Improving Quality Control in Oscillometry: Repeatability, Efficiency, Feasibility and Accuracy

by

Anas Abufardeh

Submitted in partial fulfilment of the requirements
for the degree of Master of Applied Science

at

Dalhousie University
Halifax, Nova Scotia
December 2023

To Mom, Dad, Siblings and Yara

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

RATIONALE: The current oscillometry acceptance criteria for measurement of respiratory impedance (*Zrs*) requires a minimum of three repeated measurements with a coefficient of variation (CV) of ≤ 10% in adults or ≤ 15% in young children at the lowest frequency resistance (R5). However, this acceptability criteria ignores all the other frequencies and the significant reactance (*Xrs*) component of *Zrs*. This thesis assessed novel algorithms that include variability in resistance (*Rrs)* and *Xrs* over a range of frequencies to improve the repeatability, efficiency, feasibility, and accuracy of oscillometry. It also explored if machine learning can be used to predict Chronic Obstructive Pulmonary Disease (COPD) severity using combinations of spirometry and oscillometry measures.

METHODS: This thesis explored different automated weighted combination sums of *Rrs* or *Zrs* CVs across frequencies and sought the first three measurements out of all measurements with a CV ≤15% for young children and ≤10% for adults. Three different data sets were used, each including five to as many as 12 measurements per subject: 1) 550 five years old population representative children in Toronto (CHILD5Y), 2) 110 three to five years old children with wheeze (WESER) and 3) 818 adult clinic subjects with predominantly COPD (West Island Cohort, WIC). The repeatability, efficiency, and feasibility of the proposed Quality Control (QC) algorithm was first optimized using CHILD5Y and validated using WESER and WIC. Physiological variability and artifact distributions from CHILD5Y were also used to generate a computational model, which was employed to assess the accuracy of the proposed algorithm. Machine learning algorithms including Single Decision Tree, Bagged Decision Trees, Support Vector Machines and Gradient Boosting were assessed to predict COPD Assessment Test (CAT) scores based on oscillometry and spirometry inputs.

RESULTS: It was found that using the proposed QC algorithm, *Early*, with an inverse frequency weighted sum of the *Zrs* outperformed current recommended criteria for CV, reducing the CV of the important outcome measures and achieving the best feasibility. Feasibility improved compared to no QC when restricting analysis to the first 5 measurements, from 64%, 61%, and 49% to 85%, 80% and 81%, while using all available measurements with QC improved feasibility to 94%, 91% and 82% with CHILD5Y, WESER and WIC, respectively. *Early* also improved efficiency by reducing the number of required measurements from 5.4±1.7 to 3.7(0.9). Accuracy was maintained when applying the *Early* algorithm, resulting in comparable Root Mean Square Error (RMSE) to no QC and compared to QC method based on two standard deviations. Accuracy was also maintained for the important oscillometry measures R5-19 and AX. It was found that using machine learning with combined spirometry and oscillometry measures outperforms the use of spirometry or oscillometry measures separately.

CONCLUSION: Optimizing an automated CV algorithm based on Zrs across frequencies provided improved repeatability, efficiency and feasibility, while maintaining the measurement accuracy. Additionally, machine-based algorithms using combinations of spirometry and oscillometry measures show potential for patient screening and monitoring.

LIST OF ABBREVIATIONS AND SYMBOLS USED

COPD	Chronic Obstructive Pulmonary Disease
PFT	Pulmonary Function Test
QA	Quality Assurance
QC	Quality Control
SCM	Single-Compartment Model
Rrs	Respiratory System Resistance
Irs	Respiratory System Inertance
Ers	Respiratory System Elastance
P	Pressure
V	Volume
$\dot{V}$	Air Flow
$\ddot{V}$	Air Acceleration
Raw	Airway Resistance
Rt	Tissue Resistance
Crs	Respiratory System Compliance
mMRC	Modified British Medical Research Council questionnaire
CAT	COPD Assessment Test
FEV-1	Forced Expiratory Volume
FVC	Forced Vital Capacity
Xrs	Respiratory System Reactance
Zrs	Respiratory System Impedance
FFT	Fast Fourier Transform
AX	Area Under the Reactance Curve
SD	Standard Deviation
CV	Coefficient of Variation
μ	Mean
SOP	Standard Operating Procedure
ERS	European Respiratory Society
FSIF	Flow Shape Index Filter
STFT	Short Time Fourier Transform
WT	Wavelet Transform
FT	Fourier Transform
PRO	Patient Reported Outcomes
WIC	West Island Cohort
ζ_R	Combinations of Resistance CVs
ζ_Z	Combinations of Impedance CVs
R_{fi}	Frequency Resistance at Frequency i
Z_{fi}	Frequency Impedance at Frequency i
N_f	Number of Frequencies
W_i	Weight at Frequency i
RMS	Root Mean Square
R5	Resistance at 5 Hz
X5	Reactance at 5Hz

Z5	Impedance at 5Hz
RMSE	Root Mean Square Error
Z(1/f)	Weighted Impedance Cost Function
K-S	Kolmogorov-Smirnov
EDF	Empirical Distribution Function
CDF	Cumulative Distribution Function
μ_{OR}	Test mean calculated after outlier rejection by Grubbs test
R_{nor}	Normalized outlier resistance
X_{nor}	Normalized outlier reactance
n_{GO}	Number of Outliers detected by Grubb's test
R_T	True Simulated Resistance
X_T	True Simulated Reactance
R_{vwp}	Normalized Standard Deviation for Resistance
X_{vwp}	Normalized Standard Deviation for Reactance
R_{rn}	Random Noise Factor (between zero and one) for Resistance
X_{rn}	Random Noise Factor (between zero and one) for Reactance
R_m	Modeled Test Resistance
X_m	Modeled Test Reactance
R_o	Simulated Outlier Resistance
X_o	Simulated Outlier Reactance
AX_{vwp}	Within Patient Test Variability for AX
AX_n	Normalized Noise for AX
AX_m	Modeled AX for Patient Test
AX_o	AX Outliers
$R5-19_m$	Modeled R5-19 for Patient Test
$R5-19_o$	R5-19 Outliers
$Z_T(^1/_f)$	True Weighted Impedance Cost Function Value (*Truth*)
$Z_P(^1/_f)$	Predicted Weighted Impedance Cost Function Value
MICD	Minimally Important Clinical Difference
SDT	Single Decision Tree
BDT	Bagged Decision Trees
SVM	Support Vector Machine
GB	Gradient Boosting
ROC	Receiver Operator Characteristic
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
ECE	Electrical and Computer Engineering

GLOSSARY

Measurement	One of three recordings, separated by a brief gab, with a standard duration of 16 seconds.
Test	A minimum of three repeated recordings.
tremoflo	Software used to collect measurements, provided by Thorasys Medical Systems, Montreal, Canada.
CV criteria / CV threshold	Coefficient of Variation (CV) which is the ratio of Standard Deviation (SD) to the mean (μ), to be less than or equal to 15% for young children and 10% for adults.
ζ_R	A cost function that capture the variability between repeated oscillometry measurements using different combinations of resistance CVs at multiple frequencies.
ζ_Z	A cost function that capture the variability between repeated oscillometry measurements using different combinations of impednace CVs at multiple frequencies.
Repeatability	The short-term (within-test) variability, typically expressed as the CV [16].
Reproducibility	The long-term (between tests) variability [16].
Efficiency	The required number of measurements to pass the acceptability criteria, either the CV, ζ_R or ζ_Z to be less than or equal to 15% for children and 10% for adults.
Feasibility	The percentage of subjects that were able to achieve valid data by obtaining a minimum of three repeated measurements with a CV, ζ_R or ζ_Z less than or equal to 15% for children and 10% for adults.
Accuracy	The percentage Root Mean Square Error (%RMSE) between true and predicted cost function *Zrs(1/f),* where true *Zrs(1/f)* is obtained using R_T and X_T and predicted *Zrs(1/f)* is obtained using R_m and X_m either with no QC or after applying spectral QC.
Outlier	A full measurement detected by the Grubb's test when performed on a test.
Quality Control (QC)	Process comprising a series of activities required to meet the quality standards and verify the safety and effectiveness of a product. It aims to identify errors and addresses issues to ensure the production of high-quality products before reaching customers. [1], [2]
Quality Assurance (QA)	Pre-planned process comprising of a series of activities designed to prevent errors and defects in products and ensure compliance with the requirements. [1], [2]

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my thesis supervisor, Dr. Geoffrey Maksym[1], for his persistent guidance, encouragement, and patience. Indeed, your feedback and insights were invaluable to the success of this thesis. I would like to extend my warmest thanks to my external supervisor, Dr. Thomas Schuessler[2], and committee members, Dr. Robert Adamson[1] and Dr. Shahrokh Valaee[3], for their continuous support and insightful feedback. I would also like to thank Dr. Padmaja Subbarao[4], Ruixue Dai[4], Myrtha Reyna-Vargas[4] and Dr.Ronald J. Dandurand[5] for providing the required data to execute this work. To add, I would like to acknowledge and express my sincere appreciation to the two groups of IEEE students that I had the pleasure to co-supervise for their great efforts and contributions towards the machine learning project.

Lastly, I would like to dedicate this work to my family for their unconditional love and support, this would not have been possible without you. Mom and Dad, thank you for giving me the courage and strength to shoot for the stars and chase after my dreams. My siblings and wife also deserve my wholehearted thanks for their continuous support and encouragement. Thank you for always having my back during difficult situations and for pushing me to excel in both my professional and personal life.

[1]School of Biomedical Engineering, Dalhousie University, Halifax, NS, Canada
[2]Thorasys Thoracic Medical Systems Inc., Montreal, Quebec, Canada
[3]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada
[4]Hospital for Sick Children, Toronto, Ontario, Canada
[5]Lakeshore General Hospital, Meakins-Christie Labs, Oscillometry Unit of the Centre for Innovative Medicine of the McGill University Health Centre, and McGill University, Montreal, Quebec, Canada

CHAPTER 1: INTRODUCTION

This chapter provides a broad review of the published literature relevant to this thesis. It begins with a brief overview of the respiratory system and its breathing mechanics. It then provides a general review of asthma and COPD and how they are diagnosed using spirometry; the current gold standard and most commonly performed Pulmonary Function Test (PFT). Finally, this chapter provides a thorough explanation of oscillometry, covering its principles, key measures, advantages and highlighting gaps in the current standards and Quality Control (QC) algorithms.

1.1 Anatomy of the Lungs

The respiratory system consists of the nose, pharynx (throat), larynx (voice box), trachea (windpipe), bronchi and lungs. Structurally, the lungs are a complex organ made up of thousands of tree-like branching airways. These airways start at the trachea and divide into a right and left primary bronchi. The airways continue to divide to form the secondary (lobar) bronchi, then a smaller bronchus called the tertiary (segmental) bronchi, followed by the bronchioles. The bronchioles then branch repeatedly, forming terminal bronchioles that subdivides into microscopic branches called the respiratory bronchioles which extends to alveolar ducts and sacs; where gas exchange takes place [3]–[5].

To Add, the respiratory system can be divided into two major areas depending on its structure and function. Structurally, the respiratory system can be divided into *upper* and *lower* respiratory systems. The *upper respiratory system* includes the nose, nasal cavity and pharynx, whereas the *lower respiratory system* includes the larynx, trachea, bronchi and lungs. Functionally, the respiratory system can be divided into *conducting* and *respiratory* zones. The

conducting zone consists of a series of interconnecting tubes and cavities found within and outside the lungs, including the nose, nasal cavity, pharynx, larynx, trachea, bronchi, and terminal bronchioles. The main function of the *conducting zone* is to filter, warm, and moisten air before it is conducted into the lungs. On the other hand, the *respiratory zone* consists of tubes and tissues, including the respiratory bronchioles, alveolar ducts, alveolar sacs and alveoli located within the lungs that are directly involved in gas exchange [3], [5].

1.2 Breathing Mechanism

When breathing, air flows into and out of the lungs because of alternating pressure differences created by the contraction and relaxation of respiratory muscles. These mechanics are often described using a simple mathematical model referred to as the Single-Compartment Model (SCM). In this model, a pipe is used to represent the resistance of the lungs (*Rrs*) and lung inertance (*Irs*), accounting for the energy required to accelerate air within the airways. On the other hand, a balloon is used to represent the lung elastance (*Ers*), which reflects the tissue and chest wall stiffness [6], [7]. This is illustrated in Equation (1):

$$P = Rrs\dot{V} + ErsV + Irs\ddot{V} \qquad (1)$$

Where P represents the difference in pressure between the airway opening and the alveoli, whereas V represents the volume with its derivatives flow ($\dot{V}$) and acceleration ($\ddot{V}$). *Rrs* represents the total respiratory resistance and is often considered composed of two main components: airway resistance (*Raw*) and tissue resistance (*Rt*) [6]. Raw is the dominant contributor to lung resistance and represents the viscous energy losses as air moves through the

airways. During laminar flow, *Raw* is inversely proportional to the fourth power of the airway diameter. Factors such as airway smooth muscle contraction, airway wall thickening, decreased parenchymal tethering, or excess mucus production can increase *Raw* and the pressure required to maintain the same airflow [6], [8]. In contrast, *Rt* is thought to account for about 20% of the total resistance in healthy individuals and represents the viscous losses or friction caused by the movement of lung tissue and the chest wall during breathing [9]. While *Rt* can change in disease, it usually has a very modest contribution to the changes in lung mechanics compared to *Raw* [6].

Ers is the respiratory elastance or stiffness, the inverse of respiratory compliance (*Crs*), which provides an estimate of how easily the lungs and chest wall expand and contract in response to pressure changes [4], [6]. A decrease in *Ers* indicates reduced elastic recoil of the lungs, making exhalation more difficult. Conversely, an increase in *Ers* can be attributed to increased tissue stiffness or a loss in available volume for breathing. *Irs* corresponds to gas inertance and accounts for the energy required to accelerate the respiratory gases within the system. Since *Ers* and *Rrs* are key outcome measures in oscillometry, they will be discussed in more detail in section 1.4.2.

1.3 Pulmonary Diseases

Pulmonary diseases are commonly classified as restrictive or obstructive diseases. Restrictive lung diseases are characterized by a decrease in lung volume and consequently an increase in *Ers* or stiffness and a decrease in *Crs*. This increase in lung stiffness restricts the lungs from expanding fully, making it difficult for individuals with restrictive lung diseases to fill their lungs with air [10]. Pulmonary fibrosis, interstitial lung disease and sarcoidosis are among the most common restrictive pulmonary diseases. Conversely, obstructive diseases are characterized by airflow limitation due to an excess in mucus production or narrowing of the

airways. As such, individuals with obstructive lung diseases have difficulty exhaling all the air from their lungs, which results in shortness of breath [10]. Examples of obstructive pulmonary diseases include; asthma, COPD and cystic fibrosis. While there are many respiratory diseases, the next two sections will only provide an overview of asthma and COPD as the analyzed data in this thesis comes from a cohort dominated by these two obstructive diseases.

1.3.1 Asthma

Asthma is a chronic inflammatory disease that causes narrowing and inflammation of the bronchial tubes, limiting airflow and causing breathing difficulty. The lung airways contain mucus glands and are surrounded by muscles that are normally relaxed. In asthma, however, these muscles constrict, tighten and become inflamed when exposed to triggering factors such as allergens and pollen, while the mucus glands increase mucus production. These in turn result in narrowing of the airways and difficulty breathing. Some of the common symptoms of asthma include; chest tightness, shortness of breath, coughing and wheezing. [4], [9]

In Canada, asthma is the third-most common chronic disease and the most common reason for children hospitalization. It affects more than 3.8 million Canadians, including 850 thousand children under the age of 14. On average, there are over 300 Canadians that are diagnosed with asthma daily, and an estimate of 250 that tragically die from an asthma attack every year. While asthma cannot be cured, regular follow-ups, close monitoring, proper use of medication and avoiding triggering factors can help control the disease. [11], [12]

1.3.2 COPD

COPD is an inflammatory lung disease characterized by airflow limitations and persistent respiratory symptoms due to alveolar and airway abnormalities. It is typically caused by a

significant and/or long-term exposure to noxious particles or irritating gasses [9]. COPD consists of two major breathing diseases; chronic bronchitis, caused by excessive mucus release in the airways, and emphysema, caused by an increase in alveoli stiffness. Cigarette smoking is the most common risk factor for COPD, as smokers have a higher risk for lung function abnormalities. Other environmental exposures such as air pollution and biomass fuel exposure may also contribute to this disease. Hence, the long-term exposure to such risk factors may cause individuals to exhibit symptoms such as chronic cough, wheezing, difficulty breathing, and/or mucus production [13].

Currently, COPD is the third leading cause of deaths worldwide, and Canada's fourth leading cause of death, resulting in more than 3 million deaths in 2019 alone. Although COPD cannot be cured, *Early* diagnosis, close monitoring and the initiation of proper treatment is essential to reduce symptoms, slow the progression of the disease, and improve quality of life [14]. The severity of COPD is commonly classified using the Modified British Medical Research Council questionnaire (mMRC) and COPD Assessment Test (CAT). The mMRC questionnaire is used to assess the severity of COPD based on breathlessness, in which the patient selects a statement that best describes their breathlessness while performing daily life activities on a scale of zero to five, Figure 1[15]. On the contrary, the CAT test is an eight questions assessment with a total score of 40, and is designed to assess the impact of COPD on patients' health status and life, Figure 2 [13] [13].

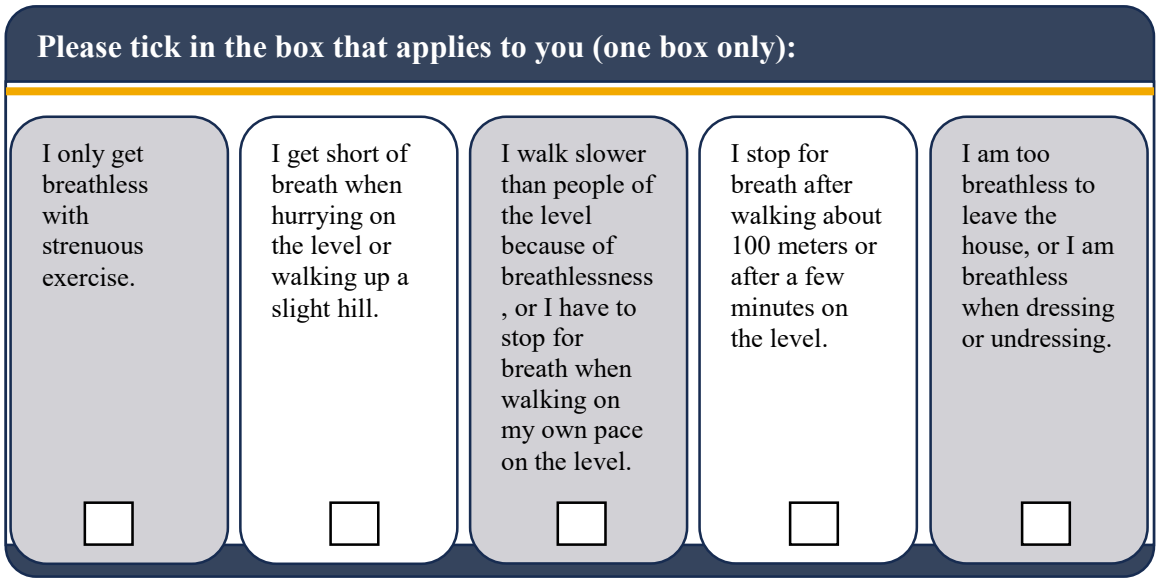


Figure 1: mMRC questionnaire used to assess COPD severity. (Reproduced with permission from [13])

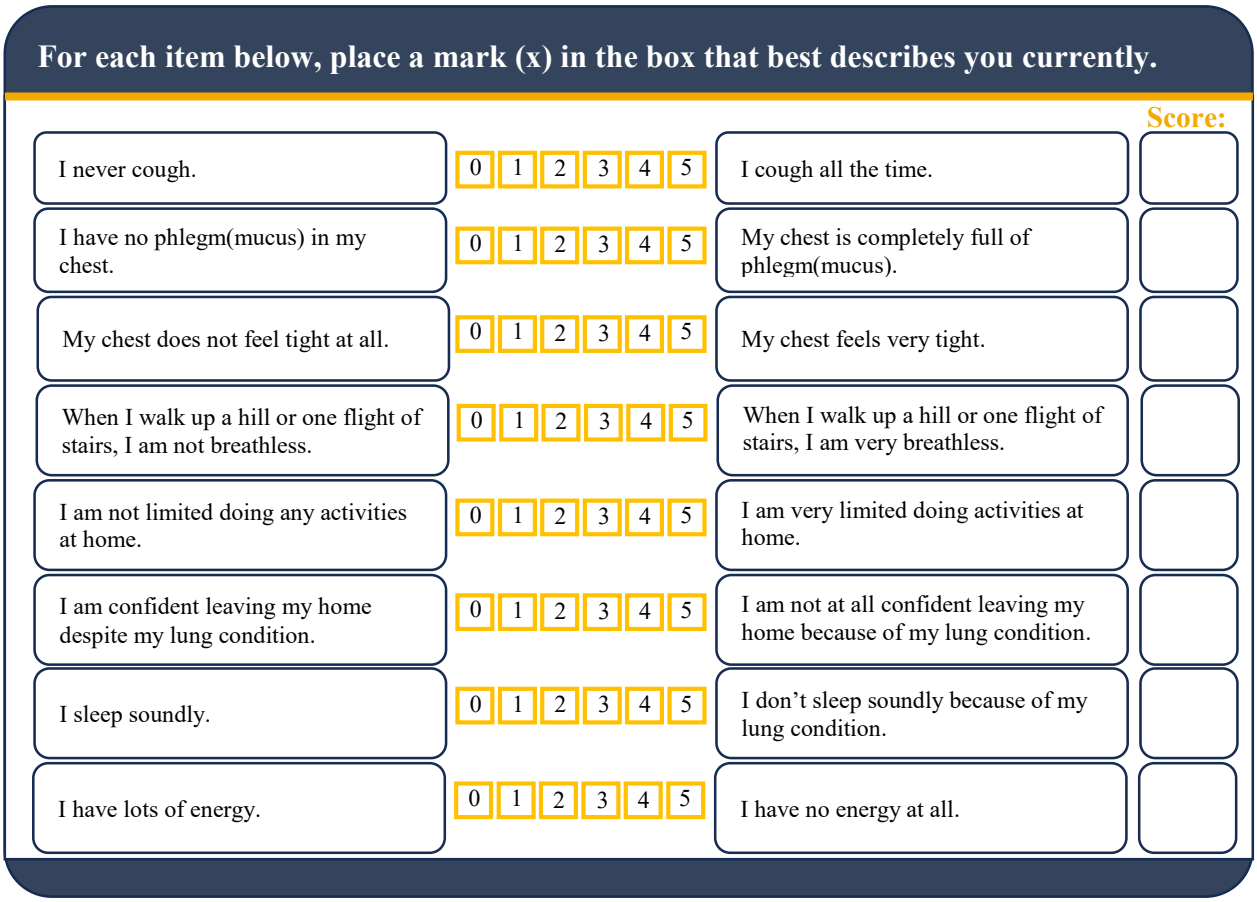


Figure 2: CAT questionnaire used to assess COPD severity. (Reproduced with permission from [13])

1.4 Pulmonary Function Tests

Pulmonary Function Tests (PFTs) provide an objective measure of how well the lungs are ventilated by measuring the volume and flow associated with normal and forced breathing [16]. While there are various pulmonary function test measurements such as plethysmography and diffusion capacity, spirometry remains the most commonly used PFT clinically. Nevertheless, the focus of this thesis is the forced oscillation technique, also known as oscillometry. Hence, the following sections will provide an overview of spirometry and oscillometry, covering their advantages and shortcomings.

1.4.1 Spirometry

Spirometry is the current gold standard PFT and the most frequently performed and widely accepted measure of pulmonary function [10], [16], [17]. In spirometry, the patient is trained and instructed to take a deep breath, reaching total lung capacity, then forcefully exhale into the spirometer. Two main measures are recorded in this test; Forced Expiratory Volume (FEV-1), which is the amount of air one can forcibly exhale in one second, and the Forced Vital Capacity (FVC), which is the total amount of air exhaled during the Forced Expiratory Volume test. The ratio FEV1/FVC is used as a measure of lung function in healthy and diseased subjects, where airway obstruction is indicated by a reduction in FEV1/FVC ratio [10]. Spirometry has readily available normative values and is widely standardized and adopted, which is an advantage over the other PFTs [10], [16].

Despite its sensitivity to pulmonary function changes, spirometry is highly effort dependent [17]. Consequently, reliable results can only be achieved with maneuver training and active patient cooperation. Resultantly, spirometry depends on the technician's ability to

properly train and guide, and the patient's ability to understand and correctly perform the test. In fact, it is estimated that about 10% of patients are unable to achieve reliable results even when trained and guided by an experienced respiratory technician [10]. This rate is even higher among seniors older than 80 years, preschool children younger than six years, and individuals with pulmonary diseases and cognitive limitations. Spirometry is also insensitive to changes in the small airways, which is where most lung diseases like asthma and COPD originate. As such, recent studies have suggested that spirometry could potentially mislead clinicians when used as the only tool in clinical decision making, as many patients with respiratory diseases have normal results [10], [17]. Given these limitations, it is of high importance to include other PFTs, such as oscillometry, in assessing, diagnosing, and monitoring pulmonary lung diseases.

1.4.2 Oscillometry

While spirometry provides a measure of the maximum lung inflation (Total Lung Capacity) and maximum exhalation (Residual Volume), oscillometry measures the resistance (*Rrs*) and reactance (*Xrs*) of the respiratory system, usually at normal breathing volume [18][19][20]. In oscillometry, sound or pressure waves are generated and superimposed on the patient's spontaneous tidal breathing at predetermined frequencies. The resultant changes in pressure and flow are then measured and used to estimate the mechanical properties of the respiratory system by calculating the respiratory impedance (*Zrs*) [19][21][22]. Standard oscillometry, sometimes referred to as spectral oscillometry, uses a periodic waveform that includes frequencies ranging from 4 - 8 Hz to 30 - 50 Hz depending on the device used. The focus is on lower frequencies in oscillometry as they provide more information about the small airways and the viscoelasticity of the respiratory system, while higher frequencies (> 50 Hz) reflect the acoustic instead of the tissue properties. However, frequencies below 4 Hz are too

close in range to the frequencies of the subject's own breathing pressure and flow, which would interfere with the accuracy of oscillometry measurements. Using multiple frequencies simultaneously via the oscillometry waveform provides useful information about the mechanical properties of the respiratory system, including the lung periphery and regional inhomogeneity. This is because changes in the diameters of the large and small airways can affect impedance differently over the frequency range, and the elastic properties dominate the respiratory mechanics at the lower frequency range while inertive properties dominate the respiratory mechanics at the higher frequency range.

1.4.2.1 Important Oscillometry Measures

The respiratory impedance (*Zrs*), which represents the mechanical properties of the respiratory system, is defined as the ratio of the difference in pressure to the changes in flow at a particular oscillatory frequency [7][16]. *Zrs* is typically calculated using the windowed periodogram technique by applying Hanning windows with up to 95% overlap to obtain an estimate for each window. The Fast Fourier Transform (FFT) of each of the windowed pressure and flow signals from each window are divided to calculate the impedance as per Equation (2), which can provide estimates of impedance as a function of time from each window, or the impedance estimates from all windows are averaged to produce a single measure of the subject's impedance. Since the pressure and flow are not measured at the airway opening, a correction particular to each device is applied so that the reported impedance is that of the subject [7].

$$Zrs(f) = \frac{P(f)}{\dot{V}(f)} \qquad (2)$$

Zrs is represented using complex numbers; the real part is the portion accounting for the pressure response in-phase with flow, and is thus the respiratory system resistance, *Rrs*. While the imaginary part accounts for the reactive forces of the respiratory system and is the pressure response out-of-phase with flow, but in phase with volume changes, and is denoted reactance, *Xrs*.

As mentioned earlier in section 1.2, *Rrs* assesses the dissipated mechanical energy required to move air through the airways and the resistance due to viscous or frictional deformations of the respiratory tissues caused by the lung and the chest wall. While *Rrs* is typically frequency-independent in healthy subjects, it becomes inversely dependent with frequency, largely attributed to the onset of heterogeneous airway narrowing throughout the airway tree [6]. The frequency dependence of resistance can be represented by taking the difference between the resistance at the low range of the oscillometry frequencies, *Rrs* at 5 Hz, and a higher frequency after which the frequency dependence in *Rrs* levels off, typically at 20 Hz although 19 Hz is also used, giving R5-20 or R5-19 (Figure 3) [7][23]. The R5-20 or R5-19 are sensitive to diseases that cause alterations to the structure and function of the lungs, such as mucus plugging, airway narrowing, or changes to lung tissue. This in turn contributes to the development of heterogeneity of airflow in the lungs, and hence R5-20 or R5-19 can be used as an predictor of treatment response and a marker for disease severity [24].

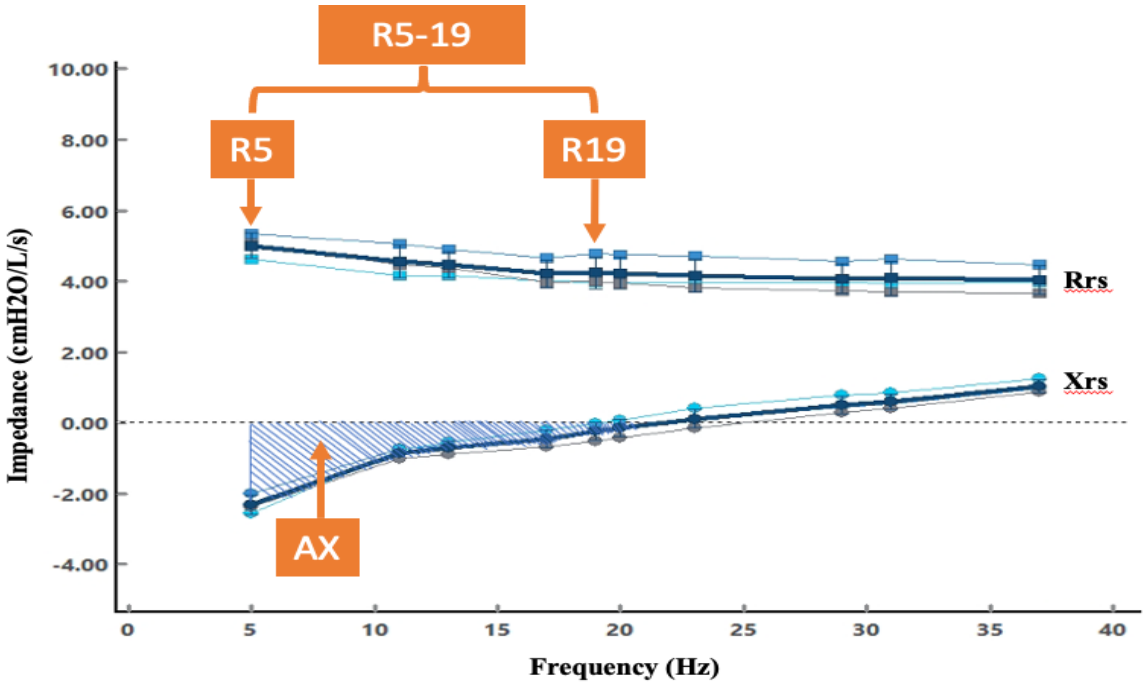


Figure 3: An oscillogram depicting key oscillometry measures: Resistance (Rrs), Reactance (Xrs), Resistance at 5Hz (R5), difference between resistance at 5Hz and 19Hz (R5-19), and the area under the reactance curve (AX). The solid line is the average of 3 measurements, and is called a 'test' with standard error bars shown, while the thin lines represent the individual measurements (more details on the measurements are found in section 1.5.10)

Other mechanisms can lead to frequency dependence, such as tissue viscoelasticity, upper airway shunting, or potentially time variation in mechanical elastance [25][26]. Frequency dependence from tissue viscoelasticity is observed at very low frequencies in healthy adults 0.1 to 0.5 Hz, but potentially may impact higher frequencies in disease or may be a contributing factor in very small children or infants [27][28]. Frequency dependence from upper airway shunting arises at high respiratory system impedances, where some of the energy of the oscillations goes into the parallel impedance path of the soft tissues of the central or upper airways (cheeks). This is minimized by holding the cheeks during oscillometry measurement [20]. Time variation of mechanical properties is a novel discovery but is thought to largely occur at lower frequencies than the oscillometry range for adults [29].

On the other hand, *Xrs* is normally frequency dependent dominated by elastic properties from volume deformation at low frequencies and from inertive properties at the higher frequency range, which reflect the acceleration of the air column during oscillation [21][22]. While the elastic properties of the respiratory system are dominated by the very distensible and large available volume of the oscillations to the tissues of the airspaces, some of the elastic properties are from airway distensibility, chest wall, or can come from the upper airways. For example, not holding the cheeks lowers the stiffness of the upper airway compartment and some energy goes into the movement of these tissues, lowering the elastance of the respiratory system. On the other hand, the majority of the inertial properties arise from the motion of the air column in the larger airways, as the amplitude of air oscillation is much larger than any motion of the tissues during oscillation. As such, the differences in pressure phase arising from the flow and the elastic and inertive forces results in a negative *Xrs* at the lower frequencies and positive *Xrs* at the higher frequencies. This transition from elastic to inertial forces dominance is defined as the resonance frequency, where the impedance is only due to resistance.

The resonance frequency typically ranges between 7-12 Hz in adults and is higher in children being inversely dependent on lung size (Figure 3) [21]. Low frequency *Xrs* reflects the elastic properties of the tissue observable by the oscillations. Any derecruitment of airways due to airway closure or narrowing is sufficient to obstruct the oscillations from the distal airspaces, increasing the elastance and resulting in a more negative *Xrs*. Also, while *Xrs* is very sensitive to changes in obstructive disease, *Xrs* does not well differentiate between restrictive or obstructive diseases, as they both result in more negative reactance due to increased stiffness [22]. The area of reactance (*AX*) is another important measure defined by the area under the reactance curve from the lowest frequency to the resonance frequency (Figure 3). Similar to *Xrs* at low

frequencies, *AX* largely reflects the elastic properties of the lungs, and indeed has the same units as elastance (kPa/l). *AX* has been found to be correlated with resistance at the lower frequencies in COPD and asthmatic children [6][21][22].

1.5 Limitations And Current Quality Control Measures in Oscillometry

An ideal pulmonary function test would be simple to perform, safe, repeatable, reproducible, and sensitive enough to distinguish between health and disease, as well as detect changes with growth [16]. Repeatability is the short-term (within-test) variability, which measures the consistency of the technical expertise of the patient, as well as the stability of the measuring instrument [16]. The repeatability of a test is typically expressed as the Coefficient of Variation (CV), which is the ratio of Standard Deviation (SD) to the mean (μ) [16][21][22]. On the other hand, reproducibility is the long-term (between tests) variability, which is influenced by the instrument's stability and a patient's ability to perform the required maneuvers consistently during each measurement session. Reproducibility can also be influenced by the technologist if performance is dependent on instruction, as well as the disease and biological variations in lung function [16]. While oscillometry is fast and easy to perform, like most measures of lung function it is affected by within- and between- test variability, and to obtain accurate measures of impedance, it requires quality control. Here, a test is the average from multiple measurements taken sequentially, typically a minimum of 3 measurements, and thus the within test variability is a short term assessment of variability, while between test variability is over a longer period such as days or weeks.

1.5.1 Quality Assurance and Quality Control

Quality Assurance (QA) and Quality Control (QC) are inseparable parts of quality management that are often used interchangeably [30]. However, while these processes feature some overlap, they are separate and take place at different times. According to ISO 9000, a set of international standards developed by the International Organization for Standardization (ISO), QA aims to provide assurance that quality requirements will be fulfilled, while QC aims to actually fulfill quality requirements [1], [2]. ISO 9000 defines QA as "part of quality management focused on providing confidence that quality requirements will be fulfilled" and QC as "part of quality management focused on fulfilling quality requirements"[1], [2]. Hence, QA is a pre-planned process comprising a series of activities designed to prevent errors and defects in products and provide confidence that the requirements are satisfied [2]. QC, on the other hand, is the inspection stage of the QA process, comprising a series of activities required to meet the quality standards and verify the safety and effectiveness of a product. Hence, QC aims to identify errors and address issues to ensure the production of high-quality products before reaching customers [28]–[30]. While both QA and QC are critical to ensure reliable and accurate measurements in oscillometry, this thesis focuses on assessing and improving the current QC criteria.

1.5.2 Standard Operating Procedures

There are several QC strategies that are currently implemented for oscillometry to help reduce the impact of variability of impedance measurements and the influence of factors that may arise apart from the respiratory system mechanics. To help minimize variability in performing an oscillometry measurement, it is essential to provide a Standard Operating

Procedure (SOP) and ensure that operators are trained according to the SOP. The SOP which combines the manufacturer's manual and standard guidelines that patients and operators need to follow to optimize repeatability of oscillometry measurements [20]. In brief, the operator conducting the test should explain the duration and number of replicates to be recorded. They should also describe the sensation caused by pressure oscillations and run trials before data acquisition [16]. During the test, patients should be instructed to sit upright with head in natural position and legs uncrossed, and to support their cheeks and mouth floor with the palm and fingers to minimize upper airway wall vibration [16][20]. Patients should also wear a nose clip, breathe calmly through a mouthpiece, and avoid swallowing with the tongue maintained forward [16][20]. Hence, the operator must be reliable, well-trained, and able to work under minimal supervision to ensure that they are capable of training patients and following this SOP. Implementing a SOP and providing hands-on training for operators can result in a significant improvement in repeatability, as demonstrated by Wu et al. [20].

1.5.3 Calibration and Verification

Since *Zrs* is measured using signals from a pressure transducer and a flowmeter at or near the mouthpiece, it is recommended to calibrate and periodically verify the calibration of these sensors. This includes static calibration to ensure correct gain and zero offset, while considering possible position or temperature drifts. Dynamic calibration during oscillatory stimuli should also be performed to compensate for the sensor's frequency response and to ensure that the signals are unaffected by the mechanical vibrations caused by the oscillation frequencies. Yet, unlike static calibrations, dynamic calibrations are not periodically required as it mainly depends on the physical dimensions of sensors and tubing. It is also recommended that end users perform periodic verifications using test loads with an impedance magnitude higher than what is expected

for any given patient in which the oscillometry device is to be used, as the use of insufficient impedance may potentially lead to errors in measurement [21]. As such, test loads of ~15 $hPa \cdot s \cdot L^{-1}$ and ~40 $hPa \cdot s \cdot L^{-1}$are recommended to be used for adult and children testing, respectively. While sensor calibration is usually performed by manufacturers, end users are also required to perform daily, or each day the device is used, verifications using the impedance test loads to ensure an acceptable tolerance of $\leq \pm 10\%$ or $\pm 0.1\ hPa \cdot s \cdot L^{-1}$, whichever is met first. Further technical requirements include a maximal dead space that is added by the device and filters, and a maximum resistance of the breathing pathway through the filter and oscillometry device. It is therefore recommended to use low-resistance filters of $< 1\ hPa \cdot s \cdot L^{-1}$ at $\leq 5Hz$ and to compensate for the combined resistance of the filter and oscillometric system, with a total equipment resistance $< 2\ hPa \cdot s \cdot L^{-1}$ at $\leq 5Hz$. The recommended dead space for oscillometric devices is the same as for lung volume with a value below 100mL, inclusive of the bacterial filter, for testing adults and below 70mL for testing preschool children. [21]

1.5.4 Biological Verification

In addition, Poorisrisak et al. demonstrated the utility of using biological verification to assess reproducibility of test measurements [31]. Biological calibration, commonly known as biological quality control, is an oscillometry test performed periodically with healthy non-smoking personnel, typically in the pulmonary function laboratory. A confidence interval of *Zrs* is first obtained by collecting sufficient *Zrs* data (≥10 measurements) in a relatively short time-interval (a few weeks). A subsequent measurement outside the confidence interval will thus indicate that the oscillometric system should be evaluated carefully. In fact, Wu et al. suggested that an “out of control” condition can be identified by one of the following situations: 1) four

consecutive measurements that exceeds the mean 1SD, 2) two consecutive measurements that exceeds the mean 2SD, 3) a measurement that exceeds the mean 3SD, or 4) 10 consecutive measurements that fall on the same side of the mean. As such, once an "out of control" condition is observed, the device should be carefully evaluated and not used for testing until it is verified [20][21].

1.5.5 Acquisition Time

The number and duration of the obtained measurements were shown to affect the repeatability of oscillometry. The European Respiratory Society (ERS) technical standards recommends performing a minimum of three 16-seconds measurements with a CV less than or equal to 15% for young children and 10% for adults [21]. The averaged measurements with acceptable CV are typically known as a successful test. It is also recommended that the data acquisition is preceded by 30 seconds of tidal volume monitoring to allow patients to achieve stable breathing patterns [21][32]. Nonetheless, Robinson et al. suggested that tidal volume monitoring before data collection has little effect on the repeatability and quality of the test, when compared to duration of data acquired [33].

1.5.6 Coherence

Current efforts to improve QC included the use of different artifact detection tests such as coherence, statistical filters, wavelet-based and supervised machine learning methods [34][35][36][37]. Coherence is a measure of how closely flow and pressure waves are related linearly at a given frequency [21][22]. It is similar to the linear correlation coefficient and is a number between zero and one that provides a causality index between the input and output signals that is reduced from one in the presence of noise or nonlinearity. Measurements with a

coherence of less than 0.90 or 0.95 in many fields are usually recommended to be discarded [16]. Yet, in some instances measurements with lower coherence can still provide accurate results with averaging. While lower than expected coherence may be indicative of artifact or noise, high values do not necessarily ensure the absence of contamination. This is because other inputs may be present in a given measurement, such as breathing with harmonics, potentially coherent with the oscillations of the test input signal. The use of coherence can also be limited by the different coherence calculation approaches, its dependence on windowing, and its reduced magnitude in disease [16][21][38]. Due to these limitations, coefficient of variation calculated over multiple measurements, further discussed later in this section, is now preferred as primary means of quality control in oscillometry [21].

1.5.7 Manual Quality Control including Artifact Detection

Manual artifact detection is also one of the first methods used for the detection and rejection of individual measurements or breaths within a measurement. This involves visually inspecting the pressure, flow, or volume signals to assess the quality of breathing or detect signs of persistent leaks that could corrupt the measurement throughout its duration. Additionally, shorter artifacts, like coughs, swallows, or brief breaks in the seal with the mouthpiece causing momentary leaks during measurement, are observed for potential identification and removal. Manual QC can also be performed by observing disturbances within the resistance and reactance time courses during measurement. Although, manual QC with training can be done, it is time consuming, subjective and may be susceptible to bias [21][33][37]. Therefore, interests have shifted towards the automation of artifact detection to provide faster, automated and objective methods to improve quality control in oscillometry.

1.5.8 Statistical Quality Control

There are a few statistical quality control tests that have been introduced to exclude outlying data points within a measurement, such as the exclusion of values that exceed a specific SD threshold. The three and five SD (3SD and 5SD) described by Schweitzer et al. and the Brown et al., respectively, are the most common and broadly applied statistical filters [33][39][40]. The 3SD filter is typically applied three times across the entire segment of oscillometry measurement data to reject points where the *Zrs*, including the *Rrs* and *Xrs*, exceeds three times the SD (>3SD) from the mean *Rrs* and *Xrs* values [33][39]. Contrarily, the 5SD filter implemented by Brown et al included an initial rejection of points corresponding to negative *Rrs* values before rejecting any *Rrs* and *Xrs* values exceeding five times the SD from the mean *Rrs* and *Xrs* values [33][36]. However, the exclusion of individual data points likely results in an uneven data segmentation, thus may distort the relative contributions of inspiratory and expiratory portions of each breath and increasing within-session variability [33]. This limitation was later addressed by Robinson et al. by rejecting full breaths instead of individual data points. Partial or incomplete breaths at the beginning or end of recordings were also excluded to ensure a balance between the inspiratory and expiratory contributions of each breath. Hence, the use of complete breath filtering was shown to result in a lower within-session variability, when compared to either the 3SD or 5SD filtering approaches [33]. The Flow Shape Index Filter (FSIF) is one of the first attempts to provide an automated artifact detection technique. The FSIF works with the assumption that the pressure oscillations used are always fairly sinusoidal when free from contamination. Hence, the deviation of the shape of the flow signal from a pure sinusoidal wave could be used as an objective and unbiased criterion to eliminate artifacts [34]. Therefore, FSIF may have advantages over manual and statistical filtering as it can operate in

real time, without the need for post-processing, while maintaining satisfactory agreement with the repeatability.

1.5.9 Advanced Artifact Detection Techniques

Other artifact detection techniques examined the possibility of using the Short Time Fourier Transform (STFT) and Wavelet Transform (WT). STFT is a repeated application of the Fourier Transform (FT) in a sequential fixed time window for each frequency [35][41]. However, STFT imposes constraints on the representation estimate, as they require the signal to be stationary during a finite time interval and limits time-frequency resolution [41]. The discrete WT attempts to overcome some of these limitations by employing scaled length windows inversely with frequency, which may be better for nonstationary signals as it optimizes time frequency resolution in all frequency ranges [41]. Machine learning has also been attempted as an objective artifact detection tool that can be used alone or in combination with other artifact detection methods. Pham et al. has demonstrated the efficiency of using feature extraction models, as they provide a better or equivalent performance than an expert operator [37][42].

All of these within measurement methods can be automated and some systems incorporate simple rejection methods such as the 3SD method, or removing negative resistance values which are obviously corrupted from a measurement. In addition to the published QC techniques, individual manufacturers have also implemented other techniques to improve quality control and automation of measurements, but these may not be fully described and thus may not be independently validated. Manual QC as described can be applied additionally to any automated within measurement QC approach. However, for acceptability of a test, that is form repeated measurements there is a proposed standard discussed in the next section.

1.5.10 ERS Technical Standards

The ERS technical standards provide a standard method to help improve data quality using the CV between measurements, which can be applied after any manufacture-based approach to remove artifacts such as leaks and swallows. The ERS technical standards recommend recording three measurements with a CV in the low frequency resistance (typically R5) that is less than or equal to 10% in adults or 15% in young children [21]. Here, a measurement is defined as one of the three recordings, separated by a brief break, each lasting for a standard duration of 16 seconds. Introducing a short break between measurements is advised to address potential concerns like leaks, insufficient cheek support, misplaced nose clips, or improper posture. These issues may persist during a measurement, and the brief breaks offer a chance for necessary adjustments, enhancing the reliability of the measurements.

While CV is provided for all recorded measurements to ensure that the CV is less than the required threshold, it is performed manually in some systems and often in post-hoc analyses. This is performed by removing measurements that cause the CV to be greater than 15%, which is time consuming and limits ease of measurement according to the ERS standard. Moreover, the ERS standard focuses on resistance at the lowest frequency while ignoring variability from other frequencies, and does not assess variability in reactance. The fact that the current standard ignores reactance is important as reactance becoming increasingly important clinically [21]. Indeed, recently, Hantos et al. demonstrated that the variability of the lowest oscillation frequency reactance (X5) can be larger than the variability of R5. As a result, Hantos et al. recommended that reactance should be incorporated in both short-term and follow-up reproducibility due to its clinical importance, specifically at the lowest frequency oscillations [43].

Additionally, while the technical standard is expert opinion, and based on measurements of variability, the rationale the taskforce used for the standardization of R5 for test acceptability was largely opinion based and was not validated. Recently however, Therkorn et al. identified three main factors that likely contributed to the standardization of R5: 1) resistance, unlike reactance, is independent of frequency in healthy subjects 2) R5 is thought to better represent the small airways 3) using R5 from whole breaths accounts for both inspiration and expiration resistance, making it easier to calculate and implement even with devices that lack explicit inspiration and expiration partitioning [44][45]. Also, while the use of CV of three measurements from multiple measurements as a measure of variability improves repeatability, and while it may be assumed that selecting repeatable measurements is more accurate, this has not been demonstrated. Therefore, this thesis will explore the use of CV across multiple frequencies and incorporate reactance to account for variability in the full impedance to develop and validate the accuracy and efficiency of a new automated QC algorithm. The hope is that this will point to a technique to be accepted and translated into the industry.

CHAPTER 2: OBJECTIVES AND HYPOTHESES

The current ERS technical standards for acceptability are based on a CV obtained from resistance at 5 Hz. The Maksym lab has done some work exploring an automated algorithm based on the resistance measurements using multiple frequencies. Specifically, the Maksym lab explored an algorithm that examines all permutations of resistance measurements to obtain the first three measurements with an average CV at R5 and R11 to be less than or equal to 15%. While the resistance at 5 Hz (R5) is the standard measure to calculate and optimize the CV, this algorithm uses combinations of resistance at multiple frequencies to minimize the effect of breathing noise at 5 Hz. However, as mentioned, the use of resistance alone does not take into account the reactance, another important measure in oscillometry, which could result in poorer overall performance. Therefore, it is hypothesized that the use of impedance, which takes into account both resistance and reactance, can provide a better and more accurate tool for the calculation and optimization of CV, and potentially accuracy in estimating *Zrs*.

Finally, in a previous course project the author carried out, it was demonstrated that machine learning can be used with oscillometry data to predict Patient Reported Outcomes (PRO), namely, to classify low and high values of the mMRC and CAT scores[46], [47]. The aim of this project was to determine the extent to which an objective measure, specifically spirometry and oscillometry, correlates with a subjective score. The goal was to investigate this relationship using machine learning, considering its potential to outperform traditional regression methods. However, it is important to note that this study does not seek to replace subjective PROs, as a strong correlation between spirometry, oscillometry, and PROs is not necessarily expected. Nevertheless, machine learning depends on the quality of the data on which it is trained and tested. Therefore, the implementation of the spectral oscillometry QC algorithm have

the potential to minimize the variability of the oscillometry signals and improve the performance of the previously developed algorithms. Hence, if possible, this thesis will test the hypothesis that either a higher model accuracy can be achieved with the same number of training samples, or the same accuracy can be achieved with a lower number of samples in the training data set. As such, this thesis has three main principal objectives:

1) To extend the automated QC method using different combinations of resistance and impedance, including reactance over the range of frequencies, to improve the variability of spectral oscillometry.
2) Next, a computational model will be developed to simulate time-varying resistance and reactance, which will be used to test and validate the accuracy of the spectral oscillometry QC algorithms developed by the Maksym lab. This will also help validate the accuracy of using the CV as a measure that reflects variability and quality of measurements.
3) The last objective will focus on improving the performance of the previously developed machine learning model and, if possible, compare the performance of the developed ML models before and after applying QC.

The first two objectives will improve repeatability and thus, the translation of oscillometry for clinical use and disease monitoring or effectiveness of therapy. On the other hand, the third objective aims to optimize the developed machine learning model used to classify the severity of COPD using the patient's objective reported outcomes.

CHAPTER 3: OPTIMIZING CV USING SPECTRAL QC

This chapter presents the methods and results of the spectral QC algorithms aiming to improve the variability, feasibility, and efficiency in oscillometry. It also provides an overview of the development of a computational model to validate the accuracy of these algorithms, which has never been attempted before. There are three main algorithms, *Early*, *Variant* and *LowestCV*, which will be discussed in more details in the upcoming chapters.

3.1 Methods

3.1.1 Data Used for the Thesis

As illustrated in Table 1, three data sets were used to develop and validate the performance of the spectral QC algorithms: CHILD5Y, WESER, and West Island Cohort (WIC). The CHILD5Y and WESER data sets were collected with informed consent, which was obtained in accordance with the approval of the research ethics board for the Hospital for Sick Children Toronto (REB # 1000060128 and #1000041089, respectively). Similarly, the WIC data set was collected with informed consent, approved by the McGill University Health Centre Research Ethics Board (MUHC-RI REB# 14-467-BMB). The measurements were collected using a commercially available device (Tremoflo C-100, Thorasys Medical Systems, Montreal, Canada) and the manufacturer' software (tremoflo 1.0.43 build 44). CHILD5Y includes measurements from a group (n=550) of five years old healthy and asthmatic children in Toronto. This data set is excellent since performing an oscillometry test on children is more challenging and thus the data sets include several subjects with more than three repeated measurements. This makes it ideal for developing and evaluating the proposed QC algorithms. The data set contains an average of six measurements per subject, allowing for the study of different permutations of three

measurements to meet the CV threshold. The WESER data set includes a group (n=110) of three to five years old children with wheeze that were admitted to emergency, with an average of eight measurements per subject. The data set also contains the notes from a well-trained operator who performed manual QC on the data set measurements, which will be compared to the performance of the developed spectral QC algorithms. The WIC data set contains 818 adult clinic data with an average of five measurements per subject. Approximately 80% of the WIC are COPD subjects, while the other 20% include normal and diseased subjects; mainly asthma, sarcoidosis, and interstitial lung disease. The algorithms were first developed and optimized using CHILD5Y and validated using the WESER and WIC. The CHILD5Y data set was finally employed to generate a computational model to assess the accuracy of the developed algorithms.

Table 1: Demographics of data sets used to develop and validate the spectral QC algorithm.

	CHILD5Y	**WIC**	**WESER**
N	550	818	110
AGE	5	45-90	3-5
HEIGHT	$111.1 \pm 4.8\ cm$	$165.6 \pm 11.6\ cm$	$105.5 \pm 7.2\ cm$
WEIGHT	$19.3 \pm 2.7\ Kg$	$78.6 \pm 19.7\ Kg$	$17.8 \pm 3.3\ Kg$

3.1.2 Early, Variant and LowestCV

Unlike the standard way of calculating the CV, which uses the resistance at 5 Hz from all available measurements, the spectral QC algorithms proposed in this thesis calculate the variability using different combinations of resistance CVs, ζ_R, and impedance CVs, ζ_Z, at different frequencies from a permutation of three measurements. Here, ζ_R or ζ_Z will be used to calculate the within-test variability as summarized in Equations 3 and 4:

$$\zeta_R = \frac{\sum_{i=1}^{N_f} W_i \cdot CV(R_{fi})}{\sum_{i=1}^{N_f} W_i} \qquad (3)$$

$$\zeta_Z = \frac{\sum_{i=1}^{N_f} W_i \cdot CV(|Z_{fi}|)}{\sum_{i=1}^{N_f} W_i} \qquad (4)$$

Where ζ_R and ζ_Z are cost functions that capture the variability between repeated oscillometry measurements, R_{fi} and $|Z_{fi}|$ are the resistance and magnitude of impedance over the measured frequencies, N_f is the number of measured frequencies, and W_i is the weight at frequency i which is used to attribute more importance to specific frequencies or correcting for an unevenly distributed frequency spectrum. As such, configuring W_i can obtain an average over multiple *Rrs* or |*Zrs*| values or control the contribution of selected frequencies to the weighted average to obtain ζ_R and ζ_Z that vary the contributions of the different variabilities (CV) at each frequency over the entire frequency spectrum.

The *Early* algorithm looks at the combinations of three measurements from the first 3,4, 5...N measurements to find the first three measurements with ζ_R or ζ_Z less than or equal to 15% (Equation 3 and 4). Therefore, the *Early* algorithm is able to meet the threshold using fewer measurements, making it the fastest algorithm to perform and the most appropriate for clinical use. That is when using early, the operator can stop repeated measurements as soon as the desired CV condition is met. However, this algorithm is potentially more subjected to a learning effect, particularly in children, where the first measurement might be more commonly the least reliable. To address this concern, a *Variant* algorithm that follows the same steps as *Early* was investigated, but it modifies the approach by initiating the assessment of variabilities among any three measurements only after the first four measurements have been obtained. In contrast, the

LowestCV algorithm looks at the permutations from all available measurements in our data sets to find the lowest ζ_R or ζ_Z possible of any three of the recorded measurements. Hence, the *LowestCV* algorithm will yield an equal or lower ζ_R and ζ_Z compared to *Early* and *Variant,* as our data generally had more than the 3 or 4 measurements used by *Early* and *Variant* respectively. The *LowestCV* was calculated to demonstrate the lowest achievable ζ_R or ζ_Z within the data sets, considering that the CHILD5Y and WESER data sets contained a larger number of measurements compared to what is typically available. It is not intended as a clinically applicable method as some subjects had many more measurements than needed to meet the CV criteria. There are four main performance measures that were studied to evaluate and compare the spectral QC algorithms to the current standards: 1) repeatability, 2) efficiency, 3) feasibility and 4) accuracy. These are each described in the sections below.

3.1.3 Repeatability

The repeatability of the spectral QC algorithms was evaluated using ζ_R and ζ_Z, the Root Mean Square (RMS) of ζ_R and ζ_Z, average SD(R5)/μ(|Z5|), SD(X5)/μ(|Z5|), SD(R5-19)μ/(|Z5|), and SD(AX). Since R5, X5, R5-19 and AX are the outcome measures of clinical interest, the variability of these measures were used as performance measures. However, the SD of R5, X5 and R5-19 divided by the mean of |Z5| was used to allow for a direct comparison of the relative variability between these measures. Contrarily, since AX can have a mean close to zero, which results in high CV values, SD(AX), rather than CV(AX), was also used as an alternative measure to represent the variability of the measurements when using different Rrs and Zrs combinations. Furthermore, the RMS average of ζ_R and ζ_Z provides a general guideline on how the CV changes across all frequencies. However, this is not an accurate representation of the overall performance,

as a lower CV of the cost function does not necessarily correspond to a good performance in terms of the other performance measures. Hence, the main focus of the analysis will be on the average SD(R5)/μ(|Z5|), SD(R5-19)/(|Z5|) and SD(AX), as it is felt that they give a very understandable representation of the overall performance. They also provide physicians with relatable information to the outcome measures to evaluate the repeatability of their repeated measurements of the mechanical properties of the respiratory system. Similarly, plots of CV(|Z|), SD(R)/μ(|Z|), SD(X)/μ(|Z|), as well as the actual Rrs and Xrs values, as a function of frequency, were also evaluated to study the variability before and after QC.

3.1.4 Efficiency

The second performance measure is the efficiency of data collection, which is defined here as the required number of measurements to pass the acceptability criteria, either the CV, ζ_R or ζ_Z to be less than or equal to 10% and 15% for adults and children, respectively. While the main focus is on improving the repeatability and reducing variability to meet the acceptability criteria defined by ERS, it is critical that the proposed QC algorithms does not improve repeatability at the cost of lowering the efficiency. Therefore, it is important to compare the efficiency before and after QC and validate that it was maintained, or better still improved, which plays a key role in implementing and adopting the proposed QC algorithms.

3.1.5 Feasibility

The third performance measure is feasibility, which is the percentage of subjects that were able to achieve valid data by obtaining a minimum of three repeated measurements with a CV, ζ_R or ζ_Z less than or equal to15% [21]. The feasibility of the proposed spectral QC algorithms and the ERS technical standards were compared using different combinations of R (ζ_R) and |Z|

(ζ_Z). Seven different methods were compared to select a minimum of three valid and artifact-free measurements, where the CV, ζ_R or ζ_Z for each patient test is calculated using: 1) the first three measurements and no QC, 2) the first four measurements and no QC, 3) the first five measurements and no QC, 4) all the available measurements and no QC, 5) a combination of three measurements from the first four, with QC, 6) a combination of three measurements from the first five, with QC and 7) a combination of three measurements from all the available measurements, with QC.

The CHILD5Y data set contains a total of 3572 measurements from 550 patient tests, with 885 user-excluded measurements and 589 tremoflo excluded measurements. The WIC contains a total of 9143 measurements from 818 patient tests, with 709 user-excluded measurements and 1132 tremoflo excluded measurements. A measurement can be automatically excluded by the tremoflo software when it contains less than 10 seconds of valid recording. Although this can be due to early termination of the measurement during recording, it is usually from the rejection of negative *Rrs* values or *Rrs* values that are more than two SD from the mean, as can occur with artifacts such as coughs, brief tongue occlusions, or swallows during recording. User exclusion, on the other hand, is performed manually based on visual inspection and is usually done by a trained technician or clinician.

The seven outlined methods used to assess the feasibility were performed after excluding tremoflo excluded measurements but included the user-excluded measurements. This was done to test algorithms under conditions that relied fully on algorithmic methods which is typical in clinical use and there were no notes available describing the rationale for any measurement exclusions in CHILD5Y and WIC. However, the WESER data set went through a careful overreading of the data by a well-trained operator who examined the pressure and flow tracing

for irregular shaped breaths, such as very large or very small breaths, or evidence of coughs, excluding measurements including notes. Thus, the WESER data was used to compare the effects of inclusion versus exclusion of manually rejected measurements. This analysis involved 86 patient tests from the WESER data, as no manual QC was conducted on the remaining data. The effect of including or excluding user excluded data on this algorithmic approaches was also analyzed in the CHILD5Y and WIC data sets. Although the user-excluded measurements in the CHILD5Y and WIC data are not being performed by a well-trained operator, they were still used to represent a form of manual QC that is common in the clinical settings.

3.1.6 Accuracy and Model Development

Since accuracy in the presence of patient variability cannot be determined with measured data, the accuracy of the spectral QC algorithms was evaluated using a computational model. Instead of using values chosen from an arbitrary random distribution, it was decided to use the exact *Rrs* and *Xrs* values that were measured in CHILD5Y as the model values and distribution, to which was added artifacts and variability based on estimates of artifacts and variability as described later. To estimate as best as possible a group of 'true' values uncorrupted by artifacts, tremoflo and user-excluded measurements were excluded from all subjects in the CHILD5Y data set. Subjects with less than with five or more repeated measurements each including their expected measurement variability amongst the five. These five measurements were used as a simplified simulation of the *true* physiological variability and hence, used to model, and then to test the accuracy in estimating the *true* resistance (R_T) and reactance (X_T) values.

Therefore, the simplified model consists of 209 subjects, each with five artifact free measurements, providing a useful range of normal physiological variabilities that mimics the variation in both impedance values and variation in values found in a group like the CHILD5Y.

Using the exact *Rrs* and *Xrs* values from CHILD5Y provided a simplified simulation while preserving the relationship between an individual's *Rrs* and *Xrs*, which can become a challenge to model independently, and easiest to use existing real-world data. Additionally, the use of five measurements enabled the experimentation with different measurement permutations using the *Early* algorithm under realistic order of measurement conditions. This approach was also used to replicate a clinical scenario like the data from Dr. Ronald J. Dandurand clinic (WIC), where five measurements are a common measurement protocol.

To model noise, this was taken from our measurements as well, where there were recognizable outliers in measurements that often could not be physiological, that is very unlikely to be produced by a change in lung mechanics, such as a measurement more than double the remaining measurements which appeared groups together. It's not known the cause of these outlier measurements, and the individual measurements data do pass the device detection and rejection of any short within measurement artifacts such as might occur due to a cough, thus these outliers must arise from persistent deviations from the average behavior of a typical measurement with its expected variability. These potentially could arise from unobserved improper holding of the cheeks, missing nose-clip, improper posture, holding the tongue back narrowing the airway for an entire measurement, but the reasons are unknown. Given that there can bias the mean values, particularly if they produce outlying values of impedance, these measurements should ideally be detected and removed from the average of the test. Outlier measurements that occur but otherwise do not affect the mean values are here not taken to represent an issue harmful to accuracy. Random measurement noise such electrical transducer noise can also corrupt the measurements but these are quite small and are likely less problematic. Hence, to provide a more accurate simulation of clinical scenarios, it was decided to add outliers

based on our data and add these to the true values. The development of this outlier distribution and the methods used to introduce these outliers in the simulation is described in next subsections. Once outliers were introduced, the accuracy obtained before applying any QC algorithms was finally compared against the accuracy obtained after applying the proposed spectral QC algorithm. In this evaluation, accuracy was defined as the percent Root Mean Square Error (%RMSE), as elaborated in subsection 3.1.6.5.

3.1.6.1 Generating the Outlier Distribution

A distribution of outliers was estimated using data from the CHILD5Y subjects and was applied it to introduce outliers to the model of normal subject physiological variability. This would enable the assessment of the effects of the QC algorithms on accuracy by the difference with the true values. Thus, similar to the development of model of normal variability, a distribution of outliers was developed from the measured data. The outliers distribution was identified by applying a method known as the Grubbs test (described below) to all the collected measurements for each subject test in the CHILD5Y data set. This was done after including user- and tremoflo- excluded measurements as they provide more data points and realistic events for the Grubbs test. Here, an outlier is a full measurement detected after applying the Grubb's test to repeated measurements.

The Grubbs test is a statistical test designed to detect a single outlier from limited measures. It is based on the assumption that the data follows a normal distribution, but this can be difficult to assess for small number of samples, which is typically where the test is applied, and an approximately normal distribution is sufficient. The Grubbs test is also known as the maximum normed residual test and is defined by:

$$G = \frac{\max|Y_i - \bar{Y}|}{s} \qquad (5)$$

Where $\bar{Y}$ and s represent the sample mean and standard deviation, respectively. The Grubbs test calculates a test statistic (G) that is the largest absolute deviation from the sample mean in units of the sample standard deviation. Thus, the test compares the difference between the suspected outlier and the sample mean to the sample standard deviation. If the calculated G value exceeds a critical value the suspected outlier is considered significant and can be removed from the data set [48][49]. The critical value is based on critical values of a t-distribution with N-2 degrees of freedom where N is the number of values, and a significance value alpha such as alpha = 0.10 is chosen (described further below on page 36)

Normality Test of CHILD5Y Data

Since normality is desired to perform Grubb's test, a normality test was first performed on the CHILD5Y data assessing the distributions *Z(1/f)* and log(*Z(1/f)*), since initial observation indicated the distribution of *Z(1/f)* could be more log-normally distributed. *Z(1/f)* is a weighted impedance cost function with an ability to identify and reduce the variability of impedance, as demonstrated in section 3.2. The CHILD5Y data set was visually compared to a normal and log-normal distribution with the same means and standard deviations (Figure 4). The skewness and kurtosis were also computed since these should be zero for normally distributed data. Objectively, the normality of individual patient test measurements were also assessed using the Lilliefors test. The Lilliefors test is a variation of the Kolmogorov-Smirnov (K-S) test for normality. The test compares the Empirical Distribution Function (EDF) of the sample with the Cumulative Distribution Function (CDF) of the expected distribution. The maximum absolute difference (D) between the EDF and the CDF is then compared to a sample size adjusted critical

value [50]. The Lilliefors test is thought to yield the best results for small sampled data [51], and thus is suitable for CHILD5Y subjects, containing 4-12 measurements in each test. The Lilliefors test was applied using the MATLAB function *lillietest* on *Z(1/f)* and $\log(Z\left({}^{1}/_{f}\right))$ per subject, with parameter ('Alpha' = 0.1) for a 10% significance level.

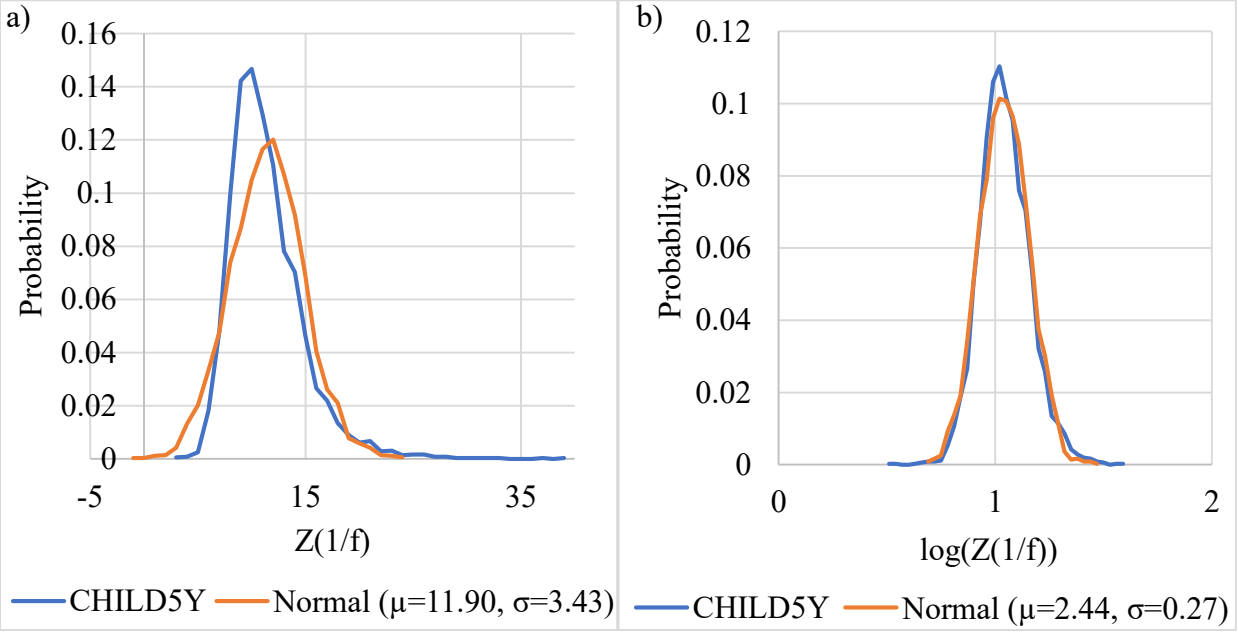


Figure 4:Probability distribution of a) Z(1/f) and b) log(Z(1/f)) from CHILD5Y(Blue) and a normal distribution with the same mean and SD obtained using the MATLAB function normrnd.

Observation of the *Z(1/f)* distribution from the CHILD5Y data set demonstrated a closer resemblance to a log-normal distribution as shown in Figure 4. Indeed, the distribution of log(*Z(1/f)*) exhibited a skewness of 0.33 and a kurtosis of 3.8, indicating closer to normality than the unadjusted Z(1/f), which had a skewness of 1.6 and a kurtosis of 8.3. While the difference between the percentage of patient tests that passed the Lilliefors test using log(*Z(1/f)*) and *Z(1/f)* was small, 90% compared to 89%, these percentages were notably high. Taken together, it was decided it is better to choose log(*Z(1/f)*) with the Grubb's test for outlier detection.

Applying the Grubbs Test

As mentioned earlier, an outlier is a full measurement that was detected and rejected by the Grubb's test performed using the MATLAB function *isoutlier* with parameters (method = 'Grubbs') and ('threshold' = 0.1) for a 10% significance level. A significance level, often denoted by alpha (α), is the probability of rejecting the null hypothesis when it is actually true. In the Grubb's test, the null hypothesis is that there are no outliers in the data set. Hence, a significance level of 10% means that there is a 10% change that a detected outlier is actually a false positive. While other confidence levels were analyzed, a 10% significance level was used as it is a common threshold and it provided a sufficient number of outliers to generate a better estimate of the artifact distribution. It was straightforward to apply the Grubbs test using the impedance rather than *Rrs* and *Xrs* independently, and easiest to use a composite measure of impedance across frequency, like the cost functions used previously in section 3.1.2, rather than detecting and modeling individual outliers at each frequency. A similar weighted impedance cost function was used, as illustrated in Equation (6), because of the ability to identify and reduce the variability of impedance using an inverse frequency weighted cost function, as demonstrated in section 3.2, and because of the variability in Zrs that is dominated at the lowest frequency.

Thus, the use of $Z_o(1/f)$ simplified the detection of outliers to a single value for a measurement that considers both the resistance and reactance across multiple frequencies.

$$Z_o(1/f) = \frac{CV(|Z_5|) + \frac{CV(|Z_{11}|)}{11} + \frac{CV(|Z_{13}|)}{13} + \frac{CV(|Z_{17}|)}{17} + \frac{CV(|Z_{19}|)}{19}}{1 + \frac{1}{11} + \frac{1}{13} + \frac{1}{17} + \frac{1}{19}} \qquad (6)$$

The Grubbs test was applied using all the collected measurements for each individual patient test (n=550) in CHILD5Y. Since the Grubbs test detects a single outlier at a time, the test was repeated for each patient until no more outliers were detected from that patient. This defined an outlier distribution specific to each subject. However, since a distribution was required to randomly add outliers across all subjects, the detected outliers were normalized to the test mean of the subject (μ_{OR}). These normalized outliers were calculated using the remaining measurements after the rejection of the outlier detected by Grubbs test, yielding normalized *Rrs* (R_{nor}) and *Xrs* (X_{nor}) values at each frequency k. Figure 5 illustrates the probability distribution of the cost function Z_o*(1/f)* before and after the detection and exclusion of outliers from all patients detected by Grubbs test, as well as the probability distribution of the detected normalized outliers. These results highlight the effectiveness of the Grubbs test in detecting outliers (n_{GO}= 105) that deviated from the test mean, resulting in shorter distribution tails after their removal. The presence of two peaks in Figure 5b reflects outliers below and above the mean values of the respective patient. Peaks at around 1.5 and 0.5 represent peaks in the distribution of outliers that were larger and smaller than the mean, respectively. Normalized outlier values greater than 1.5 or less than 0.5 indicate more extreme outlier values while values closer to zero come from patients who had narrower distributions for their non-outlier values, such that the prospective outlier was detected by the Grubb's test.

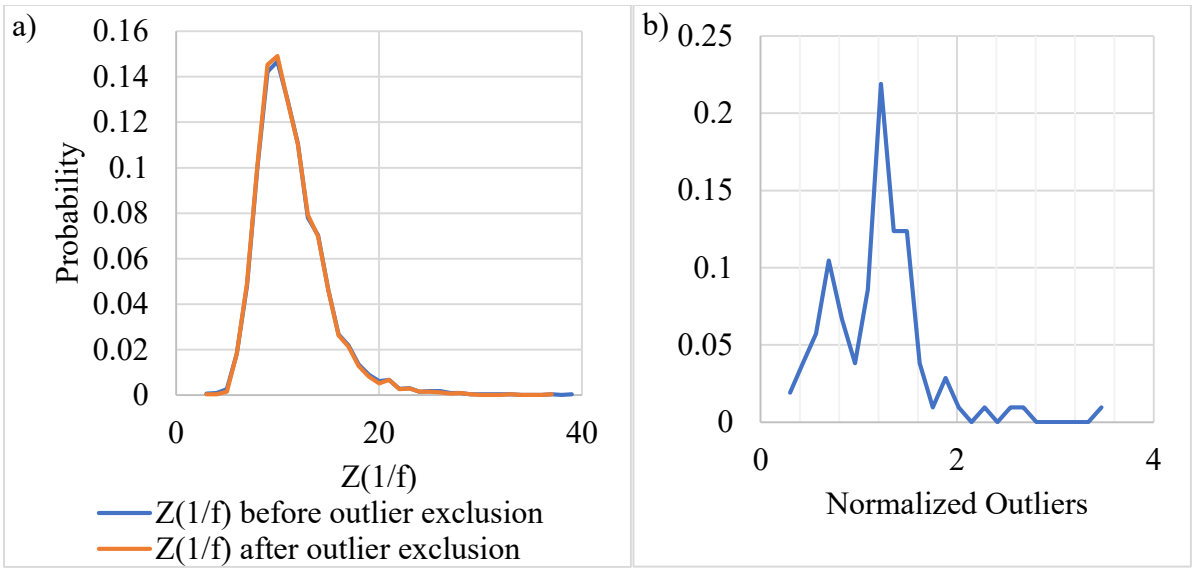


Figure 5: Probability distribution of a) all collected measurements from the CHILD5Y data set, before and after the detection and rejection of outliers using the Grubbs test (n=100), and b) the detected outliers, normalized to the mean of each test after outlier

3.1.6.2 Generating the True Impedance Values

As introduced earlier, the mean *Rrs* and *Xrs* values were used from the five artifact-free measurements in each test from CHILD5Y (n = 209) to provide the true values for the model. Different amounts of randomly chosen outliers were then added as described in the next section, and this represented the simulated data with artifacts. The accuracy was then calculated from the %RMSE between the simulated data with noise added and the true values without noise. Each simulated test has a true *Rrs* (R_T) and *Xrs* (X_T) values which were taken to be the mean *Rrs* and *Xrs* values from the five artifact-free measurements. Hence, without outliers, the %RMSE value would be zero, which can never be the true accuracy of an oscillometry test from multiple individual measurements unless an infinite number of measurements was collected. This is because of the normal variation of human impedance which is largely from breathing. Therefore, using the mean of the five measurements as the truth leads to an artificially low estimate of the %RMSE which is not a realistic simulation of the real-world situation. To overcome this

limitation, it was decided to add some realistic noise to the modeled R_T and X_T that is meant to match the within test variability observed in CHILD5Y. To do that, the within test variability was first estimated in the modeled tests in both *Rrs* and *Xrs*, normalized to the mean impedance *Zrs* for each test. This would be from SD of the five artifact-free measurements from each test normalized to each respective mean *Zrs* within a test, at each frequency. This is because the variability of *Rrs* and *Xrs* tends to be very nearly proportional to the mean impedance [52], but is still different from subject to subject. Hence, this provided a distribution of normalized SD for the within test variability of *Rrs* (R_{vwp}) and *Xrs* (X_{vwp}) at each frequency *k* as illustrated in Equations 7 and 8 for *Rrs* and *Xrs*, respectively.

$$R_{vwp}(k) = \frac{SD(R_{rs}(k))}{\mu(|Z_{rs}(k)|)} \qquad (7)$$

$$X_{vwp}(k) = \frac{SD(X_{rs}(k))}{\mu(|Z_{rs}(k)|)} \qquad (8)$$

With this, two randomly chosen values can be produced from a normal distribution with a mean of zero and SD of one, R_{rn} for *Rrs* and X_{rn} for *Xrs*. It was decided that using a normal distribution is sufficient as this is just used as a random number generator to add the simulated variation based on the SD of the within test variability. Since only one random value was used, R_{rn} for *Rrs* and one X_{rn} for *Xrs*, this was implemented similarly for all frequencies without adding additional noise across frequencies. Therefore, multiplying these random numbers by the normalized SD for the within test variability of *Rrs* and *Xrs* created noise across frequency for *Rrs* (R_n) and *Xrs* (X_n) and provided realistic sampling (Equations 9 and 10).

$$R_n(k) = R_{rn}R_{vwp}(k) \qquad (9)$$

$$X_n(k) = X_{rn}X_{vwp}(k) \qquad (10)$$

This noise was then correctly re-scaled before finally adding it to R_T and X_T at frequency k as shown in Equations 11 and 12. This generated a modeled test resistance, R_m and reactance X_m, which is shifted either up or down according to a SD randomly chosen from a distribution of within patient variabilities defined by R_{vwp} and X_{vwp}. As such, settings R_{rn} and X_{rn} to zero allow for the verification of the modeled patient tests by checking if they were reproducing R_T and X_T obtained from the five artifact-free measurements.

$$R_m(k) = R_T(k) + R_T(k) \cdot R_n(k) \qquad (11)$$

$$X_m(k) = X_T(k) + X_T(k) \cdot X_n(k) \qquad (12)$$

3.1.6.3 Combining Values and Generating the Model Data

An outlier is introduced to a list of randomly selected patient tests from the developed model to represent artifacts as described by Equations 13 and 14,

$$R_o(k) = \mu(R_m(k)) + \mu(R_m(k)) \cdot R_{nor}(k) \qquad (13)$$

$$X_o(k) = \mu(X_m(k)) + \mu(X_m(k)) \cdot X_{nor}(k) \qquad (14)$$

Where R_{nor} and X_{nor} are the normalized resistance and reactance values obtained from the artifact distributions in Figure 5. R_{nor} and X_{nor} were unnormalized to the mean values of each patient at each frequency k, by multiplication, and then added to the mean values R_m and X_m giving R_o and X_o which are the resistance and reactance values of the introduced outliers. In the

CHILD5Y data set, 17% of the patient tests contain outliers, whereas 16% and 19% of the patient test contain outliers in WESER and WIC, respectively. To explore the accuracy over a wider range, outliers were introduced to 0-50% of patient tests. Two scenarios were modeled, where each represented a model patient test and contained a total of five measurements. The first scenario was an artifact free patient test, in which all the five measurements are obtained from R_m and X_m. In the second scenario, an outlier is introduced and thus, four artifact-free measurements are randomly selected from R_m and X_m to represent physiological variability, whereas the fifth measurement is replaced with an artifact represented by R_o and X_o.

3.1.6.4 Generating the Modeled Data for AX and R5-19

The effect of the proposed spectral QC algorithm on the accuracy of outcome measures R5-19 and AX was also studied. The same process was followed to model AX as shown in Equations 15 to 18 which are similar to the previous Equations for *Rrs* and *Xrs* (Equations 7 to 14 described above).

$$AX_{vwp} = SD(AX) \quad (15)$$

$$AX_n = AX_{rn} AX_{vwp} \quad (16)$$

$$AX_m = AX_T + AX_n \quad (17)$$

$$AX_o = AX_m + AX_m \cdot AX_{nor} \quad (18)$$

Where AX_{vwp} is the within patient test variability, AX_n is the normalized noise, AX_m is AX of the modeled patient test, and AX_o is the AX value of introduced outliers. As shown in Equation 15, SD of the multiple measurements was used instead of using a normalized SD to the mean, as the mean AX can be small and lead to large variability despite low absolute variability

which could lead to an artificially high variability. On the other hand, R5-19 is a difference between resistance values and was modeled by taking the difference between the modeled resistance values (R_m and R_o) at 5 and 19 Hz as illustrated in Equations 19 and 20.

$$R5 - 19_m = R_m(5) - R_m(19) \qquad (19)$$

$$R5 - 19_o = R_o(5) - R_o(19) \qquad (20)$$

3.1.6.5 Analyzing Model Performance Using QC Algorithms

As mentioned earlier, instead of calculating individual accuracies for the *Rrs* and *Xrs* at each frequency, the RMSE across all frequencies was provided to give a single metric of accuracy for a particular outcome variable. Therefore, the accuracy of the spectral QC algorithms from the modelling was assessed by calculating the %RMSE values of the mean and median true and predicted *Zrs(1/f)*. Both the mean and the median were examined, since the median is thought to be a better representation of the accuracy as the modeled data is closer to a log normal distribution than it is to a normal distribution. This is illustrated in Equation 21, where $Z_T(^1/_f)$ is the true weighted impedance cost function value (*truth*) obtained using the mean of R_T and X_T from the CHILD5Y data set, and $Z_P(^1/_f)$ is the predicted weighted impedance cost function values obtained using the mean of R_P and X_P either with no QC or after applying spectral QC. R_P and X_P are the predicted resistance and reactance values, represented by R_m and X_m when no outliers are introduced and represented by both R_m and X_m as well as R_o and X_o when introducing outliers to modeled tests.

$$Median\ \%RMSE_{Z(1/f)} = Med\left(\sqrt{\left(\frac{Z(1/f)_T - Z\left(1/f\right)_P}{Z\left(1/f\right)_T} \right)^2} \times 100\% \right) \quad (21)$$

The accuracy was also assessed for the key measures R5-19 and AX. However, since R5-19 and AX can have values close to zero as described previously, and thus can lead to an artificially high %RMSE values, the actual RMSE values were used instead as illustrated in Equations 22 and 22.

$$Median\ RMSE_{R5-19} = \quad Med\ (\sqrt{((R5-19_T) - (R5-19_P))^2}) \quad (22)$$

$$Median\ RMSE_{AX} = \quad Med(\sqrt{(AX_T - AX_P)^2}) \quad (23)$$

Therefore, the accuracy obtained after applying the *Early* algorithm was compared to no QC using four different scenarios: 1) No QA – ALL, where the %RMSE is calculated using all modeled measurements before applying QC, 2) No QA- First 3, where the %RMSE is calculated using the first three modeled measurements before applying QC, 3) *Early*, where the %RMSE is calculated using the three selected measurements obtained after applying the *Early* QC algorithm, and 4) No QA- Stop After Meeting CV Criteria (SAMC), where the %RMSE is calculated using the first nth measurements required by the *Early* algorithm to meet the acceptability criteria (ζ_R and $\zeta_Z \leq 15\%$). No QA-SAMC is used to replicate the most common clinical scenario in which operators stop collecting oscillometry measurements once the CV criteria is met. Finally, the accuracy of the *Early* algorithm was also compared to a common QC method of detecting and rejecting outliers based on two standard deviations (QA-2SD).

3.2 Results

As discussed in subsection 3.1.2, the *Early* algorithm was identified as the most efficient and suitable method representative for the clinical setting due to its fast performance. Thus, a spectral QC algorithm was developed and optimized using the *Early* algorithm. The following subsections present the key results obtained from implementing this algorithm, focusing on the performance measures outlined in section 3.1.

3.2.1 Repeatability

There are many different combinations of *Rrs* and *Zrs* CVs, ζ_R and ζ_Z respectively, that were investigated as illustrated in appendix A (Table A1). Here, ζ_R (Equation 3) and ζ_Z (Equation 4) were used as cost functions that capture the variability between repeated oscillometry measurements using averages of resistance and impedance CVs over the measured frequencies. Hence, $\zeta_R = CV(R5 + R11 + R19)$, for example, would represent the average resistance CVs at 5, 11 and 19 Hz, whereas $\zeta_Z = CV(|Z5| + |Z11| + |Z19|)$ would represent the average impedance CVs at 5, 11 and 19 Hz. The first approach was to analyze the variability using $\zeta_R = CV(R5)$, which is the standard measure that is used to calculate and optimize the CV. Figure 6 shows CV of impedances versus oscillation frequency. The graph demonstrates the impact of quality control measures on the variability of impedance before and after applying the proposed spectral QC algorithm. The data compares the CV before implementing QC from all obtained measurements (black solid line) and using the first three measurements (black dashed line) to the CV after applying the proposed QC algorithm, where each color represents the CV from combinations of *Rrs* from different frequencies.

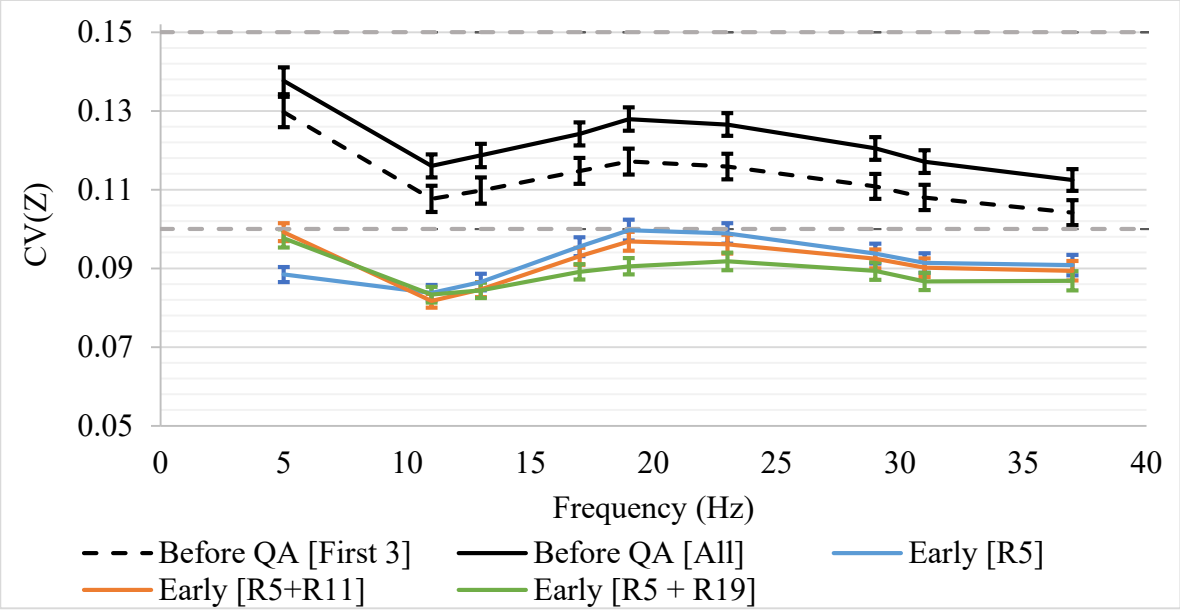


Figure 6: A plot of the mean CV(Z) vs. frequencies from the CHILD 5Y data set before and after spectral QC using ζ_R = CV(R5), CV(R5+R11), and CV(R5+R19), (Bars represent standard error).

The use of ζ_R = CV(R5) with the proposed spectral QC algorithm resulted in the lowest variability at 5 Hz but a relatively higher variability at all the higher frequencies. When adding R11 to optimize and minimize the average CV of R5 and R11 (ζ_R = CV(R5+R11)), the results tended to appear slightly improved at the higher frequencies but not significantly ($p>0.05$), at a cost of higher impedance CV at 5Hz, Figure 6. There was also higher variability assessed by the average normalized SD of R5 ($P<0.05$), R5-19 ($p<0.05$), X5 ($p>0.05$) and the average SD of AX($p>0.05$) compared to the use of ζ_R = CV(R5) as illustrated in Table 2. Important measurements from oscillometry include R19 or R20 due to the increased importance of R5-R20 as an index of heterogeneity attributed to the small airways, and as a measure of early lung disease. Some oscillometry devices are also limited to maximum frequency values between 20 and 40 Hz. Therefore, minimizing the variability using a cost function that included R5 and R19 was also investigated. Using the average CV of R5 and R19 (ζ_R = CV(R5+R19)) maintained the

same variability at R5, R5-19, X5 and yielded a similar RMS Average CV(Z) ($p>0.05$) when compared to the standard R5, with a slightly higher variability at AX ($p>0.05$), but still lower than ζ_R = CV(R5+R11) (Figure 6, Table 2).

Table 2: Summary of key performance measured used to assess the different combinations of R and |Z| CVs, obtained for ζ_R = CV of R5 alone, the average CV of R5 and R11, the average CV of R5 and R19, CV of |Z5| alone, the average CV of |Z5| and |Z19| and a weighted impedance (Z(1/F)).

ζ_R and ζ_Z		RMS Average CV(|Z|)	Average SD(R5)/ μ(|Z5|)	Average SD(R5-19)/ μ(|Z5|)	Average SD(X5)/ μ(|Z5|)	Average SD(AX)
CV(R5)	No QA [First 3]	0.12	0.12	0.09	0.1	16.32
	No QA [All]	0.13	0.13	0.09	0.1	16.79
	Early	**0.1**	**0.08**	**0.08**	**0.09**	**13.62**
	Variant	0.09	0.07	0.07	0.08	12.95
	LowestCV	0.09	0.04	0.07	0.08	12.58
CV(R5+R11)	No QA [First 3]	0.12	0.12	0.09	0.1	16.32
	No QA [All]	0.13	0.13	0.09	0.1	16.79
	Early	**0.1**	**0.09**	**0.08**	**0.09**	**14.02**
	Variant	0.09	0.08	0.08	0.09	12.91
	LowestCV	0.07	0.05	0.07	0.08	11.75
CV(R5+R19)	No QA [First 3]	0.12	0.12	0.09	0.1	16.32
	No QA [All]	0.13	0.13	0.09	0.1	16.79
	Early	**0.1**	**0.09**	**0.08**	**0.09**	**13.99**
	Variant	0.09	0.08	0.08	0.09	13.29
	LowestCV	0.07	0.06	0.06	0.08	12.59
CV(Z5)	No QA [First 3]	0.12	0.12	0.09	0.1	16.32
	No QA [All]	0.13	0.13	0.09	0.1	16.79
	Early	**0.1**	**0.08**	**0.08**	**0.08**	**12.64**
	Variant	0.09	0.07	0.08	0.08	11.86
	LowestCV	0.09	0.05	0.07	0.07	11.35
CV(Z5+Z19)	No QA [First 3]	0.12	0.12	0.09	0.1	16.32
	No QA [All]	0.13	0.13	0.09	0.1	16.79
	Early	**0.09**	**0.09**	**0.08**	**0.08**	**12.95**
	Variant	0.09	0.08	0.08	0.08	12.16
	LowestCV	0.07	0.06	0.06	0.08	11.25
Z(1/f)	No QA [First 3]	0.12	0.12	0.09	0.1	16.32
	No QA [All]	0.13	0.13	0.09	0.1	16.79

	Early	**0.1**	**0.08**	**0.08**	**0.08**	**12.77**
	Variant	0.09	0.07	0.08	0.08	11.93
	LowestCV	0.08	0.05	0.07	0.08	11.11

Next, the performance of algorithms based on CVs of the *Rrs* were compared to algorithms based on CVs of |*Zrs*| using the same frequency combinations, as this included the variation of *Xrs*. Results were similar, with lower impedance and AX variabilities, assessed by CV(|Z|) and SD(AX), obtained using $\zeta_Z = \text{CV}(|Z5|)$ and $\zeta_Z = \text{CV}(|Z5|+|Z19|)$ compared to $\zeta_R = \text{CV}(R5)$ *and* $\zeta_R = \text{CV}(R5+R19)$, respectively (Figure 7). Although the use of Rrs based algorithms appeared to lead to lower variability at R5 and R5-19, which exclusively depend on *Rrs*, this was not significant (p>0.05). Interestingly, it was discovered that $\zeta_R = \text{CV}(R5)$ or $\zeta_Z = \text{CV}(|Z5|)$ yielded the lowest variabilities in all of the key performance measures, highlighting the significant contribution of the lowest oscillation frequency (5Hz) to the variability of R5, R5-19, X5 and AX, which are principal outcome measures. Therefore, based on these results, different combinations of CVs based on *Zrs* were then investigated for their effect in decreasing the variability amongst R5, R5-19, X5 and AX with a focus on frequency dependent weights with strongly weighted |Z5| ($\zeta_R = Z(1/f)$).

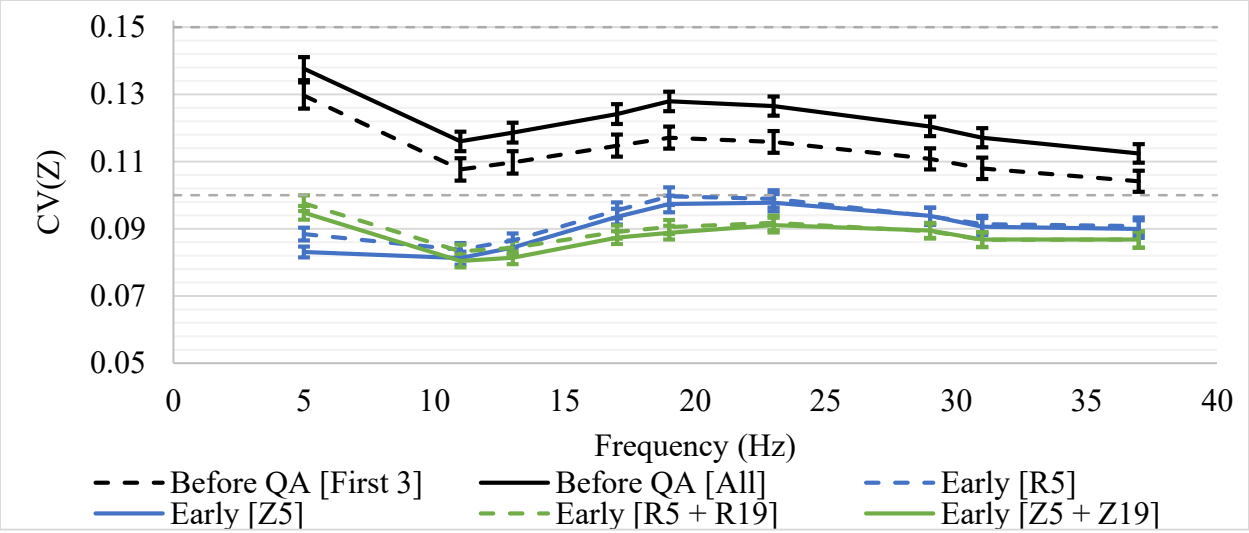


Figure 7: A plot of the mean CV(|Z|) vs. frequencies from the CHILD 5Y data set before and after spectral QC using ζ_R = CV(R5) and CV(R5+R19) , ζ_Z = CV(|Z5|) and CV(|Z5|+|Z19|), (Bars represent standard error).

As illustrated in Appendix A, Table A1, multiple combinations of weighted |*Zrs*| CVs were investigated, contributing more weight and thus higher significance to 5Hz. Amongst the investigated combinations, good improvements were achieved in R5, R5-19, X5 and AX when using a 1/f weighted CVs of impedance $\zeta_Z = Z(1/f)$. This is shown in Figure 3.5 which compares no QC to the standard R5 and two very good combinations based on CVs of *Zrs*, $\zeta_Z = |Z5|+|Z19|$ *and* $\zeta_Z = Z(1/f)$. The improved 1/f dependent cost function was also justified by the fact that the key outcome measure AX has an inverse correlation with the frequency of oscillation as demonstrated in the impedance versus frequency curve in Figure 8. This may also be useful due to the fact that in disease, resistance exhibits an inverse frequency dependence, thus weighting the cost function to minimize variability where the magnitudes of Resistance and Reactance are maximal appears to be useful for minimizing variability across frequencies. These mentioned relationships were used as a foundation to derive a novel cost function, $\zeta_Z = Z(1/f)$, which provided the best overall performance at the key outcome measures. $\zeta_Z = Z(1/f)$ provided a

significantly lower variability ($p<0.05$) at all the outcome measures compared to $\zeta_Z = |Z5|+|Z19|$, as more weight is assigned to the lowest frequency 5Hz. Nevertheless, while significant improvements were obtained for the RMS average CV(|Z|), SD(X5)/μ(|Z5|) and SD(AX) compared to the standard $\zeta_R = R5$ ($p<0.05$), no significant improvements ($p>0.05$) were achieved for SD(R5)/μ(|Z5|) and SD(R5-19)/μ(|Z5|).

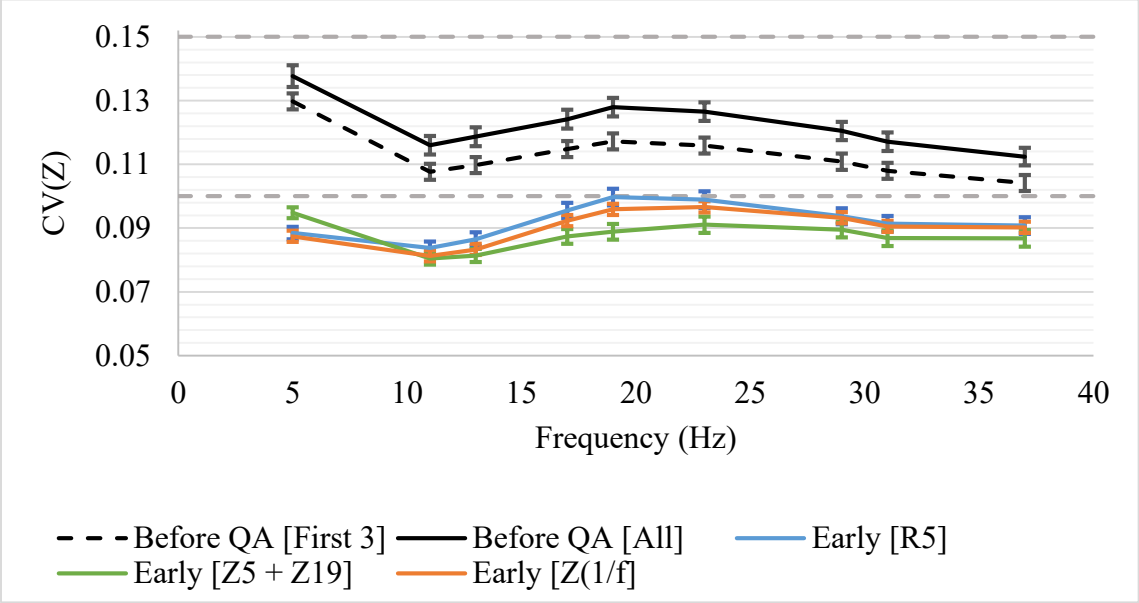


Figure 8: A plot of the mean CV(|Z|) vs. frequencies from the CHILD 5Y data set before and after spectral QC using $\zeta_R = R5$ and $\zeta_Z = |Z5|+|Z19|$, and Z(1/f) , (Bars represent standard error).

The effect of the different algorithms on mean *Rrs* and *Xrs* values across subjects was also examined. It was found that while the QC algorithms reduced the variability in individual subjects, this did not appear to affect the means across subjects (Figure 9). Specifically, *Rrs* was not altered ($p > 0.05$), and while the difference in *Xrs* is not visible, it nevertheless was significant ($0.001<p<0.05$). This significant difference in *Xrs* is likely due to the very large numbers of subject and in fact was well below group or individual variability. Overall, this

indicates, perhaps surprisingly, that while QC may be important for an individual, for large studies QC might have no effect on the mean values from oscillometry. This was also found using the WIC and WESER data sets, where $\zeta_Z = Z(1/f)$ optimized the CV across all frequencies and in fact resulted in lower CVs compared to the manual QC from the WESER data set (Figure 10), with minimal effect on the mean *Rrs* and *Xrs* values (Appendix A, Figures A1 and A2). However, QC did have substantial effects on individual outcomes despite the small change in means. The *Early* algorithm resulted in a change of more than one $cmH_2O/L/s$ in R5 for 48%, 22% and 28% of subjects and a change of more than one $cmH_2O/L/s$ in X5 for 24%, 39% and 20% of subjects in CHILD5Y, WIC and WESER, respectively.

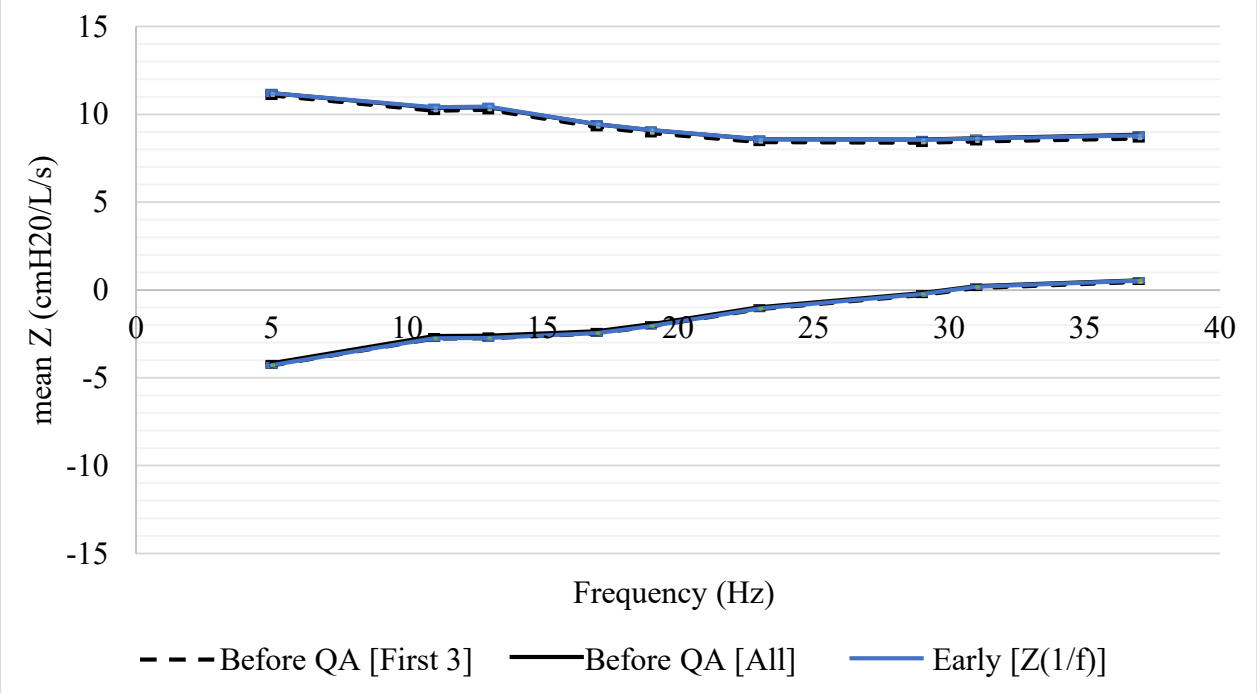


Figure 9: A plot of the mean resistances (top) and reactances (bottom) vs. frequency from the CHILD 5Y data set before and after spectral QC using $\zeta_Z = Z(1/f)$, (Bars represent standard error).

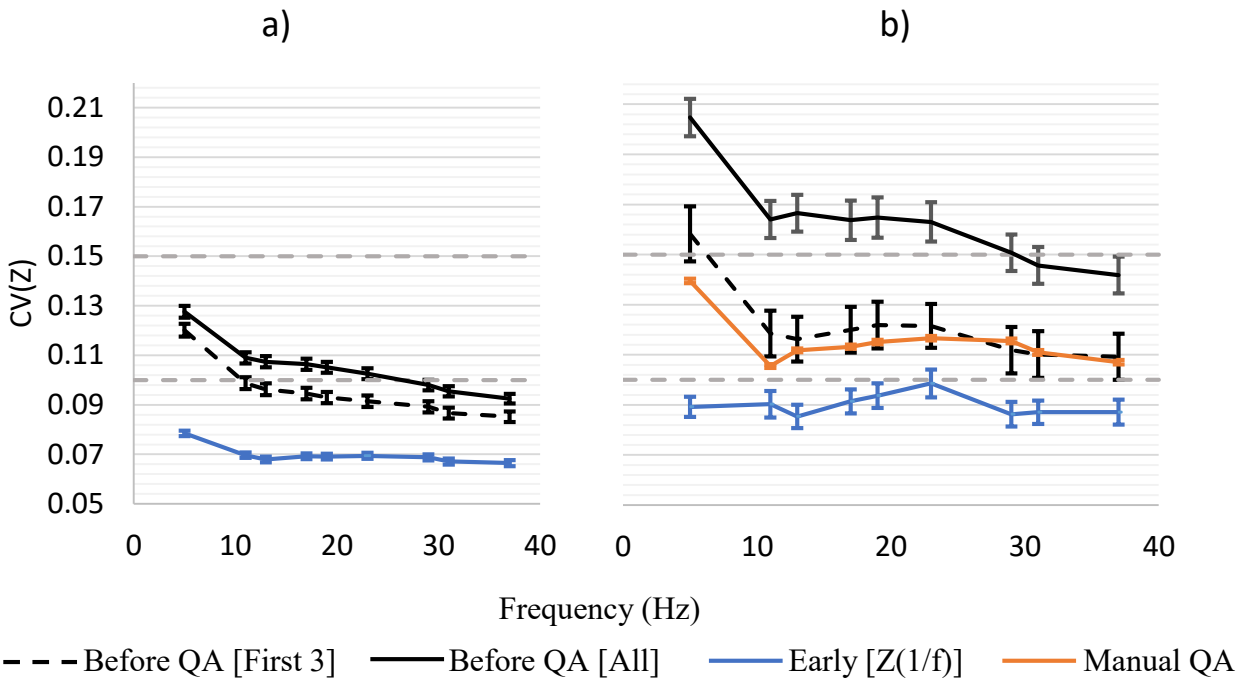


Figure 10: A plot of the mean CV(|Z|) vs. frequencies from the a) WIC and b) WESER data sets before and after spectral QC using ζ_Z = Z(1/f), (Bars represent standard error).

3.2.2 Efficiency

Once the repeatability and variability were optimized using ζ_Z = *Z(1/f)*, the efficiency was calculated and compared to no QC. It was found that the proposed spectral QC algorithm provided a significant ($p<0.05$) improvement in efficiency as it lowered the number of required measurements to meet the CV criteria from 5.4±1.7, 4.5±2.3, 8.8±3.6 to 3.7±0.9, 3.3±0.6 and 3.2±0.4 in CHILD5Y, WIC, and WESER, respectively (Table 3). The proposed spectral QC algorithm also outperformed the manual QC performed by a well-trained operator, reducing the number of required measurements from 4.4±1.9 to 3.2±0.4 ($p<0.05$).

Table 3: The number of required measurements (mean(SD)) to achieve a CV ≤ 0.15 using ζ_Z = Z(1/f) for CHILD5Y, WIC, and WESER data sets.

	QA	NO QA	MANUAL QA
CHILD5Y	3.7 (0.9)	5.4 (1.7)	3.8 (1.6)
WIC	3.3 (0.6)	4.5 (2.3)	4.2 (1.7)
WESER	3.2 (0.4)	8.8 (3.6)	4.4 (1.9)

3.2.3 Feasibility

Figure 11 illustrates the feasibility for selecting a minimum of three valid, artifact-free measurements, represented by the percentage feasibility using the seven distinct methods utilized for measurement selection. Each bar represents a different method for measurement selection, with the number of acceptable tests, achieving $\zeta_Z \leq 15\%$, indicated on the top of each bar. To aid interpretation, three arbitrary reference lines were used to give acceptability thresholds for 50%, 75%, and 90% or more of the subjects. Without applying any QC algorithms, there was an apparent slight decrease in feasibility in CHILD5Y with increasing number of measurements, possibly due to increasing probability of artifactual measurements (Figure 11). On the other hand, applying the proposed QC algorithm (*Early*) on more measurements demonstrated an increase in feasibility. This behaviour was comparable and consistent across all the investigated combinations of *Rrs* and |*Zrs*| CVs.

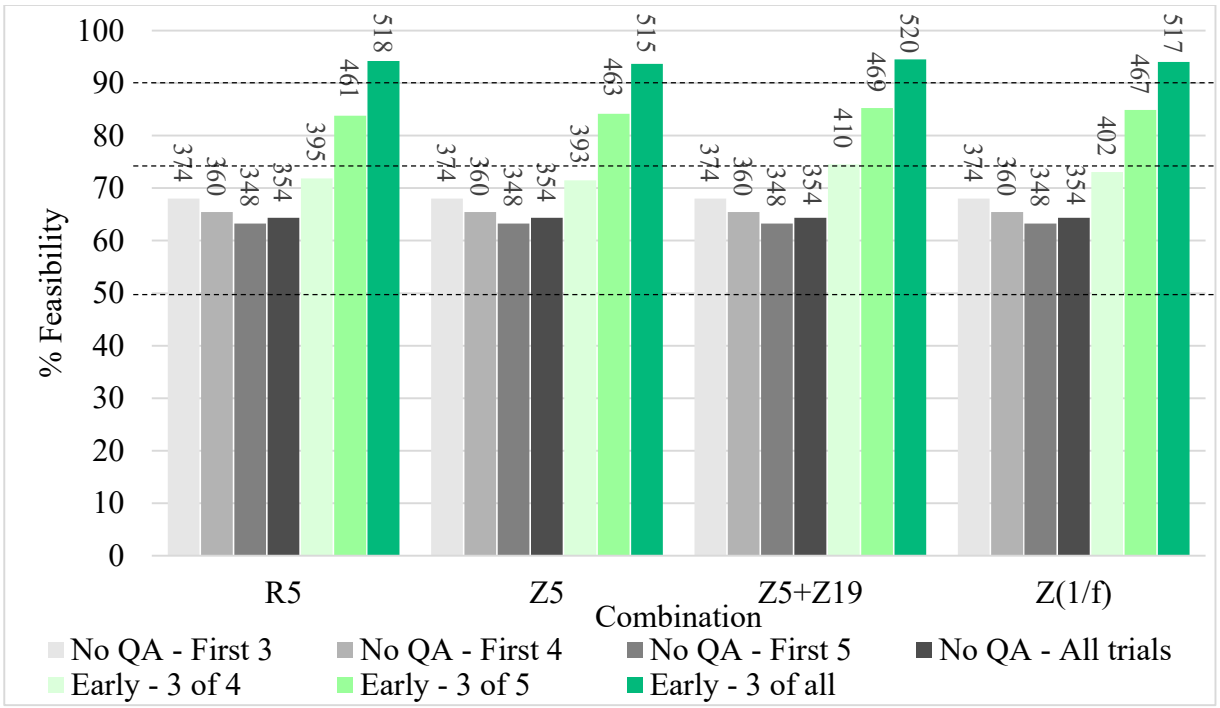


Figure 11: Percent feasibility achieved using ζ_R = CV(R5), ζ_Z = CV(|Z5|), CV(|Z5|+|Z19|), and Z(1/f). Numbers on top of each bar represent the number of feasible tests.

Using the same ζ_Z = Z(1/f) weighting function across frequencies to minimize CV, no QC from all the selection methods resulted in a feasibility ranging between 60% and 70% for CHILD5Y and WIC, with lower feasibility ranging between 20% and 60% with WESER (Figure 12). Interestingly, adding more measurements to WESER resulted in a more remarkable decrease in feasibility with no QC, decreasing from 60% using the first three measurements to 21% when using all measurements, likely due to the nature of this data set (see discussion, page 60). On the other hand, applying the proposed QC algorithm resulted in an average of 75% feasibility when only using the first four measurements across all the three data sets, outperforming the highest feasibility achieved with no QC. This feasibility is further improved, achieving 95% feasibility in CHILD5Y, 82% in WIC and 91% in WESER when using all the available measurements, which is higher than the 69%, 70% and 53% feasibility achieved with manual QC in CHILD5Y, WIC and WESER, respectively.

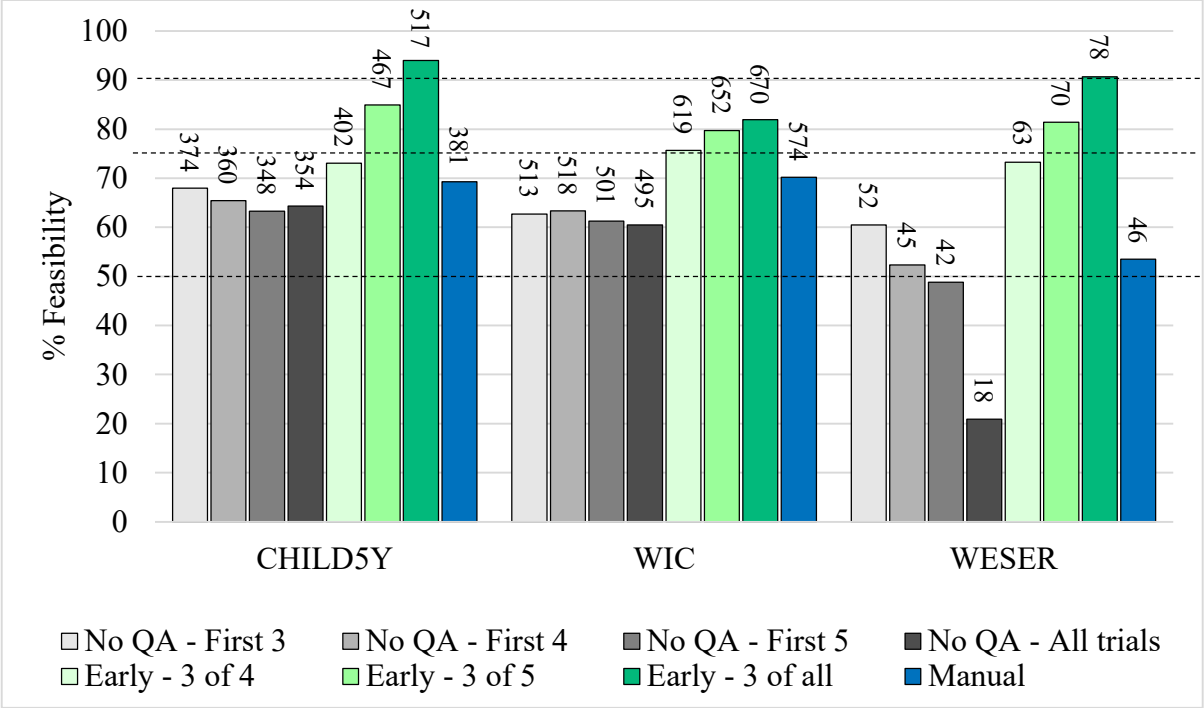


Figure 12: Percent feasibility achieved using $\zeta_Z = Z(1/f)$ in CHILD5Y, WIC and WESER compared to manual QC. Numbers on top of each bar represent the number of feasible tests.

3.2.4 Accuracy

Both the mean and median %RMSE curves in Figure 13 demonstrated similar results, with the *Early* algorithm achieving higher accuracy than no QA when introducing outliers to more than 20% of modeled tests. There was no significant difference between *Early* and No QA-All ($p>0.05$). However, a significant difference was observed between *Early* and No QA- SAMC (recall that SAMC is Stop After Meeting CV criteria) ($p<0.05$) when introducing outliers to more than 25% of modeled tests. A significant difference was also obtained between *Early* and No-QA-First 3, ($p<0.05$) when introducing outliers to more than 10% of the modeled test.

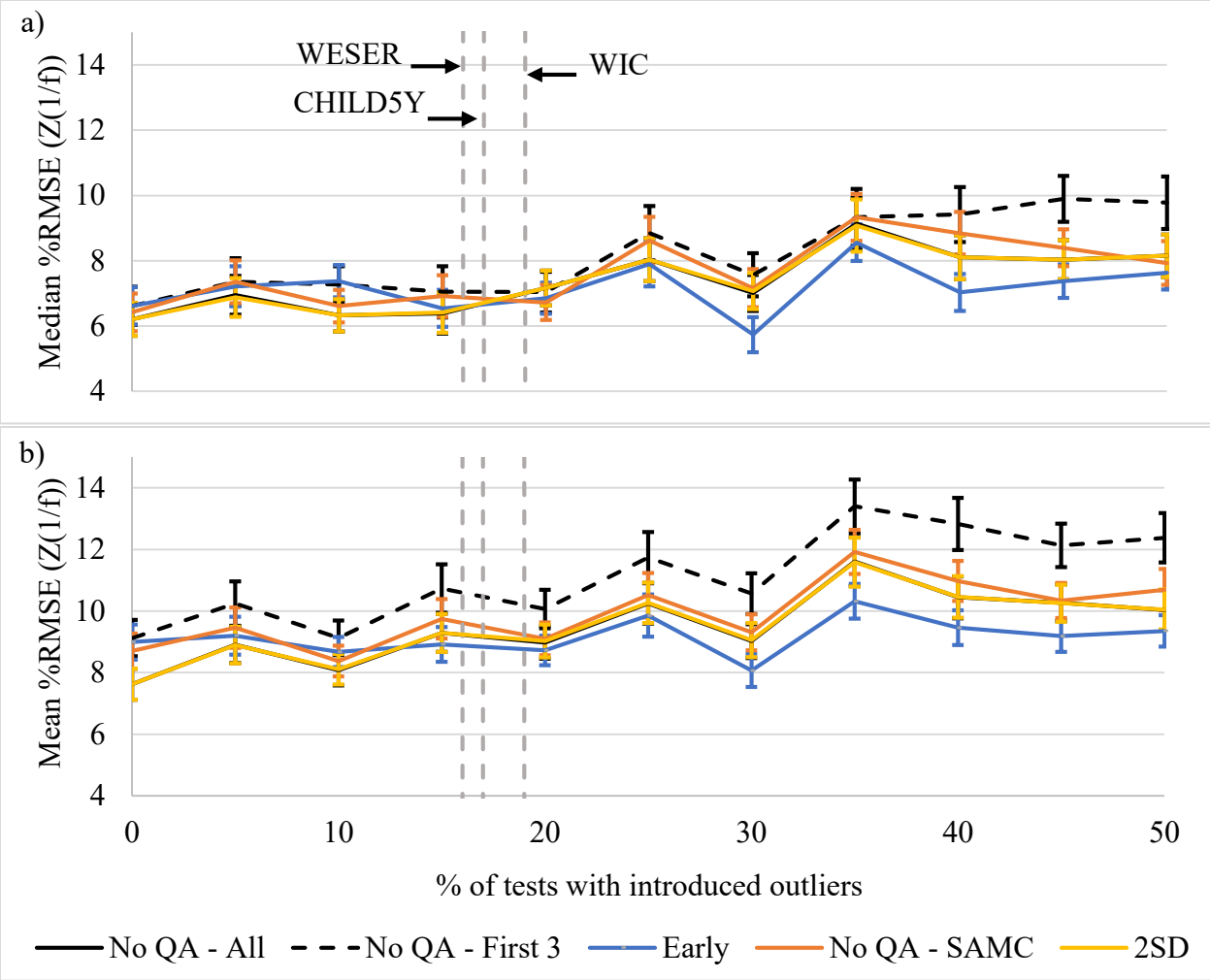


Figure 13: Plot of a) median and b) mean % RMSE values calculated after introducing outliers to different percentages of the modeled tests. Error bars represent the standard error whereas the three vertical dashed lines represent the percentage of subject tests containing outliers in the WESER (16%), CHILD5Y (17%), and WIC (19%) data sets, serving as reference points.

Figure 13 also compares the median and mean %RMSE obtained using a QC algorithm based on detecting and rejecting outliers that are 2 SDs from the mean. This technique is commonly used in most oscillometric software. For example, it is used in the tremoflo software (Thorasys) to reject outliers within a recorded measurement. However, it not used for within-test variability from multiple measurements. The comparison in Figure 13a shows that the %RMSE was low for both approaches, with no significant difference between the median %RMSE

obtained using *Early* and 2SD (P>0.05), regardless of the percentage of tests with introduced outliers. Similar results were obtained for the mean %RMSE in Figure 13b, with better accuracy (lower %RMSE) achieved using the *Early* algorithm compared to QA-2SD (p>0.05).

Although not significantly different, the *Early* algorithm in Figures 13 (in blue) appeared to have slightly better performance when introducing outliers to more than 15% of modeled tests. Thus, the *Early* algorithm, on average, maintained or appeared to possibly enhance the accuracy, compared to no QA as observed in Figure 13, and performed well compared to an SD based approach especially when dealing with outliers. Overall, as the percentage of tests with introduced outliers increased, the *Early* algorithm generally maintained the accuracy of the data set. It should be noted that despite maintaining accuracy across subjects, the *Early* algorithm had substantial effects on individual subjects. As previously noted, it did not change the mean *Rrs* values in measured data but caused a small, significant decrease in *Xrs* (Figure 9 in section 3.2.1). Similar results were also found using the modelled data. Specifically, the *Early* algorithm resulted in a change of more than one $cmH_2O/L/s$ in R5 for 42% of modeled subjects and a change of more than one $cmH_2O/L/s$ in X5 for 16% of modeled subjects. Similarly, when applied to the CHILD5Y data set, R5 changed by more than one $cmH_2O/L/s$ for 48% of subjects and more than one $cmH_2O/L/s$ in X5 for 24% of subjects.

Finally, the median and mean RMSE before and after QC were compared for two key oscillometry outcome measures, R5-19 and AX. It was observed that No-QA and QA-2SD achieved comparable accuracy for R5-19 regardless of the percentage of subjects with introduced outliers (p>0.05) (Figure 14). Compared to the *Early* algorithm, No-QA and QA-2SD achieved a higher accuracy for R5-19 when introducing outliers to less than 30% of the modeled tests (p<0.05). However, the accuracy improvement for No QA-All and QA-2SD was not significant

when introducing outliers to 30% or more of the modeled tests ($p>0.05$). Moreover, the accuracy of the *Early* algorithm was comparable to No QA-First 3 and No QA-SAMC when introducing outliers to less than 25% of the modeled tests and was slightly improved when introducing outliers to more than 25% of the modeled tests ($p>0.05$). It was also observed that, on average, the mean and median RMSE values did not illustrate any differences in the accuracy of AX between the five different investigated methods; No QA-All, No QA-First 3, No QA-SAMC, QA-2SD and *Early* (Figure 15).

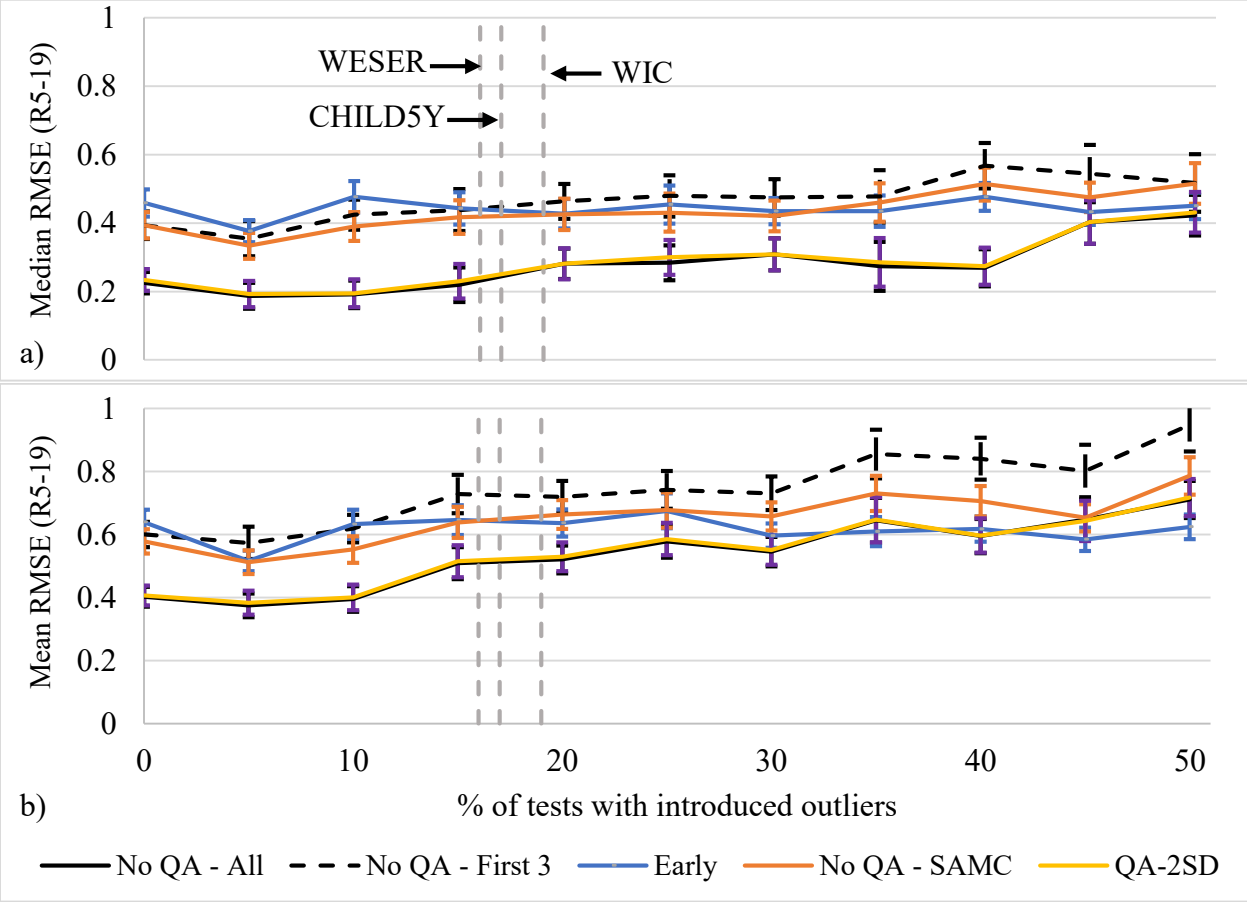


Figure 14: Plot of a) median and b) mean RMSE of R5-19, calculated after introducing outliers to different percentages of the modeled tests. Bars represent standard error.

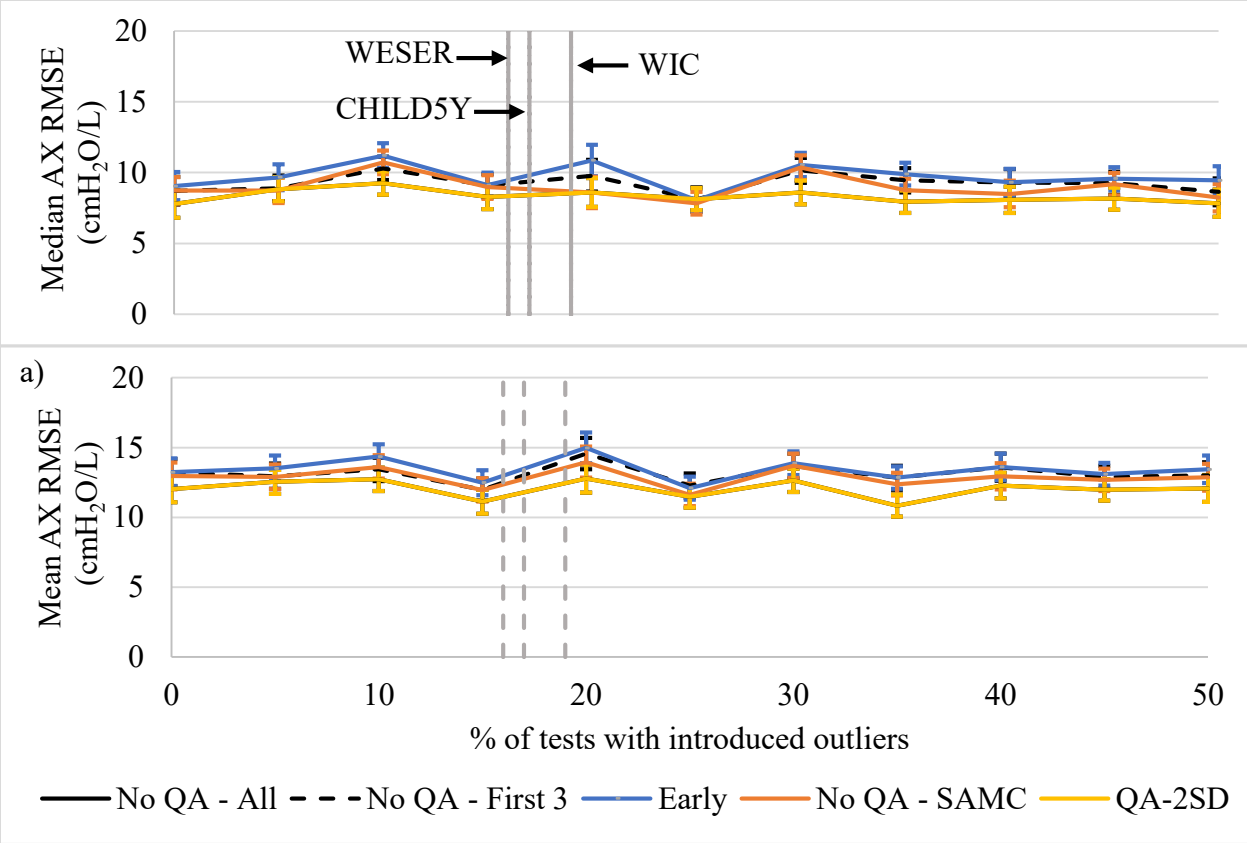


Figure 15: Plot of a) median and b) mean RMSE of AX, calculated after introducing outliers to different percentages of the modeled tests. Bars represent standard error.

3.3 Discussion

The objective of this study was to develop an improved version of the current ERS quality control recommendations to improve the repeatability, efficiency, feasibility, and accuracy of oscillometry. The principle findings are as follows: 1) using the proposed *Early* algorithm with a weighted impedance cost function, $\zeta_Z = Z(1/f)$, improved the repeatability in both Rrs and Xrs and across multiple frequencies, 2) $\zeta_Z = Z(1/f)$ improved the efficiency by reducing the number of required measurements to meet the acceptability criteria, 3) $\zeta_Z = Z(1/f)$ demonstrated significant improvements in feasibility, outperforming manual QC, and 4) the

proposed QC maintained the accuracy of the important and clinically significant oscillometry measures, R5-19 and AX.

3.3.1 Repeatability

It was found that using a weighted impedance cost function, $\zeta_Z = Z(1/f)$, performed better than the current ERS recommendation of R5 alone, improving the reported repeatability in both *Rrs* and *Xrs* at multiple frequencies. Similar to Therkon et al., it was found that optimizing the variability for a particular outcome measure yields the best variability for that very measure. For instance, Therkorn et al. found that optimizing the variability using CV(R5) improved the repeatability of R5. However, this was at the cost of a reduction in the repeatability of other key measures, including R19, R5-19, X5 and AX. Hence, Therkon et al. suggested that a variability criteria should be tailored to the oscillometric outcomes that are of most clinical interest [21]. Additionally, as emphasized by Hantos et al., X5 can exhibit greater variability than R5, highlighting the significance of incorporating the reactance in the repeatability assessment, specifically at the lowest frequency oscillations [43]. In our subjects, we found that using the weighted cost function provided improved (lowest) variability across frequencies, and thus it is not required to tailor criteria for particular oscillometric outcomes.

The choice of utilizing $\zeta_Z = Z(1/f)$ to evaluate and optimize the within-test repeatability aims to follow these suggestions while also overcoming the limitations of using R5. In fact, using $\zeta_Z = Z(1/f)$ improved the repeatability across oscillating frequencies and in all important outcome measures, including R5, R5-19, X5 and AX. Repeatability was improved significantly for X5 and AX, likely due to the inclusion of *Xrs* in the $\zeta_Z = Z(1/f)$, which cost functions based on *Rrs* alone do not have. Such improvements in repeatability are crucial, particularly in clinical settings where consistent and reliable measurements are vital for accurate diagnosis and

monitoring. Importantly, improvements in repeatability led to minimal changes in resistance and reactance means across subjects ($p>0.05$). These findings were consistent when $\zeta_Z = Z(1/f)$ was applied to WIC and WESER data sets. Improving the repeatability while introducing minimal changes to the mean values across subjects, as we found (section 3.2.1), implies that for clinical studies, (e.g. evaluating drug effectiveness or changes in treatment), that the number of required subjects is reduced to achieve the same statistical power when using this cost function approach to quality control, saving both time and costs.

3.3.2 Efficiency

The *Early* algorithm demonstrated a notable reduction in the number of measurements required to meet the acceptability criteria compared with protocols that conservatively collect additional measurements when the reported CV from all measurement is high. Specifically, when compared to no QC, the *Early* algorithm reduced the required measurements from 5.4 ± 1.7 to 3.7 ± 0.9 ($p<0.05$). It also outperformed manual QC conducted by skilled operators, decreasing the required measurements by about 1.2 in CHILD5Y. These findings were consistent in WESER and WIC data sets. This can be useful in clinical and research settings as oscillometry devices can indicate when a test has achieved acceptable repeatability, hence eliminating the need for unnecessary additional measurements. Consequently, this allows for quicker assessments, shorter patient testing times, and improved patient flow.

Similar to repeatability, improving the efficiency could also allow for more subjects to be included in a research trial, enhancing the statistical power of studies. Additionally, providing a method to reliably, and in an automated approach, indicates that the quality standards have been met could potentially save time as subjects would not have to be called back for repeated testing. The improved efficiency, however, is notable in WESER and CHILD5Y, where data is collected

meticulously, taking five measurements when possible and additional measurements in the presence of high variability. Yet, this improvement may not be evident in sites that only conduct three or four measurements and reject tests with a CV less than or equal to 15%, resulting in a small number of measurements at the cost of poor feasibility.

3.3.3 Feasibility

It was also found that feasibility using the *Early* algorithm was improved by approximately 20% compared to no QC when using any number of available measurements. Interestingly, increasing the number of measurements led to a decrease in feasibility with no QC (Figures 11 and 12). This is likely due to the increased presence of outliers or artifactual measurements in subjects that had many measurements, since these were the subjects that the operator observed to have larger variability and thus collected more measurements. This was mitigated by the use of the automated *Early* algorithm, where increasing the number of measurements resulted in a clear improvement in feasibility, achieving a feasibility that is greater than 90% when using $\zeta_Z = Z(1/f)$ with more than five measurements.

These findings were consistent with Harkness et al., as they found that using the current acceptability criteria, CV(R5) ≤ 15% for young children and ≤ 10% for adults, while including more measurements did not result in a significant improvement in feasibility. Yet, similar to the *Early* algorithm, taking more measurements increased the chance of finding three closest measurements with a CV below acceptability threshold, hence improving the feasibility. In fact, Harkness et al. demonstrated that a feasibility of more than 90% can be achieved with three closest measurements when a minimum of four, five and six measurements were collected for healthy, asthmatic and COPD subjects, respectively [53].

Moreover, the *Early* algorithm resulted in a minimum of a 10% increase in feasibility when compared to user exclusion as a form of manual QC in CHILD5Y and WIC data sets. However, it is important to note that no detailed information was provided about the operator's expertise, training, or the rationale for measurement exclusions. They were instructed to exclude measurement for observed artifacts such as cough, or deviations from the measurement protocol such as patient movement, and possibly unusual breathing patterns. While operator detected artifacts such as cough should be rejected, it may be possible that measurements that did not greatly impact the mean impedance values were rejected in some cases.

Additionally, applying the proposed *Early* algorithm to the WESER data set resulted in a 20% increase in feasibility when compared to the application of the manual QC, which was performed by a well-trained operator with documented justifications for excluded measurements. Here the rejection criteria included unusual breathing such as the presence of large breaths, which may not greatly affect the CV. This is a significant improvement in feasibility since the WESER data set consist of preschool children who had presented with wheeze, representing a challenging group with high variability. These results suggest that the *Early* algorithm can provide a more standardized and reliable approach to data quality control, even on more challenging subjects. Improving feasibility may also allow previously ineligible subjects who have difficulty obtaining successful tests, thus making oscillometry applicable to a broader population.

3.3.4 Accuracy

It was found that the *Early* algorithm outperformed No QA (All, First 3 and SAMC) when introducing outliers to over 20% of modeled tests, showcasing its effectiveness. No significant differences in accuracy were found between *Early* and No QA-All, which could be

attributed to the potential benefits of No QA-All's usage of more data for mean estimation. Nevertheless, significant differences in accuracy were found between the *Early* algorithm and the No QA-SAMC and No QA-First 3 scenarios.

The *Early* algorithm was also compared to a common method of detecting and rejecting outliers based on two standard deviations (QA-2SD). Although the *Early* algorithm reduced the %RMSE values when outliers were introduced to more than 18% of modeled data ($p>0.05$), there were no statistically significant differences in the %RMSE between *Early* and QA-2SD, regardless of the percentage of tests with introduced outliers. Nonetheless, $\zeta_Z = Z(1/f)$ has several benefits over the simpler 2SD method. Specifically, $\zeta_Z = Z(1/f)$ is based on Zrs, which effectively incorporates the variability found in both *Rrs* and *Xrs* into a single measure. It also accounts for the variability across multiple frequencies, addressing a challenge faced when using the 2SD method as the current criteria emphasize low frequency, indeed only evaluating R5.

The *Early* algorithm reduced the variability and maintained the accuracy but this perhaps surprisingly introduced a minimal change to the mean *Rrs* and *Xrs* values across subjects. This indicates that the algorithm decreased the reported *Rrs* and *Xrs* values for some subjects while it increased them for other. This is important to individual results. As described earlier in sections 3.2.1 and 3.2.4, changes in *Rrs* and *Xrs* values were important for some subjects as 48% and 24% of subjects had a change that exceeded 1 cmH2O/L/s in R5 and X5 in CHILD5Y, respectively, where a change of more than 1 cm cmH2O/L/s is likely clinically meaningful. These results were consistent for the modeled data, with 42% and 16% of modeled subjects had a change that exceeded 1 cmH2O/L/s in R5 and X5, respectively. This value of 1 cmH2O/L/s is likely clinically meaningful, but was chosen in the absence of an established minimally important clinical difference (MICD) for normal oscillometry. Thus applying QC and in particular using

the utility of the *Early* algorithm in these subjects may be thus important, as such changes may be comparable to worsening or improvement of disease over time.

3.3.4.1 R5-19 and AX

The accuracy of the key oscillometry measures R5-19 and AX was also compared. The mean and median RMSE results for R5-19 were similar, likely due to the low variability in R5-19 that is found in this modeled group, which is derived from CHILD5Y. With more challenging data, the accuracy of *Early* algorithm exceeds that of No QA- First 3 and No QA-SAMC, and lower differences in accuracy observed between the *Early* algorithm and No QA-All or QA-SD (Figure 13). This is likely due to the ability of the *Early* algorithm to try different permutations of available measurements to reduce the repeatability and hence, improving the chance of rejecting introduced outliers. The *Early* algorithm also demonstrated its ability to maintain the accuracy of AX compared to the other investigated methods. These results are important as R5-19 and AX are increasingly being used to indicate respiratory health, particularly small airways diseases [6][21]. Thus, confidence in the accuracy of these measurement is important in clinical decision-making, research outcomes, and when handling complex clinical scenarios.

3.3.4.2 The Computational Model

This study is the first to implement a computational model that is specifically designed to rigorously assess the accuracy of a QC algorithm for oscillometry. Prior to this research, this gap in the literature meant that any proposed approach for improving QC could not be quantitatively evaluated for its impact on accuracy and possibly hindered the adoption of new methods, likely presenting a challenge to confidently make recommendations. Hence, the modelling approach and its evaluation here may mark a significant advance forward, as it provides a mechanism to

assess improvements to QC in oscillometry. Nevertheless, this computational model is a model and therefore is not a perfect representation, and while it has its strengths and limitations, it can act as a basis and encouragement for future work.

The strengths of the computational model include the incorporation of the actual *Rrs* and *Xrs* values and the within-test variability obtained directly from the CHILD5Y subjects. This ensured that the computational model closely mimicked real-world values and their variation. Using the within-test variability observed in CHILD5Y also helped account for the inherent fluctuations that occur within a single testing session, which were considered to be either physiological or artifactual. To match clinical practice, the assessment of accuracy was restricted to five measurements per subject, matching current practice in Dr. Ronald J. Dandurand's clinic, where five measurements are always performed. The model included the incorporation of outliers based on outliers identified from measurements. This was to simulate artifactual changes, such as leaks or coughs or from any source that affected the impedance, apart from physiological variation. The outliers introduced in the model were detected using the Grubbs test, as it is thought to be superior to solely relying on the SD, since it features greater ability to account for specific statistical characteristics of the data. This may be especially useful when there are measurements that are significantly affected by leaks during a measurement, or factors acting to increase impedance such as obstruction by the tongue or swallowing.

It relied on estimating the presence of outlier data with actual measurements that were not confirmed to be non-physiologically related to variability of lung mechanics. It might be useful to develop a model based on purposely introduced artifactual data, although this has its challenges in that this may not mimic real-world data well. In any case, the model here allowed us to directly apply known outlier changes in impedance using measured distribution of these

deviations. While these outliers might not encompass the full spectrum of potential artifacts, using the Grubbs test was an unbiased approach to identifying an outlier distribution that closely aligned with the statistical characteristics of the outliers in our data. However, it cannot be known whether artifactual or contaminated measurements that did not greatly affect the impedance were included in the physiological distribution of impedance, or similarly if some physiological variation was counted as part of the outlier distribution, but this small overlap is unlikely to impact our results.

As mentioned above, the developed computational model used *Rrs* and *Xrs* values directly from the CHILD5Y subjects, which allowed for a better representation of the actual physiological variations observed in impedance data. The inclusion of a sufficient number of subjects and measurements also helped capture a representative range of within-test variability. The proposed algorithm's accuracy was also compared to other common QC methods as described above. This allowed for the evaluation of the algorithm's effectiveness within the context of established practices, further mitigating the potential impact of model simplification.

Moreover, the Grubbs test used to detect outliers assumes that the data follows an approximately normal distribution. Here we used log(Z(1/f)) in our CHILD data to address the normality of Z(1/f). Also using the impedance cost function rather than Rrs and Xrs at different frequency simplified and reduced the modeling complexity. It's possible the use of *Rrs* and *Xrs* at different frequencies could offer a more detailed insight about artifacts, however, using log(Z(1/f)) highlighted the impedance data that have the most significant impact on the QC algorithm's accuracy. Considering the computational efficiency and the overall goal of assessing the performance of the QC algorithm, this was believed to be a reasonable compromise that achieve these objectives effectively.

3.3.5 Strengths

This study has several strengths that contribute to its significance and impact. One notable strength that distinguishes this study from related research in this field is the large number of subjects used to test the algorithm. The use of three distinct age and disease groups for validation greatly enhanced the study's robustness and applicability, while also allowing for a broader exploration of potential clinical scenarios and variations that may impact the algorithm's performance. This is significant, as it allows for a more comprehensive understanding of how the proposed algorithm performs across diverse patient profiles. In addition, including subjects from various age groups and with different underlying health conditions allowed the study to capture a wider spectrum of potential challenges and complexities that clinicians encounter in real-world practice. This not only strengthens the study's external validity, but also underscores its relevance to a wide range of clinical settings. Moreover, validating the proposed spectral QC algorithm against data obtained from multiple real-world sources such as CHILD5Y, WIC, and WESER contributed to the study's credibility. This validation also ensured that the algorithm's performance aligned with practical challenges faced by clinicians, strengthening their confidence in adopting the algorithm to enhance the accuracy of their diagnoses.

3.4 Conclusion

In conclusion, this study extended the current recommended QC approach from the CV of R5 $\leq$ 10% in adults or $\leq$15% in young children to a method based on *Zrs* that includes *Rrs* and *Xrs* across frequencies. The proposed spectral QC algorithm, *Early*, improved the repeatability by lowering CV of Zrs, and improved the efficiency by decreasing the required number of measurements to a achieve a valid test. It also improved the feasibility and maintained

the accuracy as evaluated by a computational model. This has the potential for more efficient clinical studies, and improved performance of oscillometry in research and clinical settings with the potential to improve the reliability of oscillometry measurements.

CHAPTER 4: USING PATIENT REPORTED OUTCOMES TO DETECT COPD SEVERITY

While this thesis mainly investigates the improvement and automation of QC measures in oscillometry, the original goal here was to investigate whether improving the quality can improve the accuracy of predicting COPD severity using machine learning. As mentioned in sections 1.3.2 and 1.4, COPD is mainly diagnosed based on the spirometry measures FEV-1 and FVC [2,3], with its severity assessed using mMRC and CAT scores. Although spirometry measurements are possible in any healthcare setting, its efficiency mostly depends on the patient's cooperation in completing the spirometry tests, which is more challenging for individuals with COPD. Similarly, one of the main limitations of using mMRC and CAT scores is their dependence on the patients' subjective self-assessment. Therefore, while the use of mMRC and CAT scores can aid in diagnosis, understanding if they are related and can be used to predict COPD severity is also useful.

Machine learning offers an objective method of classifying COPD severity and is becoming increasingly popular in medical applications. This chapter assesses if machine learning using combinations of objective lung function parameters from oscillometry and spirometry could be sensitive and predict severity of COPD by predicting the mMRC and CAT scores. As mentioned above, the original goal was to compare the classification performance of the developed models before and after applying QC. However, we had difficulties obtaining a new data set that would have been large enough to effectively test this hypothesis, as it was not compiled to a spreadsheet format from paper charts. Thus this was left for future work. This chapter explores if machine learning can be used to predict COPD severity using combinations of spirometry and oscillometry measures.

4.1 Methods

4.1.1 Data Used

In this project, a data set containing spirometry and oscillometry results from 300 COPD and 21 healthy subjects (n=321) was used. The data included the average measured and predicted values, normalized for the patients' weight, gender, and age. It also included other important patient details such as age, sex, height, weight, smoking habits and PRO scores from both mMRC and CAT assessments for each subject. This data set was collected with informed consent, approved by the McGill University Health Centre Research Ethics Board (MUHC-RI REB# 14-467-BMB). The measurements were collected using a commercially available device (Tremoflo C-100, Thorasys Medical Systems, Montreal, Canada) and the manufacturer' software (tremoflo 1.0.43 build 44).

4.1.2 Machine Learning Training

Milestone 1

This project's first milestone began during a *Biomedical Signal Analysis and Modelling* course project, where machine learning models in MATLAB were trained to predict mMRC and CAT scores using spirometry and oscillometry measures. The aim was to compare the classification performance for COPD severity when training the machine learning models using combinations of either spirometry or oscillometry measures separately. Three MATLAB machine learning models were used in this milestone; Single Decision Tree (SDT), Bagged Decision Trees (BDT), and Support Vector Machine (SVM). SDT is a simple, but powerful model that mimics decision-making processes by starting at the top with a question and then,

depending on the answer (yes or no), following a path to another question until a final decision is eventually reached. These questions and decisions correspond to features that the decision tree model learns from historical data. Hence, a SDT model is easy to understand and interpret, making it great for tasks like binary classification. However, it can become overly complex if allowed to grow too deep, leading to overfitting [54][55]. The BDT model overcomes the limitation of SDT by averaging the results from multiple SDT models, each trained on a slightly different subset of the training data. This ensemble approach improves the model's overall robustness and reduces overfitting, making it a popular choice for many classification and regression tasks [54][55]. SVM is another powerful supervised machine learning algorithm used for classification and regression tasks. It works by finding the best possible decision boundary, often called a hyperplane, that separates different classes of data points with the maximum margin. This is done by identifying support vectors, which are data points closest to the boundary and are used in determining the optimal hyperplane's position [54][55].

To address the limited number of training samples, the data set was categorized into low and high severity COPD classes based on specific thresholds. mMRC cutoff values of one and two and CAT cutoff values of 10 and 17 were employed for this categorization. Any values exceeding these thresholds were labeled as high severity COPD, while those falling below were labeled as low severity COPD (Figure 16). These thresholds are commonly used in clinical settings and were established following discussions with Dr.Ronald J. Dandurand. Dividing the samples into high and low severity based on a threshold simplified the task into a binary classification problem, enhancing the ability to effectively fine-tune the model. Next, the data set was randomly divided into training and testing sets, following the common practice of 80-20 splits. This allocation reserved 80% of the data for model training, while the remaining 20%

were used for model testing and validation. The three machine learning models were then trained using different combinations of oscillometry and spirometry measures to predict the severity of COPD, with a focus on key predicted oscillometry measures including R5p, X5p, R5-19p, AXp, as well as the predicted spirometry measures FEV1p and FVCp. To reduce the random variation caused by the small number of samples, the models were trained using predicted values obtained from a larger and more representative data set described by Oostveen et al.[56].

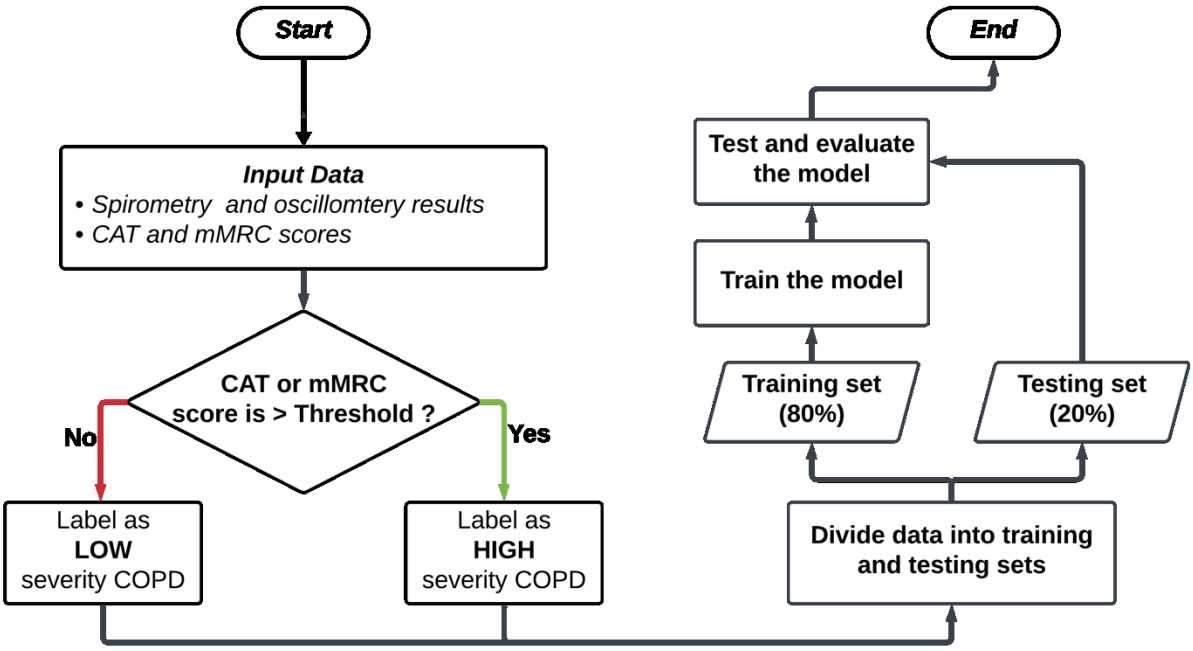


Figure 16: Flowchart of methods used to train and evaluate the performance of the different machine learning models (SDT, BDT and SVM) in the first milestone.

Milestone 2

In this project's second milestone, a group of Electrical and Computer Engineering (ECE) students who were tasked with advancing and refining the work initiated in the first milestone. Co-supervising the ECE students, they assessed the performance of the SVM and

Gradient Boosting (GB) models in Python, which was selected for its advanced capabilities in optimizing and fine-tuning parameters crucial for model training. While BDT creates multiple SDT models independently, GB model takes a different approach by building these trees sequentially in a process of optimization. This process continues iteratively, with each new tree focusing on correcting the mistakes of the previous ones. Hence, if tuned effectively, the GB model can yield an even better performance than BDT [54][55]. While the first milestone focused on training the models with spirometry or oscillometry measures separately, this milestone examined performance when combining the two measures, along with other demographics like height, age or weight. Additionally, one of the key objectives was to generate a Receiver Operator Characteristic (ROC) curve. An ROC curve is a graphical representation of the classification performance across various decision thresholds. It displays the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) as the threshold for classifying high and low severity COPD is varied. By demonstrating how the sensitivity and specificity change with different CAT thresholds, it helps determine how well the model can distinguish between high and low COPD severity. The area under the ROC curve (AUC) is another common metric used to quantify the overall performance of a classification model, with a higher AUC indicating better classification ability [54].

In this milestone the data was divided into high and low severity COPD using CAT thresholds only. This is because CAT scores have a broader range of zero to 40, when compared to mMRC's range of zero to four, allowing for the generation of a more informative ROC curve. Similar to the first milestone, the data set was divided into training and testing sets using 80-20 splits. However, this milestone maintained the ratio of healthy and COPD subjects in both the training and testing sets. This was achieved by first identifying the healthy and COPD subjects

and randomly dividing each of them into training and testing groups using the 80-20 splits. The training and testing sets from both the healthy and COPD groups were then combined, resulting in concatenated training and testing sets that maintained the ratio between healthy and COPD subjects (Figure 17). Using the training set, two machine learning models, SVM and GB, were trained to predict the severity of COPD using a mixture of oscillometry and spirometry measures, combined with demographics. Finally, ROC curves were generated after training the models using different randomizations of the training and testing sets for different CAT thresholds ranging between five and 20. This is because using CAT thresholds below five classified most of the data as high severity COPD, while thresholds above 20 classified most data as low severity COPD.

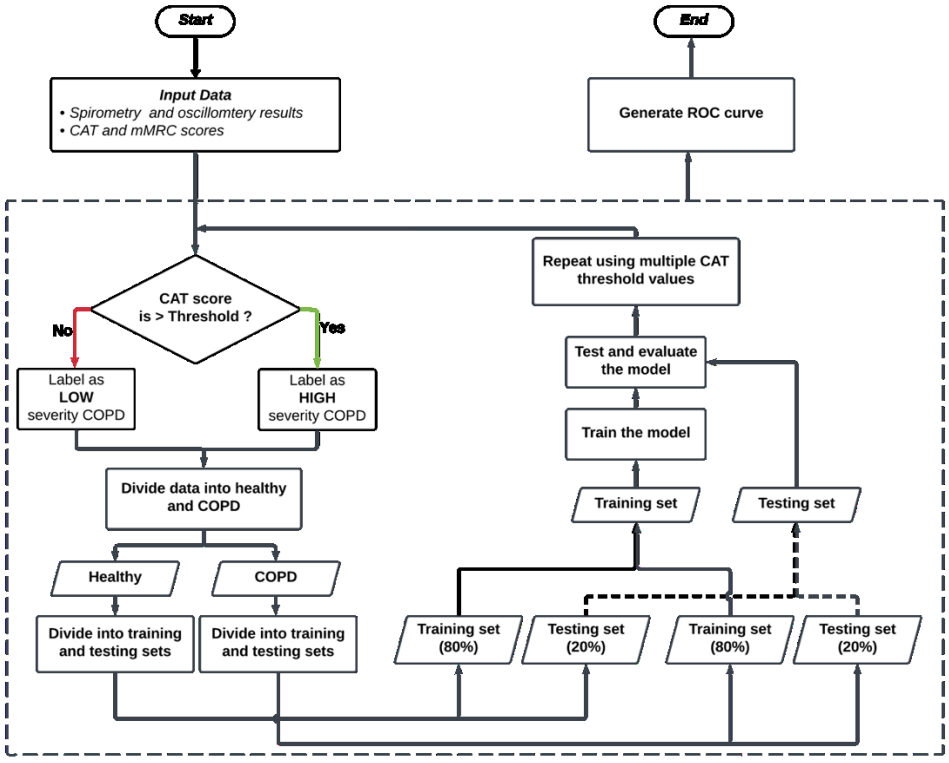


Figure 17: Flowchart of methods used to train and evaluate the performance of the different machine learning models (SVM and GB) in the second milestone.

After identifying the best-performing model and the optimal CAT threshold, the final objective of this milestone was to explore the impact of adding different oscillometry measures to spirometry measures on the enhancement of COPD severity classification. This was done by first training the machine learning model using the predicted spirometry measures FEV1p and FEV1/FVCp and then systematically introducing one oscillometry measure at a time to observe whether there is any significant improvement in the accuracy. The focus was on the important predicted oscillometry measures R5p, X5p, R5-19p, and AXp. To prevent any potential sampling bias, the same training and testing sets were consistently employed throughout these trials, ensuring that performance differences were not due to the use of different samples. The machine learning model was trained and evaluated using the same methods outlined in Figure 17.

4.1.3 Performance Outcomes

The testing set, 20% of data, was used to evaluate the performance of the developed machine learning models. This evaluation was based on a confusion matrix, which is a fundamental tool for assessing the performance of classification algorithms [54]. The confusion matrix provides a summary of the classification results, highlighting four key metrics: 1) True Positives (TP), representing correctly predicted positive instances, 2) True Negatives (TN), representing correctly predicted negative instances, 3) False Positives (FP), representing actual negatives that were predicted as positives, and 4) False Negatives (FN), representing actual positives that were predicted as negatives. Other important measures include sensitivity, specificity and accuracy. Sensitivity is used to evaluate the model's ability to predict true positives and is calculated by dividing the number of correctly predicted positives by the number of total actual positives (Equation 24).

$$Sensitivity = \frac{TP}{TP + FN} \qquad (24)$$

In contrast, specificity is used to evaluate the model's ability to predict true negatives and is calculated by dividing the number of correctly predicted negatives by the number of total actual negatives (Equation 25). Classification accuracy, on the other hand, is the percentage of total correct predictions (Equation 26). Collectively, sensitivity, specificity and accuracy offer a comprehensive evaluation of the model's classification performance.

$$Specificity = \frac{TN}{TN + FP} \qquad (25)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (26)$$

ROC was used to investigate the trade-off between sensitivity and specificity and to determine the optimal CAT threshold that resulted in the best classification performance. The ROC curve was generated using CAT thresholds ranging between five and 20 to minimize distribution bias, as described in the previous section. Finally, t-tests with 95% confidence levels were performed in Microsoft Excel to determine whether the inclusion of predicted oscillometry measures led to a statistically significant improvement in classification accuracy.

Results

Milestone 1

The SDT model was first trained using different combinations of spirometry measures and then compared the accuracy, sensitivity, specificity and the number of false negatives (FN) (Table 4). Results were comparable across all performance measures ($p>0.05$), with the highest performance achieved when using only the predicted FEV1 (FEV1p) measure, yielding $65.2\pm6.6\%$ accuracy, $74.2\pm8.1\%$ sensitivity, $42.8\pm10.7\%$ specificity, and 8 ± 2 FN (Table 4). While high sensitivity was achieved with all combinations, ranging between 72% and 75%, specificity was relatively low, ranging between 41% and 43%

Table 4: Mean (SD) classification accuracy, sensitivity, specificity and FN for SDT model when trained using combinations of spirometry measures.

	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	FN
FEV1p	65.2(6.6)	74.2(8.1)	42.8(10.7)	8(2)
FEV1p + FVCp	63.7(5.8)	72.9(7.3)	41.3(11.4)	7(2)
FEV1p + FEV1p/FVCp	64.1(6.6)	73(7.8)	42.6(14.4)	8(3)

The SDT model's performance was then compared to the classification performance of the BDT and SVM models when trained with the same spirometry measures. Compared to the SDT model, the BDT model increased the accuracy ($p<0.05$) and sensitivity ($p<0.05$), while maintaining the specificity ($p>0.05$) when training the model using the spirometry measures FEV1p and predicted FVC (FVCp). Similar results were achieved when training the BDT model using the spirometry measures FEV1p with FEV1p/FVCp, with an increase in accuracy

($p<0.05$), sensitivity ($p<0.05$) and specificity ($p<0.05$). However, training the BDT model with only FEV1p resulted in comparable accuracy ($p>0.05$), sensitivity ($p>0.05$), specificity($p>0.05$) and FN ($p>0.05$). The best overall classification performance achieved with the BDT model was using FEV1p and FEV1p/FVCp, with 68±6.4% accuracy, 76.7±7.5% sensitivity, 46.2±12.7% specificity and 8±3 FN. This is a significant improvement in all performance measures compared to the best performing combination achieved with SDT(Table 5).

Table 5: Mean (SD) classification accuracy, sensitivity, specificity and FN for BDT model when trained using combinations of spirometry measures.

	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	FN
FEV1p	65.5(6.3)	73(7.4)	46.7(9.3)	9(3)
FEV1p + FVCp	67.9(5.4)	78.7(6.3)	41(10.1)	8(2)
FEV1p + FEV1p/FVCp	68(6.4)	76.7(7.5)	46.2(12.7)	8(3)

The SVM model increased the accuracy ($p>0.05$) and sensitivity ($p<0.05$), but reduced the number of FN ($p<0.05$) for all combinations of spirometry measures, when compared to SDT. However, these improvements came at the cost of a significant reduction in the specificity ($p<0.05$) from around 41% using the SDT model to 15% using the SVM model (Table 6). The SVM model demonstrated similar increases in accuracy ($p>0.05$) and sensitivity ($p<0.05$), along with a decrease in FN ($p<0.05$) and specificity ($p<0.05$), when compared to the BDT model. Additionally, SVM results were comparable for accuracy ($p>0.05$), sensitivity ($p>0.05$), specificity ($p>0.05$) and FN ($p>0.05$) for all three combinations of spirometry measures (Table 6). The best classification performance was achieved using FEV1p and FEV1p/FVCp, yielding 69.1±7% accuracy, 88.9±6.9% sensitivity, 4±2 FN and a very low specificity of 19.5±10%.

Table 6: Mean (SD) classification accuracy, sensitivity, specificity and FN for SVM model when trained using combinations of spirometry measures.

	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	FN
FEV1p	67.6(5.6)	88.4(5.2)	15.3(5.6)	3(1)
FEV1p + FVCp	66.7(6.2)	87.1(6)	15(10.4)	3(2)
FEV1p + FEV1p/FVCp	69.1(7)	88.9(6.9)	19.5(10)	4(2)

The SVM model achieved very high sensitivity, but a very low specificity. The BDT model, on the other hand, demonstrated a comparable accuracy to SVM with a better trade-off between sensitivity and specificity. As such, the BDT model was used in the second part of this milestone, aiming to assess whether the use of oscillometry measures would yield a better classification performance than spirometry measures. The BDT model was trained using many different combinations of oscillometry measures, with the three best performing combinations summarized in Table 7. The predicted oscillometry measures R5p, R5-19, X5p and AXp yielded the best classification performance with 67.4$\pm$3.9% accuracy, 79.5$\pm$7.4% sensitivity, 38.4$\pm$11.7% specificity and 7$\pm$2 FN. This classification performance is comparable to all other oscillometry combinations ($p>0.05$) and the best performing spirometry combination FEVp and FEV1p/FVCp ($p>0.05$).

Table 7: Mean (SD) classification accuracy, sensitivity, specificity and FN for BDT model when trained using combinations of oscillometry measures.

	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	FN
R5p + X5p	66.8(4.4)	77.1(7.5)	41.7(12.9)	8(2)
R5p + X5p + AXp	65.7(5)	76.3(7.7)	39.9(12.1)	7(2)
R5p + R5-19 + X5p + AXp	67.4(3.9)	79.5(7.4)	38.4(11.7)	7(2)

The last step of this milestone was to examine how different CAT and mMRC cutoff thresholds influence the classification performance. Interestingly, there was a clear trade-off between sensitivity and specificity when changing the cutoff threshold. For example, using a CAT threshold of 10 or an mMRC threshold of one resulted in high sensitivity above 70% at a cost of low specificity below 47% (Table 8). On the other hand, using a CAT threshold of 17 and mMRC threshold of two yielded higher specificity values, above 68%, at a cost of a low sensitivity below 41%.

Table 8: Mean (SD) classification accuracy, sensitivity, specificity and FN for SVM model when trained using combinations of spirometry (FEV1p and FEV1p/FVCp) and oscillometry (R5p, X5p and AX) measure and a) CAT threshold of 10, b) CAT threshold of 17, c) mMRC threshold of one and d) mMRC threshold of two.

CUTOFF THRESHOLD	COMBINATIONS	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	FN
A) CAT 10	*FEV1p + FEV1p/FVCp*	68(6.4)	76.7(7.5)	46.2(12.7)	8(3)
	R5p +R5-19p + X5p + AXp	67.4(3.9)	79.5(7.4)	38.4(11.7)	7(2)
B) CAT 17	*FEV1p + FEV1p/FVCp*	59(7.2)	27.9(9.2)	76.2(10)	32(5)
	R5p +R5-19p + X5p + AXp	58(6.2)	23.3(9.8)	77.5(7.8)	32(3)
C) mMRC 1	*FEV1p + FEV1p/FVCp*	62.8(6.1)	73.1(7.6)	40.7(12.2)	8(3)
	R5p +R5-19p + X5p + AXp	67.7(6)	82(4.9)	37.2(12.5)	7(2)
D) mMRC 2	*FEV1p + FEV1p/FVCp*	56.7(5.1)	40.8(10)	68.2(7.3)	26(3)
	R5p +R5-19p + X5p + AXp	56.3(5)	37.8(10.4)	69.7(5.8)	26(3)

Milestone 2

The first aim of milestone two was to assess whether combining spirometry and oscillometry measures and adding demographics can improve the classification performance. Resultantly, two machine learning models were trained in Python, GB and SVM, and compared their classification performance to results obtained from milestone one (Table 9). Combining

spirometry and oscillometry measures and adding demographics demonstrated an improvement in accuracy, from 67.5±5% using BDT with spirometry measures and 68.8±5% using BDT with oscillometry measures to 76.6±5% ($p<0.05$) and 73.5±6% ($p<0.05$) using GB and SVM, respectively. Sensitivity was also improved, with lower FN and comparable specificity. The GB and SVM models were comparable, with the best classification performance obtained using the GB model, achieving 76.6±5% accuracy, 88.0±5% sensitivity, 46.9±10% specificity and 5±2% FN.

Table 9: Mean (SD) classification accuracy, sensitivity, specificity and FN for BDT model trained in MATLAB using spirometry and oscillometry measures separately, as well as the GB and SVM models trained in Python using combined spirometry and oscillometry measures with demographics.

COMBINATIONS	MODEL	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)	FN
FEV1p + FEV1p/FVCp	*BDT - MATLAB*	67.5(5)	78(6)	43.8(11)	8(2)
R5p +R5-19p + X5p + AXp	*BDT - MATLAB*	68.8(5)	82.5(6)	38(11)	7(2)
FVCp + FEV1p + R5p + X5p + AXp + HEIGHT + AGE + SMOKING YEARS	*GB - Python*	76.6(5)	88.0(5)	46.9(10)	5(2)
	SVM - Python	73.5(6)	85.5(6)	44.0(11)	6(2)

The second aim of this milestone was to generate an ROC curve to find the optimal CAT threshold value with the best trade-off between sensitivity and specificity (Figure 18, Table 10). The best classification performance and balance of sensitivity and specificity was at a cluster of CAT thresholds ranging between eight and 10. Interestingly, a CAT threshold of 10, often used in clinical settings, corresponded with a good sensitivity of 88.0 ±5%, but a low specificity of 46.9±10%. In comparison, a CAT threshold of nine improved the sensitivity to 90.4 ±4%

(p<0.05) and specificity to 51.0±12% (p<0.05), making it a potentially optimal threshold value with the best trade-off between sensitivity and specificity. Using CAT thresholds below eight resulted in very high sensitivity at a cost of low specificity, while CAT thresholds greater than 15 resulted in very high specificity at a cost of low sensitivity.

Table 10: Mean (SD) TP, FP, TN, and FN using GB model with CAT threshold of 9, 10 and 17

CAT CUTOFF THRESHOLD	TP	FP	TN	FN
9	43(3)	8(3)	8(2)	5(2)
10	40(3)	10(3)	9(2)	5(2)
17	6(2)	9(4)	32(3)	17(4)

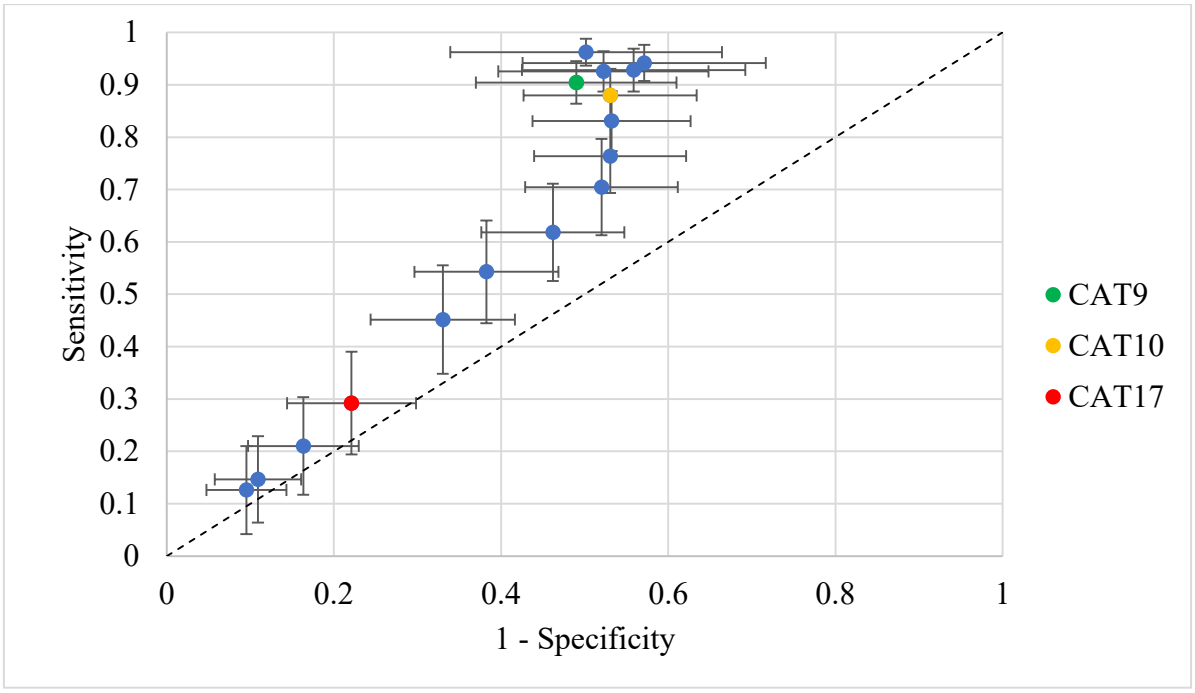


Figure 18: ROC curve obtained by training the GB model using different CAT thresholds ranging between 5 and 20. CAT thresholds of 9, 10 and 17 are highlighted in green, yellow and red, respectively.

Since the combination of oscillometry and spirometry measures improved the classification performance, the last aim of this milestone was to identify which oscillometry measure led to a significant improvement in accuracy. As such, changes in accuracy were evaluated after introducing oscillometry measures, one measure at a time, to the spirometry measures FEV1p and FEV1p/FVCp (Figure 19). While adding R5 to the spirometry measures reduced the accuracy ($p>0.05$), the oscillometry measures X5p, R5-19 and AXp all resulted in a significant increase in accuracy ($p<0.05$). Nevertheless, X5p, R5-19 and AXp resulted in comparable classification accuracy when added to the spirometry measures FEV1p and FEV1p/FVC ($p>0.05$), with the highest accuracy of 75.5 ± 6% achieved using X5p.

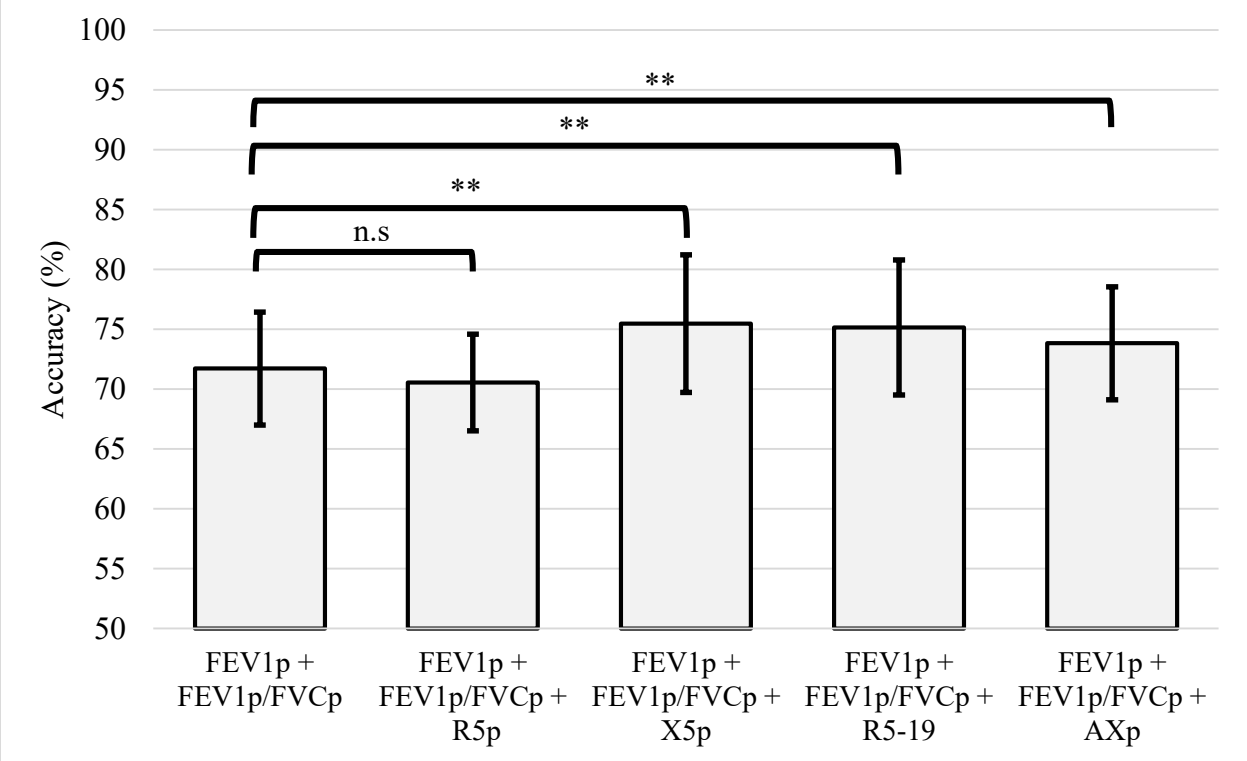


Figure 19: Mean ± SD classification accuracy when adding different oscillometry measures to the spirometry measures FEV1p and FEV1p/FVCp.

Given that X5p demonstrated the highest accuracy, it was then evaluated if adding R5-19 or AXp can further improve the classification accuracy. Interestingly, adding R5-19 to FEV1p, FEV1p/FVCp and X5p resulted in a modest improvement in accuracy from 75.5 ± 6% to 76.3± 6% ($p>0.05$) (Figure 20). Similarly, adding AXp resulted in a comparable accuracy of 75.8±6% ($p>0.05$), adding no additional significance to the machine learning algorithm.

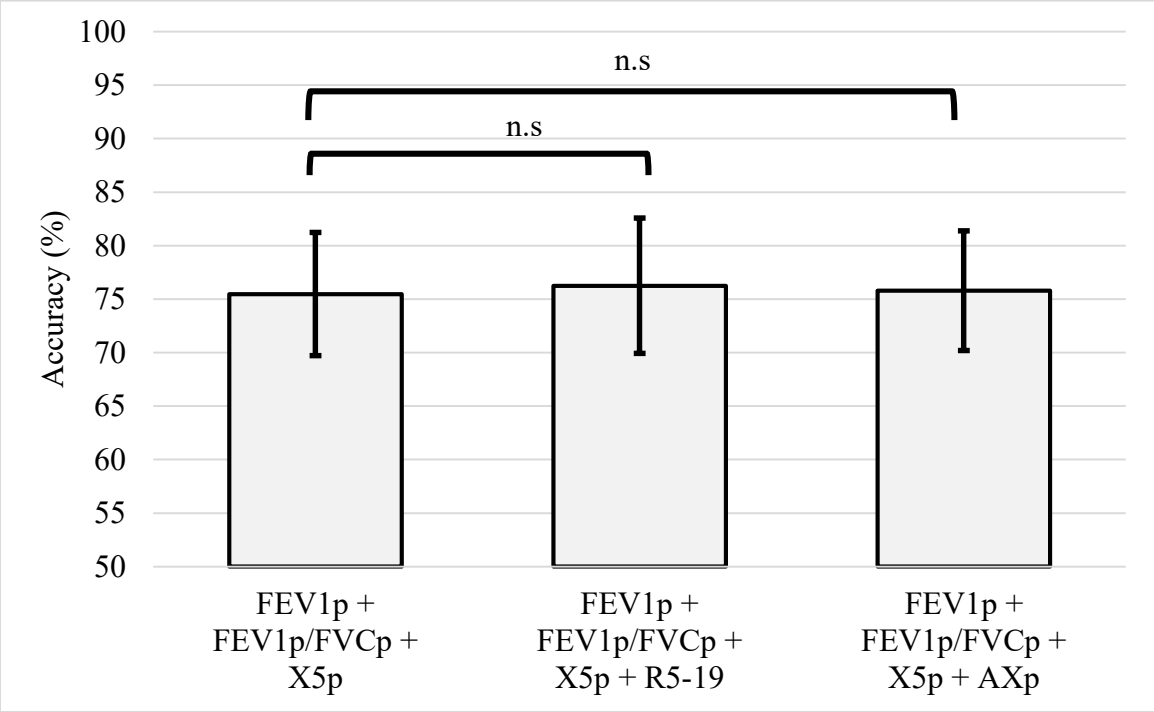


Figure 20: Mean ± SD classification accuracy when adding oscillometry measures R5-19 and AXp to combination containing FEV1p, FEV1p/FVCp and X5p.

Discussion

The objective of this study was to assess if machine learning using combinations of lung function parameters from oscillometry and spirometry could be sensitive and predict severity of COPD by predicting the mMRC and CAT scores. The principle findings are as follows: 1) comparable classification performance was achieved when training the machine learning models

using spirometry or oscillometry measures separately, 2) combining spirometry and oscillometry measures and incorporating demographics improved the classification performance, and 3) a CAT threshold ranging between eight and 10 demonstrated the best trade-off between sensitivity and specificity.

It was found that the SDT, BDT and SVM models achieved comparable results when trained with spirometry or oscillometry measures separately. The best classification performance was achieved using the SVM model when trained with the spirometry measures FEVp and FEV1p/FVCp and oscillometry measures R5p, R5-19, X5p and AXp, yielding comparable accuracies of 69.1±7% and 67.4±3.9%, respectively. Combining spirometry and oscillometry measures and adding demographics such as height, age and smoking years improved the classification performance, achieving 76.6±5% accuracy. Sensitivity and specificity also increased, with a lower number of FN. Although small, these improvements are statistically significant ($p<0.05$) and can be important, given the noisy nature of this group.

It was also found that a region on the ROC for CAT thresholds from eight to 10 had high sensitivity and good specificity, implying some robustness in performance in this range. Interestingly, it was found that a CAT of 10, the GOLD Guideline recommended cut point to distinguish between Grades A and B, and C and D in the GOLD classification of COPD, achieved very good classification performance and balance between sensitivity and specificity. The best classification performance and balance between sensitivity and specificity was obtained using a CAT of 9, achieving 90.4 ±4% sensitivity, 51.0±12% specificity and 5±2 FNs. However, using CAT thresholds below eight resulted in very high sensitivity at a cost of low specificity, while CAT thresholds greater than 15 resulted in very high specificity at a cost of low sensitivity. This is likely because a CAT threshold below eight or greater than 15 can result in a biased

distribution of the healthy and diseased subjects, resulting in a slight improvement in TP or TN rates at a cost of higher FP and FN rates.

High sensitivity models provide no information on the classification of low severity COPD patients and hence, are useful for ruling out high severity COPD when a patient test negative, while high specificity models are useful in ruling in high severity COPD when a patient test positive. To add, FNs are obtained because of the misclassification of high severity COPD as low, which may lead to deleterious effects on the patients' health. As such, a CAT threshold of nine provides a classification model with high accuracy, sensitivity and specificity and a low FN, which can help provide an accurate assessment for disease status. It was found that balancing healthy subjects in both training and testing sets as illustrated in Figure 17 played an important role in improving the classification performance. Balancing healthy subjects in both training and testing sets can reduce potential biases that can arise when for example most of the healthy subjects are included in either the training or testing sets.

Conclusion

In conclusion, this study explored if simple machine learning algorithms could predict high or low severity patient reported outcomes using spirometry and oscillometry measures. While using oscillometry and spirometry measures separately resulted in comparable accuracy, sensitivity, and specificity, combining the two measures resulted in an improved overall classification performance; indicating that machine-based algorithms using combinations of spirometry and oscillometry measures show potential for patient screening and monitoring. Interestingly this also implies that the CAT has good physiological correspondence to lung function measures near CAT = nine and 10, when including multiple lung function measurements.

CHAPTER 5: THESIS CONCLUSIONS

The first objective of this thesis was to develop an improved version of the current ERS QC recommendations that incorporates both *Rrs* and *Xrs* across frequencies. Results demonstrated that the *Early* algorithm using a weighted impedance cost function ($\zeta_Z = Z(1/f)$) was more robust and consistently performed better when dealing with challenging data containing artifacts and outliers, improving the repeatability, efficiency, and feasibility, while maintaining the accuracy. The assessment of accuracy, which was this thesis' second objective, was done using a computational model generated from physiological and artifact distributions. This is important as prior methods did not assess accuracy nor include any model assessments. The computational model used in this study created distributions of impedances with added artifactual data that were similar to observed data, both on mean and individual values. While the model provided a mean to assess accuracy, some assumptions were included in the model's development and hence may limit its interpretation. This may include the assumption of normality for repeated individual measures, which is needed for the Grubbs test. It may also include the use of the Grubbs test to identify artifactual data and to generate 'noise' and separate it from what was identified as signal. Future work could include purposely adding artifactual measures and testing the QC algorithm on individual measurements. This could be similar to the approach of Bhatawadekar et al. [41], who developed a wavelet-based filtering approach to remove artifacts within individual measurements using subjects trained to produce real-world artifacts, such as leaks or swallows. Future work could also evaluate the *Early* algorithm on patients with different or more severe diseases, as the type of artifacts may influence the computation modelling. Additionally, this work was based on the same acceptability threshold used for R5: 15% for young children and 10% for adults. These thresholds, which were based on

the ERS standards, are expert opinion-based, rather than evidence-based. Hence, future work should investigate whether this acceptability threshold is optimal when using the proposed cost function, Z(1/f), which incorporates *Rrs* and *Xrs* across multiple frequencies. This could be addressed using computational modelling, as employed in this study, with sufficiently varied data sets across multiple subjects with differing respiratory diseases.

The third objective of this thesis was to assess whether combining spirometry and oscillometry measures and adding demographics can improve the classification performance. It was found that combining spirometry and oscillometry measures with demographics resulted in an improved overall classification performance. It was also found that the CAT has good physiological correspondence to lung function measures near CAT = nine and 10, when including multiple lung function measurements. These results indicate that machine-based algorithms using combinations of spirometry and oscillometry measures show potential for patient screening and monitoring. The hope was to evaluate if applying the *Early* algorithm can improve the classification performance. However, obtaining sufficient data with both CAT and multiple measures of oscillometry took longer than the duration of the thesis, making it important that future research obtains this data to meet this objective. Additionally, the accuracy achieved using the different machine learning models was limited to no more than 80%, likely due to the limitations of the small data set (n=321). This could also be near the maximally achievable accuracy using this data set. An attempt to minimize variability amongst the small sample size was by using predicted measures that reduce variation due to impedance and spirometry dependence on age, height and sex. CAT thresholds were also used instead of attempting to predict exact CAT scores, simplifying the task into a binary classification problem. Nonetheless, future work should aim to improve the performance using a larger data set, which

could also improve the data set's power in training and testing the machine learning models. Moreover, the SDT, BDT and SVM models were used in MATLAB to predict COPD severity when using spirometry and oscillometry measures separately, while used the GB mad SVM models in Python after combining spirometry and oscillometry measures with demographics. Hence, future work should use the same models and environment to re-evaluate whether combining spirometry and oscillometry measures outperforms spirometry and oscillometry separately.

Contributions From Thesis

1. A. Abufardeh, T. Schuessler, P. Subbarao, R. Dai, M. Reyna-Vargas, R. J. Dandurand and G. Maksym, "Improving Quality Control in Oscillometry: Repeatability, Efficiency and Accuracy," ATS International Conference 2022, May 2022. Abstract was accepted and presented at the ATS 2022 conference. It was also nominated for the 10th annual CTS Research Poster Competition, May 2022.
2. A. Abufardeh, R. J. Dandurand and G. Maksym, "Machine Learning Using Combined Spirometry and Oscillometry to Predict COPD Assessment Test," ATS International Conference 2021, May 2021. Abstract was accepted and presented at the ATS 2021 conference.
3. A. Abufardeh, R. J. Dandurand and G. Maksym, "Investigating machine-based learning to predict patient reported outcomes from spirometry and oscillometry," ATS International Conference 2020, May 2020. Abstract was accepted and presented at the ATS 2020 conference.

REFERENCES

[1] 14:00-17:00, “ISO 9000:2015,” ISO. Accessed: Dec. 27, 2023. [Online]. Available: https://www.iso.org/standard/45481.html

[2] G. Sreedher *et al.*, “Magnetic resonance imaging quality control, quality assurance and quality improvement,” *Pediatr Radiol*, vol. 51, no. 5, pp. 698–708, May 2021, doi: 10.1007/s00247-021-05043-6.

[3] B. H. Derrickson and G. J. Tortora, *Tortora’s Principles of anatomy & physiology*, [15th ed.]. Danvers MA: Wiley, 2017.

[4] G. J. Tortora and B. Derrickson, *Principles of anatomy & physiology*, 13th ed. Hoboken, NJ: Wiley, 2012.

[5] A. C. Guyton and J. E. Hall, *Textbook of medical physiology*, 11th ed. Philadelphia: Elsevier Saunders, 2006.

[6] A. B. Otis *et al.*, “*Mechanical Factors in Distribution of Pulmonary Ventilation*,” *Journal of Applied Physiology*, vol. 8, no. 4, pp. 427–443, Jan. 1956, doi: 10.1152/jappl.1956.8.4.427.

[7] U. Peters, D. A. Kaminsky, S. Bhatawadekar, L. Lundblad, and G. N. Maksym, “Oscillometry for Lung Function Testing,” in *Lung Function Testing in the 21st Century*, Elsevier, 2019, pp. 25–47. doi: 10.1016/B978-0-12-814612-5.00003-8.

[8] J. H. T. Bates, *Lung mechanics: an inverse modeling approach*. Cambridge, UK ; New York: Cambridge University Press, 2009.

[9] Q. Hamid, J. Shannon, and J. Martin, Eds., *Physiologic basis of respiratory disease*. Hamilton: BC Decker, Inc, 2005.

[10] U. Peters, D. A. Kaminsky, and G. N. Maksym, “Standardized Pulmonary Function Testing,” in *Lung Function Testing in the 21st Century*, Elsevier, 2019, pp. 5–23. doi: 10.1016/B978-0-12-814612-5.00002-6.

[11] “Understanding Asthma,” Asthma Canada. Accessed: Feb. 09, 2023. [Online]. Available: https://asthma.ca/get-help/understanding-asthma/

[12] “Asthma.” Accessed: Feb. 09, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/asthma

[13] J. Vestbo *et al.*, “Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: GOLD Executive Summary,” *Am J Respir Crit Care Med*, vol. 187, no. 4, pp. 347–365, Feb. 2013, doi: 10.1164/rccm.201204-0596PP.

[14] “Chronic obstructive pulmonary disease (COPD).” Accessed: Feb. 09, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

[15] "2024 GOLD Report," Global Initiative for Chronic Obstructive Lung Disease - GOLD. Accessed: Dec. 25, 2023. [Online]. Available: https://goldcopd.org/2024-gold-report/

[16] N. Beydon *et al.*, "An Official American Thoracic Society/European Respiratory Society Statement: Pulmonary Function Testing in Preschool Children," *Am J Respir Crit Care Med*, vol. 175, no. 12, pp. 1304–1345, Jun. 2007, doi: 10.1164/rccm.200605-642ST.

[17] L. K. A. Lundblad, S. Siddiqui, Y. Bossé, and R. J. Dandurand, "Applications of oscillometry in clinical research and practice," *Canadian Journal of Respiratory, Critical Care, and Sleep Medicine*, vol. 5, no. 1, pp. 54–68, Jan. 2021, doi: 10.1080/24745332.2019.1649607.

[18] A. Lorx *et al.*, "Airway dynamics in COPD patients by within-breath impedance tracking: effects of continuous positive airway pressure," *Eur Respir J*, vol. 49, no. 2, p. 1601270, Feb. 2017, doi: 10.1183/13993003.01270-2016.

[19] K. Desiraju and A. Agrawal, "Impulse oscillometry: The state-of-art for lung function testing," *Lung India*, vol. 33, no. 4, p. 410, 2016, doi: 10.4103/0970-2113.184875.

[20] J. K. Wu *et al.*, "Development of Quality Assurance and Quality Control Guidelines for Respiratory Oscillometry in Clinic Studies," *Respir Care*, vol. 65, no. 11, pp. 1687–1693, Nov. 2020, doi: 10.4187/respcare.07412.

[21] G. G. King *et al.*, "Technical standards for respiratory oscillometry," *Eur Respir J*, vol. 55, no. 2, p. 1900753, Feb. 2020, doi: 10.1183/13993003.00753-2019.

[22] E. Oostveen *et al.*, "The forced oscillation technique in clinical practice: methodology, recommendations and future developments," *European Respiratory Journal*, vol. 22, no. 6, pp. 1026–1041, Dec. 2003, doi: 10.1183/09031936.03.00089403.

[23] O. S. Usmani, "Calling Time on Spirometry: Unlocking the Silent Zone in Acute Rejection after Lung Transplantation," *Am J Respir Crit Care Med*, vol. 201, no. 12, pp. 1468–1470, Jun. 2020, doi: 10.1164/rccm.202003-0581ED.

[24] S. Rutting, D. G. Chapman, C. S. Farah, and C. Thamrin, "Lung heterogeneity as a predictor for disease severity and response to therapy," *Current Opinion in Physiology*, vol. 22, p. 100446, Aug. 2021, doi: 10.1016/j.cophys.2021.05.009.

[25] H. H. Alamdari, K. El-Sankary, and G. N. Maksym, "Time-Varying Respiratory Mechanics as a Novel Mechanism Behind Frequency Dependence of Impedance: A Modeling Approach," *IEEE Trans Biomed Eng*, vol. 66, no. 9, pp. 2433–2446, Sep. 2019, doi: 10.1109/TBME.2018.2890055.

[26] B. Suki, A. L. Barabási, and K. R. Lutchen, "Lung tissue viscoelasticity: a mathematical framework and its molecular basis," *J Appl Physiol (1985)*, vol. 76, no. 6, pp. 2749–2759, Jun. 1994, doi: 10.1152/jappl.1994.76.6.2749.

[27] K. R. Lutchen and H. Gillis, “Relationship between heterogeneous changes in airway morphometry and lung resistance and elastance,” *Journal of Applied Physiology*, vol. 83, no. 4, pp. 1192–1201, Oct. 1997, doi: 10.1152/jappl.1997.83.4.1192.

[28] F. Peták, M. J. Hayden, Z. Hantos, and P. D. Sly, “Volume Dependence of Respiratory Impedance in Infants,” *Am J Respir Crit Care Med*, vol. 156, no. 4, pp. 1172–1177, Oct. 1997, doi: 10.1164/ajrccm.156.4.9701049.

[29] D. M. Gray *et al.*, “Intra-breath measures of respiratory mechanics in healthy African infants detect risk of respiratory illness in early life,” *Eur Respir J*, vol. 53, no. 2, p. 1800998, Feb. 2019, doi: 10.1183/13993003.00998-2018.

[30] M. Amurao, D. A. Gress, M. A. Keenan, P. H. Halvorsen, J. A. Nye, and M. Mahesh, “Quality management, quality assurance, and quality control in medical physics,” *J Applied Clin Med Phys*, vol. 24, no. 3, p. e13885, Mar. 2023, doi: 10.1002/acm2.13885.

[31] P. Poorisrisak, C. Vrang, J. M. Henriksen, B. Klug, B. Hanel, and H. Bisgaard, “Accuracy of whole-body plethysmography requires biological calibration,” *Chest*, vol. 135, no. 6, pp. 1476–1480, Jun. 2009, doi: 10.1378/chest.08-1555.

[32] J. C. Watts *et al.*, “Measurement duration impacts variability but not impedance measured by the forced oscillation technique in healthy, asthma and COPD subjects,” *ERJ Open Res*, vol. 2, no. 2, pp. 00094–02015, Apr. 2016, doi: 10.1183/23120541.00094-2015.

[33] P. D. Robinson *et al.*, “Procedures to improve the repeatability of forced oscillation measurements in school-aged children,” *Respir Physiol Neurobiol*, vol. 177, no. 2, pp. 199–206, Jul. 2011, doi: 10.1016/j.resp.2011.02.004.

[34] F. Marchal, C. Schweitzer, B. Demoulin, C. Choné, and R. Peslin, “Filtering artefacts in measurements of forced oscillation respiratory impedance in young children,” *Physiol. Meas.*, vol. 25, no. 5, pp. 1153–1166, Oct. 2004, doi: 10.1088/0967-3334/25/5/006.

[35] M. R. Canal, “Comparison of wavelet and short time Fourier transform methods in the analysis of EMG signals,” *J Med Syst*, vol. 34, no. 1, pp. 91–94, Feb. 2010, doi: 10.1007/s10916-008-9219-8.

[36] Y. Zhang, Z. Guo, W. Wang, S. He, T. Lee, and M. Loew, “A comparison of the wavelet and short-time fourier transforms for Doppler spectral analysis,” *Med Eng Phys*, vol. 25, no. 7, pp. 547–557, Sep. 2003, doi: 10.1016/s1350-4533(03)00052-3.

[37] T. T. Pham *et al.*, “Automated quality control of forced oscillation measurements: respiratory artifact detection with advanced feature extraction,” *Journal of Applied Physiology*, vol. 123, no. 4, pp. 781–789, Oct. 2017, doi: 10.1152/japplphysiol.00726.2016.

[38] H. Lorino, C. Mariette, M. Karouia, and A. M. Lorino, “Influence of signal processing on estimation of respiratory impedance,” *J Appl Physiol (1985)*, vol. 74, no. 1, pp. 215–223, Jan. 1993, doi: 10.1152/jappl.1993.74.1.215.

[39] C. Schweitzer, C. Chone, and F. Marchal, “Influence of data filtering on reliability of respiratory impedance and derived parameters in children,” *Pediatr Pulmonol*, vol. 36, no. 6, pp. 502–508, Dec. 2003, doi: 10.1002/ppul.10359.

[40] N. J. Brown *et al.*, “A comparison of two methods for measuring airway distensibility: nitrogen washout and the forced oscillation technique,” *Physiol Meas*, vol. 25, no. 4, pp. 1067–1075, Aug. 2004, doi: 10.1088/0967-3334/25/4/022.

[41] S. A. Bhatawadekar *et al.*, “A study of artifacts and their removal during forced oscillation of the respiratory system,” *Ann Biomed Eng*, vol. 41, no. 5, pp. 990–1002, May 2013, doi: 10.1007/s10439-012-0735-9.

[42] T. T. Pham, C. Thamrin, P. D. Robinson, A. L. McEwan, and P. H. W. Leong, “Respiratory Artefact Removal in Forced Oscillation Measurements: A Machine Learning Approach,” *IEEE Trans Biomed Eng*, vol. 64, no. 8, pp. 1679–1687, Aug. 2017, doi: 10.1109/TBME.2016.2554599.

[43] Z. Hantos, J. K. Y. Wu, R. J. Dandurand, and C.-W. Chow, “Quality control in respiratory oscillometry: reproducibility measures ignoring reactance?,” *ERJ Open Res*, vol. 9, no. 3, pp. 00070–02023, May 2023, doi: 10.1183/23120541.00070-2023.

[44] A. B. Fisher, A. B. DuBois, and R. W. Hyde, “Evaluation of the forced oscillation technique for the determination of resistance to breathing,” *J. Clin. Invest.*, vol. 47, no. 9, pp. 2045–2057, Sep. 1968, doi: 10.1172/JCI105890.

[45] E. D. Michaelson, E. D. Grassman, and W. R. Peters, “Pulmonary mechanics by spectral analysis of forced random noise.,” *J. Clin. Invest.*, vol. 56, no. 5, pp. 1210–1230, Nov. 1975, doi: 10.1172/JCI108198.

[46] A. Abufardeh, G. n. Maksym, and R. j. Dandurand, “Investigating Machine-Based Learning to Predict Patient Reported Outcomes from Spirometry and Oscillometry,” in *TP123. TP123 LUNG FUNCTION: FLOW, VOLUME, AND HETEROGENEITY*, 167 vols., in American Thoracic Society International Conference Abstracts. , American Thoracic Society, 2021, pp. A4605–A4605. doi: 10.1164/ajrccm-conference.2021.203.1_MeetingAbstracts.A4605.

[47] A. Abufardeh, M. Wright, R. MacDonald, T. Thorne, G. N. Maksym, and R. J. Dandurand, “Machine Learning Using Combined Spirometry and Oscillometry to Predict COPD Assessment Test,” in *B108. ORIGINS AND OUTCOMES OF COPD*, American Thoracic Society, May 2022, pp. A3645–A3645. doi: 10.1164/ajrccm-conference.2022.205.1_MeetingAbstracts.A3645.

[48] F. E. Grubbs, “Sample Criteria for Testing Outlying Observations,” *Ann. Math. Statist.*, vol. 21, no. 1, pp. 27–58, Mar. 1950, doi: 10.1214/aoms/1177729885.

[49] F. E. Grubbs, “Procedures for Detecting Outlying Observations in Samples,” 2023.

[50] H. W. Lilliefors, “On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown,” 2023.

[51] H. W. Lilliefors, “On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown,” 2023.

[52] C. L. Que, C. M. Kenyon, R. Olivenstein, P. T. Macklem, and G. N. Maksym, “Homeokinesis and short-term variability of human airway caliber,” *J Appl Physiol (1985)*, vol. 91, no. 3, pp. 1131–1141, Sep. 2001, doi: 10.1152/jappl.2001.91.3.1131.

[53] L. M. Harkness *et al.*, “Within-session variability as quality control for oscillometry in health and disease,” *ERJ Open Res*, vol. 7, no. 4, pp. 00074–02021, Oct. 2021, doi: 10.1183/23120541.00074-2021.

[54] V. T. Inc, “Fundamentals of Machine Learning | 9780198828044, 9780192563095,” VitalSource. Accessed: Oct. 30, 2023. [Online]. Available: https://www.vitalsource.com/en-ca/products/fundamentals-of-machine-learning-thomas-trappenberg-v9780192563095

[55] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN COMPUT. SCI.*, vol. 2, no. 3, p. 160, Mar. 2021, doi: 10.1007/s42979-021-00592-x.

[56] E. Oostveen *et al.*, “Respiratory impedance in healthy subjects: baseline values and bronchodilator response,” *Eur Respir J*, vol. 42, no. 6, pp. 1513–1523, Dec. 2013, doi: 10.1183/09031936.00126212.

APPENDIX A: SUPPLEMENTARY RESULTS

Table A1: List of weights used to generate the different combinations of R and Z CVs (ζ) to investigate and optimize the performance of the proposed spectral QC algorithm.

Combination of R or Z	*Weight (w) at used frequencies (Hz)*								
	5	***11***	***13***	***17***	***19***	***23***	***29***	***31***	***37***
R	1	0	0	0	0	0	0	0	0
Z	1	0	0	0	0	0	0	0	0
R	0	1	0	0	0	0	0	0	0
Z	0	1	0	0	0	0	0	0	0
R	0	0	0	0	1	0	0	0	0
Z	0	0	0	0	1	0	0	0	0
R	1	1	0	0	0	0	0	0	0
Z	1	1	0	0	0	0	0	0	0
R	1	0	0	0	1	0	0	0	0
Z	1	0	0	0	1	0	0	0	0
R	1	1	0	0	1	0	0	0	0
Z	1	1	0	0	1	0	0	0	0
Z	1	0	0	1	0	1	0	0	0
Z	1	1	1	1	1	1	1	1	1
Z	1	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
Z	3	0.5	0	1	0	1	0	0	0
Z	3	0.5	0	0.75	0	0.75	0	0	0
Z	3	0.5	0	0.25	0	1	0	0	0
Z	3	0.75	0	0.5	0	0.75	0	0	0
Z	3	0.5	0	0.5	0	0.75	0	0	0
Z	3	0.5	0	0.125	0	1	0	0	0
Z	3	0.5	0	0.125	0	0.75	0	0	0
Z	3	0.5	0	0.25	0	0.75	0	0	0
Z	3	0.5	0	0.5	0	0.5	0	0	0
Z	2	0.5	0.25	0.125	1	0	0	0	0
Z	1	0.33	0.33	0.33	1	0	0	0	0
Z	3	0.5	0.25	0.125	2	0	0	0	0
Z	2	0.5	0.25	0.125	2	0	0	0	0
Z	2	0.33	0.15	0.05	1	0	0	0	0
Z	2	0	0	0	1	0	0	0	0
Z	3	0.5	0.25	0.125	1	0	0	0	0
Z	2	0.75	0.25	0.125	1	0	0	0	0
Z	1	0.091	0.077	0.059	0.053	0.043	0.034	0.032	0.027
Z	1	0.091	0.077	0.059	0.053	0	0	0	0

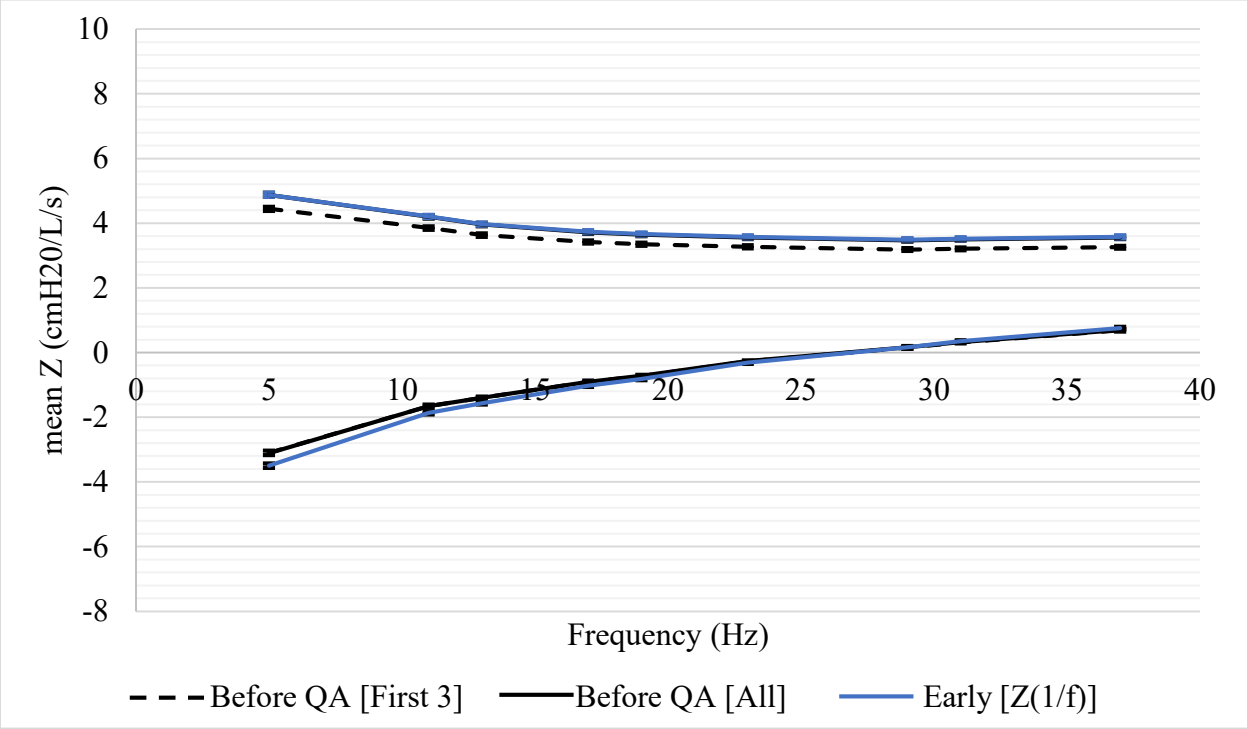


Figure A1: A plot of the mean resistances (top) and reactances (bottom) vs. frequency from the WIC data set before and after spectral QC using ζ = Z(1/f), (Bars represent standard error).

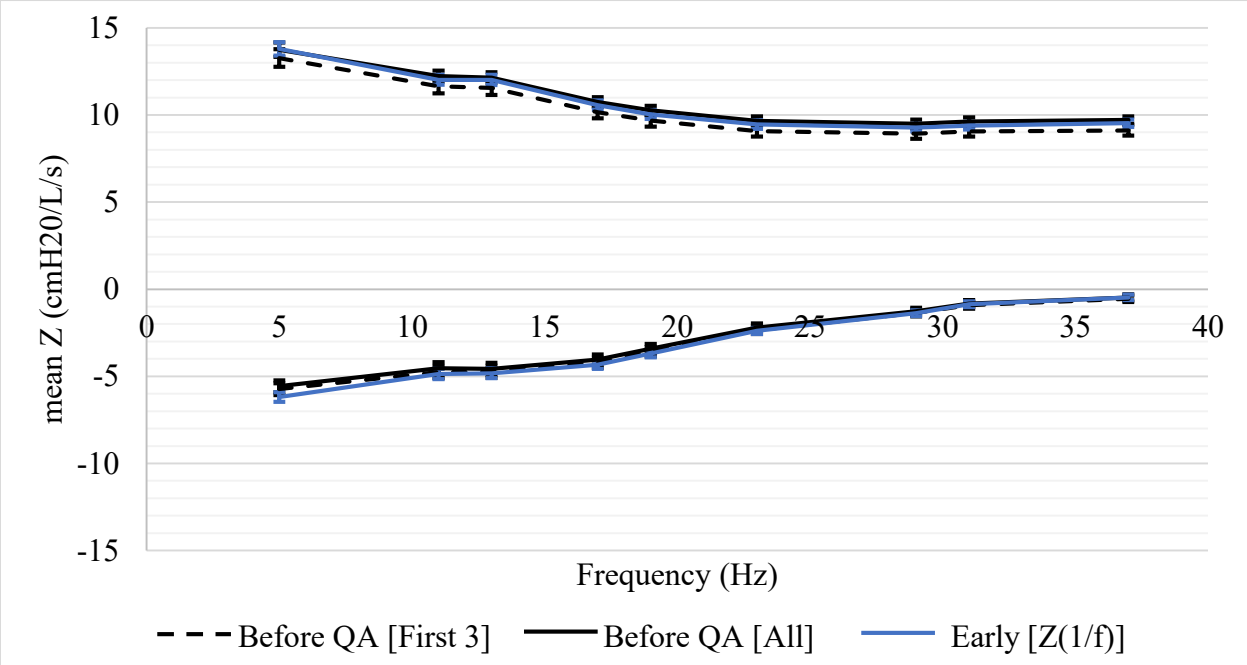
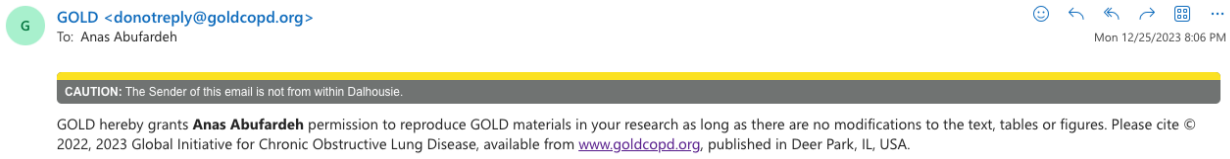


Figure A2: A plot of the mean resistances (top) and reactances (bottom) vs. frequency from the WESER data set before and after spectral QC using ζ = Z(1/f), (Bars represent standard error).

APPENDIX B: COPYRIGHT RELEASE REQUESTS

B.1 Figure 1.1 and 1.2 Permission



G

GOLD <donotreply@goldcopd.org>
To: Anas Abufardeh

Mon 12/25/2023 8:06 PM

CAUTION: The Sender of this email is not from within Dalhousie.

GOLD hereby grants **Anas Abufardeh** permission to reproduce GOLD materials in your research as long as there are no modifications to the text, tables or figures. Please cite © 2022, 2023 Global Initiative for Chronic Obstructive Lung Disease, available from www.goldcopd.org, published in Deer Park, IL, USA.