Long-term PM2.5 exposure and the clinical application of machine learning for predicting incident AF among the general population: a nationwide cohort study

In-Soo Kim
Pil-Sung Yang
Eunsun Jang
Hyunjean Jung
Seng Chan You
Hee Tae Yu
Tae-Hoon Kim
Jae-Sun Uhm
Jong-Youn Kim
Hui-Nam Pak
Moon-Hyang Lee
Boyoung Joung

Introduction: Several studies have linked elevations in fine particulate matter (PM2.5) air pollution to air pollution-related incident atrial fibrillation (AF), however, its impact had not been well studied. We used integrated machine learning (ML) to build incident AF prediction models using PM2.5 from the Korean general population.

Methods: We included 432,587 subjects who received national health examinations from the Korean National Health Insurance Service-based National Sample Cohort. We matched subjects' residential ZIP codes with average of hourly-measurements of PM2.5 concentration and meteorological data during study period. Based on supervised learning methods to construct prediction models, we used a sequential method of feature construction and automated selection by information gain ranking to identify predictive risk factors from the various health examination parameters. We built prediction models using boosted ensemble algorithm with 10-fold cross-validation. We compared the incident AF prediction models using c-index, net reclassification improvement index (NRI) and integrated discrimination improvement index (IDI).

Result: ML using boosted ensemble method exhibited a higher c-index (0.85[0.84-0.86]) compared with existing logistic regression models using CHA2-DS2-VASc (0.80[0.79-0.81]), CHADS2 (0.78[0.77-0.79]), or HATCH (0.79[0.78-0.80]) score (each p<0.001) for predicting incident AF. With application of PM2.5, prediction performances are significantly increased compared to models without PM2.5 (c-indices: boosted ensemble ML 0.95(0.94-0.96); CHA2-DS2-VASc 0.84(0.83-0.85); CHADS2 0.82(0.81-0.83); or HATCH score 0.83(0.82-0.84); each interaction p<0.001). Compared to CHA2-DS2-VASc score, NRI of ML with/without PM2.5 were 1.31 (1.29–1.34) and 0.79 (0.76-0.83), and the IDI were 0.39 (0.38–0.41) and 0.02 (0.01-0.03), respectively.

Conclusion: ML combining clinical and PM2.5 data was found to predict incident AF better than models without PM2.5 or even established risk prediction approaches in the Korean general population exposed to high-air pollution. Further research is warranted to elucidate the role of PM2.5 for atrial remodeling.