moderate symptoms. Both implementations perform above 0.80 AUROC, both in terms of the performance on pooled exhalations as well as by-record/subject. This performance is consistent when applied to a randomly selected hold-out selection of records.

The univariate model that uses only the Duration at Maximum $PeCO_2$ is a straightforward extension of the more qualitative, descriptive assessment of the time-based capnogram in patients with asthmatic exacerbation. In severe cases, subjects breathe in short, rapid breaths. In very this rapid breathing, the plateau/Phase III duration visibly shortens on the capnograph readout. The capacity of this individual feature to relate to overall asthma symptom severity as it does in this dataset suggests that it may be used as a simple and intuitive measure of severity.

Due to the fact that many of the subjects in the pediatric asthma dataset are in distress, the overall quality of the recordings is poorer than that of the methacholine dataset. There is a much higher incidence of artifacts in the capnography recordings caused by vocalizations such as talking and crying, as well as movement and partial connection of the nasal cannula. This necessitated the use of manual annotation versus any automated tools, and thus severely limited the number of records available for analysis. As asthma severity increased, the quality of the records decreased, an effect that limited the review of many recordings from subjects with severe asthma exacerbation.

Chapter 6

Summary of Contributions and Future Work

Capnography provides an appealing avenue of research in the development of an effort-independent diagnostic aid for asthma severity due to the physically demanding requirements of current standard diagnostic tools. By using the shape of the capnogram waveform, it is possible to derive information that describes the physiological state of the subject, particularly features such as those explored in this work such as duration, plateau slope and duration, and $ETCO_2$, and interpret this information with the use of a trained classifier.

6.1 Classification Results

We lay the groundwork for developing such a diagnostic aid through the implementation of several variations of a simple, logistic regression-based classification model. This model is capable of distinguishing between subjects with no, or only mild asthma symptoms, and those with severe asthmatic exacerbation.

Using data from a pulmonary function laboratory in which subjects are undergoing metered bronchial provocation, it is possible to distinguish whether individual exhalations, as measured by capnography, are produced by a subject under baseline or normal/healthy conditions, and those from a subject experiencing significant asth-

matic exacerbation and bronchoconstriction. Even in a carefully controlled diagnostic environment such as a pulmonary function laboratory, having an effort-independent, passive method of flagging periods of respiration as potentially exhibiting asthmatic exacerbation would be useful to the clinician. Naturally, the utility of such a diagnostic tool is greater in more challenging clinical environments or triage conditions, particularly when the patient cannot endure or cooperate with spirometry testing. With the addition of automatic exhalation segmentation and extraction, the classification methods described in this work could be made real-time and completely autonomous.

While marginally less performant, this logistic regression-based classification model is also able to capture differences in asthma severity in subjects in a pediatric emergency department. This environment poses a number of challenges as patients often present in physical and/or emotional distress, and cooperation with clinicians (and study staff) is often not guaranteed. Even with the increased number of artifacts present in this dataset, manual exhalation annotation and careful feature extraction was able to result in acceptable by-exhalation and by-record classification performance. Indeed, it is the very challenges imposed by data collection in this environment that makes an effort-independent, passive asthma severity monitoring tool an attractive prospect.

6.2 Challenges Presented

The development of the classification methods described in this work presented a number of challenges. Larger-scale data analysis of the pediatric emergency department dataset was impeded due to a large number of artifacts among the capnography recordings. Such artifacts appeared to include vocalizations and crying. In younger children, there were often artifacts present due to motion or attempted removal of the capnograph nasal cannula. Often these artifacts were present in subjects with higher triage asthma symptom severity, and therefore severely limited the amount of data available from this subpopulation. Previously developed automatic exhalation

extraction that was able to work well with adult patients was not able to properly extract the shorter, more frequent and artifact-laden exhalations present in the pediatric dataset. Manual annotation of these recordings were time-intensive and required careful multi-party review to minimize the introduction of bias or error, and restricted the overall feasible number of records available for analysis.

Additional challenges posed by both the methacholine challenge/pulmonary function testing dataset and the pediatric asthma dataset include metadata reconciliation. Both studies took place over the course of multiple calendar years, and as such, data collection methods changed and often useful contextual notes were missing (such as subjects' HASS). Consequently, some otherwise high quality recordings were unusable in analysis. In both clinical environments, the priority is to the patient's evaluation and treatment. Automation of the data collection process for any future data acquisition will help promote consistent data acquisition and the completeness of records without putting undue burden on assisting clinical staff.

6.3 Future Work

The development of a database of capnography recording from different clinical settings will be instrumental to future work in the development of an effort-independent diagnostic tool for asthma. The current datasets, while having useful measures of underlying symptom severity, are limited in the descriptive power of the labels associated with a subject or recording. In particular, spirometry measurements taken near in time to the capnography recordings would provide detailed, objective measures of asthma symptom severity. The HASS has an element of subjectivity in rubric categories such as auscultation and muscle retractions, as these require qualitative assessment of the patient. Evaluating specific, quantitative spirometry performance measures against exhalation features would serve to improve understanding the "ground truth" underpinning the physiological state behind a capnography recording.

Modifications to the logistic regression model may be made to improve the classification performance. The inclusion of record-wide respiratory rate as an additional

feature in by-record classification may improve performance, when accounting for the age-dependence of normal respiratory rate range. Further, implementing systematic up-selection of up to two or three features, rather than strictly implmenting univariate or four-feature multivariate models, may improve performance in some cases.

More sophisticated classification techniques pose a compelling follow-up of this work that would serve to improve and refine classification accuracy. Expanding into the physiological model-based approach such as that described in [19] would provide additional insight into the underlying disease process, and would compliment this feature-based approach.

Bibliography

1. Centers for Disease Control and Prevention. *Most Recent Asthma Data* `https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm` (2020).
2. Centers for Disease Control and Prevention. *Asthma as the Underlying Cause of Death* `https://www.cdc.gov/asthma/asthma_stats/asthma_underlying_death.html` (2021).
3. National Asthma Education and Prevention Program, Third Expert Panel on the Diagnosis and Management of Asthma. *Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma* Clinical Practice Guidelines (2007).
4. Centers for Disease Control and Prevention. *EXHALE: A Technical Package to Control Asthma* `https://www.cdc.gov/asthma/pdfs/EXHALE_technical_package-508.pdf` (2021).
5. The World Health Organization. *Asthma* `https://www.who.int/news-room/q-a-detail/asthma` (2020).
6. Chung, K. F. & Adcock, I. *Asthma. Mechanisms and Protocols* (eds Chung, K. F. & Adcock, I.) (Humana Press, 2000).
7. Thompson, B., Borg, B. & O'Hehir, R. *Interpreting Lung Function Tests: A Step-by-Step Guide* (2014).
8. Graham, B. L. *et al.* Standardization of Spirometry 2019 Update. An Official American Thoracic Society and European Respiratory Society Technical Statement. *American Journal of Respiratory and Critical Care Medicine* **200,** e70–e88 (2019).
9. Aurora, P. *et al.* Quality Control for Spirometry in Preschool Children with and without Lung Disease. *American Journal of Respiratory and Critical Care Medicine* **169,** 1152–1159 (2004).
10. Crenesse, D., Berlioz, M., Bourrier, T. & Albertini, M. Spirometry in children aged 3 to 5 years: Reliability of forced expiratory maneuvers. *Pediatric Pulmonology* **32,** 56–61 (2001).
11. Pezzoli, L. *et al.* Quality of spirometric performance in older people. *Age and Ageing* **32,** 43–46 (2003).
12. Brigham, E. P. & West, N. E. Diagnosis of asthma: diagnostic testing. *International Forum of Allergy & Rhinology* **5,** S27–S30 (2015).

13. Guidelines for Methacholine and Exercise Challenge Testing—1999: This Official Statement of the American Thoracic Society Was Adopted by the ATS Board of Directors, July 1999. *American Journal of Respiratory and Critical Care Medicine* **161,** 309–329 (2000).

14. Smalhout, B. & Kalenda, Z. *An Atlas of Capnography* (1981).

15. Colman, Y. & Krauss, B. Microstream Capnography Technology: A New Approach to an Old Problem. **15,** 403–409 (1999).

16. Jaffe, M. B. in *Monitoring Technologies in Acute Care Environments: A comprehensive Guide to Patient Monitoring Technology* (eds Ehrenfeld, J. M. & Cannesson, M.) 179–191 (Springer, New York, 2014).

17. You, B., Pelsin, R., Duvivier, C., Dang Vu, V. & Grilliat, J. Expiratory capnography in asthma: evaluation of various shape indices. *European Respiratory Journal* **7,** 318–323 (1994).

18. Mieloszyk, R. J. *et al.* Automated Quantitative Analysis of Capnogram Shape for COPD–Normal and COPD–CHF Classification. *IEEE Trans. Biomed. Eng.* **61,** 2882–2890 (2014).

19. Abid, A., Mieloszyk, R. J., Verghese, G. C., Krauss, B. S. & Heldt, T. Model-Based Estimation of Respiratory Parameters from Capnography, With Application to Diagnosing Obstructive Lung Disease. *IEEE Trans. Biomed. Eng.* **64,** 2957–2967 (2017).

20. Gravenstein, J. S., Paulus, D. A., Feldman, J. & McLaughlin, G. Capnography and the bain circuit I: A computer model. *Journal of Clinical Monitoring* **1,** 103–113 (1985).

21. Abecassis, L. *et al.* Validation of the Hospital Asthma Severity Score (HASS) in children ages 2–18 years old. *Journal of Asthma,* Preprint. (2020).

22. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

23. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27,** 861–874 (2006).