medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Deep Learning Utilizing Suboptimal Spirometry Data to Improve Lung Function and Mortality Prediction in the UK Biobank

Davin Hill[1,2], Max Torop[1], Aria Masoomi[1], Peter J. Castaldi[2,3,4], Edwin K. Silverman[2,4,5], Sandeep Bodduluri[6], Surya P. Bhatt[6], Taedong Yun[7], Cory Y. McLean[7], Farhad Hormozdiari[7], Jennifer Dy[1*†], Michael H. Cho[2,4,5*†], Brian D. Hobbs[2,4,5†]

[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA.
[2]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA.
[3]Division of General Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA, USA.
[4]Harvard Medical School, Boston, MA, USA.
[5]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA.
[6]Division of Pulmonary, Allergy and Critical Care Medicine, University of Alabama at Birmingham, Birmingham, AL, USA.
[7]Google Research, Cambridge, MA, USA.

*Corresponding author(s). E-mail(s): jdy@ece.neu.edu; remhc@channing.harvard.edu;
Contributing authors: dhill@ece.neu.edu; torop.m@northeastern.edu; masoomi.a@northeastern.edu; repjc@channing.harvard.edu; ed.silverman@channing.harvard.edu; sbodduluri@uabmc.edu; sbhatt@uabmc.edu; tedyun@google.com; cym@google.com; fhormoz@google.com; rebdh@channing.harvard.edu;
†These authors contributed equally to this work.

Abstract

Background: Spirometry measures lung function by selecting the best of multiple efforts meeting pre-specified quality control (QC), and reporting two key metrics: forced expiratory volume in 1 second (FEV_1) and forced vital capacity (FVC). We hypothesize that discarded submaximal and QC-failing data meaningfully contribute to the prediction of airflow obstruction and all-cause mortality.

Methods: We evaluated volume-time spirometry data from the UK Biobank. We identified "best" spirometry efforts as those passing QC with the maximum FVC. "Discarded" efforts were either submaximal or failed QC. To create a combined representation of lung function we implemented a contrastive learning approach, **Spiro**gram-based **C**ontrastive **L**earning **F**ramework (Spiro-CLF), which utilized all recorded volume-time curves per participant and applied different transformations (e.g. flow-volume, flow-time). In a held-out 20% testing subset we applied the Spiro-CLF representation of a participant's overall lung function to 1) binary predictions of $FEV_1/FVC < 0.7$ and FEV_1 Percent Predicted (FEV_1PP) $< 80\%$, indicative of airflow obstruction, and 2) Cox regression for all-cause mortality.

Findings: We included 940,705 volume-time curves from 352,684 UK Biobank participants with 2-3 spirometry efforts per individual (66.7% with 3 efforts) and at least one QC-passing spirometry effort. Of all spirometry efforts, 24.1% failed QC and 37.5% were submaximal. Spiro-CLF prediction of $FEV_1/FVC < 0.7$ utilizing discarded spirometry efforts had an Area under the Receiver Operating Characteristics (AUROC) of 0.981 (0.863 for FEV_1PP prediction). Incorporating discarded spirometry efforts in addition to the best effort improved AUROC for FEV_1PP prediction by 0.011 (0.874 to 0.885) and reduced the Cox mortality log-likelihood by 0.021 (0.5% relative improvement).

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

of 0.654, which exceeded the c-indices from FEV_1 (0.590), FVC (0.559), or FEV_1/FVC (0.599) from each participant's single best effort.

Interpretation: A contrastive learning model using raw spirometry curves can accurately predict lung function using submaximal and QC-failing efforts. This model also has superior prediction of all-cause mortality compared to standard lung function measurements.

Funding:
MHC is supported by NIH R01HL137927, R01HL135142, HL147148, and HL089856.
BDH is supported by NIH K08HL136928, U01 HL089856, and an Alpha-1 Foundation Research Grant.
DH is supported by NIH 2T32HL007427-41
EKS is supported by NIH R01 HL152728, R01 HL147148, U01 HL089856, R01 HL133135, P01 HL132825, and P01 HL114501.
PJC is supported by NIH R01HL124233 and R01HL147326.
SPB is supported by NIH R01HL151421 and UH3HL155806.
TY, FH, and CYM are employees of Google LLC

1 Introduction

The measurement of lung function via spirometry is one of the most common diagnostic tests used in clinical practice. Spirometry is required for diagnosis of COPD, and is a key diagnostic test used in assessing asthma, interstitial lung disease, neuromuscular disease, preoperative pulmonary risk, and monitoring for drug toxicity. While the number of spirometry tests performed is not known, spirometry at least annually is recommended for asthma and COPD, diseases that have a combined prevalence of over nearly 500 million people. Lung function is described using summary measurements from raw volume-time spirometry curves. Spirometry is performed by an individual first taking a maximum inhalation and then performing a maximum forced expiratory effort into a spirometry device. Spirometry measurements are often noisy with repeated efforts typically resulting in variable summary measures of lung function including the forced expiratory volume in 1 second (FEV_1) and the forced vital capacity (FVC). Spirometry guidelines, as they can depend on effort, and in contrast to most other biomedical measurements, require repetition. In clinical practice, individuals are required to perform a minimum of two high-quality and reproducible (volume variability less than a specified threshold) spirometry efforts, and often perform three or more. The "best" FEV_1 and FVC are then selected as the highest measured value for each measure across all high-quality and reproducible efforts. Spirometry quality control (QC) guidelines include factors such as minimal back-extrapolated volume (i.e. assuring the forced exhalation effort is maximal from the very start of the exhalation maneuver), sufficient amount of time performing the forced exhalation, and a sufficient volume plateau at the end of the forced exhalation effort. Any spirometry efforts that are not reproducible or do not meet these QC guidelines are discarded, since FEV_1 and FVC are only extracted from the maximal efforts.

In addition to omitting submaximal and QC-rejected spirometry efforts, traditional spirometry measurements only summarize a few features of the spirometry curve, usually the FEV_1 (the volume of forcibly exhaled air in the first second of an effort) and the FVC (the total forcibly exhaled volume), which may not capture a comprehensive representation of lung function. Alternative methods have been investigated using different aspects of the spirometry curve (Vandevoorde et al., 2008; Simon et al., 2010) to see if alternatives to the FEV_1 and FVC are better associated with important clinical outcomes and disease progression. Machine learning and deep learning methods have been applied to raw spirometry curves for specific tasks, such as prediction of Chronic Obstructive Pulmonary Disease (COPD) (Bhattacharjee et al., 2022), COPD subtyping (Bodduluri et al., 2020), prediction of upper airway obstruction (Wang et al., 2022a), or acceptability criteria (Das et al., 2020; Wang et al., 2022b). In these applications, the models are trained exclusively on a single representative spirometry effort from each individual.

We hypothesize that the discarded submaximal and QC-rejected spirometry efforts contain useful information related to lung function and overall health outcomes, which can be captured using deep learning methods trained on the totality of the spirometry curves. While the information carried by discarded efforts may be degraded compared with the maximal effort, that information may still be relevant to understanding overall lung function. Further, the combined information from the discarded

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Characteristic	Training Set	Validation Set	Test Set
Participants, n	211,603	70,534	70,535
Spirometry Efforts, n	564,509	188,105	188,091
Participants with 3 Efforts, n (%)	141,279 (66.8)	47,037 (66.7)	47,021 (66.7)
Age, yr, mean (SD)	56.3 (8.1)	56.3 (8.1)	56.3 (8.1)
Sex, F, n (%)	118,064 (55.8)	39,528 (56.0)	39,454 (55.9)
Ethnicity, Self-Reported			
White, n (%)	201,074 (95.3)	66,968 (95.2)	66,963 (95.2)
Black or Black British, n (%)	2,752 (1.3)	939 (1.3)	938 (1.3)
Asian or Asian British, n (%)	3,564 (1.7)	1,266 (1.8)	1,276 (1.8)
Chinese, n (%)	606 (0.3)	209 (0.3)	207 (0.3)
Ever Smoke, n (%)	96,277 (45.5)	32,039 (45.4)	32,171 (45.6)
FEV1 / FVC, mean (SD)	0.76 (0.06)	0.76 (0.06)	0.76 (0.06)
Death Events, n (%)	7,376 (3.5)	2,536 (3.6)	2,478 (3.5)
Time to Event, yr, mean (SD)	11.2 (1.5)	11.2 (1.5)	11.2 (1.5)

Table 1: Characteristics of participants in the study.

efforts may be sufficient to reconstruct the information from the maximal effort, which would reduce the number of efforts required during a clinical visit. Additionally, the discarded efforts also contain information related to the variance and reproducibility of the reported spirometry values. The magnitude of variance and the existence of failed efforts may be correlated to underlying lung function and similarly, to pulmonary disease. Thus, information mined from discarded spirometry efforts may provide a more comprehensive metric of an individual's lung function.

Therefore, the goal of this project was to determine whether the information from these discarded efforts would provide comparable or even superior information on lung function and mortality. To do this, we first investigated the relationship between information captured from discarded spirometry efforts and the information classically reported from the maximal or "best" spirometry effort. We used deep learning methods on the raw spirograms from the UK Biobank to extract a vector representation of lung function for each individual.

In order to assess the quality of the deep-learning derived vector representation of lung function, we applied it in two different prediction applications: 1) incidence and severity of airflow obstruction and 2) mortality prediction. In each of these prediction tasks, the vector representation of lung function is used as a replacement for alternative spirometry metrics including the commonly used FEV_1 and FVC.

Our results showed that including the discarded spirometry efforts in the vector representation of lung function led to improved prediction of lung function impairment and mortality prediction compared to both utilizing only maximal efforts and classic summary measures of lung function from spirometry.

2 Methods

2.1 Data Quality Control and Preprocessing

We extracted raw spirometry values from the UK Biobank field 3066. We used only spirometry data from the initial visit for each participant, thus excluding multiple visits from the same participant. Raw spirometry data were recorded as total exhaled lung volumes (mL) at 10 ms intervals over a median of 7.9 seconds and a range of 0.3 to 30.0 seconds (App. D Fig. 11). Each participant had between one and three spirometry efforts.

The following quality control filtering and preprocessing steps were applied to the spirometry samples. The spirometry QC was performed in line with the American Thoracic Society (ATS) and European Respiratory Society (ERS) guidelines using the procedure previously described by Shrine et al. (2019). QC steps are detailed in Table 2. For model validation purposes, we required all participants to have at least one QC-passing and reproducible effort. Spirometry efforts were flagged as passing or failing each step of the QC process (Fig. 1). After each step, participants with less than one passing effort were removed from the dataset. After removing the participants with less than one QC-passing effort or fail the reproducibility criteria, the entirety of the efforts from remaining participants are included in the dataset. The distribution of remaining efforts is shown in Fig. 2.

After the QC process, the raw spirometry data was downsampled from 10ms to 50ms intervals and the exhalation duration was limited to 15s to improve model efficiency. Efforts were truncated after reaching FVC; we replaced the spirometry volume values after achieving FVC to be the derived FVC value for the effort.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

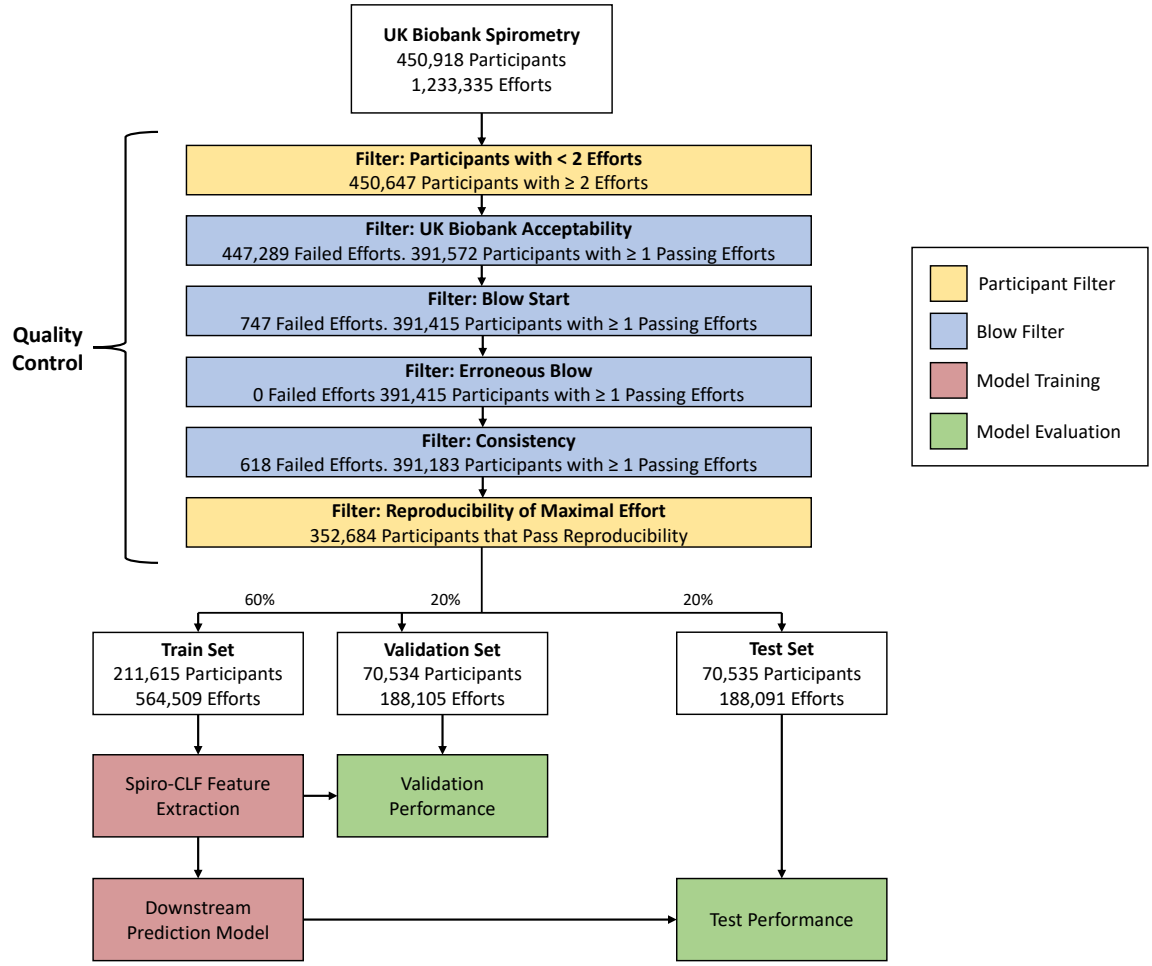


Fig. 1: Overview of the Quality Control (QC), Spiro-CLF training, and downstream prediction process. The QC process contains participant filters, which contains participant-level criteria, and blow filters, which contain effort-level criteria. In each blow filter, spirometry efforts are tested against the specified criteria and labeled as either passing or failing the given filter. Participants with less than one QC-passing efforts or failing to produce a reproducible effort are removed from the dataset. QC criteria details are provided in Table 2.

QC Filter	Description
Participants with <2 Efforts	Removes participants with less than two recorded efforts.
UK Biobank Acceptability	Fails efforts that are flagged in UK Biobank Field 3061 as START_OF_TEST (0x01), TIME_TO_PEF (0x02), END_OF_TEST (0x04), COUGHING (0x08), USER_REJECTED (0x40).
Blow Start	Fails efforts with excessive back extrapolated volume. Back extrapolated volume must be less than 5% of FVC or 150mL of total volume, whichever is greater.
Erroneous Blow	Fails efforts with invalid values for FEF25, FEF75, FEF25-75, PEF volume, back extrapolated volume, or FET volume.
Consistency	Fails efforts with inconsistency between derived FEV1, FVC values and UK Biobank recorded values. Derived FEV1 (FVC) must be within 5% of UK Biobank field 3063 (3062)
Reproducibility	Removes participants unable to reproduce the maximal effort. Participants must have at least one effort with FVC less than the maximal FVC and greater than the maximal FVC less 0.25L.

Table 2: QC filter description for preprocessing participants and spirometry efforts from the UK Biobank.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

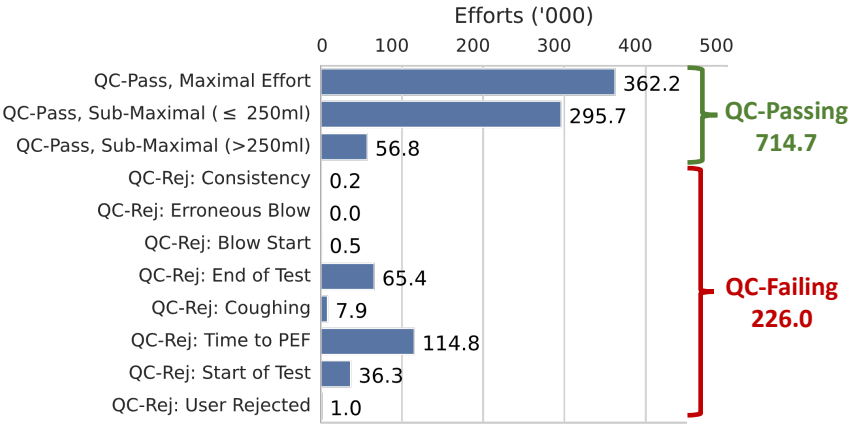


Fig. 2: Number of efforts in the dataset after preprocessing, listed by effort type. *QC-Pass, Submaximal* $\leq$ *250mL* denotes efforts that pass QC and have FVC within 250mL of the maximal FVC.

2.2 Representation Learning with Spiro-CLF

The aim of our study was to uncover the complex, non-linear relationships between discarded spirometry data and the derived spirometry metrics, i.e. FEV_1 and FVC, for UK Biobank participants. The raw spirometry efforts are often "noisy", in that the observed efforts can be characterized as noisy signals from some underlying lung function. We therefore introduced a contrastive learning approach, **Spiro**gram-based **C**ontrastive **L**earning **F**ramework (Spiro-CLF) to identify a feature representation that is invariant to noise in each individual effort, thereby capturing the underlying factors that contribute to lung function.

Contrastive learning is a self-supervised approach that has recently shown to be effective in improving image and language processing prediction tasks (Jaiswal et al., 2020; Shurrab and Duwairi, 2022). It typically involves two steps: 1) an initial prediction task, often called the *pretext* task, that learns a feature representation of the data, and 2) the *downstream* task, which uses the learned representation on the ultimate prediction goal. The pretext task is self-supervised in that the model is discriminative, i.e. it is trained to predict a given label, however the pretext labels are derived entirely from the dataset itself.

Our work is most closely related to Chen et al. (2020), which was introduced in the context of image classification, and uses different correlated transformations of a given image as input to a deep learning model. Spiro-CLF defines each data sample as an individual participant, and each spirometry effort as different perspectives of that participant's lung function. We additionally applied transformations (e.g. flow-volume, flow-time) to the spirometry efforts to further improve model performance; we define these transformations in section 2.4.

An overview of the Spiro-CLF representation learning process is shown in Figure 3. During training, Spiro-CLF sampled randomized batches of spirometry efforts. Each batch contained pairs of spirometry efforts produced from the same individual. Each effort within the batch had one positive example belonging to the same individual, and $B-2$ negative examples, where B is the batch size. The model predicted whether any two given efforts within the batch were produced by the same individual or different individuals, represented by a predicted similarity score for each pair of efforts. The contrastive loss encouraged the model to predict high similarity for efforts produced from the same individual, and low similarity for efforts from disparate individuals.

We parameterized the Spiro-CLF model using a Convolution Neural Network (CNN) architecture, which captures the temporal relationship between each time step in each sample. CNN models have been previously shown to be highly effective on time-series prediction tasks (Hoseinzade and Haratizadeh, 2019; Wang et al., 2019). The convolution layers within CNNs encode the assumption that the relationship between values at adjacent time steps are more relevant than values between non-adjacent time steps. The model architecture and training process is defined in more detail in section 2.3.

Once the model is trained, we used the output of an intermediate layer as the learned representation of a participant's overall lung function. We incorporated a feature averaging step (Foster et al., 2021) when generating Spiro-CLF features to further encourage invariance of the learned representations. During this step, we sampled the entirety of transformed efforts from each individual and averaged the respective features. The final representation can then be used in downstream prediction tasks; in particular we

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

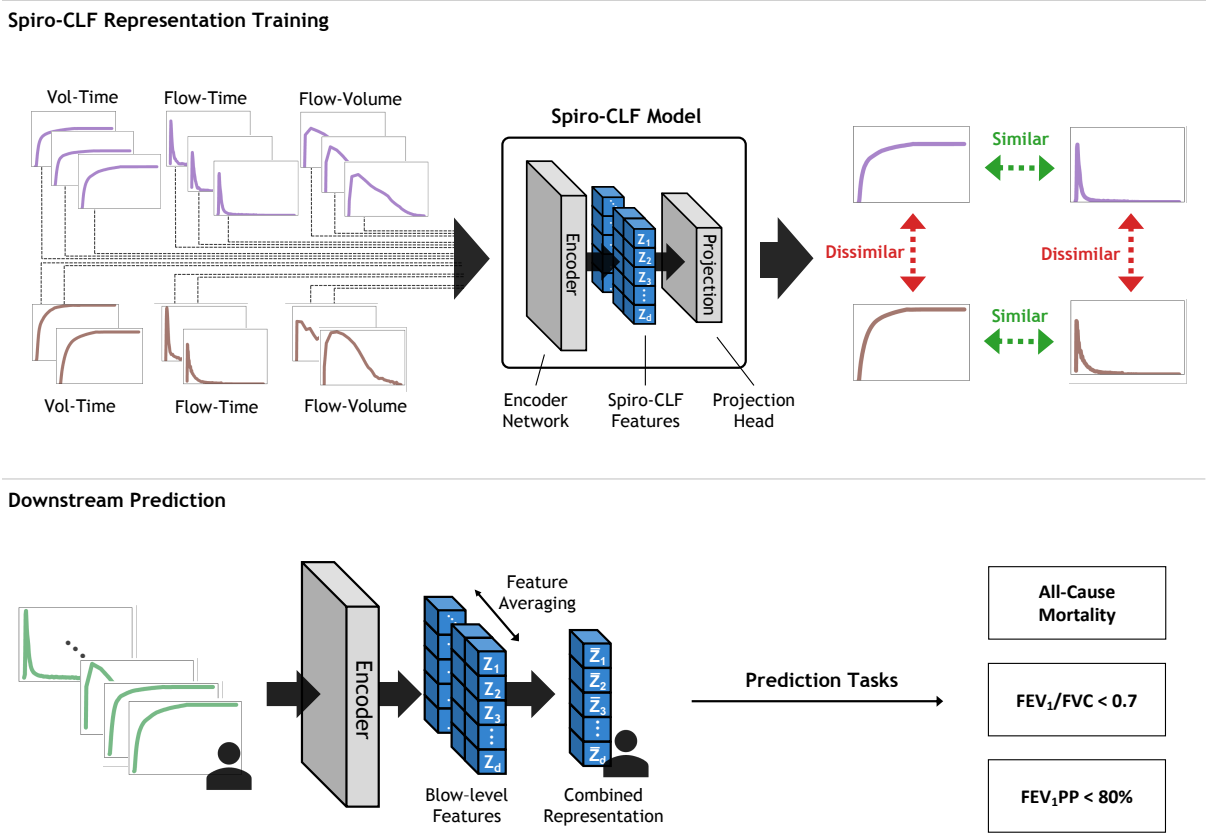


Fig. 3: **Top:** Schematic of the Spiro-CLF training process. We randomly applied flow-time, flow-volume, and identity (volume-time) transformations to each spirometry effort within the training batch. The Spiro-CLF model is trained with contrastive loss to predict high pairwise similarity between efforts from the same individual and low pairwise similarity between efforts from different individuals. **Bottom:** Once the Spiro-CLF model is trained, we applied the encoder network to the entirety of an individual's efforts, including transformation, to generate a combined feature representation. This representation can then be used in a variety of downstream predictive tasks.

evaluated 1) binary prediction for $FEV_1/FVC < 0.7$ and FEV_1 Percent Predicted (FEV_1PP) (Sec. 2.6), and 2) Cox regression for all-cause mortality (Sec. 3.3).

We used a bootstrap resampling procedure to estimate the standard deviation of the reported prediction performance. Both the Spiro-CLF model and downstream models were retrained on 24 bootstrapped training distributions and evaluated on the out-of-bag samples. The 95% confidence intervals for the performance scores are reported for each experiment.

2.3 Spiro-CLF Model

We parametrized the Spiro-CLF model with a deep neural network. Model architecture and hyperparameters were selected using a randomized grid search procedure, optimized with respect to minimum loss on the validation set. The model consisted of six convolution blocks; each block contained a 1-dimensional convolution layer with weight normalization, ReLU activation function, and skip-connection. Each convolution had a fixed kernel size of 40 with padding to avoid downsampling. After the convolution blocks, two fully-connected layers were applied to downsample to the 100-dimensional feature space. We trained an additional projection head parametrized by a multi-layer perceptron (MLP) with one hidden layer, ReLU activation, and batch normalization. The projection head maps the feature space to a 128-dimensional output to calculate contrastive loss. The detailed model architecture is shown in Appendix D Figure 10.

We trained the model for 400 epochs to minimize contrastive loss of the projection head output. During training, contrastive loss was calculated using batches of 512 data samples. For each batch, 256 unique participants were randomly selected, followed by the random selection of two transformed spirometry efforts from each selected participant. The available transformations are discussed in section

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2.4. The model was applied to the samples, and a pairwise similarity score and contrastive loss was calculated between each of the 512, 100-dimensional, vector outputs of the projection head. The model was trained using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 10^{-2} and a scale factor of 0.1 on validation loss plateau.

After training, the projection head was discarded, and the model was applied to the spirometry samples in the test set to calculate the Spiro-CLF features.

2.4 Sampling of Spirometry Efforts and Transformations

During Spiro-CLF model training, we randomly selected two efforts from each participant in a given batch of participants. The two efforts were sampled uniformly from the participant's collective spirometry efforts. If a participant had exactly two efforts, then both efforts were used in the training batch. We then applied a random data transformation to each effort.

Data transformations are non-learned mappings applied to the data samples to increase information on the underlying lung function. Each applied transformation returns different perspectives of a single participant's lung function. During the data transformation stage, we applied either a flow-time transformation, flow-volume transformation, or identity (volume-time) transformation, selected with uniform probability. These transformations are detailed below. Therefore, in each training batch, each participant had two separate spirometry efforts, which were represented by either volume-time, flow-time, or flow-volume curves.

Flow-Time. We applied a flow-time transformation to the original flow-time data by calculating the change in expiratory volume over each time interval. We similarly downsampled the flow-time curves to 50ms intervals.

Flow-Volume. We applied a flow-volume transformation to the original flow-time data. Volumes were sampled in 10mL fixed intervals. Due to the 10ms sampling interval in the original flow-time data, the exact flow at each 10mL volume interval was not always available; we used linear interpolation to estimate the flow at those missing intervals.

2.5 Time-to-Event Prediction

We used UK Biobank field 3060 to identify the date the recorded spirometry was performed and field 40000 for event time for all-cause mortality. The dataset contained 12,390 mortality events (3.5%) with a median time-to-event of 11 years.

We trained a Cox Proportional Hazards Regression model on the Spiro-CLF features extracted from the training set and evaluated the concordance index (c-index) on the Spiro-CLF test features. C-index is a goodness-of-fit measurement that evaluates the concordance between the labels and predicted outcomes. A higher c-index indicates a better model fit. We omitted any covariates in order to evaluate the predictive power of the Spiro-CLF features. Interestingly, incorporating feature averaging on the generated Spiro-CLF features reduced predictive performance for the Cox model. We therefore omitted this step in this predictive task (see Sec.4 for additional discussion).

We compared the results from using the Spiro-CLF representation with previously proposed spirometry metrics, such as FEV_1, FVC, and the forced mid-expiratory flow between 25% and 75% of the FVC (FEF25-75). These alternative metrics were calculated on the maximal QC-passing effort for each individual. The set of metrics from the training dataset were then used as predictors in the Cox Regression model.

2.6 Lung Function Impairment Prediction

We defined lung function impairment using two metrics: FEV_1/FVC and FEV_1 percent of predicted (FEV_1pp). We calculated FEV_1pp using GLI-2012 reference values (Quanjer et al., 2012) calculated using the Spiref[1] python package. We took sex, height, age at recruitment, and self-reported ethnicity values from fields 31, 12144, 21022, 21000. Self-reported ethnicities were mapped to Caucasian or Other to be consistent with the GLI-2012 classifications.

To predict lung function impairment we created binary labels for each individual based on a threshold for each metric. Specifically we defined $FEV_1/FVC < 0.7$ and $FEV_1pp < 80\%$, the Global Initiative for Chronic Obstructive Lung Disease (GOLD) spirometric criteria for moderate-to-severe airflow limitation in COPD (Vogelmeier et al., 2017) as indicative of lung function impairment. Using these definitions we trained logistic regression models for each task using the Spiro-CLF features as input. The regression

[1] https://github.com/kennethverstraete/spiref

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

models were trained with no regularization and no additional covariates. Model performance was assessed using the Area under the Receiving Operating Characteristic (AUROC) curve on the test set.

2.7 Feature Space Analysis

We use a dimension-reduction method, Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), to project the Spiro-CLF representations down to two dimensions for visualization purposes. We randomly sampled 2000 participants and calculated the Spiro-CLF feature embeddings for the entirety of their respective efforts, including flow-time and flow-volume transformations. We then applied UMAP to project the 100-dimensional embeddings to two dimensions.

We evaluated the level of invariance induced by the Spiro-CLF model by investigating the feature space. The goal of the Spiro-CLF model training is to create a representation of lung function that is invariant to the inherent noise in repeated exhalation effort. An invariant representation would have minimal variance even when the individual efforts from a participant are not identical. We evaluated the empirical performance by calculating the average Euclidean distance in the feature space for each individual's efforts. We first calculated the average Spiro-CLF representation for each individual, then averaged the Euclidean distance between each individual representation and the average representation.

2.8 Feature Importance Approximation

In order to understand how the Spiro-CLF model and subsequent downstream prediction tasks operate on the spirometry inputs, we applied feature importance methods to understand the relative importance of each individual timestamp of the raw spirometry input. We first combined the Spiro-CLF model and linear Lung Function impairment model (Section 2.6) into a single prediction pipeline. We then applied the Asymmetric Shapley Value (ASV) (Frye et al., 2020) method to calculate the feature importance values.

ASV has a number of properties that are suited for this application:

- *Post-hoc.* ASV can be applied to any prediction model after the model has been trained with no modification to the training process or model architecture.
- *Local.* ASV can be calculated for individual data samples. Traditionally, most feature importance methods are *global*, where they give a single explanation for the entire dataset and model. Using a local method allows the investigation of different subgroups within the data distribution by averaging over the relevant samples. For example, we can calculate ASV separately for flow-time, flow-volume, and volume-time transformations.
- *Asymmetric.* Due to the time-dependency in the spirometry efforts, the individual features within each sample contains asymmetric interactions. For example, in our dataset each feature represents a 50ms timestamp. Within each pair of timestamps, one timestamp occurs before the other. Therefore the relative importance of the latter timestamp is conditional on the former timestamp. We can enforce this relationship using ASV using a directed graph which we define prior to use.

We approximated ASV using a Monte Carlo sampling algorithm (Strumbelj and Kononenko, 2014), using 2000 permutations per data sample. We applied this approximation on 50 randomly sampled spirometry efforts for each of flow-time, flow-volume, and volume-time transformations. We then averaged over each transformation to obtain the subgroup-level explanation.

3 Results

All experiments were performed on a computing cluster using AMD EPYC 7302 16-Core processors and NVIDIA A100 GPUs. All source code is available at https://github.com/davinhill/Spiro-CLF.

3.1 Participant Characteristics

We included 352,684 participants in this study, with between 2-3 spirometry efforts per participant (66.7% with 3 efforts) resulting in 940,705 total efforts represented as volume-time curves. Participants were recruited across 23 assessment centers in the United Kingdom for the UK Biobank database. Additional details regarding the UK Biobank study were previously published in Sudlow et al. (2015). We utilized the spirometry records from each participant's baseline assessment, which was collected between 2006 - 2010. A summary of participant characteristics is listed in Table 1.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

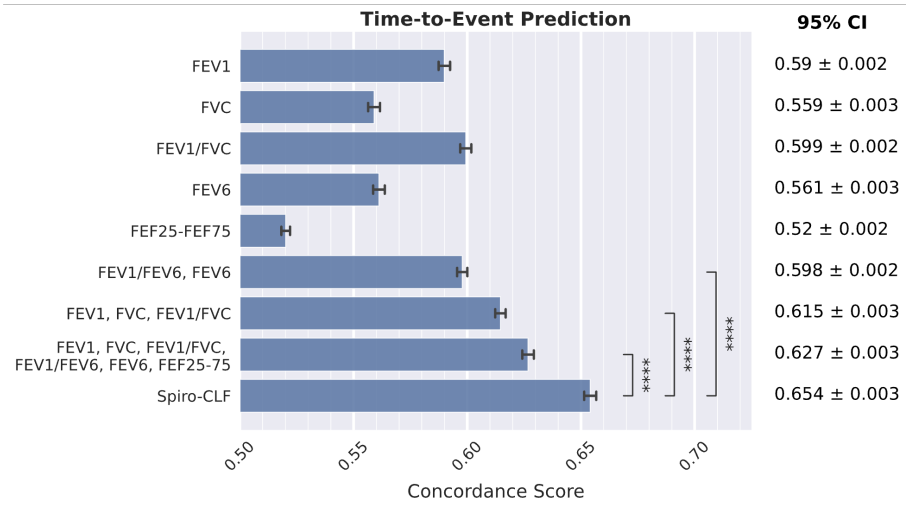


Fig. 4: Results of a time-to-event mortality prediction task, measured using c-index (higher is better). We trained the Cox regression model using different lung function representations to compare their predictive power with respect to mortality prediction. The error bars represent 95% bootstrap confidence intervals (24 iterations). **** indicates significance at the Bonferroni-corrected 10^{-4} significance level.

3.2 Spiro-CLF Training

We randomly partitioned the participants in the dataset into training (80%), validation (20%), and test (20%) sets. Data transformations (Sec. 2.4) were applied during training and test stages, which effectively tripled the number of data samples from 940,669 efforts to 2,822,007. Model training continued until convergence of the validation loss.

We applied the unsupervised Spiro-CLF representation as a proxy for lung function on a variety of prediction models in place of traditional spirometry metrics. We evaluated the quality of the learned representation on two tasks. First, we evaluated the ability of the Spiro-CLF representation, derived from any given effort, to recover the maximal FEV_1/FVC. Second, we evaluated the usefulness of the Spiro-CLF representation in predicting mortality as compared with other spirometry metrics.

3.3 Spiro-CLF Prediction of All-Cause Mortality

We used a mortality prediction task to validate that the Spiro-CLF representation contains additional relevant information for disease progression and health outcomes. We first trained a Cox time-to-event model on the Spiro-CLF representation and evaluated the model fit using the c-index. We then trained Cox regression models on alternative spirometry metrics to establish a baseline comparison. No additional covariates were included in the model training in order to compare the predictive power of each metric. Results are shown in Figure 4. The model trained on Spiro-CLF representations achieved a c-index of 0.654, which surpassed the performance of the second-highest performing single metric, FEV_1/FVC, with c-index 0.599 ($P \leq 1.4 \times 10^{-37}$). We additionally trained Cox regression models on combinations of alternative metrics. Using the combination FEV_1, FVC, and FEV_1/FVC increased c-index to 0.615. Including all competing metrics in the Cox model further improved c-index to 0.627, which was still exceeded in performance by the Spiro-CLF features ($P \leq 8.3 \times 10^{-23}$).

3.4 Spiro-CLF Prediction of Lung Function Impairment

We compared the predictive power of the maximal, QC-passing effort while restricting the model to exclusively use submaximal and QC-rejected efforts. This comparison was evaluated on two lung function impairment prediction tasks: A) $FEV_1/FVC < 0.7$ and B) $FEV_1PP < 80\%$. Results are shown in Figure 5.

We observed that the Spiro-CLF model is able to achieve similar performance when using submaximal and QC-rejected efforts to the baseline on both the FEV_1/FVC and FEV_1PP prediction tasks.

We observed that the predictive performance of the Spiro-CLF model on QC-rejected and submaximal efforts was significantly recovered as compared to the baseline of using the maximal effort on both the FEV_1/FVC and FEV_1PP prediction tasks. In the FEV_1/FVC task, the Spiro-CLF representations for submaximal and QC-rejected efforts achieved 0.981 AUROC. Exclusively using submaximal efforts

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

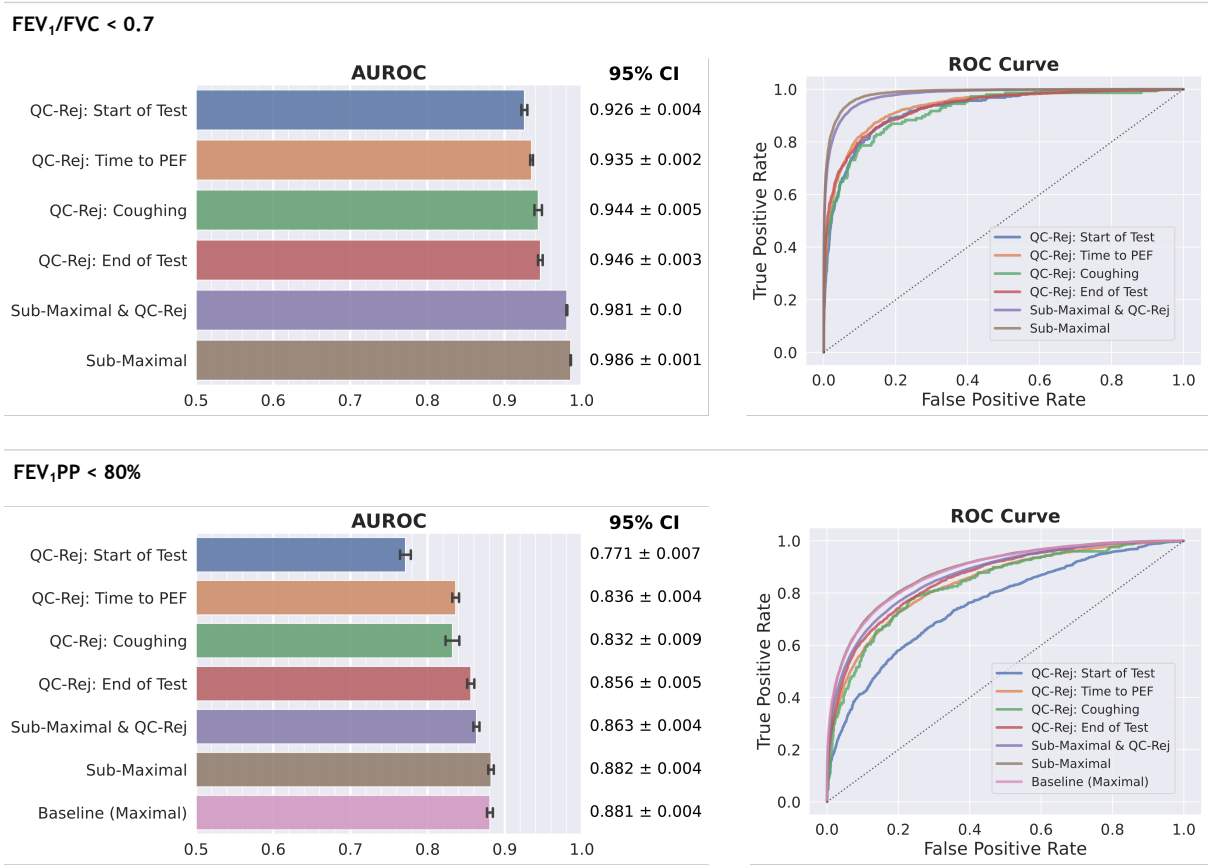


Fig. 5: Results of the lung function impairment prediction tasks: $FEV_1/FVC < 0.7$ (Top), and $FEV_1PP < 80\%$ (Bottom). The error bars represent 95% bootstrap confidence intervals (24 iterations).

yielded AUROC of 0.986. Similarly, in the FEV_1PP prediction task we compare the performance of the submaximal and QC-rejected features to a baseline performance of using only features from maximal efforts. Using the Spiro-CLF features acheived 0.863 (submaximal and QC-rejected efforts) and 0.882 (submaximal efforts) AUROC compared to the baseline of 0.881 AUROC.

We additionally evaluated the performance when using only efforts that were QC-rejected for a variety of reasons. Restricting the Spiro-CLF model to using only efforts rejected due to excessive time to peak expiratory flow (PEF) recovered an AUROC of 0.935 for task A and 0.863 for task B.

3.5 Understanding the Spiro-CLF Feature Space

We visualized the feature embeddings generated by the Spiro-CLF model using dimension reduction methods. Given that the generated features are high-dimensional (100 dimensional vectors), we applied the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) algorithm on the output of the Spiro-CLF model. The resulting feature space visualization is shown in Figure 6.

The Spiro-CLF model was trained using a contrastive loss, which pushes spirometry efforts from the same individual closer together in the feature space. In the low dimensional feature space, Spiro-CLF spirometry representations resulting from different efforts or different data transformation appeared to project to a small locality. This observation indicates that the Spiro-CLF model generated similar feature embeddings for different spirometry efforts (or different transformations) performed by the same individual. We evaluated this observation by measuring the average Euclidean distance in the feature space generated using our trained model. Figure 6 is a visualization of the clusters of each individual efforts in the feature space. Each blue circle in the figure surrounds the features for a single individual. Table 3 calculates the distance between projections given different pairwise combinations of transformations. We observed that increasing the number of transformations increased the distance between clusters of feature projections. This indicates that adding transformations is a more difficult task for the Spiro-CLF model. Between the three investigated transformations, we observed that the flow-volume transformation

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

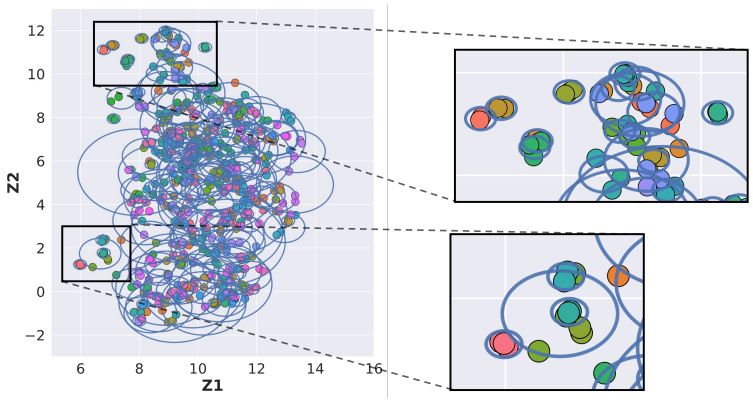


Fig. 6: Analysis of the Spiro-CLF feature space which shows that efforts from the same individual are mapped to be close together in the feature space. We projected the Spiro-CLF representations of a random subset of 500 individuals to two dimensions using the UMAP algorithm. Each individual's spirometry efforts were coded to be a different color and circled in blue.

Transformation	Volume-Time	Flow-Time	Flow-Volume
Volume-Time	0.128	0.282	0.215
Flow-Time	0.282	0.152	0.291
Flow-Volume	0.215	0.291	0.176

Table 3: Effects of different transformations on the corresponding feature space. The table shows average distance between the feature representations of each individual's efforts.

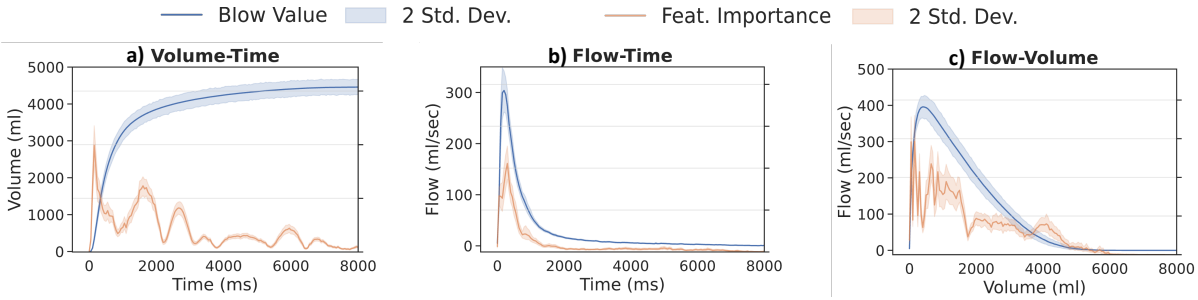


Fig. 7: Visualization of the relative importance of different sections of the spirometry curve with respect to the prediction of lung function impairment.

yielded the largest average distance, which indicates that the model had the most difficulty in minimizing contrastive loss under this transformation.

3.6 Assigning Relative Importance to Spirometry Curves

We used Asymmetric Shapley Values (ASV) to assign relative importance to the spirometry effort features for maximal efforts. We selected 50 randomized samples for each spirometry transformation (Fig. 7) and averaged ASV over each transformation. Importance measures are relative to the prediction of COPD in the downstream task. We observed that feature importance peaked within the first 1000 ms of the spirometry efforts, which is associated with peak expiratory flow (PEF) and the FEV_1. This section of the importance curve corresponds to low variance in importance estimate over the averaged samples. In the flow-volume figure (Fig. 7(c)) we observed a decrease in importance between 2000mL and 4000mL segment with another local peak at 4100mL. This section corresponds to higher variance. This section generally corresponds to FVC, with higher variance due to sampled efforts achieving FVC at different timestamps.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

4 Discussion

In this work, we trained a deep learning model on the raw spirometry curves in the UK Biobank study. We demonstrated that this Spiro-CLF model, trained on the entirety of each participant's efforts, could not only predict lung function impairment and mortality when only using suboptimal or failed quality control data, but also improve prediction of mortality when added to maximal efforts.

Spiro-CLF features are a representation of lung function that are trained to be invariant to the associated noise within the spirometry test process. Therefore the Spiro-CLF features can be calculated for any effort from a given individual, including QC-rejected and submaximal efforts. Our model is able to maintain high prediction of lung function impairment using QC-rejected and submaximal efforts, and improve prediction of mortality when added to maximal efforts. In addition, the Spiro-CLF representation is generated from an unsupervised model which can be easily transferred across datasets. The resulting vector representation can be directly used as a replacement for traditional spirometry metrics in any prediction tasks. These data suggest that Spiro-CLF or other machine learning models could augment current clinical spirometry testing in providing more accurate prediction of airflow limitation, lung function, and other outcomes not only in optimal settings, but also when a sufficient number of efforts is unable to be obtained.

The Spiro-CLF model maps complex high-dimensional data to a rich lower-dimensional feature space. Existing literature in the machine learning field has shown that the learned Spiro-CLF feature space may encode other latent representations of the underlying data (Chen et al., 2021) and has been used for clustering (Li et al., 2021; Caron et al., 2020) as well as combined with generative models (Kim et al., 2021). Applying such methods to spirometry data could be a promising future direction for disease subtyping, disease progression, or combined with genome-wide association studies.

There are a number of possible causes for the improved predictive performance gained from using the Spiro-CLF representation. First, the Spiro-CLF encoding allows access to information from QC-rejected efforts. For each individual with rejected efforts, the rejected efforts may indicate underlying lung function impairment. For instance, elderly patients or those with severe lung disease may have difficulty performing a traditional FVC measurement appropriately and thus their spirometry efforts may more often be rejected (Vandevoorde et al., 2008; Bellia et al., 2008; Jing et al., 2009; Allen et al., 2010). For efforts that are not rejected, the additional submaximal efforts provide information related to the relative variance for the maximal effort. Encoding these sources of information can provide additional information not contained within the maximal effort.

In addition, the neural network parametrization of the Spiro-CLF model encodes nonlinear effects from the raw spirometry efforts. Previous works have shown that using neural networks, especially in conjunction with convolution layers, can improve representations of lung function with respect to prediction of COPD (Bhattacharjee et al., 2022), COPD subtypes (Bodduluri et al., 2020), and gene association (Cosentino et al., 2023; Yun et al., 2023).

Our models were trained on a large set of systematically obtained flow-volume curves available through UK Biobank. A significant factor in the predictive performance improvement likely resulted from the increase in training samples from utilizing the entirety of an individual's efforts and additionally applying data transformations to obtain flow-time, flow-volume, and volume-time views of each effort. At a basic level, the combined efforts and transformations increase the number of training samples per individual, up to 9x more when compared with using the maximal effort and a single blow representation. Deep learning models are known to require a significant number of data samples during training (Vapnik, 1999) and increasing the number of samples and the variability in data views offered by volume-time transformations to flow-volume and flow-time enables the use of more complex models and reduces the likelihood of over fitting the training data.

We incorporated a feature averaging step for the Spiro-CLF features (Sec. 2.2), which improved performance for the $FEV_1/FVC < 0.7$ and the Time-to-Event prediction tasks. However, using the averaged features for the $FEV_1PP < 80\%$ reduced performance. The FEV_1PPprediction label contains demographic information which is not available to the Spiro-CLF model, resulting in a more difficult prediction task. Since the Spiro-CLF model excludes covariate information, the resulting feature space does not disentangle these additional factors, which may result in worse performance when using feature averaging for tasks that depend on these covariates. For these tasks, the user should therefore omit the feature averaging step or include additional covariate information in the downstream prediction step.

Despite the increase of training samples used in the Spiro-CLF model, we acknowledge that overfitting is still possible, and we do not know if our model is transferable to clinical spirometry or other cohorts. The spirometry efforts in UK Biobank were limited to 3 efforts per individual and the QC definitions

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

may be different from those used in a clinical setting. In order to compare our models to 'ground truth', we excluded participants that were unable to perform at least one QC-passing spirometry effort. The UK Biobank was predominantly of self-identified white race, European genetic ancestry, and - as a volunteer cohort - has a "healthy volunteer" selection bias (Fry et al., 2017). We did not evaluate the performance of our model for specific subsets of participants (age, race, chronic illness). Further validation in other cohorts and in clinical settings is needed.

Thus, for these reasons, additional training may be needed when applying the Spiro-CLF model to clinical spirometry. During training, additional validation may also be needed to ensure that model parameter and architecture choices are optimal with respect to the new dataset.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

A Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 20915.

B Funding

MHC is supported by NIH R01HL137927, R01HL135142, HL147148, and HL089856.
BDH is supported by NIH K08HL136928, U01 HL089856, and an Alpha-1 Foundation Research Grant.
DH is supported by NIH 2T32HL007427-41
EKS is supported by NIH R01 HL152728, R01 HL147148, U01 HL089856, R01 HL133135, P01 HL132825, and P01 HL114501.
PJC is supported by NIH R01HL124233 and R01HL147326.
SPB is supported by NIH R01HL151421 and UH3HL155806.
TY, FH, and CYM are employees of Google LLC

C Disclosures

BDH receives grant support from Bayer.
MHC has received grant support from GlaxoSmithKline and Bayer, consulting fees from Genentech and AstraZeneca, and speaking fees from Illumina.
EKS has received grant support from GlaxoSmithKline and Bayer.
PJC has received grant support from Bayer.
SPB has received consulting fees from Sanofi/Regeneron and Boehringer Ingelheim, and CME fees from IntegrityCE. His institute has received funds from Sanofi and Nuvaira for the conduct of clinical trials.
TY, FH, and CYM are employees of Google LLC and own Alphabet stock.

D Supplemental Figures

D.1 Comparison of FEV_1/FVC and FEV_1PP prediction using the second-highest effort

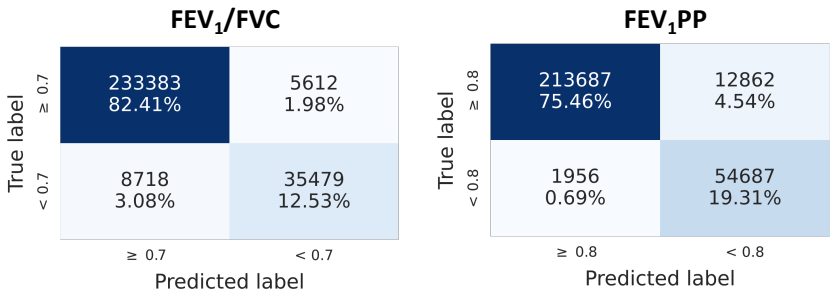


Fig. 8: Confusion matrix when using the FEV_1, FVC, and FEV_1PP of the second-highest, QC-passing effort to predict $FEV_1/FVC < 0.7$ and $FEV_1PP < 80\%$.

In Section 3.4 we tested the effectiveness of using the Spiro-CLF representation for the binary prediction of A) $FEV_1/FVC < 0.7$ and B) $FEV_1PP < 80\%$. However, the majority of QC-passing, submaximal efforts are within 250mL FVC of the maximal effort (Fig. 2). In Figure 8 we evaluated the effectiveness of using the second-highest effort (measured by FVC), in place of the maximal effort for the two binary prediction tasks. Specifically, we defined the ground-truth FEV_1/FVC and FEV_1PP values on the maximal effort and compared the agreement with the same values calculated on each individual's second-highest, QC-passing, effort. The results are shown in a confusion matrix. We observe that using the second-highest effort leads to misclassification rates of 5.06% (0.832 F1) for task A and 5.23% (0.881 F1) for task B.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

D.2 Comparison of $FEV_1/FVC < 0.7$ prediction using QC-rejected and submaximal efforts

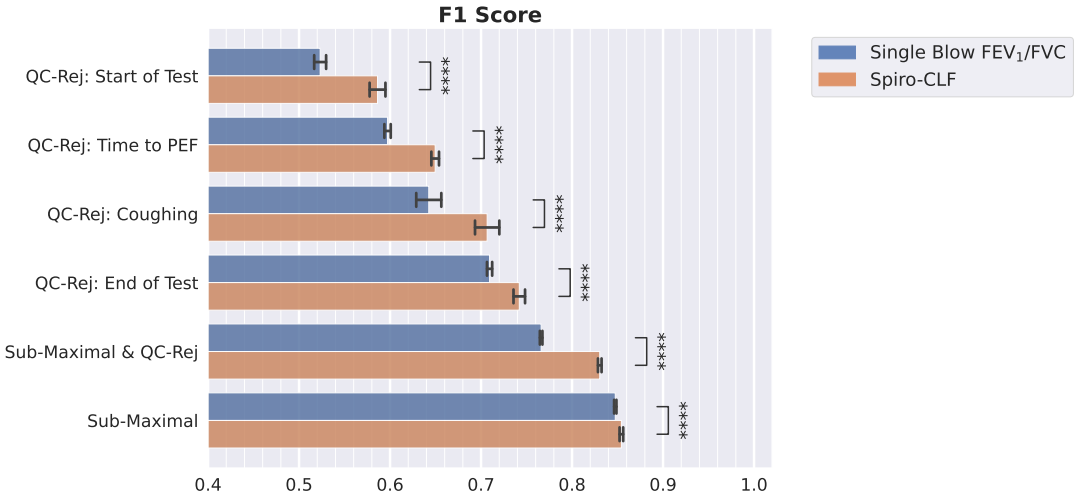


Fig. 9: F1 score for $FEV_1/FVC < 0.7$ prediction task. We compared the performance of the Spiro-CLF representation and the raw single-effort FEV_1/FVC of each subset of efforts. The error bars represent 95% bootstrap confidence intervals (24 iterations). **** indicates significance at the Bonferroni-corrected 10^{-4} significance level.

In this experiment, we compared the results of the Spiro-CLF prediction of $FEV_1/FVC< 0.7$ with the FEV_1/FVCvalues of each subset of efforts. Specifically, we calculated the $FEV_1/FVC$for individual efforts in each QC-rejected and submaximal effort subset and used these values as predictions for the $FEV_1/FVC< 0.7$ prediction task. We evaluate results using F1 score in place of AUROC due to the non-probabilistic property of the predictions. We observe in Figure 9 that using the Spiro-CLF representation improves prediction performance of all spirometry effort subsets ($P \leq 2.8 \times 10^{-6}$).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

```
CNN(
  (conv_encode): Sequential(
    (0): ConvBlock(
      (conv): Conv1d(1, 200, kernel_size=(40,), stride=(1,), padding=(39,))
      (dropout): Dropout(p=0, inplace=False)
      (downsample): Linear(in_features=1, out_features=200, bias=True)
    )
    (1): ConvBlock(
      (conv): Conv1d(200, 200, kernel_size=(40,), stride=(1,), padding=(39,))
      (dropout): Dropout(p=0, inplace=False)
      (downsample): Linear(in_features=200, out_features=200, bias=True)
    )
    (2): ConvBlock(
      (conv): Conv1d(200, 200, kernel_size=(40,), stride=(1,), padding=(39,))
      (dropout): Dropout(p=0, inplace=False)
      (downsample): Linear(in_features=200, out_features=200, bias=True)
    )
    (3): ConvBlock(
      (conv): Conv1d(200, 200, kernel_size=(40,), stride=(1,), padding=(39,))
      (dropout): Dropout(p=0, inplace=False)
      (downsample): Linear(in_features=200, out_features=200, bias=True)
    )
    (4): ConvBlock(
      (conv): Conv1d(200, 200, kernel_size=(40,), stride=(1,), padding=(39,))
      (dropout): Dropout(p=0, inplace=False)
      (downsample): Linear(in_features=200, out_features=200, bias=True)
    )
    (5): ConvBlock(
      (conv): Conv1d(200, 200, kernel_size=(40,), stride=(1,), padding=(39,))
      (dropout): Dropout(p=0, inplace=False)
      (downsample): Linear(in_features=200, out_features=200, bias=True)
    )
  )
  (fc1): Linear(in_features=200, out_features=10, bias=True)
  (fc2): Linear(in_features=3000, out_features=100, bias=True)
  (g): Sequential(
    (0): Linear(in_features=100, out_features=25, bias=False)
    (1): BatchNorm1d(25, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (2): ReLU(inplace=True)
    (3): Linear(in_features=25, out_features=128, bias=True)
  )
)
```

Fig. 10: Model Architecture for Spiro-CLF model.

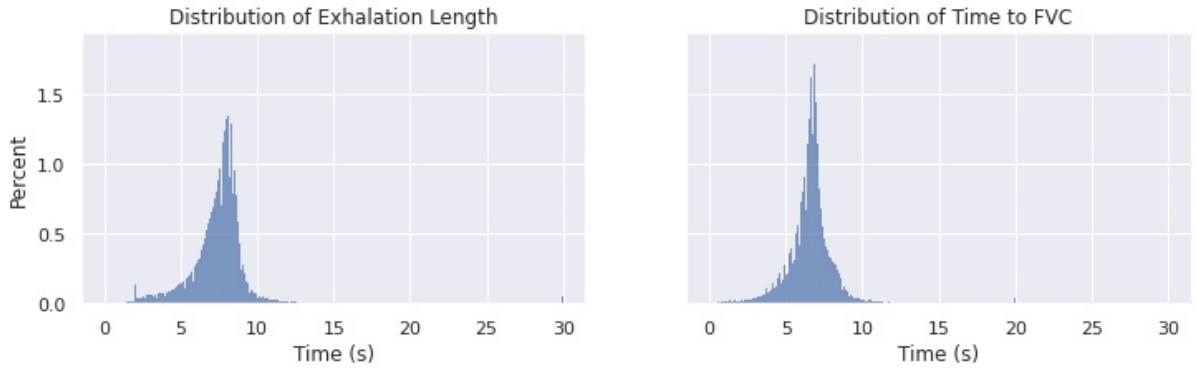


Fig. 11: Distribution of exhalation length and time to FVC in the dataset.

References

S. C. Allen, C. Charlton, W. Backen, M. Warwick-Sanders, and P. Yeung. Performing slow vital capacity in older people with and without cognitive impairment–is it useful? *Age and ageing*, 39(5):588–591, September 2010. ISSN 1468-2834 0002-0729. doi: 10.1093/ageing/afq084.

V. Bellia, C. Sorino, F. Catalano, G. Augugliaro, N. Scichilone, R. Pistelli, C. Pedone, and R. Antonelli-Incalzi. Validation of FEV6 in the elderly: Correlates of performance and repeatability. *Thorax*, 63 (1):60–66, January 2008. ISSN 1468-3296 0040-6376. doi: 10.1136/thx.2007.080572.

Sudipto Bhattacharjee, Banani Saha, Parthasarathi Bhattacharyya, and Sudipto Saha. Classification of obstructive and non-obstructive pulmonary diseases on the basis of spirometry using machine learning

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

techniques. *Journal of Computational Science*, 63:101768, 2022. ISSN 1877-7503. doi: 10.1016/j.jocs.2022.101768.

Sandeep Bodduluri, Arie Nakhmani, Joseph M. Reinhardt, Carla G. Wilson, Merry-Lynn McDonald, Ramaraju Rudraraju, Byron C. Jaeger, Nirav R. Bhakta, Peter J. Castaldi, Frank C. Sciurba, Chengcui Zhang, Purushotham V. Bangalore, and Surya P. Bhatt. Deep neural network analyses of spirometry for structural phenotyping of chronic obstructive pulmonary disease. *JCI Insight*, 5(13), July 2020. doi: 10.1172/jci.insight.132781.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, July 2020.

Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11834–11845. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/628f16b29939d1b060af49f66ae0f7f8-Paper.pdf.

Justin Cosentino, Babak Behsaz, Babak Alipanahi, Zachary R. McCaw, Davin Hill, Tae-Hwi Schwantes-An, Dongbing Lai, Andrew Carroll, Brian D. Hobbs, Michael H. Cho, Cory Y. McLean, and Farhad Hormozdiari. Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. *Nature Genetics*, April 2023. doi: 10.1038/s41588-023-01372-4. URL https://doi.org/10.1038/s41588-023-01372-4.

Nilakash Das, Kenneth Verstraete, Sanja Stanojevic, Marko Topalovic, Jean-Marie Aerts, and Wim Janssens. Deep-learning algorithm helps to standardise ATS/ERS spirometric acceptability and usability criteria. *The European respiratory journal*, 56(6):2000603, December 2020. ISSN 1399-3003 0903-1936. doi: 10.1183/13993003.00603-2020.

Adam Foster, Rattana Pukdee, and Tom Rainforth. Improving transformation invariance in contrastive representation learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NomEDgIEBwE.

Anna Fry, Thomas J. Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E. Allen. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American journal of epidemiology*, 186(9):1026–1034, November 2017. ISSN 1476-6256 0002-9262. doi: 10.1093/aje/kwx246.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 978-1-71382-954-6.

Ehsan Hoseinzade and Saman Haratizadeh. Cnnpred: Cnn-based stock market prediction using a diverse set of variables. *Expert Systems with Applications*, 129:273–285, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.03.029. URL https://www.sciencedirect.com/science/article/pii/S0957417419301915.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Ji-Yong Jing, Tian-Cha Huang, Wei Cui, Feng Xu, and Hua-Hao Shen. Should FEV1/FEV6 replace FEV1/FVC ratio to detect airway obstruction? A metaanalysis. *Chest*, 135(4):991–998, April 2009.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

ISSN 1931-3543 0012-3692. doi: 10.1378/chest.08-0723.

Saehoon Kim, Sungwoong Kim, and Juho Lee. Hybrid Generative-Contrastive Representation Learning. *arXiv e-prints*, art. arXiv:2106.06162, June 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=KmykpuSrjcq.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2018.

Philip H. Quanjer, Sanja Stanojevic, Tim J. Cole, Xaver Baur, Graham L. Hall, Bruce H. Culver, Paul L. Enright, John L. Hankinson, Mary S.M. Ip, Jinping Zheng, and Janet Stocks. Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. *European Respiratory Journal*, 40(6):1324–1343, 2012. ISSN 0903-1936. doi: 10.1183/09031936.00080312. URL https://erj.ersjournals.com/content/40/6/1324.

Nick Shrine, Anna L. Guyatt, A. Mesut Erzurumluoglu, Victoria E. Jackson, Brian D. Hobbs, Carl A. Melbourne, Chiara Batini, Katherine A. Fawcett, Kijoung Song, Phuwanat Sakornsakolpat, Xingnan Li, Ruth Boxall, Nicola F. Reeve, Ma'en Obeidat, Jing Hua Zhao, Matthias Wielscher, Stefan Weiss, Katherine A. Kentistou, James P. Cook, Benjamin B. Sun, Jian Zhou, Jennie Hui, Stefan Karrasch, Medea Imboden, Sarah E. Harris, Jonathan Marten, Stefan Enroth, Shona M. Kerr, Ida Surakka, Veronique Vitart, Terho Lehtimäki, Richard J. Allen, Per S. Bakke, Terri H. Beaty, Eugene R. Bleecker, Yohan Bossé, Corry-Anke Brandsma, Zhengming Chen, James D. Crapo, John Danesh, Dawn L. DeMeo, Frank Dudbridge, Ralf Ewert, Christian Gieger, Amund Gulsvik, Anna L. Hansell, Ke Hao, Joshua D. Hoffman, John E. Hokanson, Georg Homuth, Peter K. Joshi, Philippe Joubert, Claudia Langenberg, Xuan Li, Liming Li, Kuang Lin, Lars Lind, Nicholas Locantore, Jian'an Luan, Anubha Mahajan, Joseph C. Maranville, Alison Murray, David C. Nickle, Richard Packer, Margaret M. Parker, Megan L. Paynton, David J. Porteous, Dmitry Prokopenko, Dandi Qiao, Rajesh Rawal, Heiko Runz, Ian Sayers, Don D. Sin, Blair H. Smith, María Soler Artigas, David Sparrow, Ruth Tal-Singer, Paul R. H. J. Timmers, Maarten Van den Berge, John C. Whittaker, Prescott G. Woodruff, Laura M. Yerges-Armstrong, Olga G. Troyanskaya, Olli T. Raitakari, Mika Kähönen, Ozren Polašek, Ulf Gyllensten, Igor Rudan, Ian J. Deary, Nicole M. Probst-Hensch, Holger Schulz, Alan L. James, James F. Wilson, Beate Stubbe, Eleftheria Zeggini, Marjo-Riitta Jarvelin, Nick Wareham, Edwin K. Silverman, Caroline Hayward, Andrew P. Morris, Adam S. Butterworth, Robert A. Scott, Robin G. Walters, Deborah A. Meyers, Michael H. Cho, David P. Strachan, Ian P. Hall, Martin D. Tobin, and Louise V. Wain. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nature Genetics*, 51(3):481–493, March 2019. ISSN 1546-1718. doi: 10.1038/s41588-018-0321-7.

Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022.

Michael R. Simon, Vernon M. Chinchilli, Brenda R. Phillips, Christine A. Sorkness, Robert F. Lemanske, Stanley J. Szefler, Lynn Taussig, Leonard B. Bacharier, and Wayne Morgan. Forced expiratory flow between 25% and 75% of vital capacity and FEV1/forced vital capacity ratio in relation to clinical and physiological parameters in asthmatic children with normal FEV1 values. *Journal of Allergy and Clinical Immunology*, 126(3):527–534.e8, 2010. ISSN 0091-6749. doi: 10.1016/j.jaci.2010.05.016.

Erik Strumbelj and Igor Kononenko. Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowl. Inf. Syst.*, 41(3):647–665, December 2014. ISSN 0219-1377. doi: 10.1007/s10115-013-0679-x.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman,

medRxiv preprint doi: https://doi.org/10.1101/2023.04.28.23289178; this version posted April 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10, 03 2015. doi: 10.1371/journal.pmed.1001779. URL https://doi.org/10.1371/journal.pmed.1001779.

J. Vandevoorde, S. Verbanck, D. Schuermans, L. Broekaert, D. Devroey, J. Kartounian, and W. Vincken. Forced vital capacity and forced expiratory volume in six seconds as predictors of reduced total lung capacity. *European Respiratory Journal*, 31(2):391–395, 2008. ISSN 0903-1936. doi: 10.1183/09031936.00032307.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

Claus F. Vogelmeier, Gerard J. Criner, Fernando J. Martinez, Antonio Anzueto, Peter J. Barnes, Jean Bourbeau, Bartolome R. Celli, Rongchang Chen, Marc Decramer, Leonardo M. Fabbri, Peter Frith, David M. G. Halpin, M. Victorina López Varela, Masaharu Nishimura, Nicolas Roche, Roberto Rodriguez-Roisin, Don D. Sin, Dave Singh, Robert Stockley, Jørgen Vestbo, Jadwiga A. Wedzicha, and Alvar Agustí. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. *American journal of respiratory and critical care medicine*, 195(5):557–582, March 2017. ISSN 1535-4970 1073-449X. doi: 10.1164/rccm.201701-0218PP.

Kang Wang, Kenli Li, Liqian Zhou, Yikun Hu, Zhongyao Cheng, Jing Liu, and Cen Chen. Multiple convolutional neural networks for multivariate time series prediction. *Neurocomputing*, 360:107–119, 2019. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.05.023. URL https://www.sciencedirect.com/science/article/pii/S092523121930685X.

Yimin Wang, Yicong Li, Wenya Chen, Changzheng Zhang, Lijuan Liang, Ruibo Huang, Wenhua Jian, Jianling Liang, Senhua Zhu, Dandan Tu, Yi Gao, Nanshan Zhong, and Jinping Zheng. Deep Learning for Automatic Upper Airway Obstruction Detection by Analysis of Flow-Volume Curve. *Respiration; international review of thoracic diseases*, 101(9):841–850, 2022a. ISSN 1423-0356 0025-7931. doi: 10.1159/000524598.

Yimin Wang, Yicong Li, Wenya Chen, Changzheng Zhang, Lijuan Liang, Ruibo Huang, Jianling Liang, Dandan Tu, Yi Gao, Jinping Zheng, and Nanshan Zhong. Deep learning for spirometry quality assurance with spirometric indices and curves. *Respiratory Research*, 23(1):98, April 2022b. ISSN 1465-993X. doi: 10.1186/s12931-022-02014-9.

Taedong Yun, Justin Cosentino, Babak Behsaz, Zachary R McCaw, Davin Hill, Robert Luben, Dongbing Lai, John Bates, Howard Yang, Tae-Hwi Schwantes-An, Anthony P Khawaja, Andrew Carroll, Brian D Hobbs, Michael H Cho, Cory Y McLean, and Farhad Hormozdiari. Unsupervised representation learning improves genomic discovery for lung function and respiratory disease prediction, 2023. Preprint.