Genome-wide association study of emphysema- and airway-predominant deep learning subtypes

Davin Hill[1,2], Kangjin Kim[2,4], Matthew Moll[2,3,4], Max Torop[1], Aria Masoomi[1], Sandeep Bodduluri[5], Peter J. Castaldi[2,4], Brian D. Hobbs[2,3,4], Jennifer Dy[1], Surya P. Bhatt[5] , Michael H. Cho[2,3,4]

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

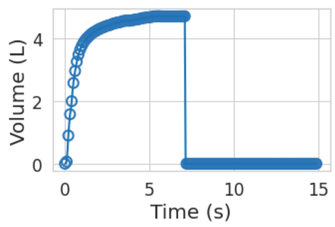
Introduction

- Chronic obstructive pulmonary disease (COPD) may be characterized by a combination of two different structural diseases: emphysema and airway disease.
- The relative contribution from these two diseases can be used to categorize patients into emphysema-predominant and airway-predominant phenotypes, which is established with CT measures of emphysema and airway remodeling.
- While CT can differentiate between these two phenotypes, CT results are often unavailable due to expense and entail radiation exposure.
- In Bodduluri et al [1], a neural network (SpiroNet) trained on spirometry volume-time curves in COPDGene [2] (N=8,980) was effective in differentiating emphysema-predominant and airway-predominant subtypes on CT scans.
- Whether these phenotypes have specific genetic susceptibility, differing from summative measures of lung function, is unknown.

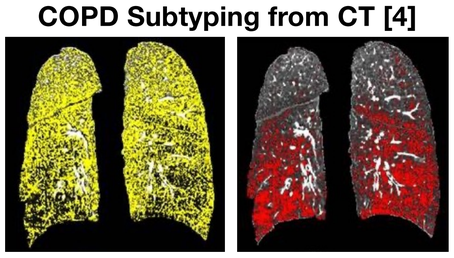
Method

- We used SpiroNet [1], which is trained on COPDGene [2] to predict COPD subtype using volume-time exhalation curves.

Spirometry Exhalation Curve



COPD Subtyping from CT [4]



- We applied SpiroNet to predict emphysema-predominant and airway-predominant phenotypes from the UK Biobank [3] (N=352,684)
- We performed a GWAS on each phenotype as a binary indicator. We adjusted for age, sex, pack-years, ever smoking status, batch, and ancestry principal components using plink with Firth fallback. We considered $< 1x10^{-10}$ as genome-wide significant ($5x10^{-9}$, two traits), and loci as associations > 500kb apart.
- We compared results to previously published GWAS using COPD and measures of lung function (forced expiratory volume in one second, forced vital capacity, and the ratio of the two measures).

Conclusion

- We attempted to establish Emphysema-predominant and airway-predominant phenotypes derived in COPDGene in the UK Biobank by using a neural network model applied to spirometry.
- We observed a class imbalance in the Emphysema-predominant subtype in SpiroNet, which is exacerbated when applying to UK Biobank, evidenced through sex and FEV_1.
- Nevertheless, we identified genome-wide significant associations overlapping with previously described lung function and COPD loci.
- Future Work: We will investigate transfer learning methods to improve phenotype generalization. We also plan to investigate other deep learning spirometry phenotypes and cell types.

Results / Figures

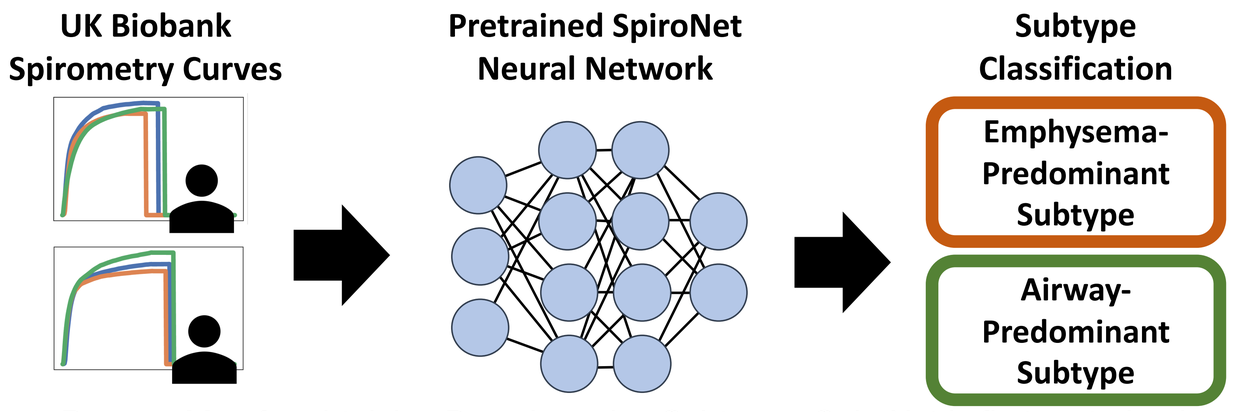


Figure 1: Process Map for obtaining Deep Learning Subtypes. SpiroNet is first pretrained on the COPDGene dataset. Next, UK Biobank participants are categorized as Emphysema-predominant or Airway-predominant subtypes based on raw spirometry curves.

	COPDGene		UK Biobank			
	Emphysema	Airway	Emphysema	Airway	Control (E)	Control (A)
Participants, n	1,390	3,207	30,677	101,849	322,007	250,835
Incidence, %	15.5	35.7	8.7	28.9		
Age, yr, mean (SD)	63.1 (8.2)	57.3 (8.5)	52.7 (7.9)	58.7 (7.6)	56.6 (8.0)	55.3 (8.0)
Sex, F, n (%)	438 (31.5)	1,938 (60.4)	668 (2.2)	73,104 (72.0)	196,378 (61.0)	123,942 (49.4)
(Non-Hispanic) White, n (%)	1,160 (83.5)	1,766 (55.1)	30,256 (99.0)	95,553 (94.1)	304,743 (94.9)	239,446 (95.8)
(Non-Hispanic) Black/Black British, n (%)	230 (16.5)	1,441 (44.9)	1,266 (1.8)	1,276 (1.8)	4,569 (1.4)	2,960 (1.2)
Current Smokers, n (%)	413 (29.7)	2,155 (67.1)	3,681 (12.0)	13,190 (13.0)	31,945 (9.9)	22,436 (8.9)
FEV_1, L, mean (SD)	2.1 (0.9)	2.2 (0.7)	4.1 (0.6)	2.2 (0.5)	2.7 (0.7)	3.1 (0.7)
FEV_1 / FVC, mean (SD)	0.56 (0.1)	0.73 (0.1)	0.73 (0.1)	0.72 (0.1)	0.76 (0.1)	0.78 (0.1)

Table 1: Participant characteristics. Airway- and Emphysema-predominant subtypes were derived using COPDGene and then applied to the UK Biobank to derive deep learning subtypes.

- SpiroNet identified Airway and Emphysema phenotypes in UK Biobank, but with substantial class imbalance for Emphysema.
- We identified 83 loci between the two phenotypes (3 for emphysema and 80 for airway) at genome-wide significance.
- Seventy-nine of these loci were within 500kb of previously described regions associated with lung function or COPD (including the 3 identified for emphysema), with the exception of 4 airway-associated variants in the *HLA* region (spanning *HLAA* and *ZKSCAN4*), and one airway-associated variant > 500kb away from the lead variant near *HHIP*).

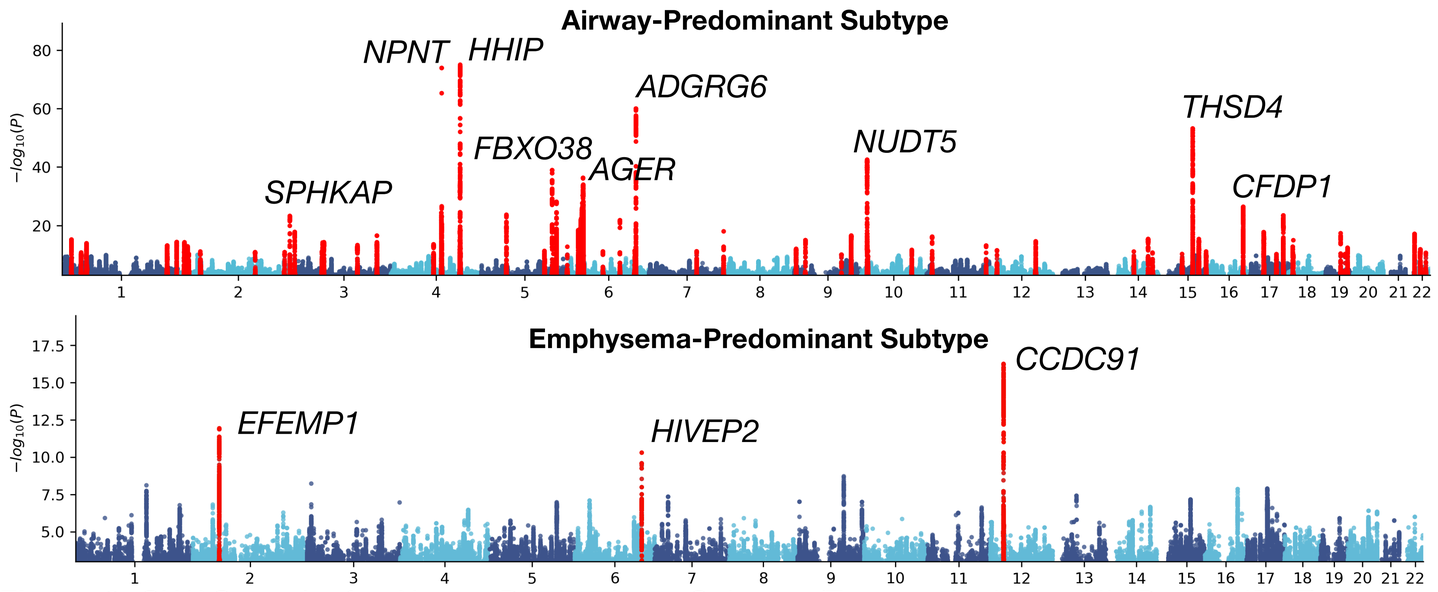


Figure 2: GWAS results for Airway-Predominant Subtype (Top, top loci near *HHIP* and *NPNT*) and Emphysema-Predominant (Bottom, top loci near *EFEMP1* and *CCDC91*).

References / Funding

[1] Bodduluri et al. *Deep neural network analyses of spirometry for structural phenotyping of chronic obstructive pulmonary disease*. JCI Insight, 5(13), July 2020.
[2] Regan et al. *Genetic epidemiology of COPD (COPDGene) study design*. COPD. 2010 Feb;7(1):32-43.
[3] Sudlow et al. *UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age*. PLoS Med 12(3): e1001779.
[4] Galbán et al. *Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression*. Nat Med. 2012 Nov;18(11):1711-5.

DH was supported by NIH 2T32HL007427-41. MHC was supported by R01HL162813, R01HL153248, R01HL149861, and R01HL147148. PJC is supported by NIH R01HL124233 and R01HL147326. SPB is supported by NIH R01HL151421 and UH3HL155806. BDH is supported by NIH K08HL136928, U01 HL089856, and an Alpha-1 Foundation Research Grant. The COPDGene study (NCT00608764) is supported by NHLBI grants R01 HL089897 and R01 HL089856, and the COPD Foundation through contributions made to an Industry Advisory Board composed of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, GlaxoSmithKline, Siemens, and Sunovion. MHC has received grant support from GSK and Bayer, consulting or speaking fees from Genentech, AstraZeneca, and Illumina. BDH receives grant support from 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.