SpiroSmart: Using a Microphone to Measure Lung Function on a Mobile Phone

Eric C. Larson[1†], Mayank Goel[2†], Gaetano Boriello[2], Sonya Heltshe[3], Margaret Rosenfeld[3], Shwetak N. Patel[1,2]

[1]Electrical Engineering,
[2]Computer Science & Engineering,
DUB Institute, University of Washington
Seattle, WA 98195
{eclarson, mayankg, gaetano, shwetak }@uw.edu

[3]Seattle Children's Hospital
Center for Clinical and Translational Research
4800 Sand Point Way NE
Seattle, WA 98105
{sonya.heltshe, margaret.rosenfeld}@seattlechildrens.org

ABSTRACT

Home spirometry is gaining acceptance in the medical community because of its ability to detect pulmonary exacerbations and improve outcomes of chronic lung ailments. However, cost and usability are significant barriers to its widespread adoption. To this end, we present SpiroSmart, a low-cost mobile phone application that performs spirometry sensing using the built-in microphone. We evaluate SpiroSmart on 52 subjects, showing that the mean error when compared to a clinical spirometer is 5.1% for common measures of lung function. Finally, we show that pulmonologists can use SpiroSmart to diagnose varying degrees of obstructive lung ailments.

Author Keywords

Health sensing, spirometry, mobile phones, signal processing, machine learning.

ACM Classification Keywords

H5.m. Information interfaces and presentation (*e.g.*, HCI): Miscellaneous.

INTRODUCTION

Spirometry is the most widely employed objective measure of lung function [37] and is central to the diagnosis and management of chronic lung diseases, such as asthma, chronic obstructive pulmonary disease (COPD), and cystic fibrosis. During a spirometry test, a patient forcefully exhales through a flow-monitoring device (a tube or mouthpiece), which measures instantaneous flow and cumulative exhaled volume (Figure 1). Spirometry is generally performed in medical offices and clinics using conventional spirometers, but home spirometry with portable devices is slowly gaining acceptance [6,26]. Measurement of spirometry at home allows patients and physicians to more regularly monitor for trends and detect changes in lung function that may need evaluation and/or treatment. Home spirometry has the potential to result in earlier treatment of exacer-
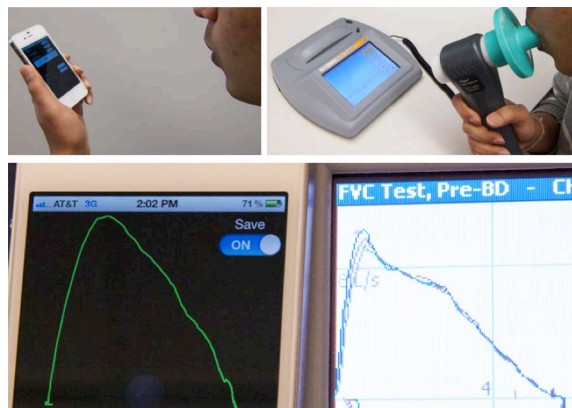
Figure 1. Subjects using SpiroSmart (left) and a clinical spirometer (right) and example curves from each device.

bations, more rapid recovery, reduced health care costs, and improved outcomes [15,23,34,35]. However, challenges currently faced by home spirometry are cost, patient compliance and usability, and an integrated method for uploading results to physicians [9,12]. Importantly, while office-based spirometry is coached by a trained technician, current home spirometers have no coaching, feedback, or quality control mechanisms to ensure acceptable measurements.

In this paper, we present *SpiroSmart*, a smartphone-based approach that measures lung function using the phone's built-in microphone (*i.e.*, a complete software-enabled solution). SpiroSmart requires the user to hold the smartphone at approximately arm's length, breathe in their full lung volume, and forcefully exhale at the screen of the phone until the entire lung volume is expelled. The phone's microphone records the exhalation and sends the audio data to a server, which calculates the exhaled flow rate by estimating models of the user's vocal tract and the reverberation of sound around the user's head. Flow rate is estimated by calculating the envelope of the sound in the time domain; performing resonance tracking in the frequency domain; while measuring white noise gain through linear prediction. SpiroSmart is able to compute and provide flow rates and graphs similar to those found in home or clinical spirometers (Figure 1).

[†] The first two authors are equal contributors to this work.

There are important advantages to developing a smartphone-based solution as compared to current commercial spirometers. Firstly, the low-cost and inherent portability of the smartphone allows much greater uptake of home spirometry. The relative low cost of smartphones as compared to spirometers can also help in lowering access barriers to medical devices in the developing world. Secondly, a smartphone spirometer can have built-in coaching and feedback—mechanisms to maximize measurement acceptability that are critically lacking in current home spirometers. Thirdly, smartphones provide the capability of easy data uploading, enabling longitudinal tracking of results and instantaneous alerts. Finally, with the smartphone, spirometry can easily be coupled with evaluations such as symptom scores, cough sensing, or oximetry to provide a comprehensive disease self-management tool.

Our research team closely collaborated with experienced pulmonologists from two different hospitals to help inform the design of SpiroSmart and to critically compare the accuracy of SpiroSmart to a clinical spirometer (endorsed by the American Thoracic Society or ATS). Through a study that included 52 subjects, we show that SpiroSmart has a mean error of 5.1% for the most common measures of lung function and that SpiroSmart can be used directly out-of-the-box, without any user-specific training or calibration. However, our results indicate that the performance of SpiroSmart improves if we calibrate this generic system for a particular user, decreasing the mean error in estimation of lung function to 4.6%. Lastly, we compare the differences in diagnosis from five pulmonologists using measures and graphs generated from SpiroSmart and from a clinical spirometer. We show that SpiroSmart is effective for diagnosing not only abnormal lung function but also the degree of obstruction.

EXPLANATION OF SPIROMETRY

A standard spirometer measures flow rate of air as it passes through a mouthpiece. This flow can be integrated from the mouthpiece to achieve Flow *vs.* Time (FT), Volume *vs.* Time (VT), or Flow *vs.* Volume (FV) plots of the expiration. An example FV plot can be seen in Figure 2. From the plot, several quantities are measured:

(1) **Forced Vital Capacity (FVC)** is the total expelled volume during the expiration,
(2) **Forced Expiratory Volume** in **one second** (FEV_1) is the volume exhaled in the first second,
(3) **FEV_1/FVC** is simply the ratio of the aforementioned two measures, and,
(4) **Peak Expiratory Flow (PEF)** is the maximum flow velocity reached during the test.

The most common clinically-reported measures are FEV_1, FVC, and FEV_1/FVC, as they are used to quantify the degree of airflow limitation in chronic lung diseases such as asthma, COPD, and cystic fibrosis. In general, a healthy result is >80% of the predicted value based on height, age, and gender [16]. Abnormal values are [22]:
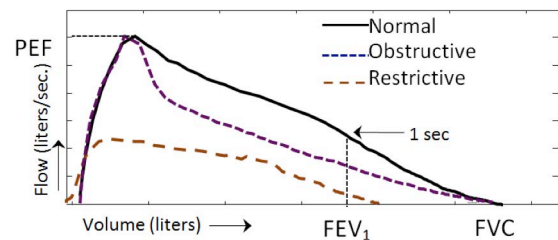
Figure 2. Example of different Flow *vs.* Volume curves and example lung function measures.

- Mild Lung Dysfunction: 60-79%
- Moderate Lung Dysfunction: 40-59%
- Severe Lung Dysfunction: below 40%

However, diagnosis from spirometry is more complicated than simple benchmarking—the shape of the flow curve is subjectively evaluated by a pulmonologist, who examines the descending limb of the Flow *vs.* Volume curve (*i.e.*, after PEF, Figure 2). A linear slope indicates the absence of airflow limitation (*i.e.,* normal lung function, solid line in Figure 2). A concave or "scooped" slope signifies airflow limitation (*e.g.*, asthma or COPD, purple dashed line in Figure 2) due to differing time constants of exhaled air in different parts of the lung. The last curve (orange dashed line in Figure 2) is suggestive of restrictive lung disease such as that caused by respiratory muscle weakness or pulmonary fibrosis; it can be seen as a slight bowing of the curve, a plateau, and/or a decreased FVC.

Modern Spirometry Devices

Modern spirometers are generally flow based, measuring the instantaneous exhaled flow (liters/sec.). There are four prevalent types of flow-based spirometers: pneumotachographs, turbines, anemometers, and ultrasounds. Pneumotachs measure the pressure differential across a membrane as the subject exhales. These devices are affected by humidity and temperature and require daily calibration. Pneumotachs are the most prevalent spirometers in medical offices and clinics because of their accuracy.

High-end clinical spirometers can cost upwards of $5000 USD and be comparable in size to a small refrigerator. The patient sits inside an enclosure that controls humidity, temperature, and oxygen levels. Portable, ATS-endorsed spirometers (about the size of a laptop) generally cost between $1,000-$4,000 USD, and although they are relatively portable compared to their counterparts, they are still bulky, complicated devices (Figure 1).

Low cost peak flow meters exist, which can only measure PEF ($10-$50 USD). They are about the size of a baseball and typically use a mechanical apparatus without any electronics. PEF in isolation, however, is generally considered to be a poor indicator of lung function [28]. Digital home spirometers that report only FEV_1 are also commercially available ($50-$200 USD). These meters vary widely in their reporting and archiving of results—some require patients to manually write down the values in journals or have a USB desktop connection. The newest models (*e.g.*, Spi-

roTube) can connect to a mobile phone or laptop via Bluetooth, but cost considerably more, from $900-$3500 USD. In 2011, Gupta *et al.* presented mobileSpiro, a $100-$200 pneumotach that connects to an Android smart phone [13]. This is one of the first systems that reduce cost by leveraging the mobile phone to present feedback and process data.

We note that a number of applications (*e.g.*, mySpirometer, Spirometer Pro) have recently appeared on the iPhone and Android platforms that claim to measure aspects of lung function. However, these applications are advertised as games and have disclaimers warning not to use them for medical assessment.

RELATED WORK

Our work draws motivation from prior research exploring solutions that leverage sensing and computing capabilities of smartphones as well as technologies that leverage audible sensing for improvement in healthcare.

Mobile Phone Based Health Sensing

There are a number of healthcare sensing systems in which external sensors are connected to smartphones. Poh *et al.* developed a system containing electro-optic sensors worn on the earlobe to provide photoplethysmography (PPG) data on a smartphone [30]. A number of researchers have also evaluated how multiple sensors could be connected to a smartphone via an external board to collect physiological information [7,21]. Bishara *et al.* have successfully modified the existing on-device camera to perform lens-free holographic microscopy [4]. Pamplona *et al.* have developed NETRA, a system that combines simple optical components, like lenses, with high-resolution LCD screens of smartphones to detect human eye impairments [27].

Researchers have also been exploring solutions that require no hardware modification. Grimaldi *et al.* employed a smartphone's camera and LED flashlight to measure pulse from the fingertip using photoplethysmography [11]. While this requires users to be in contact with the device, Poh *et al.* use a tablet's camera and blind source separation of color channels to measure pulse at a distance [29].

Audio Based Health Sensing

Like SpiroSmart, there are several technologies that sense medically relevant quantities using a microphone. Using an in-ear microphone, researchers have shown that one can detect when (and sometimes what) a person is eating [3,25]. Wheeze detection with in-air and throat microphones has shown promising results in diagnosing the severity of asthma [14]. Respiratory rate is another vital sign typically sensed with body worn [2] or bedside microphones [17].

A few systems have leveraged simple, low-cost microphones to analyze signals, such as heart rate and cough. Many systems exist that extract heart rate using a mobile phone [24,40] and, with higher-end microphones, some systems can actually be used to detect certain audible manifestations of high blood pressure referred to as Korotkoff sounds [1]. In 2011, we presented a solution that uses the microphone on the mobile phone to detect and count

Subject Demographics (N=52)	
Males (n, %)	32 (61.5%)
Age (yrs) (mean, range)	32 (18 – 63)
Height (cm) (mean, range)	172 (152 – 196)
Reported Lung Ailments (n, %)	
Mild Asthma, 9 (17.3%)	Chronic Bronchitis, 2 (3.8%)
Cystic Fibrosis, 1 (1.9%)	Collapsed Lung, 1 (1.9%)
Abnormal Curves (n, %)	12 (23.1%)
Wheeze Present (n, %)	26 (50.0%)
Never Performed Spirometry (n, %)	29 (55.8%)

Table 1. Demographic information for participants.

coughs [20]. Similar to SpiroSmart—this was a tool for measuring a medically accepted quantity without the need for specialized equipment.

DATA COLLECTION PROCEDURE

To evaluate and inform the design of SpiroSmart, we created a dataset of audio samples. In all, 52 volunteers participated in a 45-minute study session (Table 1). All participants self-identified themselves as having none or only mild lung conditions. Our custom data collection application for the iPhone 4S recorded subjects' exhalation sounds using the built-in microphone (at 32 kHz) and provided feedback to the user, coaching them through the spirometry maneuver. We also obtained measurements during the same session using an ATS certified standard clinical spirometer, the nSpire KoKo Legend, as the "gold standard." The KoKo is a pneumotach spirometer and was calibrated with a 3L syringe before each session.

Spirometry measurements are completely effort-dependent and patients are coached through this maneuver by a trained technician. While using the clinical spirometer, participants were coached both orally and with gestures. With SpiroSmart, participants were coached with textual prompts on the screen and only with gestures—oral prompts would have interfered with the audio recording (Figure 3).

SpiroSmart also calculated a real-time estimate of flow (using LPC gain, discussed in the next section), and displayed the measure as a real-time visualization. This also provided an incentive graphic; namely, a ball displaced vertically in a cylinder proportionally to the strength of the exhalations. After the initial burst, the ball dropped slowly to the bottom of the cylinder, signifying the end of the test. Like the KoKo Legend Spirometer, SpiroSmart displayed an estimated Flow *vs*. Volume curve at the end of the effort (Figure 1). An estimate of exhaled volume was calculated by integrating estimated flow with respect to time.

The forced expiratory maneuver was explained to participants and they were asked to practice using the spirometer.
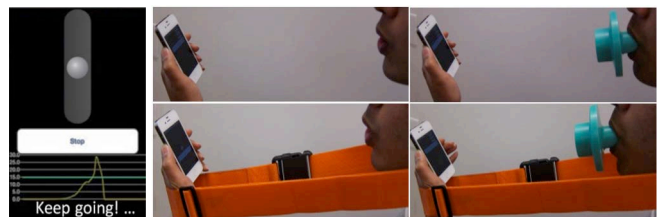

Figure 3. Interface and configurations used to test SpiroSmart.

Once the participants were able to perform an acceptable maneuver according to ATS criteria for reproducibility, three efforts were recorded using the spirometer [22]. The raw flow and volume measurements from the KoKo were obtained using a USB connection and custom software. Next, participants were introduced to SpiroSmart.

In our pilot study, we observed that participants unintentionally varied the distance at which they held the phone as well as lip posture, potentially introducing unwanted variability. We therefore had participants use SpiroSmart in four configurations, in random order: with a mouthpiece (to maintain lip posture), with a sling (to maintain distance), with neither attachment; and with both attachments (Figure 3). Note that it was impossible to collect data from SpiroSmart and the KoKo Legend at the same time so explicit ground truth is unknown. Instead, each effort from SpiroSmart was associated with one randomly selected acceptable curve from the KoKo device during that same session. The signals were aligned using PEF for the KoKo and the maximum amplitude in the audio stream from SpiroSmart as reference points. The audio stream was segmented automatically starting one second before and ending six seconds after the maximum audio amplitude.

We also selected 10 participants to return for two more data collection sessions (2 days up to two weeks apart), allowing us to look at the consistency of measurements from SpiroSmart over longer periods. The participants were asked back based on specific demographics—an equal number of men and women, and equal number of normal and abnormal subjects. We refer to "abnormal" subjects as those with abnormally shaped curves, not necessarily reduced lung function measures. In total, we collected data from 248 clinical spirometer uses and 864 SpiroSmart uses.

Interestingly, 6 subjects were found to have abnormally shaped curves from ailments that they were unaware of and 8 of the 13 subjects who reported lung ailments produced normally shaped curves—albeit with less than expected lung function measures.

ALGORITHM AND THEORY OF OPERATION

Our data collection resulted in a dataset of digitized audio samples from a smartphone. These audio samples are uncalibrated, AC-coupled measures of pressure, $p(t)$. However, we want to convert them into measures of airflow at the lips, $u_{lips}(t)$. Our main goals, then, are (1) to compensate for pressure losses as the sound travels from mouth to microphone, (2) convert the pressure values to an approximation of flow, and (3) remove the effects of AC-coupling. Pressure losses can be approximated using an inverse model of the sound reverberation around the head. Turbulent airflow, as it passes through a fixed opening (*i.e.*, the mouth), has a characteristic pressure drop that we can use for converting pressure into flow. Lastly, we use signal power, frequency characteristics, and models of the vocal tract to remove the effects of AC-coupling and refine the measurement.

Finally, we use regression to combine these approximations and remove non-linearity. Our methodology is broken into two block diagrams: compensation and feature extraction (Figure 4), and machine learning regression (Figure 5).

Distance and Flow Compensation

The first stage in the processing pipeline (Figure 4) is inverse radiation modeling, which compensates for pressure losses sustained over the distance from mouth to microphone, and reverberation/reflections caused in and around the subject's body. The transfer function from the microphone to the mouth is approximated by a spherical baffle in an infinite plane and is given by [10]:

$$H(e^{j\omega}) = \frac{P(e^{j\omega})}{P_{lips}(e^{j\omega})} \sim \frac{j\omega C_{head}}{D_{arm}} \exp\left(-\frac{j\omega D_{arm}}{c}\right)$$

where D_{arm} is the arm length, C_{head} is the head circumference (both approximated from the patient's height), and c is the speed of sound. The transfer function inverse is applied by converting it to the time domain, $h_{inv}(t)$, and using FIR filtering with the incoming audio. Once applied, the output is an approximation of the pressure at the lips, $p_{lips}(t)$.

This pressure value is then converted to a flow rate. For turbulent airflow, the non-linear equation converting pressure drop across the lips to flow rate through the lips is given by (ignoring viscous losses) [10]:

$$u_{lips}(t) \sim 2\pi\, r_{lips}^2 \sqrt{2p_{lips}(t)}$$

where r_{lips} is the radius of the mouth opening (a constant resistance across frequency). Note that some scaling constants from each equation have been removed and the equations are only proportional. This is done because $p(t)$ is not calibrated, so $u_{lips}(t)$ is only proportional to the actual flow rate. Moreover, it is unclear how well these equations perform when using approximations of D_{arm}, C_{head}, and r_{lips} and how non-linearity in the electret microphone affects inverse modeling. Therefore, we use each measure $p(t)$, $p_{lips}(t)$, and
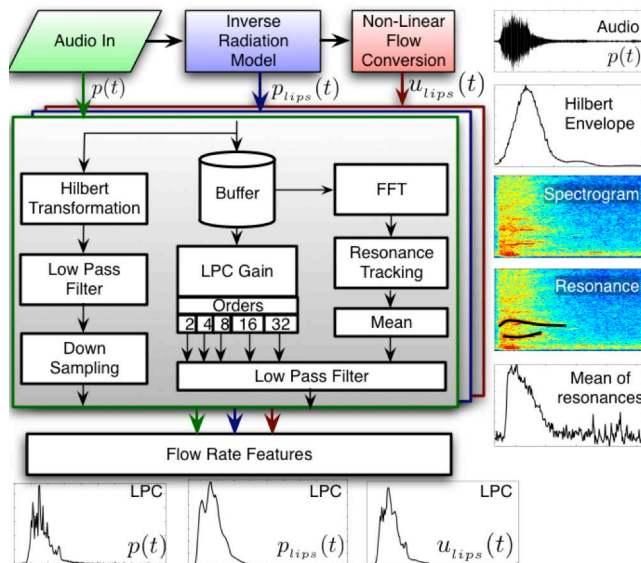

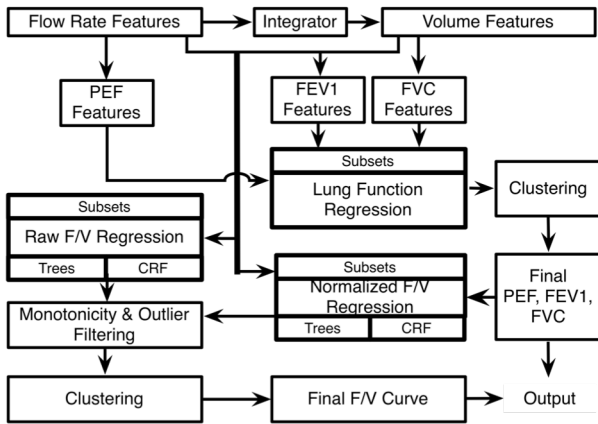**Figure 4. Block diagram of SpiroSmart's feature extraction.**

Figure 5. Block diagram of regression used in SpiroSmart

$u_{lips}(t)$ for feature extraction and let the regression decide which features are most stable.

Feature Extraction

At this point, each measure, $p(t)$, $p_{lips}(t)$, and $u_{lips}(t)$, is a high frequency, AC-coupled signal (Figure 4), from which we want to approximate the volumetric flow rate. We achieve this conversion using three transformations of the signals: (1) envelope detection, (2) spectrogram processing, and (3) linear predictive coding (LPC). The envelope of the signal can be assumed to be a reasonable approximation of the flow rate because it is a measure of the overall signal power (or amplitude) at low frequency. In the frequency domain, resonances can be assumed to be amplitudes excited by reflections in the vocal tract and mouth opening—and therefore should be proportional to the flow rate that causes them. Finally, we can use linear prediction as a flow approximation. Linear prediction assumes that a signal can be divided into a source and a shaping filter and it estimates the source power and shaping filter coefficients. The "filter" in our case is an approximation of the vocal tract [38]. The "source variance" is an estimate of the white noise process exciting the vocal tract filter—in our case this is an approximation of the power of the flow rate from the lungs. The implementation of each stage is explained below.

Envelope Detection: The time domain envelope is taken using the Hilbert envelope. The Hilbert transform of the signal is taken and added back to the original signal, then low pass filtering is used to extract the envelope (an example envelope is shown in Figure 4, callout). Each signal is down-sampled to have the same sampling rate as the spectrogram and linear prediction models.

Spectrogram Processing: During the forced exhalation, the audio from the phone is buffered into 30ms frames (with 50% overlap between frames). Most tests last from 4 to 7 seconds, resulting in 250-500 frames per exhalation. Each frame is then windowed using a hamming window and the $|FFT|_{dB}$ is taken to produce the magnitude spectrogram of the signal. We extract the resonances using local maxima in each FFT frame, calculated over a sliding window (callout in Figure 4). Any maxima that is greater than 20% of the global maximum is saved. After all frames have

been processed, in order to preserve only large and relatively long resonances, any resonance less than 300ms is discarded as noise. Finally, the average resonance magnitude in each frame is calculated and saved (callout in Figure 4).

Linear Prediction Processing: The audio signal is again windowed into 30ms overlapping frames. For each frame a number of LPC models are taken with filter orders of 2, 4, 8, 16 and 32 (increasing vocal tract complexity). The approximated "source power" that excites the filter is saved for each frame as an approximation of the flow rate. Examples of the LPC from using $p(t)$, $p_{lips}(t)$, and $u_{lips}(t)$ are shown at the bottom of Figure 4.

Post Feature Processing: Once the approximated flow rates are returned they are denoised using a Savitsky-Golay polynomial filter of order 3 and size 11 [33]. This operation fits a 3rd order polynomial inside a moving window and is robust to many types of noise while keeping the relative shape of the most prominent signal intact. The filtered and non-filtered signal are both fed as features to the subsequent regression stage.

Machine Learning Regression

The feature extraction results in a number of uncalibrated approximations of the flow rate. These features are used in two different regressions (Figure 5): one to attain specific lung function measures and a second to attain the relative shape of the curve.

Folding: The participants in the dataset are folded into several training subsets, providing a number of diverse models that can be combined to create a global model. For example, one subset randomly divides the participants into ten folds equally. Another subset divides participants with wheezes together into ten folds. Another subset divides the dataset into ten folds, but ensures there are equal numbers of abnormal and normal curves to train on. Each subset is used to create a different regression model and the ensemble can be clustered together to form one decision. Note that for any subset a participant in a testing fold is never used in the training fold. Moreover, to investigate "personalizing" the models, we also create augmented folds that contain data from repeat sessions (for the 10 subjects who performed three sessions spanning multiple days). In this way, "personalized" models are trained using data from the same participant (but on different days) mixed with data from the general model. We evaluate general and personalized models separately.

Lung Function Regression: The output of the feature extraction creates a number of features at 15ms time steps that approximate flow rate over time. If we treat each feature as a flow rate, we can regress directly to PEF, FEV_1, and FVC. For example, PEF is defined as the maximum flow reached in a single effort. Thus, for a curve in a given fold, we can take the max of each feature and regress to the PEF. Note that integrating flow with respect to time gives us the volume of air exhaled. Thus for FVC, the integration of each
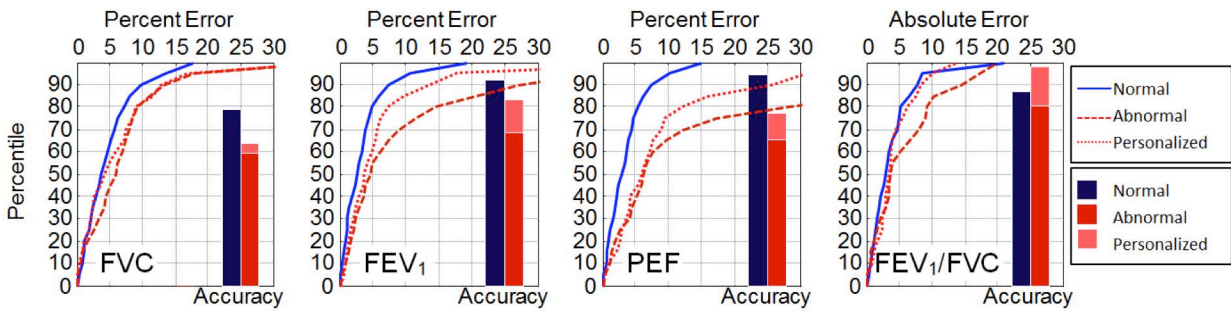
Figure 6. Cumulative error plot of the percentage of the results (y-axis) within the percent error of the x-axis, shown for "normal," "abnormal" and "personalized abnormal" groups. Also shown are the accuracies for each measure, defined as the percentage of measures within accepted clinical limits (*i.e.,* the variation one would expect on a traditional spirometer).

feature is taken. For FEV_1, the integration of the features during the first second is used (Figure 5).

Regression is implemented using bagged decision trees and mean square error; 100 trees are used in each forest. Each training subset is used to predict lung function for a given test instance, resulting in an ensemble of predictions. The final decision is made by clustering the ensemble using k-means (k=2). The centroid of the cluster with the most instances is the final prediction of PEF, FEV_1, or FVC.

Curve Shape Regression: The shape of the curve is a more difficult and involved regression. Instead of a single measure to regress to for each curve, we want to measure the flow rate and volume for each 15ms frame of the curve. The ideal regression should use not only the feature value, but also the correlations across time (as flow rates that are close in time should also be close in value). To exploit this, we employ a conditional random field (CRF) [19] and a bagged decision tree regression. In order to reduce the complexity for CRF based regression, we also generate normalized flow-volume curves aimed at obtaining the correct shape of the curve. We also normalize each feature. Once the shape is obtained, the curve is scaled by measures from the lung function regression stage.

This process is then repeated for the volume (*i.e.*, each feature is integrated and the volume curve is regressed to rather than the flow curve). This results in separately calculated flow and volume curves. Much like the regression for lung function measures, the curves from different subsets are clustered using k-means with k=2 (in this case the area under the curve is used to cluster). Before clustering, outliers are removed and curves that are physically impossible are discarded (*i.e.*, the volume is not monotonically increasing). Finally, the average of the curves in the largest cluster is taken as the predicted flow-volume loop.

RESULTS AND DISCUSSION
In this section, we discuss the performance of SpiroSmart when compared to a clinical spirometer in terms of the accuracy of estimated lung function measures and false positive *vs.* false negative readings. We then discuss the ability of participants to use SpiroSmart without the need for a mouthpiece or sling to control distance. Finally, we discuss the accuracy of the curves SpiroSmart generates and compare different diagnoses from pulmonologists using Spi-

roSmart and a clinical spirometer. Based on our evaluation we conclude that SpiroSmart will meet the needs of home lung function monitoring.

Estimate of Lung Function Measures
We breakdown the comparison of measurements from SpiroSmart and a clinical spirometer by how the percent error is distributed and how well this conforms to accepted clinical variances in each measure.

Distribution of Percent Error in Lung Function Measures
The curves in Figure 6 present the cumulative percentage error plots for FVC, FEV_1, PEF, and FEV_1/FVC. **Explanation:** The horizontal axis on the top shows the percent error between the actual and predicted value. The vertical axis shows the percentile of the lung function measures that are within that percent error. Hence, curves that approach the top left quickly are more accurate for a greater percentage of subjects. The results are categorized by normal subjects, abnormal subjects, and abnormal subjects whose models have been personalized. **Result:** For all lung function measures the algorithms perform best on normal subjects, and abnormal distributions tend to have longer tails. The mean percent errors are 5.2%, 4.8%, 6.3%, and 4.0% for FVC, FEV_1, PEF, and FEV_1/FVC, respectively. When personalization is used, the means improve to 5.0%, 3.5%, 4.6% and 3.6%. The personalized models significantly improve FEV_1, PEF, and FEV_1/FVC for abnormal subjects (based on an F-test of the residual variance, $p<0.05$), but do not significantly improve FVC. Personalized models for normal subjects (not shown) are not statistically significant. **Implication**: SpiroSmart produces results that are consistent with other handheld spirometers, even without personalization [31,39]. The long tails for abnormal subjects require further investigation, but it appears that personalization brings the error distributions much closer to that of normal subjects.

Accuracy of Lung Function Measures
Bar graphs are also shown in Figure 6, displaying the "accuracy" of each measure categorized by normal, abnormal, and personalized, as before. **Explanation:** For FVC, FEV_1, and PEF the accuracy is calculated by finding the number of measures that fall within a certain clinically relevant range. A range is used because the "actual value" of the measure is not specifically known. As previously discussed,
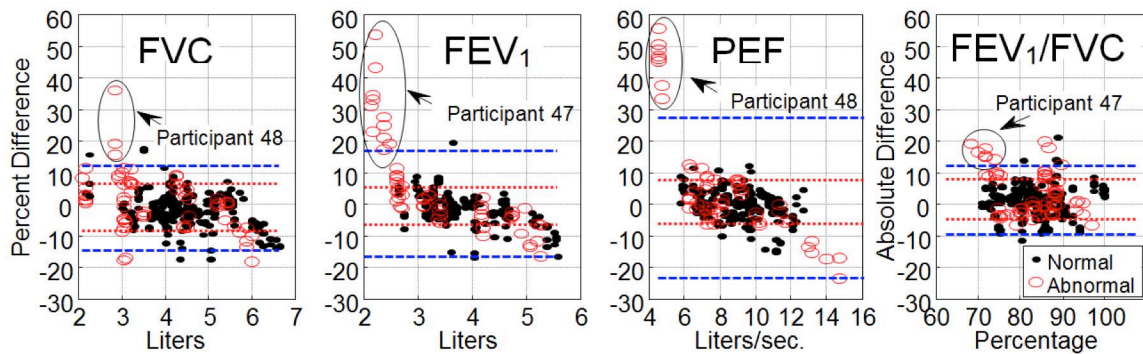
Figure 7. Bland Altman plots of percent error between SpiroSmart and a clinical spirometer versus the value obtained from the clinical spirometer. ±2σ (long dashes) and the 10th and 90th percentiles (short dashes) are also shown. Outliers in each plot are the result of two participants, as shown.

a subject cannot simultaneously use a spirometer and SpiroSmart, so actual ground truth is unattainable. The range is calculated using ATS criteria for what constitutes a "reproducible" lung function measure [22], and published limits of variability for that measure of lung function [8,18,31,32]. For example, a subject's FVC values can consistently be within 0.05 L or within 7% over short durations [22,32]. **Result:** From the accuracies, it is apparent that, for normal individuals, FVC is within the range of expected variability almost 80% of the time and FEV_1 and PEF over 90% of the time. However, there is a significant drop in accuracy for patients with abnormal lung function. **Implication**: Most subjects will almost always produce similar FEV_1 and PEF whether using SpiroSmart or a spirometer. When abnormality is detected, personalization can be performed at the clinic to increase performance. FVC has the least accuracy and it appears personalization has little effect. We discuss this result more in the limitations section.

Explanation: The accuracy of FEV_1/FVC is calculated differently than the other measures. Notice that FEV_1/FVC is already a percentage and the x-axis on top is absolute error (not percent error). FEV_1/FVC is the most common measure of lung function used in diagnosis [36] and there are a number of standards for interpreting the value. For example, if the FEV_1/FVC is less than 80%, the subject is generally considered to have obstructive lung function [36]. If greater than 80%, other measures are used to interpret lung function. The primary motive of this measure is to classify between obstructive, non-obstructive dysfunctional, or normal lungs. Hence we consider estimation accurate (a) if both, spirometer and SpiroSmart estimate FEV_1/FVC values on same side of 80% (either lower or higher); or (b) the two estimates are within ±3 percentage points [22,36]. **Result:** Both normal and abnormal subjects have similar accuracy (80-90% of all results), and personalization increases this accuracy to near 100%. **Implication**: SpiroSmart can be effectively used to diagnose normal vs. obstructed spirometry using FEV_1/FVC, especially when a personalized model is used.

False Positives *vs.* False Negatives
In previous analysis, we did not investigate whether the lung function measure is under-estimated or over-estimated.

Explanation: To elucidate the direction of the bias, Figure 7 shows modified Bland-Altman plots [5] of each lung function measure, showing percentage difference between the spirometer and SpiroSmart versus the spirometer measure. Measures taken from normal subjects are shown as black dots and abnormal subjects are circles. Lines indicating the ±2σ (long dashes) and the 10th and 90th percentiles (short dashes) are also shown. **Result:** From these plots it can be seen that SpiroSmart tends to over-estimate actual value for abnormal subjects (a false negative). However, the most prominent outliers in each plot are from a single individual, which also accounts for the long tails in Figure 6. **Implication**: SpiroSmart generalizes well across normal *and* abnormal participants. However, the real problem is that for some participants (2 of 52 in our dataset), SpiroSmart gives completely erroneous values. We note that for these subjects, SpiroSmart returns values with large dynamic ranges that could be used to detect when a subject is an outlier.

Confounding Factors and Trends
We performed an 8-way ANOVA to determine if other variables significantly contributed to the magnitude of the residual variance (difference between SpiroSmart and spirometer). We used grouping variables of height, weight, gender, recent illness, if a mouthpiece was used, if distance was controlled using a sling, whether the subject had performed spirometry tests before, and how often the subject used a spirometer. The only significant variable was if the subject had used a spirometer in the past ($p<0.001$). The level of agreement and reproducibility between spirometry measures is highly correlated with how experienced the subject is at performing the tests [22]. The only other variable suggestive of being associated with bias magnitude ($p<0.1$) was mouthpiece use.
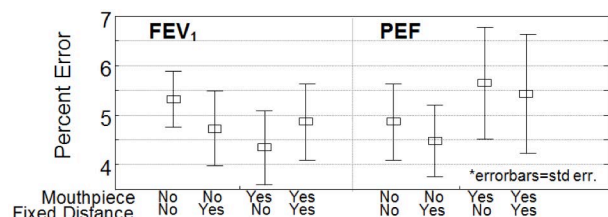
Figure 8. Average percent error for each configuration.

We investigate this further in Figure 8. **Explanation:** Figure 8 shows the average percent error in the lung function measures, categorized by the four different configurations. **Result:** For FVC (not shown), there are no statistically significant differences. For FEV_1, there is a small but significant decrease in the percent error when both a mouthpiece and sling are used—about 0.5% decrease on average. For PEF, there is a significant reduction in the percent error when no mouthpiece is used, about 1% on average. This may at first seem surprising, but recall that our measure of flow rate is derived from a model of the lip reverberation—a factor that is diminished when using the mouthpiece. **Implication:** For general spirometry, the mouthpiece and fixed distances are not indicated. However, for clinical studies that report on a specific spirometry measure, a mouthpiece might be indicated for FEV_1 or contraindicated for PEF.

Participant Feedback

It is interesting that a mouthpiece and fixed distance did not significantly improve measures of lung function. This implies that the subject does not need to carry any additional equipment for performing spirometry outside of the clinic. However, some subjects expressed a preference for using the mouthpiece during the test. In an exit survey, we asked the participants to rate the ease of use and comfort level with each of the configurations. 54% chose the "mouthpiece only" as their strongest preference. Many subjects expressed that it *"felt awkward without the mouthpiece"* (Subj. 22) or that *"I felt like I had to concentrate on the position of my mouth when it was just me and the phone"* (Subj. 35). Thus, it may be more comfortable for some individuals to use a mouthpiece during the tests, even though measures of lung function are unaffected. Some participants felt "*the phone is so much more mobile without the mouthpiece*" (Subj. 43) and "*it made the test more simple*" (Subj. 20). In these scenarios, carrying a mouthpiece (that also needs to be cleaned) is perceived as an impediment.

Evaluating Curves Generated by SpiroSmart

To investigate whether the shapes of curves and lung function measures could be used for diagnosis, we designed an online survey for pulmonologists to interpret curves from SpiroSmart. Three subjects with abnormal curves were selected at random and the two abnormal subjects that made up the outliers in Figure 7, participants 47 and 48. The outlier subjects were chosen to investigate the worst case performance of SpiroSmart. These and five randomly selected subjects with normal curves (a total of 10 subjects) were used in the survey. SpiroSmart curves from the no mouthpiece, fixed distance configuration were used (again, chosen at random). One representative curve from each system was selected for each subject. Per ATS criteria, curve with the largest FVC value was selected [22].

The curves were compiled into an annotated sheet that pulmonologists regularly see summarizing a spirometry session. Along with the resultant volume *vs.* time and flow *vs.* volume plots, the sheets identified subject's age, gender, height, and weight along with the expected lung function

Curve	Range of Diagnoses		Within Rater	
	Spirometer	SpiroSmart	Agree	One-off
1	Norm-Obs. (Mild)	Norm-Obs. (Min.)	3	1
2	Norm-Obs. (Mild)	Norm-Obs. (Mild)	3	1
3	Normal	Norm-Obs. (Mild)	3	1
4	Normal	Normal	5	0
5	Normal	Normal	5	0
6	Obs. (Min.-Mod.)	Restrictive	0	0
7	Obs. (Mild-Mod.)	Norm-Obs. (Mild)	0	2
8	Insuff.-Norm	Insuff.-Norm	3	0
9	Normal	Normal	5	0
10	Normal	Normal	5	0
Total	N/A	N/A	32/50	5/18

Table 2. Summary of survey responses from 5 pulmonologists. Also shown are the number of times a pulmonologist's diagnoses matched (within rater agreement) and were similar to within one degree (within rater one-off).

values [16]. The measured FVC, FEV_1, and FEV_1/FVC were also shown along with the percentage of expected (*i.e.*, measured/expected $\times$ 100). The summary sheet configuration was based on an example from a local clinic. Two digital sheets (from the clinical spirometer and from SpiroSmart) were created for each of the ten subjects.

Five pulmonologists completed the survey over a two-week period. The survey took 20-30 minutes. Four respondents worked primarily in pediatric pulmonology and one in adult pulmonology. Two respondents reported that they use these types of summaries 5-10 times per week, two use them 10-20 times per week, and one used them >20 times per week. The pulmonologists were shown each of the 20 summary sheets in the online survey in random order and were not aware which came from SpiroSmart and which came from the clinical spirometer (they were told that some curves may be from a cell-phone based spirometer).

For the survey, the pulmonologists were asked to view each sheet and select whether the curve suggested normal lung function, obstructed, restrictive, or whether the information presented was insufficient. If obstructed, classification into minimal, mild, moderate, or severe was sought (*i.e.*, a total of 7 possible diagnoses). These categories are the exact categories used on standard spirometry sheets used by pulmonologists [22]. There was also an optional comments block, where the doctors could enter any additional information about the diagnosis.

Explanation: Table 2 summarizes the results of the survey. For each curve, the range of responses is shown for the spirometer and SpiroSmart. Also shown are the "within rater agreement" (the number of times the ratings from one pulmonologist agreed exactly) and the "within rater, one off" (the number of times ratings from the same pulmonologist are off by one degree). For example, "Mild" and "Moderate" obstructions would be considered one away from each other and so would "Normal" and "Minimal Obstruction". "Restrictive" and "Insufficient" were not considered 'one off' from any other response. Note that curves 6 and 7 are from the subjects whose measures were considered outliers in Figure 7. **Result:** Overall, the pulmonologists had strong

agreement with one another. For normal curves (3, 4, 5, 9 and 10) there was a general agreement from the pulmonologist regardless of whether they used SpiroSmart or a traditional spirometer. 23 of the 25 responses matched identically and 2 responses were false positives rated as Minimal or Mild obstructions. Curves 1, 2, and 8 are rated with 9 of 15 responses matching exactly and 2 being one degree off. Of the four remaining, three were false positives, indicating worse obstruction or insufficient effort, and one was a false negative (mild obstruction categorized as normal). **Implication**: For these eight subjects the range of diagnoses for each subject is nearly identical. It appears SpiroSmart can be effective for diagnosing not only if the patient is obstructed, but also the degree of the obstruction. More data is still needed from a larger survey and database to establish clinical significance. For the data presented here, the initial results appear very promising.

As expected most disagreement occurred in curves 6 and 7 (the outliers marked in Figure 7). Result: For curve 6, all respondents thought the curve was restrictive because of a falsely high FEV_1 value. The low FVC indicates the abnormality is not from obstruction, but from restriction. Even so, it is encouraging that the curve is judged as abnormal on both the Spirometer and SpiroSmart. For curve 7, the FEV_1/FVC is falsely high in SpiroSmart, resulting in three of the five responses being false negatives. Implication: In this worst-case scenario, the inflated lung function values caused diagnostic disagreement in pulmonologists, despite the curves having a similar "scooping" shape.

LIMITATIONS AND FUTURE WORK

We conclude that SpiroSmart will meet the needs of home lung function measurement. However there are some limitations of SpiroSmart. First, inhalation measures are also recorded in clinical spirometry, but are seldom used in diagnosis. This inhalation is inaudible and hence, cannot be measured with SpiroSmart. Another limitation is that the setting where the test is performed must remain relatively quiet. This also precludes the coaching that normally occurs during spirometry for children and adolescents. A technician motivates the user with high-energy explanations like "really push it out" and "keep going." With SpiroSmart, these motivations would need to be handled from a user interface or embedded into the playing of a blowing game. It will be interesting to analyze how incentives and motivations could be used with SpiroSmart to improve compliance, reproducibility, and accessibility to children with chronic lung function ailments.

Another limitation of SpiroSmart is that there is a decreased accuracy in FVC compared to other measures (even when personalization is used). This may be a limitation of audible flow rate sensing. Only about the first three seconds of the test is audible. The first two seconds make up 90-100% of the flow-volume loop and includes $FEV_1$and PEF (which could explain why these measures are so accurate). However, the remainder of the test is mostly inaudible. We believe that the regression is extrapolating what the FVC should be,

rather than regressing from the audible signal. Thus, it is worth exploring if FVC accuracy can be increased by creating an interface that requires users to bring the phone closer to their mouth as the test progresses, and thereby boosting the signal-to-noise ratio at the end of the test.

It is unclear from the current study how well SpiroSmart tracks trends in lung function. SpiroSmart consistently predicted lung function for subjects over repeat sessions, but none of the subjects' lung function changed from session to session. Although there is no indication that SpiroSmart will poorly track trends, larger clinical studies are currently underway to verify such an important outcome. Additionally, these studies also aim to evaluate SpiroSmart outside of a lab setting.

Lastly, current smartphones cannot simultaneously calculate all the features of SpiroSmart in real time and need to do the analysis offline or in the cloud. Our current implementation uses the cloud. The lung function measures are each calculated through a bagged decision tree, and therefore should be straightforward to implement on a phone, rather than in the cloud. However, the calculation of the flovolume loop requires complex computation and is more appropriate for implementation on a server. Even so, a laptop with 4GB memory can calculate the flow volume loop in less than 30 seconds. As such we believe such computational capabilities are not too far away to be available on smartphones.

Given these limitations, it is useful to examine the context and utility of this technology. The current implementation of SpiroSmart is most useful for replacing home spirometers of individuals with chronic lung ailments. These patients are already familiar with the maneuver, can easily be in a quiet environment somewhere in their home, and likely have access to a smartphone. Smartphones are, at times, accused of widening the "digital divide" and push people in low socio-economic environments further away from facilities. However, the cost of current home spirometers has already done that. SpiroSmart lowers this cost, helping devices to pervade such environments. As smartphones become cheaper, technologies like SpiroSmart will help in improving access to medical devices and, rather, narrow the digital divide. Another group that may initially encounter problems adopting SpiroSmart is individuals who have never used a spirometer. Once we better understand ways to train individuals to perform a spirometry maneuver with SpiroSmart, we can lower the adoption- and access-barrier. This can then lead to wider adoption of spirometry in other areas like non-chronic disease management, air quality effects, allergic reactions, *etc*.

CONCLUSION

In conclusion, we presented SpiroSmart, an accurate mobile phone spirometer that uses the built-in microphone to infer flow rate during a spirometry test. Our evaluation implies that, although SpiroSmart cannot replace a clinical spirometer, it may be used to replace many handheld monitors and

home spirometers in a variety of cases. SpiroSmart is not intended as a substitute for clinical spirometry, but rather a home-based solution that can increase compliance and monitoring through the convenience of a mobile phone. SpiroSmart can be used as an out-of-the-box system (without calibration), but certain individuals may benefit from personalized models.

REFERENCES

1. Allen, J. and Murray, A. Time-frequency analysis of Korotkoff sounds. *IEE Seminar Digests 1997*, 6 (1997).

2. Alshaer, H., Fernie, G.R., and Bradley, T.D. Phase tracking of the breathing cycle in sleeping subjects by frequency analysis of acoustic data. *International Journal of Healthcare Technology and Management 11*, 3 (2010).

3. Amft, O. and Lukowicz, P. Analysis of chewing sounds for dietary monitoring. *UbiComp'05*, (2005).

4. Bishara, W., Su, T.-W., Coskun, A.F., and Ozcan, A. Lensfree on-chip microscopy over a wide field-of-view using pixel super-resolution. *Opt. Express 18*, 11 (2010).

5. Bland, J.M. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet 327*, 8476 (1986).

6. Brouwer, a F.J., Roorda, R.J., and Brand, P.L.P. Home spirometry and asthma severity in children. *The European Respiratory Journal 28*, 6 (2006).

7. Brunette, W., Sodt, R., Chaudhri, R., et al. The Open Data Kit Sensors Framework : Application-Level Sensor Drivers for Android. *MobiSys*, (2012).

8. Cochrane, G.M., Prieto, F., and Clark, T.J. Intrasubject variability of maximal expiratory flow volume curve. *Thorax 32*, 2 (1977).

9. Finkelstein J, Cabrera MR, H.G. internet-based home asthma telemonitoring: can patients handle the technology. *Chest 117*, 1 (2000).

10. Flanagan, J. *Speech Analysis, Synthesis, and Perception*. Springer-Verlag, Berlin - Heidelberg - New York, 1972.

11. Grimaldi, D., Kurylyak, Y., Lamonaca, F., and Nastro, A. Photoplethysmography detection by smartphone's videocamera. *IDAACS*, (2011).

12. Grzincich, G., Gagliardini, R., and Bossi, A. Evaluation of a home telemonitoring service for adult patients with cystic fibrosis: a pilot study. *J. of Telemedicine*, (2010).

13. Gupta, S., Chang, P., Anyigbo, N., and Sabharwal, A. mobileSpiro: accurate mobile spirometry for self-management of asthma. *Proceedings of the First ACM Workshop on Mobile Systems, Applications, and Services for Healthcare*, ACM (2011), 1:1--1:6.

14. Homs-Corbera, A. and Fiz, J. Time-frequency detection and analysis of wheezes during forced exhalation. *IEEE Transactions 51*, 1 (2004).

15. Kessler R, Stahl E, Vogelmeier C, Haughney J, Trudeau E, Lofdahl CG, et al. Patient understanding, detection, and experience of COPD exacerbations: an observational, interview-based study. *Chest 130*, (2006).

16. Knudson, R.J., Slatin, R.C., Lebowitz, M.D., and Burrows, B. The maximal expiratory flow-volume curve. Normal standards, variability, and effects of age. *The American review of respiratory disease 113*, 5 (1976).

17. Kroutil, J. and Laposa, A. Respiration monitoring during sleeping. *ISABEL'11*, (2011).

18. Künzli, N., Ackermann-Liebrich, U., Keller, R., Perruchoud, A.P., and Schindler, C. Variability of FVC and FEV1 due to technician, team, device and subject in an eight centre study: three quality control studies in SAPALDIA. *European Respiratory Journal 8*, 3 (1995).

19. Lafferty, J., McCallum, A., and Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proc. Int. Conf. on Machine Learning*, (2001).

20. Larson, E.C., Lee, T., Liu, S., Rosenfeld, M., and Patel, S.N. Accurate and Privacy Preserving Cough Sensing using a Low-Cost Microphone. *Proceedings of the 13th ACM international conference on Ubiquitous computing*, (2011).

21. Majchrzak, T. and Chakravorty, A. Improving the Compliance of Transplantation Medicine Patients with an Integrated Mobile System. *International Conference on System Sciences*, (2012).

22. Miller, M.R., Hankinson, J., Brusasco, V., et al. Standardisation of spirometry. *The European Respiratory Journal 26*, 2 (2005).

23. MiravitallesS, M., Murio, C., Guerrero, T., and Gisbert, R. Pharmaco-economic evaluation of acute exacerbations of chronic bronchitis and COPD. *Chest 121*, 5, 1449-1455.

24. Neuman, M.R. Vital Signs: Heart Rate. *Pulse, IEEE 1*, 3 (2010).

25. Nishimura, J. and Kuroda, T. Eating habits monitoring using wireless wearable in-ear microphone. *ISWPC 2008*, (2008).

26. Otulana, B., Higenbottam, T., Ferrari, L., Scott, J., Igboaka, G., and Wallwork, J. The use of home spirometry in detecting acute lung rejection and infection following heart-lung transplantation. *Chest 97*, 2 (1990).

27. Pamplona, V.F., Mohan, A., Oliveira, M.M., and Raskar, R. NETRA: interactive display for estimating refractive errors and focal range. *SIGGRAPH'10*, ACM (2010), 77:1--77:8.

28. Pesola, G., O'Donnell, P., and Jr, G.P. Peak expiratory flow in normals: comparison of the Mini Wright versus spirometric predicted peak flows. *Journal of Asthma*, 4 (2009), 845-848.

29. Poh, M.-Z., McDuff, D.J., and Picard, R.W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express 18*, 10 (2010).

30. Poh, M.-Z., Swenson, N.C., and Picard, R.W. Motion-Tolerant Magnetic Earring Sensor and Wireless Earpiece for Wearable Photoplethysmography. *Information Technology in Biomedicine, IEEE Transactions on 14*, 3 (2010).

31. Rebuck, D. a., Hanania, N. a., D'Urzo, a. D., and Chapman, K.R. The Accuracy of a Handheld Portable Spirometer. *Chest 109*, 1 (1996).

32. Rubini, A., Parmagnani, A., Redaelli, M., Bondì, M., Del Monte, D., and Catena, V. Daily variations of spirometric indexes and maximum expiratory pressure in young healthy adults. *Biological Rhythm Research 41*, 2 (2010).

33. Savitzky, A. and Golay, M.J.E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry 36*, 8 (1964).

34. Seemungal, T. a, Donaldson, G.C., Bhowmik, A., Jeffries, D.J., and Wedzicha, J. a. Time course and recovery of exacerbations in patients with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine 161*, 5 (2000).

35. Sevick, M., Trauth, J., Ling, B., et al. Patients with Complex Chronic Diseases: Perspectives on Supporting Self-Management. *Journal of General Internal Medicine 22*, 0 (2007).

36. Swanney, M.P., Ruppel, G., Enright, P.L., et al. Using the lower limit of normal for the FEV1/FVC ratio reduces the misclassification of airway obstruction. *Thorax 63*, 12 (2008).

37. Townsend, M.C. Spirometry in the occupational health setting. *Journal of occupational and environmental medicine / American College of Occupational and Environmental Medicine 53*, 5 (2011).

38. Wakita, H. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *Audio and Electroacoustics, IEEE Transactions on 21*, 5 (1973), 417–427.

39. Walters, J., Woodibaker, R., and Walls, J. Stability of the EasyOne ultrasonic spirometer for use in general practice. *Respirology 11*, 3 (2006).

40. Ölmez, T. and Dokur, Z. Classification of heart sounds using an artificial neural network. *Pattern Recognition Letters 24*, 1–3 (2003).