Automatic Characterization of User Errors in Spirometry

Andrew Z. Luo, Eric Whitmire, James W. Stout, Drew Martenson, Shwetak Patel

***Abstract*— Spirometry plays a critical role in characterizing and improving outcomes related to chronic lung disease. However, patient error in performing the spirometry maneuver, such as from coughing or taking multiple breaths, can lead to clinically misleading results. As a result, spirometry must take place under the supervision of a trained specialist who can identify and correct patient errors. To reduce the need for specialists to coach patients during spirometry, we demonstrate the ability to automatically detect four common patient errors. Creating separate machine learning classifiers for each error based on features derived from spirometry data, we were able to successfully label errors on spirometry maneuvers with an F-score between 0.85 and 0.92. Our work is a step toward reducing the need for trained individuals to administer spirometry tests by demonstrating the ability to automatically detect specific errors and provide appropriate patient feedback. This will increase the availability of spirometry, especially in low resource and telemedicine contexts.**

I. Introduction

Spirometry is the most commonly used test of lung function available in primary care and specialty settings. The test plays a critical role in identifying and managing chronic lung diseases such as chronic obstructive pulmonary disease (COPD) and asthma [1]. During a spirometry maneuver, a patient rapidly exhales into a monitoring device, which tracks the flow and volume of air exhaled from a patient, as shown in Fig. 1. Commonly, physicians interpret flow vs. volume (FV) and volume vs. time curves (VT) (Fig. 2) to understand the respiratory health of the patient.

Increasing access to spirometry enables more frequent monitoring of patient health, which can lead to improved outcomes and lower health care costs [2]. To achieve this, recent efforts have focused on creating and evaluating cheaper and more portable devices that can be used in homes and low resource settings [3, 4, 5, 6].

Normally, trained professionals administer and monitor testing, coaching the patients through the maneuver to ensure clinically useful spirometry measures. Despite the increase in the availability of portable spirometry outside of traditional clinical settings, patients still need coaching and feedback on the maneuver. Due to the difficulties in training personnel to administer and monitor testing, this presents a problem in delivering spirometry to new markets. Even in contexts where spirometry is widespread, there is a shortage of trained professionals. A study of primary care spirometry practices in Australia found that 64% of professionals administering tests had received less than 14 hours of training in conducting spirometry [7]. This can have serious ramifications for the effectiveness of spirometry as a clinical tool; professionals with inadequate training could only produce acceptable spirometry maneuvers from 60% of their patients nine months after training [8].

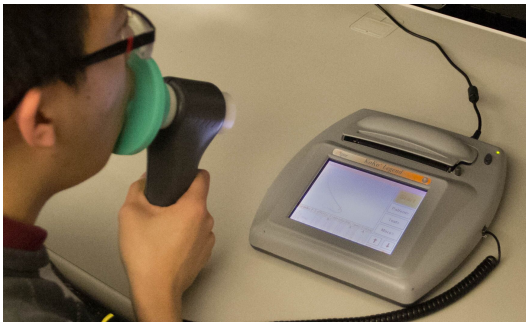

Fig. 1: An example of a spirometry device

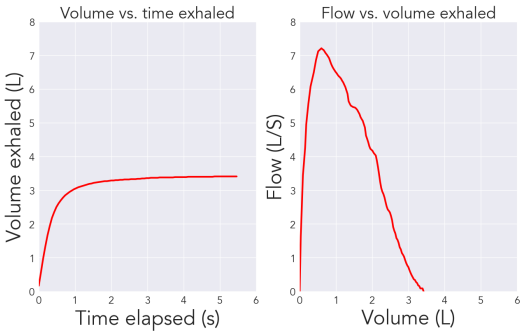

Fig. 2: Example output flow vs volume (FV) and volume vs time (VT) curves from spirometry.

Automatic feedback to patients after the incorrect execution of a spirometry maneuver provides an attractive solution to reducing the need for trained professionals. Many spirometers currently give automated feedback based upon numeric guidelines published by the American Thoracic Society (ATS) and European Respiratory Society (ERS) [9, 10, 11]. These features include the amount of volume exhaled in the first second of the test, and the total length of the test. Current ATS/ERS guidelines, which examine a limited number of features derived from spirometry curves, suffer from poor performance in classifying the clinical quality of spirometry curves, achieving a correct classification only 80.6% of the time with 90% sensitivity and 56% specificity [11]. As a result, visual inspection by a professional remains the gold

A. Luo, E. Whitmire, and S. Patel are with the Department of Computer Science and Engineering at the University of Washington, Seattle, WA 98195; email: luoa@cs.washington.edu, emwhit@cs.washington.edu, shwetak@cs.washington.edu

J. Stout, MD MPH, is with the Department of Pediatrics and Department of Health Services at the University of Washington, Seattle, WA 98195; email: jstout@uw.edu

D. Martenson, RRT, is with the Glendale Adventist Medical Center, Glendale, CA 91206; email: drewrrt@uw.edu

978-1-5090-2809-2/17/$31.00 ©2017 IEEE

standard for evaluating spirometry curves. Recently, Umberto et. al. improved quality assessment by adding additional features to ATS/ERS guidelines, categorizing spirometry curves as of acceptable, unacceptable, or unknown (requiring an expert to evaluate) clinical quality using a variety of hand-picked features in a manually constructed decision tree model [12].

We improve spirometry by detecting specific patient errors with actionable outcomes: variable flow throughout the maneuver, coughing during the maneuver, extra breaths taken through the maneuver, and early termination of the maneuver. Although some patient errors (e.g. slow start and sub-maximal flow) have well-defined numeric guidelines [13], the errors we chose to classify are more subjective and still often rely on visual inspection for detection. Furthermore, although past work has used carefully chosen thresholds on manually-tuned features for classification, we use machine learning methods to automatically learn a classification model. Each of these common errors has prescribed actions to help patients ensure they do not make the same error during later spirometry maneuvers [14]. For example, if a patient is inadvertently taking extra breaths through their nose near the end of the maneuver, nose clips might be applied as a solution to improve the quality of the maneuver [14]. Our results for automatically detecting patient errors therefore have actionable outcomes for improving the clinical quality of spirometry maneuvers. Overall classification F1-scores of 0.92024, 0.86498, 0.85515, and 0.84629 were achieved for the respective errors of early termination, cough, variable flow, and extra breath.

II. Background

Below is a brief description of each error detected [14, 9] and a visual representation (Fig. 3).

1) Early Termination: Early termination of the spirometry maneuver occurs when the patient has not exhaled all the air from their lung. Additionally, if a spirometry maneuver is not of sufficient length, a maneuver may be labeled with the early termination error. On a VT curve, this is characterized by the lack of a plateau in total recorded volume. On a FV curve, this is characterized by an abrupt drop-off in flow near the end of the test. Common coaching for this error entails encouraging the patient to keep blowing through the test until they empty their lungs.

2) Cough: Coughing during the test can disrupt clinically useful metrics such as the volume of air exhaled in the first second of the maneuver. On a VT curve, this is characterized by a period where the recorded exhaled volume plateaus for a short period of time. On a FV curve, this is characterized by a sharp drop in flow followed by recovery. A common solution to deal with this error is to give the patient a glass of water.

3) Variable Flow: Variable flow occurs when the flow of exhaled air varies substantially throughout the maneuver. This can affect clinically useful measurements such as the total volume of air the patient can exhale in the first second or the maximum flow of air a patient can exhale. On a FV curve, this is characterized by dips in flow smoother than those of a cough. A common solution to deal with this error is to coach the patient to blow air out harder and keep blowing through the maneuver.

4) Extra Breath: Extra breaths may be accidentally taken in through the nose or the edges of a patient's mouth surrounding the spirometry mouthpiece. This can lead to falsely reported statistics related to the patient's lung capacity. On a VT curve, this can be characterized by smaller versions of the standard VT curve shape appearing after volume readings plateau. On a FV curve, this can be characterized by smaller peaks in flow that appear at the end of the maneuver. A common solution to deal with this error is to use nose clips or instruct the subject to keep a tighter seal around the spirometry mouthpiece.

III. Materials and Methods

A. Data Source

TABLE I: Overview of Data Set

Metric	Value
Total curves in data set	19880 curves
Sample rate FV/VT curve	0.06 L/16.67 Hz
Mean age	19.17 years
Std. age	17.29 years
Age in years	3 - 95 years

Curves from clinical spirometers collected since 2011 as part of a training effort were uploaded onto an online labeling system[1] for offline evaluation and feedback. An overview of the data can be found in Table I. Of note is most of the collected curves (72.2%) came from patients between the ages of 6 and 18. Six trained respiratory therapists annotated the curves with the appropriate error labels and entered feedback on the quality of the maneuver. For each curve used in training classifiers, a single expert annotated the presence of relevant errors. For each curve used in testing classifiers, a head annotator decided upon the final labels for the curve. Separate classifiers were trained to detect the presence of each error. For this analysis, separate training and testing sets for each error were created by sampling data from a larger set of curves that was labeled with the pertinent error and only that error as positive cases. At no point was the entire data set utilized for training classifiers to avoid bias issues arising from error imbalances in the dataset. An equally-sized random sampling of curves with no errors was used as the negative case for each classifier.

B. Training and Testing

Because our dataset was annotated by multiple experts, this introduces the potential for noisy labels. Since ensemble methods tend to be outperform other classifiers in the presence of noise, we use AdaBoost classifiers with decision trees as the base estimator. For training, 90% (N=17882) of the data was used and the remaining curves were held out for testing. Features for detecting the presence of errors were created by feedback from doctors, spirometry training

[1] http://www.spirometry360.org

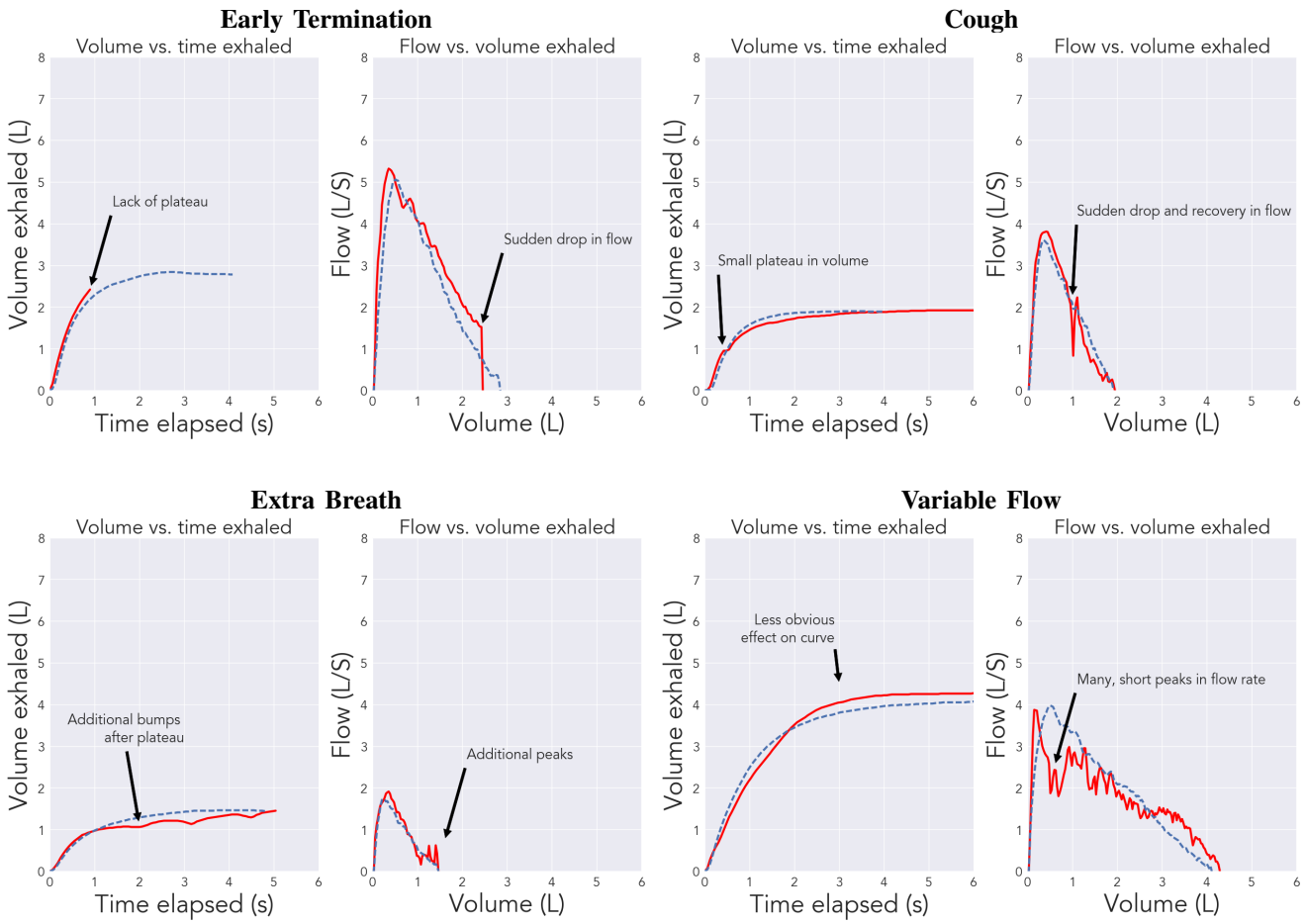

Fig. 3: Example curves demonstrating features which indicate each error. The red line represents a maneuver with the specified error while the dotted blue line represents a similar maneuver without the specified error

materials, [14, 9], and features derived from earlier work on determining spirometry quality [12]. In total, 68 different features were extracted for each error. After class balancing, a total of 5728, 1344, 5614, and 1314 curves were used to train for the early termination, cough, variable flow, and extra breath errors drawing from the 17882. Each classifier was then tested on a completely separate set of curves whose labels had been annotated by a single respiratory therapist with over a decade of experience. In total, 1998 curves were part of the overall testing set. For each error, a minimum of 486 curves were used for testing utilizing the partitioning process used in training. The number of error and error-free curves were balanced in each set.

C. Evaluation

Metrics used for evaluating the efficacy of classification include the precision, recall, and F1-score of each classifier when applied to a testing set. Consider spirometry curves which are labeled with the presence of a relevant error being detected by a classifier. Let such spirometry curves be considered a positive case for classification. Let t_p, f_p, t_n, f_n be the number of true positives, false positives, true negatives, and false negatives in classifying the testing set for each classifier. Precision P, recall R , and F1-score $F1$ are defined as:

$$P = \frac{t_p}{t_p + f_p} \qquad R = \frac{t_p}{t_p + f_n} \qquad F1 = 2\frac{PR}{P + R}$$

IV. Results

A. Features

Though all features were used in the AdaBoost classifier, below are the two most significant features for classification as decided by utilizing the Gini importance metric in the decision tree models used.

1) Early Termination:

- The total time elapsed during the maneuver
- The volume exhaled in the last second of the maneuver

2) Cough:

- A heuristic for the total amount of time the slope in the VT is relatively flat. This is obtained by examining the period of volume exhaled where the slope of the FV curve is less than 10% of the maximum slope.
- The maximum slope in the FV curve after peak flow.

3) Extra Breath:

- The minimum slope in the VT curve.
- The maximum slope in the FV curve after peak flow.

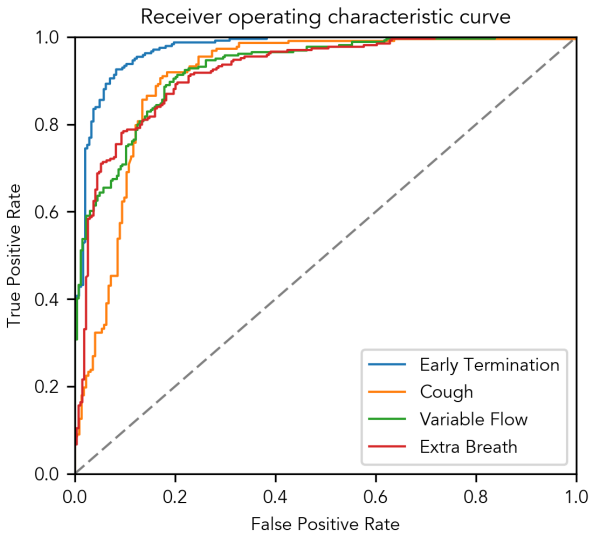

Fig. 4: Receiver operating characteristic curves demonstrating the performance of each error classifier

4) Variable Flow:

- The maximum slope in the FV curve after peak flow.
- The sum of the total first derivative whose values were positive after the area of highest flow in the FV curve.

B. Classification Error

Table II shows the performance of each classifier on the held out dataset. Each error was evaluated against on a set of curves that had been annotated with either a single error, or no error. Using the weighted confidence scores of base estimators in each classifier, we construct receiver operating characteristic (ROC) curves, as shown in Fig 4.

TABLE II: Classifier Performance

Error	Precision	Recall	F1-Score
Early Termination (n=486)	91.5%	92.6%	0.920
Cough (n=446)	81.7%	91.9%	0.865
Variable Flow (n=528)	79.3%	92.8%	0.855
Extra Breath (n=538)	82.4%	87.0%	0.846

V. Discussion

Our work provides a baseline for future work in automatic error detection in spirometry. As there are significant issues in the quality of spirometry being performed in clinical practice due to a lack of training [15, 7], our results suggest that automated error detection could be a promising solution. Future areas of work could include automatic diagnosis of chronic lung diseases such as COPD and asthma from spirometry data.

However, there is still room for significant improvement. For example, in our analysis, we considered only spirometry curves with one labeled error (e.g. only extra breath), ignoring the effects of compounded errors (e.g. extra breath and variable flow). We leave an analysis of these effects to future work. Furthermore more complex feature engineering may be needed to improve classification performance on chosen errors. A promising area of future research is to investigate the use of recurrent neural networks (RNNs) which can be adept at recognizing patterns in time series data, removing the need for manually constructed features.

There are also limitations related to the dataset used in our work. A majority of the curves used came from those between the ages of 6 and 18. In the future, classification should be done across a more even distribution of ages to ensure proper generalization of models to all ages. Furthermore, labels in the training dataset were provided by 6 labelers. As a result, it is possible that variance between labelers in labeling specific errors may exist. Compounding this, the training set was annotated only by a single professional rather than a group or pair. Further validation should be done using data whose labels have been verified by multiple experienced professionals. In addition, as the data set was heavily anonymized analysis and control of patient information, including the number of unique patients could not be done. Nevertheless, our initial results are promising and shows potential in automating coaching and feedback in spirometry.

References

[1] Mary C Townsend. "Spirometry in the Occupational Health Setting-2011 Update". In: *Journal of Occupational and Environmental Medicine* 53.5 (2011), pp. 569–584.

[2] Terence A. R. Seemungal et al. "Time Course and Recovery of Exacerbations in Patients with Chronic Obstructive Pulmonary Disease". In: *American Journal of Respiratory and Critical Care Medicine* 161.5 (May 2000), pp. 1608–1613.

[3] Babatunde A. Otulana et al. "The Use of Home Spirometry in Detecting Acute Lung Rejection and Infection Following Heart-Lung Transplantation". In: *Chest* 97.2 (1990), pp. 353–357.

[4] A. F. J. Brouwer, R. J. Roorda, and P. L. P. Brand. "Home spirometry and asthma severity in children". In: *European Respiratory Journal* 28.6 (Jan. 2006), pp. 1131–1137.

[5] Mayank Goel et al. "SpiroCall". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016).

[6] Eric C. Larson et al. "SpiroSmart". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12* (2012).

[7] David P. Johns et al. "National survey of spirometer ownership and usage in general practice in Australia". In: *Respirology* 11.3 (2006), pp. 292–298.

[8] Brigitte M Borg et al. "Spirometry Training Does Not Guarantee Valid Results". In: *Respiratory Care* 55.6 (2012), pp. 689–694. ISSN: 0020-1324.

[9] M. R. Miller. "Standardisation of spirometry". In: *European Respiratory Journal* 26.2 (Jan. 2005), pp. 319–338.

[10] Felip Burgos et al. "Telemedicine enhances quality of forced spirometry in primary care". In: *European Respiratory Journal* 39.6 (Oct. 2011), pp. 1313–1318.

[11] Christine Muller-Brandes et al. "LUNOKID: can numerical American Thoracic Society/European Respiratory Society quality criteria replace visual inspection of spirometry?" In: *European Respiratory Journal* 43.5 (2014), pp. 1347–1356. ISSN: 0903-1936.

[12] Umberto Melia et al. "Algorithm for Automatic Forced Spirometry Quality Assessment: Technological Developments". In: *PLoS ONE* 9.12 (2014).

[13] Paul L Enright et al. "Quality of Spirometry Performed by 13,599 Participants in the World Trade Center Worker and Volunteer Medical Screening Program". In: *Respiratory Care* 55.3 (2012), pp. 303–309. ISSN: 0020-1324.

[14] NIOSH, Lu-Ann F. Beeckman-Wagner, and Diana Freeland. *Spirometry quality assurance: common errors and their impact on test results*. Department of Health et al., 2012.

[15] "Spirometry Use Among Pediatric Primary Care Physicians". In: *Pediatrics* 127.2 (2010).