Hindsight is 20/20: Retrospective Lessons for Conducting Longitudinal Wearable Sensing Studies

Salaar Liaqat
University of Toronto
Toronto, Canada
sliaqat@cs.toronto.edu

Daniyal Liaqat
University of Toronto
Toronto, Canada
dliaqat@cs.toronto.edu

Tatiana Son
University Health Network
Toronto, Canada
Tatiana.Son@uhnresearch.ca

Andrea Gershon
Sunnybrook Health Sciences Centre
Toronto, Canada
Andrea.Gershon@sunnybrook.ca

Moshe Gabel
University of Toronto
Toronto, Canada
mgabel@cs.toronto.edu

Robert Wu
University Health Network
Toronto, Canada
Robert.wu@uhn.ca

Eyal de Lara
University of Toronto
Toronto, Canada
delara@cs.toronto.edu

***Abstract*—Pervasive sensing using wearables for health monitoring presents a promising and unique opportunity to widely manage illnesses and conditions. To better understand the capabilities and limitations of using wearable devices for health monitoring, systems need to be developed and studies conducted. We conducted one such study for monitoring patients with Chronic Obstructive Pulmonary Disease (COPD), in which we aim to understand the disease and predict patient outcomes. However, despite a carefully well-planned and well-conducted study that resulted in a very large dataset, some non-obvious design oversights meant the data was much less useful. We analyze the shortcomings of our study to construct lessons and concrete actions to avoid these pitfalls. We ratify these lessons by briefly discussing a second iteration of our study, in which we apply these lessons and obtain much better outcomes. Real-world sensing studies are time consuming and expensive investments, for a promising research area. By sharing our failure and proposing actionable lessons, we hope to minimize the risk for others aiming to run such studies.**

***Index Terms*—Pervasive computing, Study design, Health care, Sensors, Smartwatches**

I. Introduction

Pervasive monitoring is a promising area for improving how healthcare is delivered to large populations. Modern smart devices have powerful sensors capable of collecting continuous information, providing insights about the user [1]–[3]. Remote monitoring can be conducted in many forms, ranging from unobtrusive smartwatch sensing of chronic conditions, to contact tracing based monitoring to detect exposure to a contagious disease [4]. Regardless of form of remote monitoring, widespread adoption requires that devices and systems be validated in real world environments [5]–[8].

Chronic Obstructive Pulmonary Disease (COPD) is a respiratory illness that is most common in the older population. People with COPD experience respiratory symptoms such as coughing, wheezing and difficulty breathing. Additionally, due to environmental or physiological factors, people with COPD can sometimes experience episodes of worsened symptoms and health, called *exacerbations*. If gone unnoticed and untreated, exacerbations can require hospitalizations to address. Thus, early detection and intervention for people with COPD is crucial for improved quality of life [9].

In order to enable remote patient monitoring for early detection of COPD exacerbations, we designed, developed and deployed WearCOPD. WearCOPD is a longitudinal study designed to explore the use of smartwatch based sensing for long-term remote monitoring of people with COPD to detect exacerbations episodes before they occur [10]. The study consisted of a smartwatch which continuously collects heart rate, movement and audio, and a smartphone with a daily questionnaire to track patient symptoms and establish occurrences of exacerbations. The study followed patients for 3 months, after which the devices were returned.

However, upon completion of this study and the start data analysis, we discovered oversights in the design of our study and collection of the data. This rendered our data and potential analysis severely limited for achieving our goal of predicting COPD exacerbations.

Our most significant pitfall was waiting until the study ended before looking at the data. Once we started the analysis, the consequences of all our other oversights became apparent, at a point where it was too late to address them. We found that participants were not always diligent with using the devices, oftentimes, not using the system for several days or in some cases, the entire study. There were also software and device issues which caused missing data or data anomalies, such as heart rate readings when the device was not worn. Despite the technical issues in the data collection, we still collected plenty of data, however, we ran into another issue. We lacked any form of ground truth for our data. The sensors for the watch had not been validated for accuracy for our population of people with COPD. We found the readings could vary based on the participant and we had no way of identifying events of interest in our data. Additionally, due to the design of the daily questionnaire, the responses from our participants were unreliable. This led to no way of establishing the occurrence of exacerbations. Overall these issues resulted in poor data collection and little possible analysis with our data.

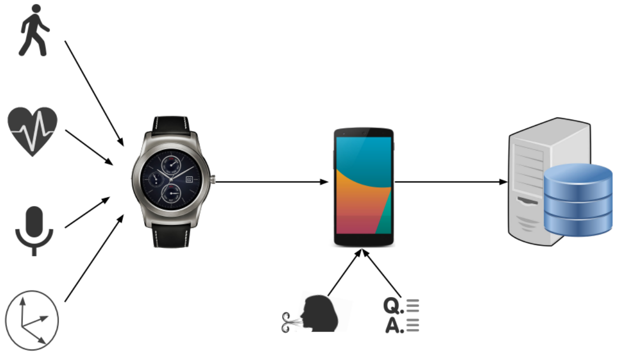


Fig. 1. WearCOPD data collection architecture design

In this paper, we describe the intended goals and design of WearCOPD. We describe in detail the repercussions of our choices and why we were unable to perform robust analysis. Based on the flaws in WearCOPD, we contribute a list of considerations when designing real-world studies. We justify these considerations by describing a second iteration of WearCOPD in which we implement these changes and collect a more insightful dataset. We found these recommendations will limit the risk of poor data collection, validate the collected dataset and ensure a smooth deployment of the study. Running wearable sensing studies presents a valuable and unique opportunity for developing continuous health monitoring. By understanding the potential pitfalls, we hope to help others avoid them and run successful studies.

II. The WearCOPD Study

The goal of our research was to use smartwatches to detect exacerbations of COPD before they occur. To do this, we designed a longitudinal study to monitor COPD patients during their every day lives. We used a smartphone-based daily questionnaire to establish the condition of the patient's COPD for the day, and a smartwatch to continuously collect sensor data from the patient. We planned on using the daily questionnaire data as ground truth for establishing when exacerbations occurred and then exploring the smartwatch data for indicators.

From the smartwatch microphone, we foresaw building a cough detection algorithm [6] or a speech analysis system. Additionally, using the accelerometer and gyroscope, we envisioned specialized activity monitoring algorithms and behavioural analysis. These metrics and analysis could be used in the overall system to detect and predict the onset of exacerbations of COPD.

A. Design

WearCOPD is an REB approved longitudinal study that monitored patients with COPD for 3 months (REB 15-9068). During this period, patients were provided two devices. A smartwatch, worn every day, collected continuous heart rate, movement, and audio data. This allowed us to measure and observe the user's physiological and behavioural patterns. Additionally, a smartphone was provided for the patients. The smartphone was pre-loaded with a questionnaire asking about the patient's symptoms related to their COPD, which they completed every day. During the nights, the patients were instructed to charge the devices in order to have them ready for the following day. When devices were charging, the data for the day was uploaded to a remote server. The architectural design for the data collection can be seen in Figure 1.

The questionnaire we used was a clinically-validated list of questions used by doctors to establish the occurrence of COPD exacerbations [11]. Traditionally, this was a paper questionnaire conducted by doctors to their patients, which we adapted to be delivered asynchronously on a smartphone. The questionnaire, seen in Table I, asks patients a series of Yes/No questions about changes in various symptoms from the patient's baseline. These questions can be classified into minor symptoms which have a score of 1, or major symptoms which have a score of 5. When a participant reports symptoms whose scores add up to 6 or greater for two consecutive days, that is classified as an exacerbation. We also included an additional question asking whether the participant was in the hospital. The response to these questions would indicate if the patient was in an episode of exacerbation, or in the hospital, and we aimed to use the data collected from the smartwatch to identify patterns and build prediction models.

Before starting the study, we conducted a short beta test to evaluate the various components of the system. The testers included the development team, as well as other research colleagues. Testers were given the smartwatch and smartphone for a short period of time, and asked to follow the study procedure of wearing the smartwatch all day and answering the questionnaire daily. The technical components and data collection of the system were tested and feedback from the testers was incorporated into the system.

The selection criteria for WearCOPD required participants to have moderate to severe COPD and be comfortable using mobile devices. Additionally, during onboarding, they provided informed consent, demonstrating they understood the goals and intentions of the study and agree to participate. Participants were then given instructions on how to use the application and how to care for their device. This concluded the onboarding session, and they were then sent on their way, with the devices, to continue with their everyday life.

Though patients could reach out during the study if they had questions or concerns, in practice they settled into their life had little contact with the research team during the study. If participants wished to stop participation in the study, they could contact the research team. Otherwise, after 3 months, the patients returned their devices and their participation in the study ended.

III. Results and Issues

Over the course of the WearCOPD study, 16 patients were recruited and monitored for a duration of 3 months. The study yielded 52 gigabytes of data, containing 1173 days of data and 897 symptom reports. Meaning, on average participants

TABLE I
DAILY SYMPTOM QUESTIONNAIRE

Question	Score
1. Did you experience any increased breathlessness?	5
2. Did you experience any increased sputum colour?	5
3. Did you experience any increased sputum amount?	5
4. Do you have a cold?	1
5. Did you experience any increased wheeze or chest tightness?	1
6. Do you have a sore throat?	1
7. Do you have an increased cough?	1
8. Do you have a fever?	1
9. Are you in the hospital?	N/A

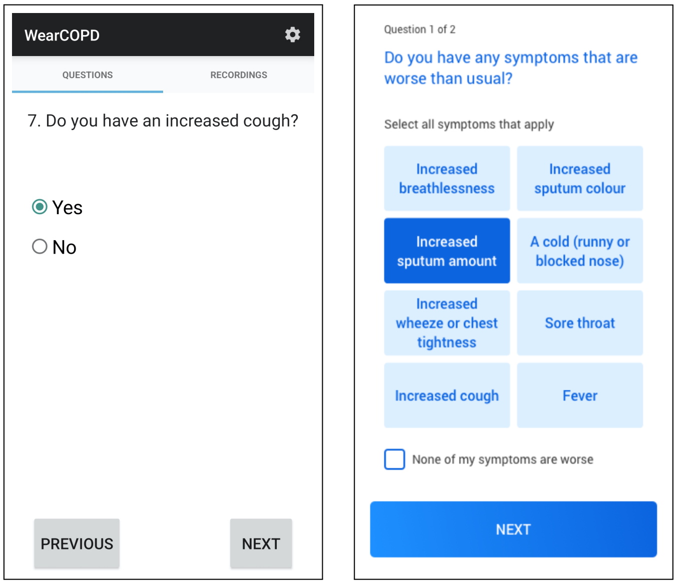
provided 73 days of data, of which 76% had a symptom report. Unfortunately, despite our initial excitement at the wealth of collected data, we quickly found several issues with our study design that prevented robust analysis.

A. Missing data

Missing data was a common and pervasive occurrence in our dataset for a number of reasons. Most commonly, patients would not use some component of the system, such answering the questionnaire, or in some occasions, stop using both the smartwatch and smartphone. This could occur sporadically for a few days throughout the study, or sometimes for the entire duration of the study. Another cause for missing data was patients inadvertently using the devices incorrectly, such as wearing the smartwatch while it is not powered on. Lastly, due to device or connectivity issues, the phone would occasionally stop uploading data. In these cases, the data needed to be retrieved once the devices were returned. Although in our case, all data was retrieved successfully, there was an added risk of data loss if devices malfunctioned or were not returned.

B. Poor Ground Truth Data

A shortcoming of our questionnaire was the variability in interpretation among participants. As a result, when we started analyzing this data, we saw that many patients were in a constant state of exacerbation (defined as score of 6 or higher for two consecutive days). This was caused by participants misinterpreting the question. For instance, some participants answered the question “Do you have an increased cough?” with “Yes” every day, since, with COPD, they always have an increased cough compared to when they were healthy. However, our intended interpretation of “increased cough” was a change from their baseline. Yet another interpretation was an increase in cough since the last questionnaire, which may or may not have been in the previous day. Additionally, there was ambiguity about the time frame the response covers. For instance, if patients answered the questionnaire first thing in the morning, the patient’s response could be referring to the previous night, overnight, or for the morning. These two ambiguities rendered the questionnaire data completely unusable for establishing the ground truth for COPD exacerbations.



7. Do you have an increased cough?

● Yes
○ No

Question 1 of 2

Do you have any symptoms that are worse than usual?

Select all symptoms that apply

□ None of my symptoms are worse

Fig. 2. Screenshots of the questionnaire from the original WearCOPD study (left) and the revised study based the lessons learned (right). The revised questionnaire is less ambiguous, and requires only a handful of taps to answer.

Beyond the ambiguous wording of the questionnaire, we found the questionnaire UI was was too burdensome. Each question of the questionnaire had a separate page, where the user had to select a “Yes” or “No” button, and then select the “Next” button. This meant, to answer 9 questionnaires, users needed to traverse 9 pages, and press 18 separate buttons. Some patients found the UI to burdensome to use, resulting in fewer questionnaires answered or questionnaires being answered with less accuracy. A screenshot of the questionnaire can be seen in Figure 2 (left).

C. Sensor Data Quality

We had collected hundreds of hours of audio, with the intention of analyzing coughs and speech using automatic, state-of-the-art tools. However, given the unique sensing environment of the smartwatch, which is often in movement, and can be covered by clothing, we found that detecting the coughs and speech itself was a challenge [12]. Our original intention of using state-of-the-art algorithms and machine learning models, did not generalize to our audio dataset [6]. Additionally, we had no examples of such events of interest available to develop in-house algorithms. Ultimately, we overcame this through an expensive and time-consuming manual annotation process to identify events of interest in the audio data.

Since COPD arises later in life, most of our participants were older adults, which is not the target audience for smartwatches. This often meant the sensors and algorithms developed for smartwatches weren’t designed with this population in mind. For instance, step count may struggle with the slower movement of our population, or if a participant requires assistance from a cane, step count could become inaccurate. Furthermore, depending on skin characteristics of the patient, the amount and quality of reported heart rate data can vary [13]. Because the smartwatch was not validated for our population of interest, we were unable to gauge the reliability of the readings we observed.

An aspect of the study we didn't anticipate was the amount of anomalies that would occur in the data over the span of 3 months. For example, in one instance a participant went on vacation in a different timezone, while the uploaded data still maintained the timestamp of the original timezone. This caused issues with some data analysis, as we would have data corresponding to unexpected times. For instance, elevated heart rate at 4 a.m. raises concerns whereas 10 a.m is less worrying. We also observed heart rate data when the device was being charged on the dock, overnight, along with variations in the reported confidence from the smartwatch of these readings. These types of anomalies in sensor readings differed based on smartphone manufacturer, and their specific implementation of the WearOS API [13]. Addressing these inconsistent sensors required testing and time to identify and incorporate adjustments in our analysis. Additionally, if anomalies like these go unnoticed, then downstream models and algorithms may be noisy, biased or nonsensical.

IV. Lessons

Having identified the flaws in WearCOPD, we now describe what we could have done differently to avoid these pitfalls, structured as actionable lessons. These lessons are summarized in Table II. Additionally in each subsection, we describe how we implemented these lessons in a second iteration of WearCOPD [6], and the benefits we observed.

A. Perform Active Data Analysis and Monitoring

The first lesson we found was to conduct the intended data analysis once the data starts being collected. Running data analysis will quickly identify flaws in the study design, such as incorrect data collection or missing components to the study. If this is done while the study is being run, these fixes can be implemented and deployed.

Alongside data analysis, we learnt studies need active monitoring for compliance. Which in turn requires that the study design, from the beginning, incorporates: (1) a person who's responsible for active monitoring of compliance; (2) a dashboard that enables the simple viewing of patients and their data and (3) a way to contact patients during the study. This person should be able to identify poor compliance, or data anomalies shortly after they occur. Then, using the communication channels established during enrollment, contact the participant to address any issues. Additionally, the reason for the data anomaly can be recorded and considered during data analysis.

When we conducted our second iteration of WearCOPD, we built a dashboard that would display all patients, the data uploaded, and a summary of each patient's data. We hired a patient coordinator to monitor the dashboard everyday to ensure data was being uploaded as expected. Additionally, the dashboard incorporated automatic alerts to indicate when patients were missing data or when COPD exacerbations reported from the questionnaire. The patient coordinator could contact these patients to resolve any issues and log explanations into the system. Throughout the study, we observed many events which caused data anomalies, including hospital visits, personal injuries, changing timezones and device issues. In all these cases, we were able to quickly identify the issue, log an explanation into the dashboard and provide a fix if applicable.

B. Collect Instances of Clean Data

Smartwatches are generally not validated data collection devices even for in-lab settings [14], and in-wild data tends to be noisy. We learnt that running experiments with our participants in a controlled environment at the start of the study would provide valuable information for later data analysis. These experiments should collect events of interest, in our case coughs and speech, as well as regular activities, such as walking, sitting and lying down. Collecting this information in a controlled environment with our participants would allow for a comparison with the same participant, in the wild. We recommend doing an in-lab session at the start of the study during enrollment. Participants can get introduced to the study, get familiar with the devices, and provide a baseline of their condition.

Collecting data at start of the study provides baseline information for the devices and participants. However, because the participants have a chronic lung condition, their baseline can evolve with the disease. For instance, their speech may change [15] or they may start to use a cane to assist in walking. For this, we learnt that routine in-lab checkups, in which they repeat the initial controlled experiments would allow us to track these changes.

Additionally, allowing for patients to run semi-controlled experiments on their own can increase the amount of labelled data. This can look like a mobile application in which the participant does breathing exercises, reads passages, or does voluntary forced coughs. Increasing the amount of labelled information being collected for participants would help manage the abundance of unlabelled data that is collected in the wild. We recommend ensuring these experiments be optional, as to minimize the burden on participants.

To summarize, we suggest:

- Run in-lab experiments when enrolling patients to establish their sensor footprint in a noise-free environment
- Run in-lab experiments routinely throughout the study to capture any changes or variations in their sensor data
- Incorporate a method for patients to actively provide labelled data during in an out-of-lab setting

In our second iteration of WearCOPD, we included 3 in-lab components, in which we would run experiments to capture various physical such as walking tests. We would also run auditory tasks, such as scripted speech, voluntary cough, spirometry, and tidal breathing. These would be run during onboarding, offboarding and halfway through the study. These experiments provided us with clean data for each patient, and allowed us to explore the use of patient-specific models for speech analytics.

TABLE II
ACTIONABLE LESSONS FROM THE WEARCOPD STUDY

Lesson	Discussion
Start data analysis early, and continue analysis throughout the duration of the study	Sec. IV-A
Build a dashboard and have a dedicated person actively monitor data collection throughout the study	Sec. IV-A
Collect clean data from participants during onboarding	Sec. IV-B
Perform routine in-lab check-ups with participants	Sec. IV-B
Provide optional exercises for participants to provide clean data in the wild	Sec. IV-B
Validate sensors and the sensor platform	Sec. IV-C
Minimize effort required of study participants	Sec. IV-D
Test with representatives of the target population	Sec. IV-E

C. Validate Sensors and Platforms

We also recommend validating the accuracy of the sensors and smartwatch for each participant. This can be included in the in-lab experiments, where medically validated devices are also provided to the participant during each activity. These gold standard devices will provide accurate readings, and allow comparison with the corresponding values received from the smartwatch. In turn, this establishes a level of confidence on the smartwatch reported readings, once the participant starts using them in the wild.

In WearCOPD version 2, we used a Zephyr Bioharness [16], which is a medically validated chest band for measuring heart rate, oxygen saturation and movement. We used it in all our in-lab experiments, and were able to establish the accuracy of the smartwatch reported heart rate and oxygen saturation.

D. Minimize Active Patient Participation

We found having a burdensome tasks during the study caused for poor adherence. This was most evident in the questionnaire, in which a minimum of 18 interactions throughout 9 screens was required daily. Thus, we recommend that when designing longitudinal studies, minimize the effort required from the patient. By making processes as streamlined and easy as possible, patients will be more likely to complete the required tasks.

When deployed for a second time, we redesigned the questionnaire user interface to improve the patient experience and reduce time. The questionnaire fit on a single page, and each question had a large toggle button for “Yes” and “No”, as well as a “Submit” button on the bottom. This design of the questionnaire reduced the amounts of required clicks by half and had no transitions to new pages. Additionally, we phrased the questions more clearly to remove ambiguity and improved the appearance of the user interface. From this study, we found better upload rate for the questionnaire, as well as fewer misinterpretations of the questions. The improved questionnaire can be seen in Figure 2 (right).

E. Testing

During our initial beta tests, we involved fellow researchers outside of the development team, as their non-technical background would provide a fresh perspective. However, we would instead recommend, using a sample from the target population, as their perspective represents the issues and confusions that will be faced in the study. As our participants are generally older and less tech-savvy, these issues can look like tech confusions, or they may not be familiar with terminology that is standard within the research team. Alternatively, a slow roll out of the study with close communication with the participant can serve a similar purpose. This can be utilized alongside data analysis to test the application and data collection.

When deploying the second iteration of WearCOPD, we initially launched the system with a single participant. We closely monitored the use of the system by this single participant, and slowly incorporated more participants, as our confidence in the system grew. This allowed us to detect and fix any issues in the system without the overhead of managing many patients.

V. CONCLUSION

In this work, we discuss WearCOPD, our longitudinal in-wild wearable sensing study to explore the use of smartwatches to monitor COPD patients. The study consisted of months of effort dedicated to design, development and deployment. However, upon completion of the the data collection phase, we found that due to our choices in study design, the quality of data we collected was not sufficient for realizing our goal of studying COPD. Participant adherence to the system varied greatly, the quality of data was not reliable, and we lacked a source of high accuracy ground truth data. From these shortcomings, we construct a set of lessons and recommendations to consider when running longitudinal sensing studies. Additionally, we describe a second iteration of WearCOPD where our lessons resulted in improved results. The shortcomings of WearCOPD created for a valuable learning experience and can provide guidelines for designing future studies.

REFERENCES

[1] D. Kumar, S. Jeuris, J. E. Bardram, and N. Dragoni, “Mobile and wearable sensing frameworks for mhealth studies and applications: a systematic review,” *ACM Transactions on Computing for Healthcare*, vol. 2, no. 1, pp. 1–28, 2020.

[2] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, “A review of wearable sensors and systems with application in rehabilitation,” *Journal of neuroengineering and rehabilitation*, vol. 9, no. 1, pp. 1–17, 2012.

[3] C. R. Ahn, S. Lee, C. Sun, H. Jebelli, K. Yang, and B. Choi, "Wearable sensing technology applications in construction safety and health," *Journal of Construction Engineering and Management*, vol. 145, no. 11, p. 03119007, 2019.

[4] H. Canada, "Download covid alert: Canada's exposure notification app," Aug 2021. [Online]. Available: https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/covid-alert.html

[5] D. Liaqat, M. Abdalla, P. Abed-Esfahani, M. Gabel, T. Son, R. Wu, A. Gershon, F. Rudzicz, and E. de Lara, "WearBreathing: Real world respiratory rate monitoring using smartwatches," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, pp. 1–22, 06 2019.

[6] D. Liaqat, S. Liaqat, J. L. Chen, T. Sedaghat, M. Gabel, F. Rudzicz, and E. de Lara, "Coughwatch: Real-world cough detection using smartwatches," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8333–8337.

[7] T. Sedaghat, S. Liaqat, D. Liaqat, R. Wu, A. Gershon, T. Son, T. H. Falk, M. Gabel, A. Mariakakis, and E. de Lara, "Unobtrusive monitoring of COPD patients using speech collected from smartwatches in the wild," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 2022.

[8] A. Tiwari, S. Liaqat, D. Liaqat, M. Gabel, E. de Lara, and T. H. Falk, "Remote copd severity and exacerbation detection using heart rate and activity data measured from a wearable device," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 7450–7454.

[9] P. M. Calverley, "COPD: early detection and intervention," *Chest*, vol. 117, no. 5, pp. 365S–371S, 2000.

[10] R. Wu, D. Liaqat, E. de Lara, T. Son, F. Rudzicz, H. Alshaer, P. Abed-Esfahani, A. S. Gershon *et al.*, "Feasibility of using a smartwatch to intensively monitor patients with chronic obstructive pulmonary disease: prospective cohort study," *JMIR mHealth and uHealth*, vol. 6, no. 6, p. e10046, 2018.

[11] S. Aaron, G. Donaldson, G. Whitmore, J. Hurst, T. Ramsay, and J. Wedzicha, "Time course and pattern of COPD exacerbation onset," *Thorax*, vol. 67, pp. 238–43, 03 2012.

[12] D. Liaqat, R. Wu, A. Gershon, H. Alshaer, F. Rudzicz, and E. de Lara, "Challenges with real-world smartwatch based audio monitoring," in *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*, 2018, pp. 54–59.

[13] I. Ray, D. Liaqat, M. Gabel, and E. de Lara, "Skin tone, confidence, and data quality of heart rate sensing in wearos smartwatches," in *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 2021, pp. 213–219.

[14] D. R. Bassett Jr, A. V. Rowlands, and S. G. Trost, "Calibration and validation of wearable monitors," *Medicine and science in sports and exercise*, vol. 44, no. 1 Suppl 1, p. S32, 2012.

[15] E. E. Mohamed *et al.*, "Voice changes in patients with chronic obstructive pulmonary disease," *Egyptian Journal of Chest Diseases and Tuberculosis*, vol. 63, no. 3, pp. 561–567, 2014.

[16] "Zephyr performance systems." [Online]. Available: https://www.zephyranywhere.com/