Interpretable Machine Learning Prediction Model for Prenatal Ambient Air Pollution Exposure and Its Impact on Small-for-Gestational-Age Births: The Korean CHildren's ENvironmental health Study (Ko-CHENS)

Payam Hosseinzadeh Kasani [1,2], Eun Sun Kim [1,2], Heui Seung Jo [1,2], Kee Hyun Cho [1,2]*, Woo Jin Kim [3]*

[1] Department of Pediatrics, Kangwon National University Hospital, Chuncheon, Republic of Korea. [2] Department of Pediatrics, Kangwon National University School of Medicine, Chuncheon, Republic of Korea. [3] Department of Internal Medicine and Environmental Health Center, Kangwon National University, Chuncheon, Republic of Korea

Abstract

Introduction

Epidemiological studies have demonstrated an association between prenatal exposure to ambient air pollution and adverse birth outcomes. However, no predictive model assesses the risk of small-for-gestational-age (SGA) births in women exposed to specific pollutants during pregnancy, utilizing two pollution estimation methodologies: the Tele-Monitoring System (TMS) and the Kriging Interpolation Method (KRIG).

Methods:

Using data from 2,734 mothers enrolled in the Korean CHildren's ENvironmental Health Study (Ko-CHENS) cohort, we constructed two growth measure datasets (weight-based datasets (n = 2,734) 27.40% SGA, and length-based dataset (n = 2,422) with 16.52% SGA). We compared TMS and KRIG, applied by trimester and the entire pregnancy. Four machine learning models were evaluated by receiver-operating characteristic (ROC) curves, with Shapley Additive Explanations (SHAP) used to provide global model interpretations.

Results:

Our analysis showed no multicollinearity and XGBoost consistently outperformed other models across all trimesters and the entire pregnancy in both weight-based and length-based datasets, especially with the KRIG method. The highest performance was observed with KRIG for XGBoost in the first trimester for weight-based datasets (AUC 91.22%) and in the second trimester for length-based datasets (AUC 93.64%). The explainable results revealed by KRIG method the ambient pollution variables consistently appearing across all trimesters using the KRIG Method were particulate matter (PM2.5), nitrogen dioxide, and ozone in the weight-based dataset, and ozone, carbon monoxide (CO), and PM2.5 in the length-based dataset.

Conclusion:

Machine learning algorithms can create effective tools for predicting SGA in mothers exposed to ambient air pollution, potentially aiding in identifying high-risk mothers and neonates.

Results

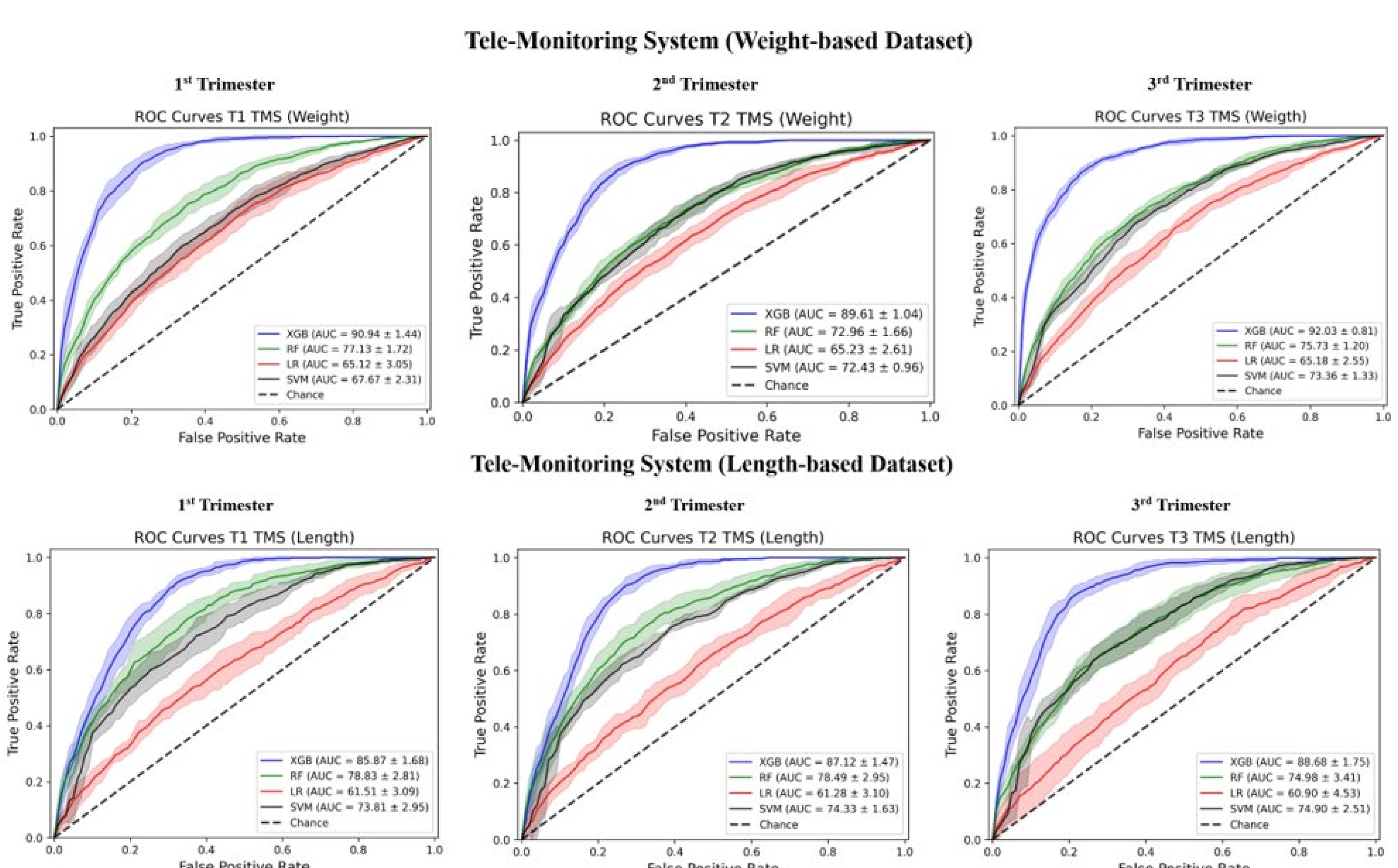

Figure 1. Receiver operating characteristic curves for 5-fold cross-validation for tele-monitoring system.

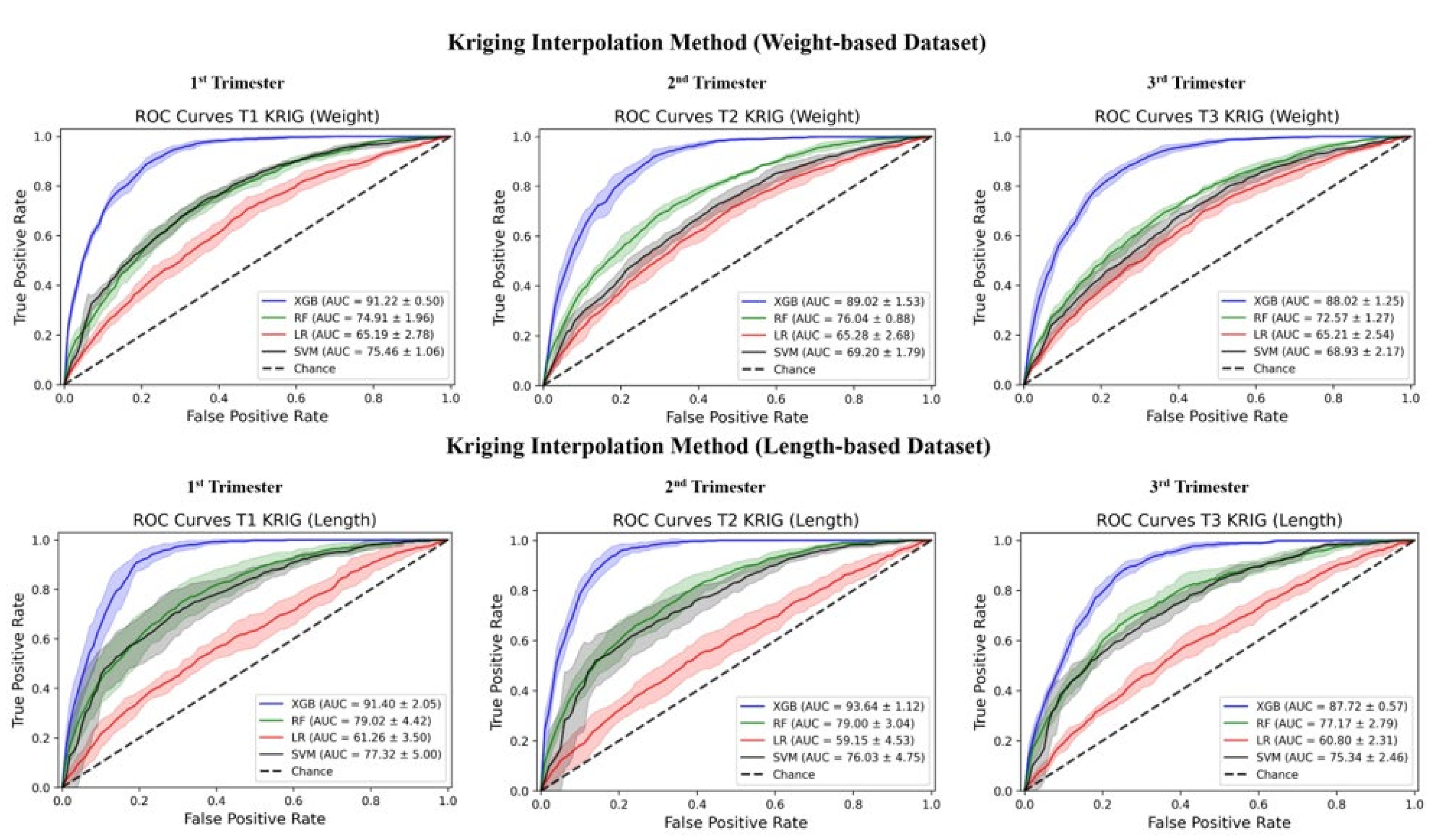

Figure 2. Receiver operating characteristic curves for 5-fold cross-validation for Kriging method.

Results

Table 1. XGBoost model interpretation using feature importance based on SHAP ranking in in weight-based datasets.

Tele-Monitoring System (Weight-based Dataset)					
1st Trimester		2nd Trimester		3rd Trimester	
Feature	Importance	Feature	Importance	Feature	Importance
TMS.NO2	2.7013	Pre_Preg Wt	4.7482	Maternal Ht	177.8706
TMS.PM2.5	2.6251	TMS.NO2	4.6149	GA	64.9483
GA	2.603	TMS.PM10	4.0815	TMS.PM10	14.9537
Pre_Preg Wt	2.4704	TMS.PM2.5	4.0653	Mother Age	14.5105
TMS.SO2	2.4204	GA	4.0618	Pre_Preg Wt	12.9402
TMS.CO	2.0808	TMS.CO	3.7697	TMS.PM2.5	12.3359
TMS.PM10	1.9937	Maternal Ht	3.3628	TMS.O3	11.6533
TMS.O3	1.986	TMS.SO2	3.0578	TMS.CO	11.3532
Maternal Ht	1.981	Mother Age	2.7385	TMS.NO2	7.1823
Mother Age	1.5599	TMS.O3	2.6085	C_section	3.0948
Parity	1.4272	C_section	1.9641	Gravidity	2.9998
C_section	1.1528	Sex	1.8353	Parity	2.7098
Gravidity	0.9603	Parity	1.4014	TMS.SO2	2.0103
Sex	0.8933	Gravidity	1.0648	Sex	1.9663
GDM	0.074	PIH	0	GDM	0.045
PIH	0	GDM	0	PIH	0

Kriging Interpolation Method (Weight-based Dataset)					
1st Trimester		2nd Trimester		3rd Trimester	
Feature	Importance	Feature	Importance	Feature	Importance
GA	2.5448	KRIG.PM10	7.5009	GA	5.8274
Pre_Preg Wt	2.3422	KRIG.NO2	6.9528	KRIG.PM2.5	5.2604
KRIG.CO.	2.2767	KRIG.O3	6.1259	Pre_Preg Wt	5.0771
KRIG.PM2.5	2.1299	KRIG.CO	5.4288	KRIG.PM10	5.0164
KRIG.NO2	2.0794	Pre_Preg Wt	5.3418	KRIG.CO	4.7744
KRIG.O3	1.8684	GA	4.8263	Maternal Ht	3.7686
KRIG.PM10	1.7798	KRIG.PM2.5	4.5774	Mother Age	3.1486
Mother Age	1.5051	Maternal Ht	4.0647	KRIG.O3	2.7708
Maternal Ht	1.4588	C_section	3.6817	KRIG.NO2	2.1783
Sex	1.4193	Mother Age	3.1899	Gravidity	1.6133
KRIG.SO2	1.2756	Gravidity	3.157	Sex	1.568
C_section	0.9625	KRIG.SO2	2.0644	C_section	1.5162
Parity	0.591	Sex	1.8445	KRIG.SO2	1.3131
Gravidity	0.5314	Parity	1.1604	Parity	1.2795
PIH	0.0218	GDM	0.0374	PIH	0.0284
GDM	0	PIH	0	GDM	0

Table 2. XGBoost model interpretation using feature importance based on SHAP ranking in in length-based datasets.

Tele-Monitoring System (Length-based Dataset)					
1st Trimester		2nd Trimester		3rd Trimester	
Feature	Importance	Feature	Importance	Feature	Importance
TMS.SO2	5.3315	TMS.NO2	6.6596	TMS.PM2.5	4.6561
TMS.PM2.5	4.2663	TMS.SO2	4.9789	TMS.PM10	3.8532
TMS.NO2	4.0852	GA	4.9747	GA	3.6595
TMS.CO	3.7858	TMS.PM2.5	4.7738	TMS.CO	3.2641
TMS.PM10	3.2106	TMS.PM10	4.0984	Pre_Preg Wt	3.0887
Pre_Preg Wt	3.1363	TMS.CO	4.0809	Mother Age	2.7376
Maternal Ht	3.009	Mother Age	3.6065	Maternal Ht	2.4393
GA	2.6731	TMS.O3	3.561	TMS.O3	1.6235
TMS.O3	2.5389	Pre_Preg Wt	3.4316	TMS.NO2	1.5398
Mother Age	2.4838	Maternal Ht	3.0902	Parity	1.5099
Gravidity	1.6239	Gravidity	2.3839	C_section	1.4758
C_section	1.4304	Sex	1.4252	Gravidity	1.2773
Parity	1.1752	C_section	0.8561	Sex	0.8999
Sex	0.9212	Parity	0.6352	TMS.SO2	0.8626
PIH	0	GDM	0.0754	GDM	0.0434
GDM	0	PIH	0	PIH	0

Kriging Interpolation Method (Length-based Dataset)					
1st Trimester		2nd Trimester		3rd Trimester	
Feature	Importance	Feature	Importance	Feature	Importance
KRIG.O3	5.0141	KRIG.O3	5.7116	KRIG.PM2.5	4.3573
GA	3.7492	KRIG.NO2	4.6596	KRIG.CO	4.2971
KRIG.CO	3.7144	KRIG.CO	4.6043	KRIG.PM10	3.8116
KRIG.NO2	3.6967	KRIG.PM2.5	4.5273	GA	3.2887
KRIG.PM2.5	3.6714	GA	4.1688	Pre_Preg Wt	3.2256
KRIG.PM10	3.3412	Pre_Preg Wt	3.5971	Mother Age	3.1029
KRIG.SO2	2.8925	KRIG.PM10	3.3025	Maternal Ht	2.6873
Pre_Preg Wt	2.6476	Mother Age	2.8424	KRIG.O3	2.0757
Mother Age	2.6248	Maternal Ht	2.7613	Parity	1.8856
Maternal Ht	2.5189	KRIG.SO2	2.0649	Gravidity	1.5811
Gravidity	1.4176	Gravidity	1.4817	C_section	1.2522
C_section	1.3574	C_section	1.3524	Sex	1.1702
Sex	0.988	Sex	1.254	KRIG.NO2	1.1097
Parity	0.7367	Parity	0.2736	KRIG.SO2	0.9832
GDM	0.0984	PIH	0	GDM	0.0157
PIH	0	GDM	0	PIH	0

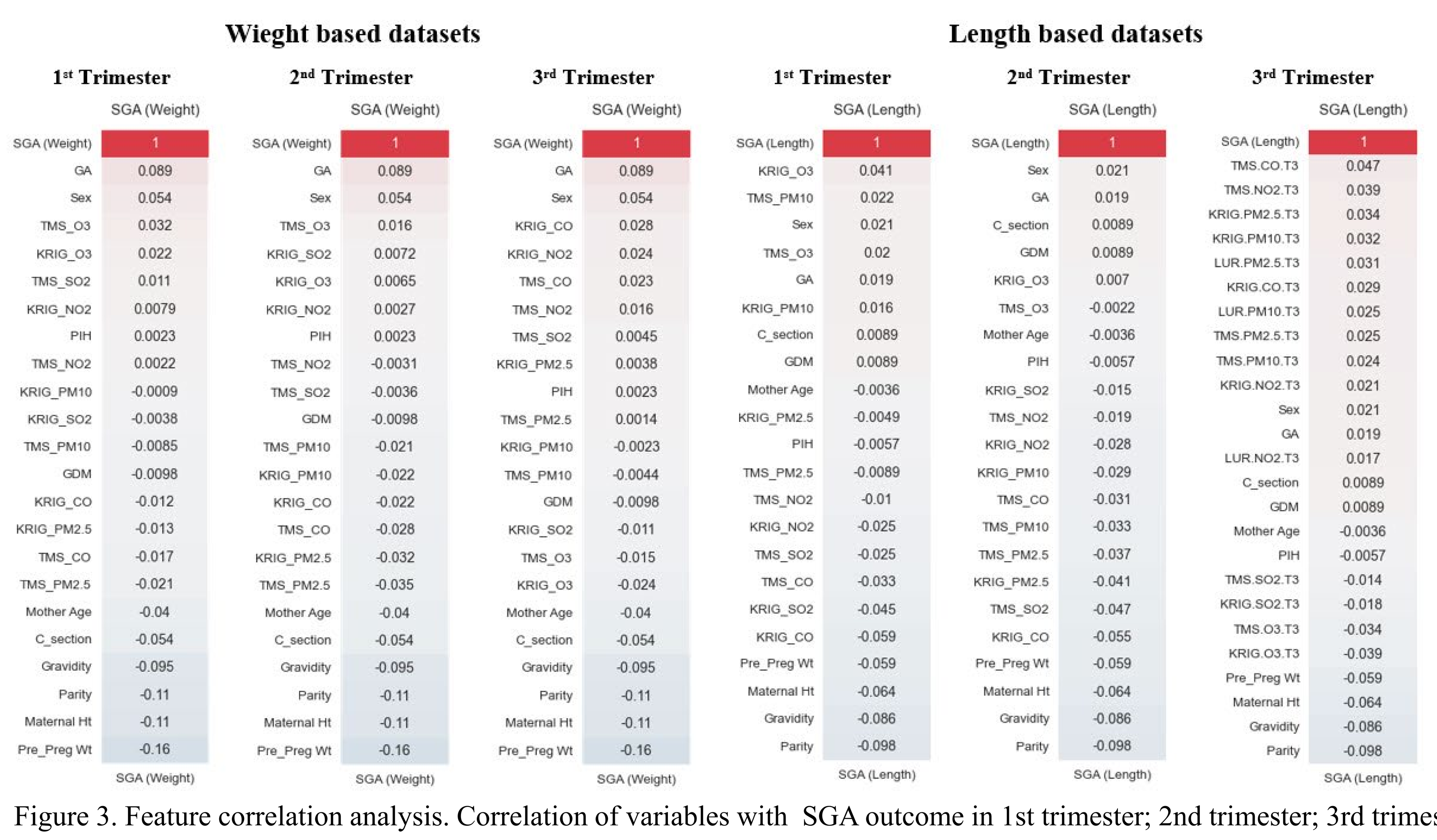

Figure 3. Feature correlation analysis. Correlation of variables with SGA outcome in 1st trimester; 2nd trimester; 3rd trimester. Positive impact sizes are represented by hues of red, while negative effect sizes are represented by shades of blue.

Discussion

- The consistent superior performance of the XGBoost model in our ROC curve analysis across all trimesters reveals its robustness and suggests that it can effectively manage the nonlinearities and interactions between various predictors in prenatal environments.
- The variation in feature importance across different trimesters and datasets highlights the dynamic nature of factors influencing fetal growth. The fact that pollutants like PM2.5 and sulfur dioxide fluctuate in influence suggests that environmental risks may need trimester-specific public health interventions. This could guide more targeted prenatal care and policy adjustments depending on local environmental conditions.
- The findings regarding gestational age reinforce its established importance in fetal monitoring but also suggest that integrating GA with real-time pollution data could refine risk assessments. Such integration could lead to the development of predictive models that are not only reactive but also proactive in adjusting to ongoing environmental and maternal health data throughout the pregnancy.

References

Needleman HL, Schell A, Bellinger D, Leviton A, Allred EN. The Long-Term Effects of Exposure to Low Doses of Lead in Childhood. N Engl J Med [Internet]. 1990 Jan 11;322(2):83–8. Available from: http://www.nejm.org/doi/abs/10.1056/NEJM199001113220203

Pedersen M, Giorgis-Allemand L, Bernard C, Aguilera I, Andersen AMN, Ballester F, et al. Ambient air pollution and low birthweight: a European cohort study (ESCAPE). Lancet Respir Med [Internet]. 2013 Nov;1(9):695–704. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2213260013701929

Jeong KS, Kim S, Kim WJ, Kim HC, Bae J, Hong YC, et al. Cohort profile: Beyond birth cohort study – The Korean CHildren's ENvironmental health Study (Ko-CHENS). Environ Res [Internet]. 2019 May;172:358–66. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0013935118306388

강원대학교 어린이병원
Kangwon National University Children's Hospital

Interpretable Machine Learning Prediction Model for Prenatal Ambient Air Pollution Exposure and Its Impact on Small-for-Gestational-Age Births: The Korean CHildren's ENvironmental health Study (Ko-CHENS)

Video Presentation

Scan the QR code to watch the video presentation!

Hot Topics in Neonatology®