ICEES – Environmental Exposures – Feature Variables

Table 1. ICEES feature variables – environmental exposures data: name, description, and binning strategy.*

Index Variables	***Description and binning strategy***
index	*Patient pseudo identifier, not available as feature variable, but allows for linkages across years when creating cohorts*
year	*Study period (calendar year, with a few exceptions)*
Active_In_Year	*Allows for selection of only those patients who are 'active' in a given study period or year, meaning that they had one or more visits with a healthcare provider*
Clinical Feature Variables	
Nomenclature format for medications: **XX**RX	*Medication XX prescribed or administered over study period, 0=no, 1=yes (one more more prescriptions/administrations over study period)*
Nomenclature format for diagnoses: **XX**Dx	*Diagnosis XX made over study period, 0=no, 1=yes (one more more diagnoses over study period)*
Nomenclature format for procedures **XX**	*Procedure XX performed over study period, 0=no, 1=yes (one more more diagnoses over study period)*
Nomenclature format for laboratory measures **XX** **XX**_first_flag **XX**_last_flag	 *Laboratory measurement (no units)* *Flag (e.g., normal, above normal, below normal) for first laboratory measurement over study period* *Flag (e.g., normal, above normal, below normal) for last laboratory measurement over study period*

US Census Bureau American Community Survey Data (years 2007-2011 & 2012-2016 survey samples)	
EstResidentialDensity	*Estimated total population [block group], binned according to US Census Bureau definitions (1=rural [0,2500), 2=urban cluster [2500,50000), 3=urbanized area [50000,inf))*
ur	*Estimated urban or rural residence, binned according to US Census Bureau definitions* *Note that u=rural, r=urban [known issue]*
EstResidentialDensity25Plus **2012-2016 sample only*	*Estimated total population aged 25 years or older [block group], binned as quintiles (pandas qcut) (1, 2, 3, 4, 5)*
EstPropNonHispWhite	*Estimated proportion of persons who are non-Hispanic white [block group], binned as quartiles (pandas qcut) (1, 2, 3, 4)*
EstPropHouseholdNonHispWhite	*Estimated proportion of households that are non-Hispanic white [block group], binned as quartiles (pandas qcut) (1, 2, 3, 4)*
EstPropHighSchoolMaxEducation	*Estimated proportion persons aged 25 y or older with a HS diploma or less at their highest level of schooling [block group], binned as quartiles (pandas qcut) (1, 2, 3, 4)*
EstPropHighSchoolDropout	*Estimated proportion persons aged 16-19 y who are neither attending school nor HS graduates [block group], binned as quartiles (pandas qcut) (1, 2, 3, 4)*
EstPropHighSchoolDropoutNoWork	*Estimated proportion persons aged 16-19 y who are neither attending school nor HS graduates and are without work [block group], binned as quartiles (pandas qcut) (1, 2, 3, 4)*
EstPropHouseholdsNoAuto	*Estimated proportion of households without an automobile [block group], binned as quartiles (pandas qcut) (1, 2, 3, 4)*

EstPropHouseholdsNoHealthIns **2012-2016 sample only*	*Estimated proportion of persons without health insurance [block group], binned as quartiles (pandas qcut) (1, 2, 3, 4)*
EstProp5PlusESL	*Estimated proportion of persons 5 years or older who sometimes speak a language other than English at home [block group], binned as quartiles (pandas qcut) (1, 2, 3, 4)*
EstMedianHouseholdIncome	*Estimated median household income [block group], binned as quintiles (pandas qcut) (1, 2, 3, 4, 5)*
EstPropMaleLittleWork	*Estimated proportion of males aged 16-64 years who worked less than 26 weeks in previous year [block group], binned as quintiles (pandas qcut) (1, 2, 3, 4, 5)*
EstPropHouseholdSSI	*Estimated proportion of households receiving Supplemental Security Income block group], binned as quintiles (pandas qcut) (1, 2, 3, 4, 5)*
EstPropHouseholdPA	*Estimated proportion of households receiving Public Assistance [block group], binned as quintiles (pandas qcut) (1, 2, 3, 4, 5)*
EstPropFemaleHouseholdNoSpouse	*Estimated proportion of family households headed by a female (no male partner present) [block group], binned as quintiles (pandas qcut) (1, 2, 3, 4, 5)*
EstPropFemaleHouseholdFamilyChild	*Estimated proportion of total households headed by a female with family children aged 18 y or less (no male partner present) [block group], binned as quintiles (pandas qcut) (1, 2, 3, 4, 5)*
EstPropFemaleHouseholdAnyChild	*Estimated proportion of total households headed by a female with any children aged 18 y or less (no male partner present) [block group], binned as quintiles (1, 2, 3, 4, 5)*
DeGauss Social Deprivation Index (2018)	
SocialDeprivationIndex	*DeGauss composite metric derived from six 2018 ACS variables: fraction_assisted_income, fraction_high_school_edu, median_income,*

	fraction_no_health_ins, fraction_poverty, fraction_vacant_housing *Not yet integrated into ICEES*
US Census* Topologically Integrated Geographic Encoding and Referencing System (TIGERline) *Roadway Data (year 2016)	
MajorRoadwayHighwayExposure	*Distance in meters from household to nearest major road/highway (1 = 0-49, 2 = 50-99, 3 = 100-199, 4 = 200-299, 5 = 300-499, 6 = >=500 meters)*
MajorRoadwayHighwayExposure2	*Distance in meters from household to nearest major road/highway (1 = 0-49, 2 = 50-99, 3 = 100-149, 4 = 150-199, 5 = 200-249, 6 = >=250 meters)*
US Department of Transportation, Federal Highway Administration, Highway Patrol Monitoring System Roadway Data (year 2016)	
RoadwayDistanceExposure	*Distance in meters from household to nearest roadway (1 = 0-49, 2 = 50-99, 3 = 100-199, 4 = 200-299, 5 = 300-499, 6 = >=500 meters)*
RoadwayDistanceExposure2	*Distance in meters from household to nearest roadway (1 = 0-49, 2 = 50-99, 3 = 100-149, 4 = 150-199, 5 = 200-249, 6 = >=250 meters)*
RoadwayType	*UNC DOT roadway classification (e.g., major highway)*
RoadwayAADT	*US DOT Annual average daily traffic estimate*
RoadwaySpeedLimit	*US DOT Roadway speed limit*
RoadwayLanes	*UNC DOT Roadway number of lanes*
NC Department of Environmental Quality Data	
CAFO_Distance	*NC Department of Environmental Quality distance in meters from household to nearest concentrated animal farming operation (1 = <500, 2 = 500-1000, 3 = 1000-2000, 4 = 2000-4000, 5 = >4000)* *Includes data on:* ● *Swines* ● *Cattle*

	• *Poulty*
LandfillDistance	*NC Department of Environmental Quality distance in meters from household to nearest landfill (1 = <500, 2 = 500-1000, 3 = 1000-2000, 4 = 2000-4000, 5 = >4000)* *Includes data on:* • *Unregulated landfills (pre-1983)* • *Regulated landfills* • *Superfund sites*
National Center for Education Statistics Public School Data (2018)	
PublicSchoolDistance	*Distance from nearest public school* *Data have not yet been integrated into ICEES*
US Environmental Protection Agency Community Multiscale Air Quality (CMAQ) model exposure estimates: PM2.5, ozone	
CMAQ exposure estimates: PM2.5, ozone (years 2010 & 2011) [§]	

Variable nomenclature format:	
AvgDaily**XX**Exposure_StudyAvg	*UNC IE average of estimated average daily PM2.5 exposure over 'study' period, binned by data values (pandas cut) (1, 2, 3, 4, 5)*
MaxDaily**XX**Exposure_StudyAvg	*UNC IE average of estimated maximum daily exposure over 'study' period, binned by data values (pandas cut) (1, 2, 3, 4, 5)*
AvgDaily**XX**Exposure_StudyMax	*UNC IE maximum of estimated average daily PM2.5 exposure over 'study' period, binned by data values (pandas cut) (1, 2, 3, 4, 5)*
MaxDaily**XX**Exposure_StudyMax	*UNC IE maximum of estimated maximum daily exposure over 'study' period, binned by data values (pandas cut) (1, 2, 3, 4, 5)*
AvgDaily**XX**Exposure_StudyAvg_qcut	*UNC IE average of estimated average daily PM2.5 exposure over 'study' period, binned by quintiles (pandas qcut) (1, 2, 3, 4, 5)*
MaxDaily**XX**Exposure_StudyAvg_qcut	*UNC IE average of estimated maximum daily exposure over 'study' period, binned by quintiles (pandas qcut) (1, 2, 3, 4, 5)*
AvgDaily**XX**Exposure_StudyMax_qcut	*UNC IE maximum of estimated average daily PM2.5 exposure over 'study' period, binned by quintiles (pandas qcut) (1, 2, 3, 4, 5)*
MaxDaily**XX**Exposure_StudyMax_qcut	*UNC IE maximum of estimated maximum daily exposure over 'study' period, binned by quintiles (pandas qcut) (1, 2, 3, 4, 5)*
US Environmental Protection Agency continental USA (conUS) CMAQ model exposure estimates: PM2.5, ozone (years 2002-2016), CO, NO, NO2, NOx, SO2, acetaldehyde, formaldehyde, benzene (2002-2016, estimates are not consistently available for all years)*	

Variable nomenclature format: AvgDaily**XX**Exposure_2 MaxDaily**XX**Exposure_2 AvgDaily**XX**Exposure_2_qcut MaxDaily**XX**Exposure_2_qcut	 *US EPA conUS CMAQ daily exposure estimates for PM2.5, CO, NO, NO2, NOx, SO2, acetaldehyde, formaldehyde, or benzene, averaged over 'study' period, binned by data values (pandas cut) (1,2,3,4,5)* *US EPA conUS CMAQ 8-hour ozone maximum exposure estimate, averaged over 'study' period, binned by data values (pandas cut) (1, 2, 3, 4, 5)* *US EPA conUS CMAQ daily exposure estimates for PM2.5, CO, NO, NO2, NOx, SO2, acetaldehyde, formaldehyde, or benzene, averaged over 'study' period, binned by quintiles (pandas.qcut) (1,2,3,4,5)* *US EPA conUS CMAQ 8-hour PM2.5 exposure estimate, averaged over 'study' period, binned by quintiles (pandas.qcut) (1,2,3,4,5)*
US EPA National Air Quality System Pollutant Data (2021-2022)	
US EPA National Emissions Inventory Data (2017)	
Variables not yet named	*Estimates of air emissions of criteria pollutants, criteria precursors, and hazardous air pollutants from air emissions sources (nonpoint wagon wheel (HAP, chromium, PM); CAP and lead emissions for*

	wildland (wild and prescribed) fires; commercial marine vehicle emission estimates; airport-related and aircraft emission estimates; locomotive and railyard emission estimates) *Not yet integrated into ICEES*
US EPA Unregulated Contaminant Monitoring Rule (UCMR) Public Water Supply Data, UCMR3 (2015-2015) & ICMR4 (2018-2020)	
Variables not yet names	*Data for contaminants suspected to be present in public water supplies, but that do not have regulatory standards set under the Safe Drinking Water Act (SDWA), e.g., PFAS contaminants* *Not yet integrated into ICEES*

Abbreviations: PM2.5 = particulate matter ≤ 2.5 μm in diameter

*The feature variables listed in the table are those for the patient-level tables, which include data on each patient for each year of available data (i.e., data on individual patients are represented as rows in the table). Similar feature variables are available for the visit-level tables, although the variables are sometimes treated differently. For example, $PM_{2.5}$ exposures in visit-level tables are expressed in relation to the 24-hour and two-week period before visits, not in relation to the one-year 'study' period, as was done for the patient-level tables. Additional feature variables (e.g., laboratory measures) are available for select years. Further information can be accessed at https://github.com/NCATSTranslator/Translator-All/wiki/Exposures-Provider-ICEES.

§From first batch of CMAQ output, derived from UNC Institute for the Environment, hourly estimates, 36-km (2010) or 12-km (2011) resolutions

**From second batch of CMAQ output, derived from US EPA continental USA (conUS); US Census tract resolution, 2002-2015 for PM2.5 and ozone, 2002 only for other chemicals