Transformer-based Time-Series Biomarker Discovery for COPD Diagnosis

Soham Gadgil[*]
University of Washington

Joshua Galanter
Genentech, Inc.

Mohammadreza Negahdar
Genentech, Inc.

Abstract

Chronic Obstructive Pulmonary Disorder (COPD) is an irreversible and progressive disease which is highly heritable. Clinically, COPD is defined using the summary measures derived from a spirometry test but these are not always adequate. Here we show that using the high-dimensional raw spirogram can provide a richer signal compared to just using the summary measures. We design a transformer-based deep learning technique to process the raw spirogram values along with demographic information and predict clinically-relevant endpoints related to COPD. Our method is able to perform better than prior works while being more computationally efficient. Using the weights learned by the model, we make the framework more interpretable by identifying parts of the spirogram that are important for the model predictions. Pairing up with a board-certified pulmonologist, we also provide clinical insights into the different aspects of the spirogram and show that the explanations obtained from the model align with underlying medical knowledge.

1 Introduction

Chronic Obstructive Pulmonary Disorder (COPD) is a progressive lung condition characterized by persistent airflow limitation that impairs breathing and is typically caused by long-term exposure to harmful substances like cigarette smoke [8, 17]. It is the third leading cause of death worldwide, with a global prevalance of 3.91% (>250 million) [26]. Therefore, timely and accurate diagnosis of COPD is crucial to reducing patients' health risks.

Traditionally, the diagnosis of COPD is confirmed using a spirometry test, which measures the flow and volume of air that can be inhaled and exhaled. There are a couple of important summary measures derived from the spirogram: (i) FEV_1, or the forced Expiratory Volume in 1 second, which is the volume of air a person can forcefully exhale in the first second of exhalation (ii) FVC, or the forced vital capacity, which is the total amount of air a person can forcefully exhale (iii) PEF, or the peak expiratory flow, which is the maximum flow of air a person can forcefully exhale. A post-bronchodilatory ratio of FEV_1 to FVC less that 0.7 is considered to be diagnostic of COPD in patients with appropriate symptoms and exposures [28, 12]. However, there can be some challenges associated with using such pulmonary function test markers for COPD diagnosis. First, while they identify diagnosed COPD cases, they do not effectively detect patients at an earlier stage or those at risk of developing COPD. Second, the FEV_1/FVC ratio is the current gold standard for COPD detection, but it might not always be accurate since there is a lot of heterogeneity involved [2, 19]. Third, there can be a number of individuals with the pathological hallmarks of COPD who nonetheless fall above the arbitrary threshold, particularly patients with PRISM (preserved ratio impaired spirometry) [11]. Fourth, the FEV_1/FVC ratio is subject to effort dependence and there can be variability in measurement across different efforts for the same patient.

---

[*]Correspondence to <sgadgil@cs.washington.edu>.

Recently, some studies have explored the use of deep learning methodologies to determine whether raw spirogram values can provide a stronger prognostic signal for predicting clinically relevant COPD-related outcomes compared to traditional summary measures [20, 5]. In this work, we show that a transformer-based architectural change in one of the prior methods, DeepSpiro [20], performs better across most of the endpoints that we tested while being ~4.5 times more computationally efficient. By using the attention weights of the transformer, we are able to identify parts of the spirometry curve which are important for the different prediction tasks. Leveraging the expertise of a board-certified pulmonologist, we also provide clinical insights into the model predictions and find that the important parts of the curve are confirmed by medical knowledge.

2 Methods

2.1 Dataset and Label Generation

We use the UK Biobank dataset for this study [3], which is a large scale biomedical database with health information from around 500K participants. Specifically, we use the spirometry information from UKB field 3066, which includes the expiratory volumes in milliliters recorded every 10 milliseconds. Although the same participant may have multiple visits, we consider the Time-Volume measurements only from their first visit. After multiple preprocessing steps (details in Appendix A.2), we convert the Volume-Time curves to Flow-Volume curves which are the input to our model. The demographic features we use are age, sex, smoking status, and height (collected on the spirometry test date) since they are recognized as being predictive of COPD by our pulmonologist collaborator.

We test three clinical relevant COPD-related endpoints as part of this study: COPD risk, exacerbation, and mortality. COPD risk refers to the likelihood of developing COPD after the spirometry test date. Hospitalization is used as a proxy for exacerbation and mortality indicates a COPD-related death of the patient after the initial visit. To extract these labels from the UK Biobank dataset, we use procedure similar to the one described in [5]. Specifically, we derive binary COPD labels from the medical records available about the patients, including the primary-care reports, the hospital inpatient records, and the self-report records. Appendix A.3 provides more details. This study has been conducted using the UK Biobank Resource under Application Number 44257.

2.2 Time-Series Transformer

Here we describe the modeling procedure that we use for the preprocessed dataset obtained from UK Biobank (Figure 1). Given a patient's smoothened Flow-Volume curve $X \in \mathbb{R}^T$ with T steps, their demographic information f_d, and the correponding target label y, we first patchify the input curve into N patches, each of length P: $X \in \mathbb{R}^T \rightarrow \{p_1, p_2, \ldots, p_N\}$, $p_i \in \mathbb{R}^P$, $N = \frac{T}{P}$. Each patch p_i, which is standardized across all patients, is then passed through a linear projection layer, transforming it into an embedding space $Z_i \in \mathbb{R}^{d_{\text{embed}}}$, where d_{embed} is the embedding dimension:

$$Z_i = W p_i + b, \quad W \in \mathbb{R}^{d_{\text{embed}} \times P}, \quad b \in \mathbb{R}^{d_{\text{embed}}}$$

Next, following the BERT-style training paradigm [6], we prepend a learnable classification token, $\text{CLS} \in \mathbb{R}^{1 \times d_{\text{embed}}}$ to the sequence of patch embeddings and then add positional encodings, $PE_i \in \mathbb{R}^{d_{\text{embed}}}$, to all the tokens:

$$Z' = [\text{CLS} + PE_0; Z_1 + PE_1; Z_2 + PE_2; \ldots; Z_N + PE_N], \quad Z' \in \mathbb{R}^{(N+1) \times d_{\text{embed}}}$$

This augmented sequence is then fed into a Transformer encoder layer [29], which captures the temporal dependencies within the data, producing the output embeddings H.

$$H = \text{Transformer}(Z', M), H \in \mathbb{R}^{(N+1) \times d_{\text{embed}}}, M \in \mathbb{R}^{(N+1)}$$

Here, M represents a mask indicating which of the input patches are padding tokens. Following that, we extract the encoder embedding corresponding to the CLS token from H and pass it through a Multi-Layer perceptron to get the initial assessment: $f_{\text{initial}} = MLP(H[0])$. We then perform a feature fusion step in which the CLS token embedding is concatenated with demographic information. Finally, the fused features are processed using a CatBoost classifier to predict the target probability: $\hat{y} = \text{CatBoost}(f_{\text{initial}} \oplus f_d)$. CatBoost [25] is a gradient boosted decision tree framework that can handle categorical features. Details about the training and model selection are in Appendix A.4.

There are multiple baselines that we compare against: (i) **FEV$_1$/FVC ratio**, which is the summary metric traditionally used as a measure of COPD risk (ii) **DeepSpiro** [20], which is the prior method based on a Bi-LSTM and a temporal attention layer along with feature fusion (iii) **MLP (demographic)**, which is a MLP classifier with two hidden layers trained on the demographic features age, sex, smoking status, height (iv) **MLP (summary stats)**, which is a MLP classifier with two hidden layres trained on the summary statstics derived from the spirometry curve, including FEV$_1$, FVC, and the blow ratio FEV$_1$/FVC.
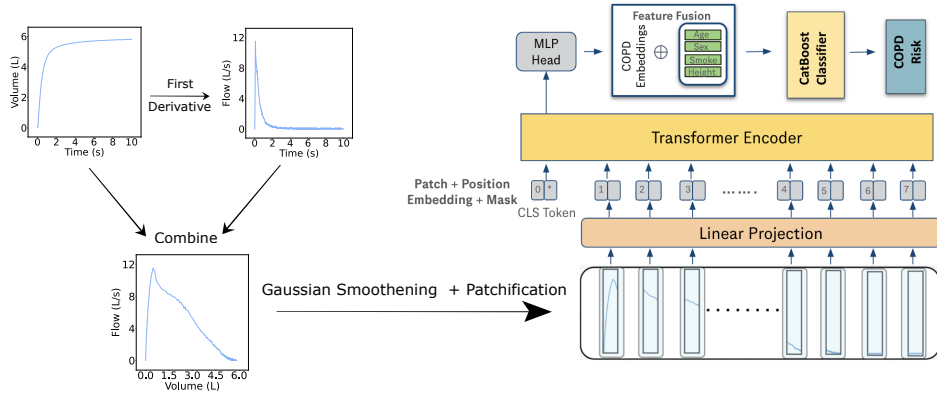

Figure 1: The overall framework that we use. We start with the raw Volume-Time curve and convert it to a Flow-Volume curve. This curve is then smoothened using a Gaussian filter and patchified before passing as an input to the time-series transformer.

3   Results

The evaluation metric we use is the area under the receiver operating characteristic curve (ROC-AUC). This metric is suitable for our tasks since it handles any imbalance that may be present in the dataset. Figure 1 shows the performance of the time-series transformer compared against the baselines along with the standard deviations across five trials. We observe that the time-series transformer outperforms the baselines across all three endpoints. The improvement compared to the strongest baseline, DeepSpiro, is marginal but the time-series transformer allows for greater parallelism during training, as the operations can be batched efficiently. As a result, it is more computationally efficient, with each epoch taking ∼40 seconds, compared to ∼180 seconds per epoch for training DeepSpiro. Results for another evaluation metric, the Brier score, are provided in Appendix A.5.

	COPD risk	Mortality	Exacerbation
FEV$_1$ /FVC Ratio	0.759 ± 0.001	0.863 ± 0.0204	0.805± 0.0075
MLP (Summary Stats)	0.779± 0.006	0.909 ± 0.0188	0.822± 0.0114
MLP (Demographic)	0.743± 0.005	0.817± 0.0133	0.801± 0.0076
DeepSpiro	0.817± 0.002	0.923± 0.0106	0.844± 0.0047
Time-Series Transformer (Ours)	**0.828± 0.004**	**0.929 ± 0.0147**	**0.851 ± 0.0027**

Table 1: Performance of the time-series transformer compared to the other baselines in terms of the area under the ROC curves (ROC-AUC score) across all endpoints. The uncertainty estimates indicate the standard deviation across five independent trials.

To understand the importance of each patch to the model predictions, we consider the mean attention weights of the CLS token across all the encoder layers with each of the other patch tokens after normalizing them using a softmax function. We then overlay these attention scores on top of the original Flow-Volume curves to visualize the importance of each patch. To better understand the model predictions for different patient cohorts, we stratify the patients based on the severity of COPD using the Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria. We consider two cohorts: (i) GOLD Stage 1 and 2; Mild to Moderate COPD (FEV$_1$ % predicted >= 50) and (ii) GOLD Stage 3 and 4; Moderate to Severe COPD (FEV$_1$% predicted < 50). For each cohort, the attention scores are averaged across all patients. We also annotate the curve with some important markers,

3

like the PEF and the FEF25/50/75 which represents Forced expiratory flow % exhaled (Figure 2). The black boxes represent the part of the curve that is the most important for that specific prediction task (due to space constraints, results for the mortality endpoint are deferred to Appendix A.6). For predicting COPD risk, different segments of the spirometry curve hold varying levels of importance depending on the GOLD stage. For GOLD stages 1 and 2, the initial segment of the curve, between the PEF and FEF25, is particularly significant, whereas for GOLD stages 3 and 4, the latter portion of the curve, between FEF50 and FEF75, becomes more relevant. When predicting exacerbation, the early segment (PEF to FEF25) is crucial for GOLD stages 1 and 2, while the middle segment (FEF25 to FEF50) is more pertinent for stages 3 and 4.
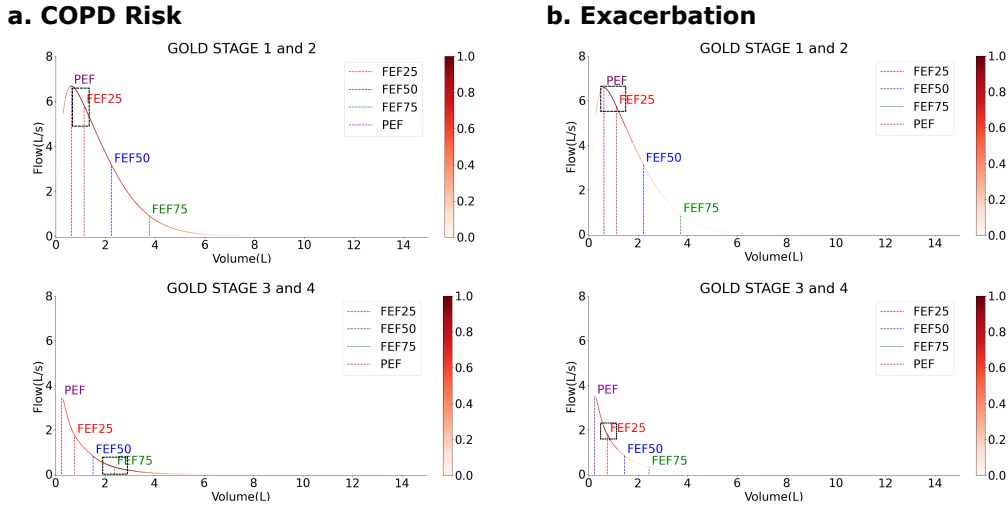

Figure 2: Overlaying the attention weights from the transformer encoder onto the Flow-Volume curve to visualize the importance of each patch for the COPD Risk and Exacerbation prediction tasks. The black rectangle in each curve represents the most important patch.

4 Clinical Insights

After identifying the significance of various curve segments for model predictions, we consulted a board-certified pulmonologist to gain clinical insights into the results. The initial portion of the flow-volume curve reflects airflow from the larger airways (e.g., main bronchi), while the tail end of the curve corresponds to airflow from the smaller airways (e.g., alveoli, bronchioles). Key features of the curve that are characteristic of COPD include: (i) a steep initial decline in flow at high lung volumes (following the PEF), (ii) a scooped or concave shape indicative of airflow limitation, (iii) significant displacement in the middle portion of the curve, and (iv) a flattening toward the end (between FEF50 and FEF75), indicative of reduced expiratory flow rate at low lung volumes, which may indicate respiratory disorders. Our findings reveal that the model prioritizes segments occurring after the PEF, with none of the pre-PEF segments identified as significant, suggesting the model is correctly focusing on curve aspects that are clinically relevant for COPD diagnosis. While further evaluation of the model's explanations is necessary, the current observations align with established clinical knowledge of COPD-related outcomes.

5 Conclusion

In this study, we investigate the advantages of using raw spirometry curves over traditional summary measures for predicting COPD-related outcomes, employing a transformer-based deep learning framework. Our model outperforms previous methods while demonstrating greater computational efficiency. By leveraging the attention weights of the transformer encoder, we incorporate interpretability into the framework, enabling us to identify critical regions of the curve that influence model predictions. The identified regions correspond with clinical insights provided by a board-certified pulmonologist, highlighting the model's alignment with established medical understanding. This work establishes a foundation for a more comprehensive exploration of therapeutic biomarkers for COPD.

References

[1] Timothy J Barreiro and Irene Perillo. An approach to interpreting spirometry. *American family physician*, 69(5):1107–1115, 2004.

[2] Surya P Bhatt, Arie Nakhmani, Spyridon Fortis, Matthew J Strand, Edwin K Silverman, Frank C Sciurba, and Sandeep Bodduluri. Fev1/fvc severity stages for chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 208(6):676–684, 2023.

[3] UK Biobank. About uk biobank, 2014.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[5] 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, et al. Inference of chronic obstructive pulmonary disease with deep learning on raw spirograms identifies new genetic loci and improves risk models. *Nature Genetics*, 55(5):787–795, 2023.

[6] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Miguel Angel Fernandez-Granero, Daniel Sanchez-Morillo, Miguel Angel Lopez-Gordo, and Antonio Leon. A machine learning approach to prediction of exacerbations of chronic obstructive pulmonary disease. In *Artificial Computation in Biology and Medicine: International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2015, Elche, Spain, June 1-5, 2015, Proceedings, Part I 6*, pages 305–311. Springer, 2015.

[8] Michael C Ferrera, Wassim W Labaki, and MeiLan K Han. Advances in chronic obstructive pulmonary disease. *Annual review of medicine*, 72(1):119–134, 2021.

[9] Nina S Godtfredsen, Tai H Lam, Trevor T Hansel, ME Leon, N Gray, C Dresler, DM Burns, E Prescott, and J Vestbo. Copd-related morbidity and mortality after smoking cessation: status of the evidence. *European Respiratory Journal*, 32(4):844–853, 2008.

[10] Kyle A Hasenstab, Nancy Yuan, Tara Retson, Douglas J Conrad, Seth Kligerman, David A Lynch, Albert Hsiao, and COPDGene Investigators. Automated ct staging of chronic obstructive pulmonary disease severity for predicting disease progression and mortality with a deep learning convolutional neural network. *Radiology: Cardiothoracic Imaging*, 3(2):e200477, 2021.

[11] Daniel H Higbee, Raquel Granell, George Davey Smith, and James W Dodd. Prevalence, risk factors, and clinical implications of preserved ratio impaired spirometry: a uk biobank cohort analysis. *The Lancet Respiratory Medicine*, 10(2):149–157, 2022.

[12] Firdaus AA Mohamed Hoesein, Pieter Zanen, and Jan-Willem J Lammers. Lower limit of normal or fev1/fvc< 0.70 in diagnosing copd: an evidence-based review. *Respiratory medicine*, 105(6):907–915, 2011.

[13] Ali Hussain, Ikromjanov Kobiljon Komil Ugli, Beom Su Kim, Minji Kim, Harin Ryu, Satyabrata Aich, and Hee-Cheol Kim. Detection of different stages of copd patients using machine learning techniques. In *2021 23rd International Conference on Advanced Communication Technology (ICACT)*, pages 368–372. IEEE, 2021.

[14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[15] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Kate Ann Levin, David Anderson, and Emilia Crighton. Prevalence of copd by age, sex, socioeconomic position and smoking status; a cross-sectional study. *Health Education*, 120(5/6):275–288, 2020.

[17] William MacNee. Pathology, pathogenesis, and pathophysiology. *Bmj*, 332(7551):1202–1204, 2006.

[18] Hannaneh Mahdavi, Saeideh Rahbarpour, Seyed Mohsen Hosseini-Golgoo, and Hamidreza Jamaati. A single gas sensor assisted by machine learning algorithms for breath-based detection of copd: A pilot study. *Sensors and Actuators A: Physical*, 376:115650, 2024.

[19] Astri Medbø and Hasse Melbye. Lung function testing in the elderly—can we still use fev1/fvc< 70% as a criterion of copd? *Respiratory medicine*, 101(6):1097–1105, 2007.

[20] Shuhao Mei, Yuxi Zhou, Jiahao Xu, Yuxuan Wan, Shan Cao, Qinghao Zhao, Shijia Geng, Junqing Xie, and Shenda Hong. Deep learning for detecting and early predicting chronic obstructive pulmonary disease from spirogram time series: A uk biobank study. *arXiv preprint arXiv:2405.03239*, 2024.

[21] Martin R Miller, JATS Hankinson, Vito Brusasco, F Burgos, R Casaburi, A Coates, R Crapo, Pvd Enright, CPM Van Der Grinten, P Gustafsson, et al. Standardisation of spirometry. *European respiratory journal*, 26(2):319–338, 2005.

[22] Shigeo Muro, Masato Ishida, Yoshiharu Horie, Wataru Takeuchi, Shunki Nakagawa, Hideyuki Ban, Tohru Nakagawa, Tetsuhisa Kitamura, et al. Machine learning methods for the diagnosis of chronic obstructive pulmonary disease in healthy subjects: retrospective observational cohort study. *JMIR medical informatics*, 9(7):e24796, 2021.

[23] Anne O Nielsen, Peter Lange, Ole Hilberg, Ingeborg Farver-Vestergaard, Rikke Ibsen, and Anders Løkke. Copd and smoking status–it does matter: characteristics and prognosis of copd according to smoking status. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation*, 11(1):56, 2024.

[24] Tamara Alonso Perez, Elena García Castillo, Julio Ancochea, María Teresa Pastor Sanz, Pere Almagro, Pablo Martínez-Camblor, Marc Miravitlles, Mónica Rodríguez-Carballeira, Annie Navarro, Bernd Lamprecht, et al. Sex differences between women and men with copd: a new analysis of the 3cia study. *Respiratory medicine*, 171:106105, 2020.

[25] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

[26] SA Quaderi and JR Hurst. The unmet global burden of copd. *Global health, epidemiology and genomics*, 3:e4, 2018.

[27] Dimitris Spathis and Panayiotis Vlamos. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health informatics journal*, 25(3):811–827, 2019.

[28] Kjell Torén, Linus Schiöler, Anne Lindberg, Anders Andersson, Annelie F Behndig, Göran Bergström, Anders Blomberg, Kenneth Caidahl, Jan E Engvall, Maria J Eriksson, et al. The ratio fev1/fvc and its association to respiratory symptoms—a swedish general population study. *Clinical physiology and functional imaging*, 41(2):181–191, 2021.

[29] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[30] Konstantin Willer, Alexander A Fingerle, Wolfgang Noichl, Fabio De Marco, Manuela Frank, Theresa Urban, Rafael Schick, Alex Gustschin, Bernhard Gleich, Julia Herzen, et al. X-ray dark-field chest imaging for detection and quantification of emphysema in patients with chronic obstructive pulmonary disease: a diagnostic accuracy study. *The Lancet Digital Health*, 3(11):e733–e744, 2021.

A Appendix

A.1 Related Work

Since COPD is a high-prevalence disease with multiple modeling challenges, there has been prior research done on detecting COPD using deep learning techniques on different types of data modalities. One study employed a recursive feature elimination technique evaluated with several classification models to distinguish early-stage COPD patients from those in advanced stages using clinical and demographic information [13]. Another one conducted an empirical pulmonolgy study of a representative sample of 132 patients for the prediction of COPD and asthma using a random forest classifier [27]. The study utilized 22 different features about the patients, ranging from demographic and medical information to special lung summary measurements obtained from a spirometer. One method aimed at predicting the risk factors for COPD diagnosis using machine learning techniques like gradient boosted decision trees and logisitic regression models [22]. The dataset consisted of the annual medical check-up of Hitachi employees during a 20 year period. One other method demonstrated the potential of using a single commercial metal oxide gas sensor, TGS-2602STMS, to capture comprehensive breath information for COPD detection by modeling it using a Support Vector Machine (SVM) classifier [18]. There is a prior work that trained, validated and compared three different type of networks: a radial basis function (RBF), a k-means classifier, and a probabilistic neural network (PNN) to predict exacerbations of COPD from daily responses to a health questionnaire spanning 6 months [7]. Focusing on imaging modalities, one study developed a method to detect and quantify emphysema in COPD patients using x-ray dark-field images, CT images, and pulmonary function test measurements [30]. Another imaging study developed a deep learning method to stage COPD severity by quantifying emphysema and air-trapping in CT images using a co-registration and lung segmentation algorithm [10]. However, none of these prior methods focused on utilizing the raw spirometry blow values for COPD prediction.

There have been prior works that have tried modeling the raw spirogram using deep learning techniques. One method used a CNN-based ResNet-1D model to predict COPD liability scores without requiring specific domain knowledge [5]. We use the same dataset and labeling procedures used in this study. However, this work does not perform any explainability analysis to understand the reasoning processes of the trained classifier. Another method, called DeepSpiro, built on the ResNet-1D model by introducing several modifications like Bi-LSTM for modeling the long range dependencies, a temporal attention layer for providing interpretability, and incorporating demographic information [20]. This is the main method that we build on and compare against.

A.2 Data Preprocessing

From the initial Volume-Time measurements, we employ a number of preprocessing steps to curate the input to our model. To ensure that the blows are accurately measured, we consult UKB field 3061 which indicates the acceptability of each blow result. If the value in this field is either 0 or 32, the blow is considered to be valid. To control the quality of the measurements, we look at the FEV_1, FVC, and PEF values derived from the Time-Volume curve and drop any blow which is in the extreme tail of all observed values (top or bottom 0.5%). After completing these preprocessing steps, we get 356,848 patients with valid spirograms.

To process the Time-Volume curve, we first convert the measurements from milliliters to liters and apply a 1D-Gaussian smoothing with a standard deviation of 1 to enhance the stability of the curves. Following this, based on respiratory physiology and prior works [21, 1, 20, 5], we convert the Volume-Time to Flow-Time curves by taking the first derivative with respect to time using finite differences. Specifically, if $V(t)$ represents the Volume measurements at time t and Δt represents the time interval, the flow $F(t) = \frac{V(t+\delta t) - V(t)}{\delta t}$. The flow data is then linearly interpolated to ensure consistency between the Flow-Time and Volume-Time curves. The resulting volume-time and flow-time curves are then combined to generate a one-dimensional flow-volume curve which is the input to our model (Figure 1). To ensure that all the curves are of the same length, we right-pad the shorter curves by zero upto the length of the longest curve in the dataset. We also extract demographic information like age, sex, smoking status, and height, which are all collected at the same time as the date of spirometry and have been shown previously to be predictive of COPD. The prevalence of COPD has been shown to be higher in older people and females [16, 24]. Smoking status is a known confounder for COPD [9, 23] so we adjust for it by including it as a feature in our model. Patients

with different heights will have different lung capacities which can affect the COPD status so it is also included as a demographic features. During modeling, age and height are represented as integers while smoking status and sex are represented as binary variables.

A.3 Details of Label Generation

To obtain the labels for training, we use the medical records of the patients present in the UK Biobank dataset, specifically the self report records, the hospital inpatient records, and the primary care records. The self report records (UKB field 20002) contain non-cancerous illness codes which are placed in a coding tree, either by the participant or a trained nurse. Presence of codes 1112, 1113, or 1472 indicates a positive COPD-status. For the hospital inpatient records (fields 41270 and 41271) we consider the presence of COPD-related ICD-9 and ICD-10 codes. COPD status from GP clinical records (field 42040) is obtained after the read v2 and read v3 codes are mapped to the corresponding ICD-10 codes using the TRUD mappings (data codings 1835 and 1834). The exacerbation label only includes cases with COPD as primary cause of hospitalization after the spirometry test date. COPD-related Mortality information is obtained from UKB field 40000.

Following the filtering codes used in [5], the ICD-10 codes considered as being related to COPD include: J43 (Emphysema), J440 (Chronic obstructive pulmonary disease with acute lower respiratory infection), J441 (Chronic obstructive pulmonary disease with (acute) exacerbation), J449 (Chronic obstructive pulmonary disease, unspecified). The ICD-9 codes considered include: 492 (Emphysema), 496 (Chronic Airway Obstruction). For the mortality endpoint, ICD-10 code J41 (Chronic Bronchitis) is also considered.

A.4 Model Selection and Training Details

For the disease classifier after the feature fusion step, we decided to use the CatBoost framework because of its ability to handle categorical features natively. Other boosting algorithms like XGBoost [4] and LightGBM [14] require manually encoding of categorical features (e.g. one-hot encoding). CatBoost is able to encode categorical features without causing data leakage, resulting in reduced overfitting. Also, CatBoost has been shown to be robust to hyperparameter tuning [25].

For training, after doing a hyperparamater sweep, we use a patch length $P = 30$ and an embedding dimension $d_{\text{embed}} = 200$ with 2 transformer encoder layers, each with 2 multiattention heads. For the CatBoost classifier, we use the default parameter values provided as part of the library. The standard cross-entropy loss is used to train the model for 30 epochs with the Adam optimizer [15] and a learning rate of 10^{-5}, using a train/test/val split of 80/10/10. During the feature fusion step, we tested with using both the normalized and the raw values for the demographic features to concatenate with the transformer's output embedding and got similar performance in both scenarios.

A.5 Additional Evaluation Metrics

In addition to the ROC-AUC scores, we also use the Brier score as an evaluation metric. The Brier score is a metric used to assess the accuracy of probabilistic predictions. It measures the mean squared difference between predicted probabilities and the actual outcomes (typically binary, where 0 or 1 represents the observed result). A lower Brier score indicates that the model is better calibrated and has more accurate predictions, with a score of 0 representing perfect predictions.

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - y_i)^2$$

where $\hat{p}_i$ is the predicted probability and y_i is the target label for the i-th instance. We observe that our method is able to outperform the the other baselines across all endpoints except for the mortality, but the brier score is still close to zero, indicating that the model is confident in its predictions (Supplementary Table 1).

	COPD risk	Mortality	Exacerbation
FEV$_1$ /FVC Ratio	0.0776 ± 0.0001	0.065 ± 0.0000	0.0724± 0.0001
MLP (Summary Stats)	0.0335± 0.0000	0.0012 ± 0.0000	0.0209± 0.0000
MLP (Demographic)	0.0317± 0.0001	**0.0011± 0.0000**	0.0203 ± 0.0002
DeepSpiro	0.0281± 0.0002	0.0061± 0.0001	0.0202± 0.0001
Time-Series Transformer (Ours)	**0.0280± 0.0002**	0.0045 ± 0.0001	**0.0199 ± 0.0000**

Supplementary Table 1: Performance of the time-series transformer compared to the other baselines in terms of the brier score across all endpoints. The uncertainty estimates indicate the standard deviations across five independent trials.

A.6   Results for Mortality endpoint

Here we present the interpretability results for the mortality endpoint. In the attention maps for predicting mortality, the middle portion of the curve (FEF25 to FEF50) is key for GOLD stages 1 and 2, whereas the final portion (FEF50 to FEF75) holds greater importance for stages 3 and 4 (Supplementary Figure 1). This still aligns with the underlying medical knowledge as observed for the COPD risk and exacerbation endpoints.
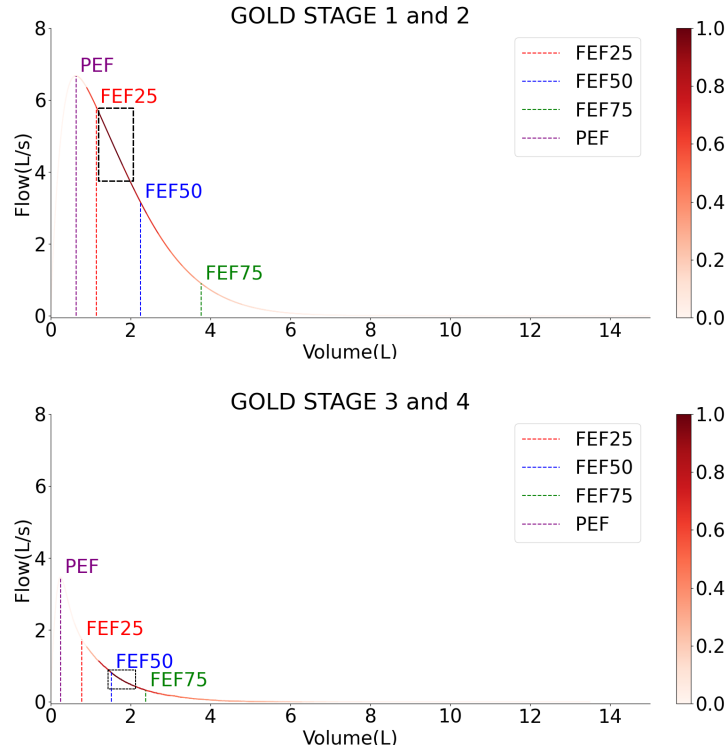
Supplementary Figure 1: Overlaying the attention weights from the transformer encoder onto the Flow-Volume curve for the Mortality endpoint. The black rectangle in each curve represents the most important patch.