Chronic Respiratory Diseases (CRDs) are among the leading causes of mortality and morbidity worldwide, especially Chronic Obstructive Pulmonary Disease (COPD) and asthma. In 2017, there were 545 million prevalent cases and 62 million incident cases of CRDs. In terms of mortality, there were 3.2 million deaths due to COPD and 495,000 deaths due to asthma. COPD was the seventh leading cause of years of life lost (YLLs).

Tobacco, ambient air pollution, airborne allergens, and occupational exposures are the main risk factors associated with a lifetime increase of respiratory diseases and symptoms. Symptoms of noninfectious CRDs are also exacerbated by changes in temperature and humidity. Despite being noninfectious in etiology, many CRDs have some aspects of their pathogenesis influenced by infections organisms. However, this is an active area of investigation.

The present study explores a novel application of machine learning analyses to forecast CRD risk and discuss its merits and limitations. We trained and tested a Random Forest regressor, with mortality rate as the target variable. We tested features from datasets of climate variables and indices, fire emissions, biosphere fluxes, and particulate matter over the United States during 2000-2016. County population data and shapefiles were used to adjust mortality rates by population density. Using recursive feature elimination with scikit-learn, we selected features that positively contribute to predicting mortality rates attributed to CRDs. Due to data limitations on the country-wide level, we did not evaluate the effects of tobacco, occupational exposures, and air conditioning overuse.

The final regressor produced R-squared values of 0.7526 and 0.7528 for cross-validation and test dataset prediction, respectively, implying that our model generalizes well out-of-sample. The selected features comprise location and temporal encoders, population density, and net primary production.

The study reveals significant potential for modeling CRD risk, but highlights setbacks due to the primarily noninfectious nature of CRDs, phenomena only identifiable on finer spatiotemporal scales, and data limitations. We also identify methods that may refine our approach and describe future developments that may improve CRD risk modeling.