JCI The Journal of Clinical Investigation

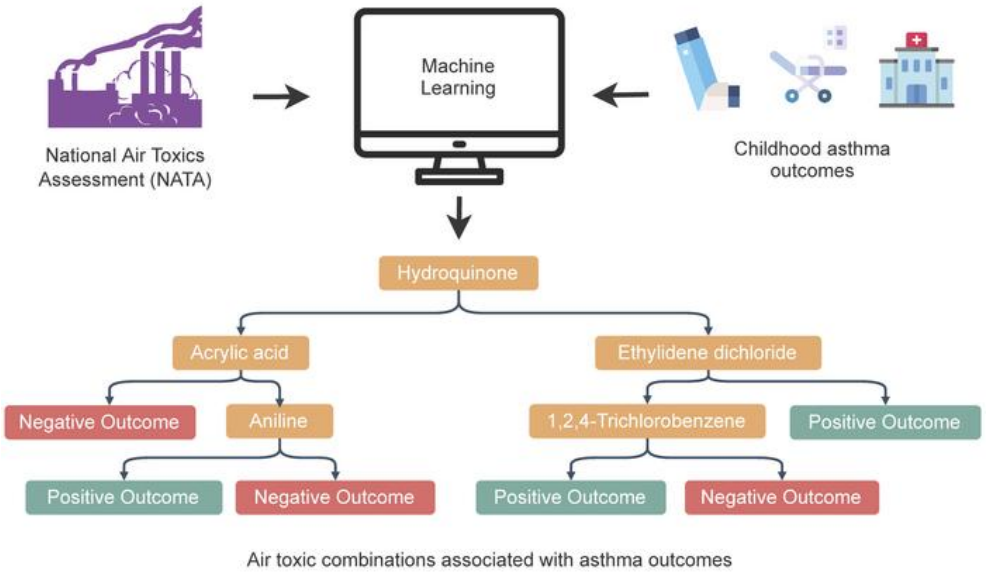
Machine learning-driven identification of early-life air toxic combinations associated with childhood asthma outcomes

Yan-Chak Li, … , Gaurav Pandey, Supinda Bunyavanich

J Clin Invest. 2021. https://doi.org/10.1172/JCI152088.

Research In-Press Preview **Pulmonology**

Graphical abstract



Air toxic combinations associated with asthma outcomes

Title: Machine learning-driven identification of early-life air toxic combinations associated with childhood asthma outcomes

Authors: Yan-Chak Li[1$], Hsiao-Hsien Leon Hsu[2,3$], Yoojin Chun[1], Po-hsiang Chiu[1], Zoe Arditi[1,4], Luz Claudio[2,3], Gaurav Pandey[1,3]*, Supinda Bunyavanich[1,4]*

Affiliations:

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai and Icahn Institute for Data Science and Genomic Technology, New York, NY, USA

[2]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[3]Institute for Exposomic Research, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[4]Division of Allergy and Immunology, Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[$]Contributed equally to this work

*Co-corresponding Authors

19 **Correspondence:** Supinda Bunyavanich, One Gustave Levy Place, Box #1498, New York, NY

20 10029, tel. +1-212-241-5548, supinda@post.harvard.edu; Gaurav Pandey, One Gustave Levy

21 Place, Box #1498, New York, NY 10029, tel. +1-212-659-8535, gaurav.pandey@mssm.edu

Abstract:

Air pollution is a well-known contributor to asthma. Air toxics are hazardous air pollutants that cause or may cause serious health effects. While individual air toxics have been associated with asthma, only a limited number of studies have specifically examined combinations of air toxics associated with the disease. We geocoded air toxic levels from the US National Air Toxics Assessment (NATA) to residential locations for participants of our AiRway in Asthma (ARIA) study. We then applied **D**ata-driven **E**xposur**E** **P**rofile extraction (DEEP), a novel machine learning-based method, to discover combinations of early-life air toxics associated with current use of daily asthma controller medication, lifetime emergency department visit for asthma, and lifetime overnight hospitalization for asthma. We discovered 20 multi-air toxic combinations and 18 single air toxics associated with at least one outcome. The multi-air toxic combinations included those containing acrylic acid, ethylidene dichloride, and hydroquinone, and they were significantly associated with asthma outcomes with odds ratios of 1.60 to 3.19. Several air toxic members of the combinations would not have been identified by single air toxic analyses, supporting the use of machine learning-based methods designed to detect combinatorial effects. Our findings provide knowledge about air toxic combinations associated with childhood asthma.

Introduction

Air toxics are hazardous air pollutants that cause or may cause serious health effects (1). They are well-established detriments to human respiratory health, especially for children (2-8). In particular, exposure to air toxics early in life predisposes children to asthma, one of the most prevalent diseases in this demographic group. Epidemiologic studies have linked prenatal and early life exposure to air toxics with childhood wheeze, asthma, and altered lung function (6-14).

Although air toxics are generally analyzed and regulated as individual chemicals (6), we are exposed to combinations of air toxics in ambient air. The specific combinations of individual air toxics that influence childhood asthma have not been studied adequately. Assessing the respiratory health effects of multiple air toxics is challenging for several reasons (7, 15). First, it is logistically difficult and expensive to collect detailed individualized exposure data for multiple air toxics using personal or local monitoring. Additionally, there are limited statistical methods to parse the effects of mixtures where individual air toxics may only contribute slightly to an adverse outcome, but have a different impact in combination with other air toxics (15). As a result, few studies have considered simultaneous exposure to air toxic mixtures and their associations with children's health, including asthma (8, 15-17).

Several studies linking air toxic mixtures and health outcomes, as well as a prior review of 57 studies that examined air pollutants and their health effects, reached no consensus on the ideal methods for multi-pollutant analyses (6, 7, 15, 16, 18). A key limitation of the studies reviewed was that most metrics assumed pure additivity of the effects of multiple air toxics, without consideration of synergistic and/or antagonistic interactions. Due to these challenges, air toxic combinations that collectively influence childhood asthma remain suboptimally characterized. Furthermore, identifying air toxic combinations associated with health outcomes is also difficult due to the exponentially large number of combination subsets in a set of air toxics, i.e., 2^N-1 combinations in a set of N air toxics. Conventional statistical methods (19-24) and feature importance assessment using machine learning algorithms (16, 25-27) have not been effective for this task since they generally assess the association of air toxics individually.

In this study, we hypothesized that exposure to combinations of air toxics during early life is associated with asthma outcomes in later childhood. These outcomes included current need for daily asthma controller medication, lifetime emergency room visit for asthma, and lifetime overnight hospitalization for asthma (**Figure 1**). Asthma-related medication use, emergency room visits, and hospitalizations are frequently studied asthma outcomes that reflect asthma severity, control, and healthcare utilization (28-30). While some studies have reported associations between particular air toxics and these asthma subphenotypes (9, 11,

12), none addressed our goal to identify combinations of air toxics from a large national assessment of air toxics associated with these asthma outcomes. We tested our hypothesis by geocoding levels of 125 air toxic from the US Environmental Protection Agency's (EPA) National Air Toxic Assessment (NATA) (31), one of the richest sources of multi-air toxic profiling across the US, to the residential addresses of children with asthma from our Airway in Asthma (ARIA) study (32) to map each child's exposure to air toxics during the first years of life. We addressed the challenges of combinatorial air toxic analysis by applying a machine learning-based algorithm called **D**ata-driven **E**xposur**E** **P**rofile extraction (DEEP), which, to the best of our knowledge, is a novel method for this problem. DEEP uses the high-performing eXtreme Gradient Boosting (XGBoost) (33) algorithm to identify air toxic combinations associated with health outcomes. The combinations identified using XGBoost were then adjusted for potential confounders, including age, gender, race/ethnicity, and family income, to identify early-life multi-air toxic combinations, statistical interactions within combinations, and demographic profiles associated with adverse asthma outcomes in later childhood. Our approach identified several combinations of air toxics associated with asthma.

Results

Characteristics of the study cohort

Table 1 shows the characteristics of the Airway in Asthma (ARIA) study (32) participants with asthma examined in this study. These 151 children with mild to severe persistent asthma were recruited from the Mount Sinai Health System, New York, NY with informed consent from their parent/guardian via an IRB-approved protocol. Participants had a mean age of 12 years (standard deviation 3.2 years) at the time of assessment, and were of diverse self-identified racial/ethnic backgrounds (**Table 1**). Their asthma was generally not well-controlled, with a mean score on the Asthma Control Test (ACT)(34) of 16.8 (maximum value 25 representing optimal control), and 96% of the cohort reporting regular use of a short-acting beta agonist rescue inhaler.

Children who used daily asthma controller medication (n=84, 56%) were younger than those who did not (n=65; p=0.048). Inhaled corticosteroids (ICS) were used most frequently, both independently and in combination with long-acting beta agonist (LABA). Children who had at least one lifetime emergency room visit for asthma (n=103, 68%) were more likely to self-identify as Black or Latino, had significantly lower ACT scores than their counterparts who had never required an emergency department visit for asthma, and were more likely to be taking combination ICS/LABA as their daily asthma controller medication. Children who had been hospitalized overnight for asthma in their lifetime (n=51, 34%) had significantly lower forced expiratory volume in 1 second percent predicted (FEV1%) on spirometry, and

114 higher rates of ICS/LABA and leukotriene receptor antagonist for daily asthma treatment
115 compared to the participants with asthma who had never been hospitalized overnight for
116 asthma.
117
118 **Air toxic characteristics**
119 Ambient annual average concentrations for over a hundred toxics based on emissions
120 inventories and computer simulation models are publicly available for each US census tract
121 in the EPA's NATA database (31). We mapped the available toxic levels to the residential ZIP
122 code for each child in our cohort. Ninety-four zip codes spanning 443 square miles across
123 New York, New Jersey, and Connecticut were represented in this cohort. We used the
124 closest calendar year of NATA data available subsequent to a child's birth date. We retained
125 only the air toxics whose levels were available for all the participants in the mapped
126 datasets, yielding 125 air toxics for analysis.
127
128 ***DEEP-enabled identification of combinations of air toxics associated with childhood***
129 ***asthma***
130 We then applied DEEP to identify air toxic combinations associated with each of the
131 three childhood asthma outcomes, namely the need for daily asthma controller medication,
132 lifetime emergency room visit for asthma, and lifetime overnight hospitalization for asthma.

In the first analytical stage of DEEP (detailed in Methods), for each outcome, the full dataset was randomly split 100 times into training and test sets in an 80:20 ratio. For each split, an XGBoost model consisting of 100 decision trees was learned from the training set and evaluated on the test set in terms of the Area Under the ROC Curve (AUC score) (35).

In the second analytical stage of DEEP, we analysed the combinations of toxics from the XGBoost models, identified as root-to-leaf paths in the constituent decision trees, for each outcome. Note that in some cases, a combination may consist of only one air toxic if it is sufficient to predict the outcome under consideration for a subset of the cohort, thus giving DEEP flexibility in discovery. Also, in cases of multiple air toxics in these combinations, their sequence of appearance on the path also indicates their relative order of relevance to the outcome being predicted. This is because variables closer to the root of a decision tree have higher predictive power than those closer to the leaves.

Next, the frequency of each combination was calculated as the number of models (out of 100) where it was included in at least one of the constituent trees. Candidate combinations were then identified as those with a frequency of at least ten. These combinations were then used in multivariable regression models to test their association with the asthma outcome of interest, while adjusting for age, gender, race/ethnicity, and income.

689 profiles of air toxics across all the asthma outcomes were discovered after the first XGBoost stage of DEEP. These sets included both individual air toxics and their combinations. 359 of these sets were then found to be significantly associated ($P \leq 0.05$) with the respective outcome in the second stage of DEEP. After multiple hypothesis correction by the Benjamini-Hochberg procedure (36), 273 air toxic profiles were found to be significantly associated (FDR$\leq$0.05) with at least one of the three outcomes. Our goal was to identify air toxic combinations whose increased levels are associated with adverse asthma outcomes. Therefore, among the significantly associated combinations, we focused on groups that included air toxics with levels higher than threshold. Among these finally determined combinations, there were 18 with only one air toxic each (**Figure 2**), and 20 multi-air toxic combinations (**Figure 3**).

Air toxic combinations associated with asthma outcomes

Twenty multi-air toxic combinations and eighteen individual air toxics were found to be significantly associated with at least one of the three asthma outcomes. The medians and interquartile ranges of the exposure levels of the 34 air toxics included in these associations are shown in **Table 2**.

Higher levels of seventeen individual air toxics were significantly associated with worse asthma outcomes (**Figure 2**). Odds ratios (ORs) for these associations ranged from 1.56 to

2.65. Several of the identified toxics are established risk factors for childhood asthma, especially the chemicals previously categorized as halogenated, ketones and ethers (8, 37-39). Among these, the air toxics most associated with the outcomes were acrylic acid (OR=2.10), mercury compounds (OR=2.65), and ethyl chloride (OR=1.87) respectively. Acetamide, pentachlorophenol, and polychlorinated biphenyls were associated with more than one asthma outcome.

A major strength of DEEP is its ability to identify multi-air toxic combinations associated with health outcomes. Indeed, here DEEP revealed significant associations between higher exposure to 20 multi-air toxic combinations and the three asthma outcomes of interest (**Figure 3**). Among these, 19 combinations included two air toxics and one included three. The associations of these combinations were generally stronger than those of the individual air toxics, with ORs ranging from 1.60 to 3.19 (**Figure 3**).

Notably, acrylic acid was not only the individual air toxic most strongly associated with daily controller medication (**Figure 2**), it was the first (i.e. primary) member of 7 of the 9 multi-air toxic combinations associated with this outcome (**Figure 3**). Acrylic acid also appeared in 3 of the other 11 combinations associated with emergency room visit and overnight hospitalization for asthma (**Figure 3**), indicating that it is a major contributor to adverse asthma outcomes among children.

Three air toxic combinations were associated with lifetime emergency room visit for asthma, all with an OR of over 2 (**Figure 3**). Acetaldehyde, acrylamide, and acrylic acid were the primary exposures in these combinations, despite the fact they were not individually significantly associated with the outcome. Several other air toxics in these combinations, namely carbon disulfide and hydroquinone, were also not individually associated with this outcome. These findings highlight the main strength of DEEP, namely its ability to identify significant multi-air toxic combinations, whose constituent air toxics may not be individually associated with the health outcome of interest.

Among the eight air toxic combinations associated with lifetime overnight hospitalization for asthma, 1,4-dioxane, carbonyl sulfide, ethylidene dichloride, hydrochloric acid, and hydroquinone were the primary exposures (**Figure 3**). Both ethylidene dichloride and hydroquinone appeared in three of these eight combinations, indicating that these two chemicals may play a role in the development of poor asthma outcomes among children. Most other air toxics in these combinations (**Figure 3**) were largely not individually associated with this outcome (**Figure 2**), again supporting DEEP's ability to identify multi-air toxic combinations that may not be inferred from single air toxic associations.

Effect sizes of multi-air toxic combinations may not be evident from the individual associations of their members

Some air toxics had relatively low effect sizes when assessed individually (**Figure 2**) compared to the larger ORs from combination analyses (**Figure 3**). For example, acrylic acid was associated with daily controller medication with an OR of 2.10 as an individual air toxic (**Figure 2**), but the ORs of its combinations with dimethyl phthalate, 1,1,1-trichloroethane, ethyl chloride, acetophenone, and cobalt were higher (OR 2.16 to 3.19; **Figure 3**). Also, none of these five air toxics were individually associated with the outcome. Similarly, hexachlorobenzene was associated with daily controller medication with an OR of 2.03 (**Figure 2**), while simultaneous exposure to the combination of hexachlorobenzene and dimethyl phthalate identified by DEEP had an OR of 2.96 (**Figure 3**). This was despite the fact that there was no significant individual association between dimethyl phthalate and the outcome. For the pair of toluene and phosphorus, neither air toxic was individually associated with daily controller medication (**Figure 2**), but their combination was associated with the outcome with an OR of 1.81 (**Figure 3**).

Similar cases of combinatorial effects were also seen for lifetime emergency room visit for asthma. For example, simultaneous exposure to polychlorinated biphenyl, acetaldehyde and carbon disulfide had 3.10-fold odds of the outcome (**Figure 3**), while polychlorinated biphenyl's individual effect size was substantially lower (OR=1.72; **Figure 2**). Similarly, the combination of acrylic acid and hydroquinone was significantly associated with emergency

225 room visit with an OR of 2.73 (**Figure 3**), but neither was associated with the outcome
226 individually (**Figure 2**).

227 We observed similar results for multi-air toxic combinations and lifetime overnight
228 hospitalization for asthma. Exposure to hydroquinone was individually associated with this
229 outcome with an OR of 1.79 (**Figure 2**), but in combination with ethylidene dichloride, the
230 association was stronger (OR=2.03; **Figure 3**). Similarly, carbonyl sulfide was not individually
231 associated with this outcome (**Figure 2**), but it was the primary member in two of the
232 multi-air toxic combinations found to be associated with OH (**Figure 3**).

233 In summary, the above comparison of the effect sizes of the individual (**Figure 2**) and
234 multi- (**Figure 3**) air toxic associations demonstrated that combinations of air toxics had
235 effects that were not fully explained by simply adding together the individual effects from
236 their constituents. Overall, DEEP identified 34 air toxics associated with the asthma
237 outcomes (**Table 2**), including 16 air toxics with significant effects only as members of
238 combinations.

239

240 **Statistical Interactions among Members of Air Toxic Combinations**

241 To assess potential synergy between members of air toxic combinations associated
242 with asthma outcomes, we conducted statistical tests for interactions. Significant statistical
243 interactions detected between air toxic members within the combinations are shown in

244 **Table 3**. Acrylic acid was the primary air toxic (i.e. primary branch point in the decision tree)
245 of all the combinations with significant statistical interactions. While other combinations did
246 not reveal significant interactions, such interactions remain possible given the limitations of
247 statistical detection of interactions. Directed experimental work could be undertaken to test
248 for additional interactions.
249
250 ***Representative Air Toxic Combinations and Demographic Risk Factors***
251 Finally, one of the advantages of DEEP is that the trees constituting its underlying
252 XGBoost models can be visualized and interpreted, which is difficult to do for several other
253 machine learning methods. However, since it is difficult to simultaneously depict all the
254 trees inferred by DEEP, we visualized sample trees that contained the most strongly
255 associated multi-air toxic combination for each childhood asthma outcome. Sample decision
256 trees inferred by DEEP for each of the outcomes are shown in **Figures 4**, **5**, and **6**
257 respectively. To provide an additional level of interpretation, we also compared the
258 demographic characteristics (age, sex, race/ethnicity, and family income) of children
259 exposed to each of these combinations to those of children who were not exposed in **Tables**
260 **4, 5, and 6**. Differences could suggest demographic risk factors that may increase a child's
261 exposure to these multi-air toxic combinations.

Acrylic acid and cobalt compounds was the air toxic combination associated with daily controller medication use with the highest odds ratio of 3.19 (**Figure 3**, **Figure 4**). Children exposed to this combination were older compared to those who were not (p=0.02; **Table 4**).

Acetaldehyde, carbon disulphide, and polychlorinated biphenyls was the air toxic combination most strongly associated with lifetime emergency room visit for asthma (OR=3.10; **Figure 3, Figure 5**). Children exposed to this combination were younger ($p=5.34x10^{-8}$; **Table 5**) and had lower family income than those who were not exposed (p=0.019; **Table 5**). Exposed children were also less likely to be White (p=0.0046; **Table 5**). These observations point to social disparities among these groups of children.

The most strongly associated combination for overnight hospitalization was hydroquinone and ethylidene dichloride (OR=2.03; **Figure 3**, **Figure 6**). Children exposed to this combination were younger (p=0.00218; **Table 6**) and had lower family incomes ($p=8.26x10^{-5}$; **Table 6**) than those who were not exposed.

Discussion

Our application of a novel machine learning-driven algorithm called DEEP to a cohort of mild to severe asthmatic children identified several individual and combinations of air toxics, to which increased exposure during early-life was associated with adverse asthma outcomes in later childhood. In particular, due to a unique ability of DEEP to examine air toxic

281 combinations, we identified 16 air toxics that were only found to be significantly associated
282 with childhood asthma outcomes in combination with other air toxics.

283 Many air toxics in the identified combinations, such as carbonyl sulfide, carbon
284 disulfide, ethyl chloride, and ethylidene chloride, are similar in structure and have analogous
285 formation, production, chemical fate, and chemical transport properties (40). Ten air toxics
286 in the combinations contained chlorine, three included heavy metal compounds, and many
287 were acidic chemicals. This aligns with prior literature implicating acidic chemicals,
288 chlorinated chemicals, and heavy metal compounds as risk factors for asthma and asthma
289 severity (8, 41-44). However, an understanding of the biological mechanisms through which
290 these combinations of air toxics can jointly affect respiratory health and asthma merits
291 further study.

292 Among the air toxics individually associated with asthma outcomes (**Figure 2**),
293 triethylamine was associated with increased overnight hospitalizations for asthma.
294 Triethylamine is a clear, colorless liquid used in waterproofing and as a catalyst, corrosion
295 inhibitor, and propellant (45). It is a respiratory irritant, to which chronic exposure even at
296 low levels can inhibit the function of organic cationic transporters, thus preventing efficient
297 uptake of inhaled bronchodilators used to control acute asthma symptoms (46, 47).

298 Acrylic acid was individually associated with daily controller medication (**Figure 2**) and
299 appeared as a member of at least one combination associated with all three outcomes

300 (**Figure 3**). Furthermore, it was found to interact with other member air toxics of three
301 combinations (**Table 3**)**.** Acrylic acid is used in the manufacture of adhesives, elastomers,
302 plastics, coatings, as well as floor paints and polishers (48). Literature has suggested that the
303 presence of water-soluble cobalt complexes increases the conversion of polyacrylic acid into
304 acrylic acid, which is more biologically viable. Acrylic acid also reacts with cobalt complexes
305 to produce organocobalt complexes (49). Additionally, hydroquinone acts as a stabilizer to
306 prevent the polymerization of acrylic acid, which keeps the latter in a form with a lower
307 molecular weight that is more biologically viable (50, 51). Our results, including evidence of
308 statistical interactions between acrylic acid and other chemical compounds, suggest further
309 investigation of mechanisms for acrylic acid's associations with adverse childhood asthma
310 outcomes.

311 Ethyl chloride, also known as chloroethane (C_2H_5Cl), and ethylidene dichloride ($C_2H_4Cl_2$)
312 are both chlorinated hydrocarbons. Ethyl chloride is used as a thickening agent and binder in
313 paints and cosmetics, and as a refrigerant, aerosol spray propellant, anesthetic, and blowing
314 agent for foam packaging (52). We found ethyl chloride to be associated with asthma
315 outcomes, both as an individual air toxic (overnight hospitalizations; **Figure 2**) and as a
316 member of multi-air toxic combinations (daily controller medication; **Figure 3**). Ethylidene
317 dichloride, which is mainly used as a solvent for plastics, oils and fats, and as a degreaser
318 and fumigant in insecticide sprays (53), appeared as a member of several combinations

associated with lifetime overnight hospitalization for asthma (**Figure 3**). Both these compounds are well-known members of the chloroethanes family, which is comprised of liposoluble chemicals that can be taken up by the lipoprotein within the alveolar film layer (AFL) (54). AFL disruption is observed in multiple pulmonary diseases, including acute respiratory distress syndrome, infant respiratory distress syndrome, emphysema, chronic obstructive pulmonary disease, asthma, chronic bronchitis, pneumonia, pulmonary infections, and idiopathic pulmonary fibrosis (55). Thus, chronic exposure to ethyl chloride and ethylidene dichloride may lead to dysfunction in the AFL, which may contribute to worse asthma control.

Hydroquinone, a commonly studied air toxic, was also identified in our analyses. While exposure to higher levels of hydroquinone alone was not associated with overnight hospitalization for asthma, DEEP identified it as a member of several multi-air toxic combinations associated with this outcome (**Figure 3**). Hydroquinone is commonly found in the indoor environment, and exposure to it has been associated with airway hypersensitivity (56-59). Hydroquinone is widely seen in cosmetic and health products, including skin creams (60). It is thought to prevent the polymerization of acrylic acid, methyl methacrylate, cyanoacrylate, and other monomers that are susceptible to radical-initiated polymerization, thus allowing them to persist in their original form (50, 51). This suggests a mechanism through which the identified synergistic combination of hydroquinone with

338 acrylic acid (emergency room visit; **Figure 3**) is associated with adverse asthma outcomes.
339 Although our analysis did not find a statistically significant interaction between
340 hydroquinone and ethylidene dichloride, potentially due to lower exposure levels,
341 hydroquinone is industrially added to shelf ethylidene dichloride as a stabilizer (61). This
342 suggests that in the presence of hydroquinone, ethylidene dichloride is less likely to react
343 with other chemicals in the environment, and thus retain its toxic form longer (similar to
344 acrylic acid). Thus, it is still possible that hydroquinone and ethylidene dichloride may act
345 synergistically, but this needs to be investigated in future studies.

346 While our study has advanced the identification of air toxic combinations associated
347 with childhood asthma outcomes, it also has limitations. We used the NATA national model
348 (31) to estimate exposures rather than personal sampling or local monitors. Collecting
349 personal or locally monitored measures for 125 air toxics at each cohort participant's
350 residence would be logistically and financially challenging. Given this, NATA is commonly
351 used for estimating ambient exposures, since it is a well-validated deterministic dispersion
352 chemical transportation model created by the EPA that accounts for sources included in the
353 NATA emission inventory (31). NATA estimates of a given air toxic may under-report a
354 personally or locally monitored value, since the latter may include emissions from indoor
355 and undocumented sources not in the EPA's inventory. For instance, one study found that
356 higher personally monitored benzene concentrations relative to NATA-predicted values

were likely due, at least in part, to indoor sources not included in the EPA's inventory (62). Other studies also found discrepancies between NATA estimates and monitored chemical concentrations, due again, in part, to local or indoor sources (63, 64).

Despite the above limitations, using NATA as the primary source of exposure estimates has several strengths over locally monitored values. First, NATA has a finer geographical prediction resolution and spread than currently available monitoring sites (31). This enabled us to include participants in our study that may not have had a monitoring site close to their residence. Second, NATA data are generated from an advanced chemical transportation model that aggregates exposure over a long period, and thus is able to capture transient exposures. Also, several factors potentially affecting air toxic estimates, such as seasonality, ambient temperature, meteorology, precipitation, and solar radiation, have already been incorporated into NATA's model (31). This level of comprehensive modelling is typically not available from personal or local monitoring. Finally, local measurements may also have detection and quantification limits, while NATA is able to estimate air toxics even at lower levels and over a longer period of time. Due to these strengths, we chose to study NATA data in this work.

We recognize that our results do not provide evidence of a causal effect of any chemical on adverse childhood asthma outcomes, which will need further investigation. Additionally, our geocoding was based on a single zip code for each participant so would not

account for potentially dynamic exposures due to residential moves. Last, our study participants were all from the New York, New Jersey, and Connecticut tri-state area. Thus, our findings may not generalize to other US regions or parts of the world. Future studies could examine combinations from other geographical regions and/or utilize direct air sampling to confirm the combinations identified in this study.

In conclusion, this study demonstrated innovative use of data science methods and data sources to identify specific combinations of early-life air toxic exposures associated with later childhood asthma outcomes. Our study suggests that chemical pollutants should be closely monitored together in combination, especially for locations with vulnerable populations.

Materials and Methods

An overview of the study approach is shown in **Figure 1**.

Study Population

The study population included children with asthma from the AiRway In Asthma (ARIA) study, a cohort recruited from the Mount Sinai Health System, New York, NY (32). The study was approved by the Mount Sinai Institutional Review Board. Children with asthma had mild to severe persistent asthma according to National Asthma Education and Prevention

Program (NAEPP)/National Heart Lung Blood Institute (NHLBI) Expert Panel Report 3 (EPR3) criteria (65) and positive bronchodilator response on spirometry or methacholine challenge with provocative challenge causing a 20% fall in forced expiratory volume (PC20) < 12.5 mg/ml. Phenotyping for all participants included detailed questionnaires about asthma-related symptoms, medication and healthcare use, and pre and post-bronchodilator spirometry following American Thoracic Society guidelines (65).

Asthma Outcomes

We focused on the following three self-reported asthma-related outcomes: (1) current use of prescribed daily asthma controller medication (daily controller medication), (2) at least one emergency department visit for asthma during subject's lifetime (emergency room visit), and (3) at least one overnight hospitalization for asthma during subject's lifetime (overnight hospitalization).

Air Toxic Exposures

The air toxic exposure profile of each participant was derived from EPA's National Air Toxics Assessment (NATA) (31). NATA estimates the annual ambient concentrations of over a hundred air toxics at each census tract in the United States based on emissions inventories and advanced computer simulation models (31). Seasonality of air toxics, ambient

414 temperature, meteorology, precipitation and solar radiation are incorporated into NATA's
415 model (31).
416 NATA data are available for 1996, 1999, 2002, 2005, 2011, and 2014 (27), and children
417 in this study were born between 1997 and 2012. To assign the most representative ambient
418 air toxic exposure levels to each participant, we mapped the residential ZIP code of each
419 child in our cohort to the geometric centroid of the closest census tract (31). We then used
420 the annual exposure data for that tract from the NATA release closest in time following the
421 child's birth year. This choice of year closest to birth was based on prior evidence that
422 early-life exposure to air pollution is associated with childhood asthma outcomes (66-68).
423 Finally, we retained the 125 air toxics that had data available, i.e., no missing data, for all
424 participants in the final dataset (15, 69).
425
426 ***Covariates***
427 We included age, sex, race/ethnicity, and family income as covariates in multivariable
428 regression models based on considerations that these variables could confound associations
429 between air toxic levels and asthma outcomes. Since the questionnaire completed by ARIA
430 participants did not include queries about family income, we used the average income of
431 each participant's residential ZIP code obtained from US Census Business Patterns data (70)
432 as a surrogate for this variable.

Data-driven ExposurE Profile extraction (DEEP)

To identify multi-air toxic exposure profiles associated with asthma outcomes, we developed a data-driven method called Data-driven ExposurE Profile extraction (DEEP) (**Figure 1**). DEEP is inspired by a simpler method that we previously used to identify multi-air toxic combinations associated with children's cognitive skills (15).

In the first stage of DEEP, exposure combinations are identified using eXtreme Gradient Boosting (XGBoost) (33), an algorithm that uses an ensemble method to iteratively learn decision trees and generally performs well at prediction tasks. XGBoost generally yields strong predictive power (71, 72) due to its use of multiple optimization methods, including regularization and gradient boosting, which reduces overfitting of models to training data. Specifically, the full exposure dataset was randomly split 100 times into training and test sets in an 80:20 ratio. For each split, an XGBoost model consisting of 100 decision trees was learned from the training set to predict the outcome under consideration. This model was then applied to and evaluated on the corresponding test set in terms of the Area Under the ROC curve (AUC score(35)). The overall predictability of the target outcome was evaluated in terms of the average value of the AUC scores across the 100 training/test splits.

The decision trees constituting each XGBoost model contain internal decision nodes, edges, and leaf nodes to represent how the value of an outcome could be predicted based

452 on air toxic levels. **Figures 4-6** show several trees derived in the current work. Each decision
453 node in these trees contains an air toxic and a threshold value for its level. It is also
454 connected by two edges representing the decisions made depending on whether an
455 individual's exposure was higher or lower than the threshold. Each of these edges is
456 connected to either the next decision node or a leaf node. A leaf node determines the value
457 of the outcome for the individual with the exposure profile represented by the decision path
458 taken to reach it. Each decision and leaf node also represents a subpopulation of the cohort
459 exposed to the air toxics on the path taken to reach it. Candidate multi-air toxic
460 combinations are then defined as the air toxics and thresholds in the decision nodes
461 constituting the paths from the root of a tree to the leaf nodes. We calculated the frequency
462 of each combination as the number of XGBoost models (out of 100) where it was included in
463 at least one of the constituent trees, and set of threshold of 10 to identify the most relevant
464 combinations. Note that, if two or more variables are highly correlated, and thus similarly
465 associated with the outcome, a key characteristic of the decision trees in the XGBoost
466 model is that they will include only one of these variables as an internal decision node. Thus,
467 unlike traditional regression models, XGBoost is not as adversely affected by collinearity
468 among the input variables. Furthermore, DEEP executes XGBoost 100 times on randomly
469 selected training sets, different selections of these variables may be included in the different
470 trees inferred, thus enhancing the coverage of the air toxic profiles.

In the second stage of DEEP, a multivariable linear regression model is built to assess the association of a candidate combination with the target outcome, adjusted for covariates. The asthma outcome is the dependent variable in this model, while the air toxic combination and covariates are its independent variables. The variable representing the combination takes a value of 1 for individuals exposed to it, determined using the threshold values of the constituent air toxics, and 0 otherwise. One model is built for each outcome and candidate combination, yielding the odds ratio (OR) denoting the strength of the association between the two. The p-values of all the associations are converting into false discovery rates (FDRs) after correcting for multiple hypothesis testing using the Benjamini-Hochberg method (36). In this study, significant associations were identified as those with FDR$\leq$0.05.

To assess potential synergy between members of air toxic combinations associated with asthma outcomes, we conducted statistical tests for interactions. Interactions between pairs of air toxics were assessed through additional multivariable regression models where the outcome was the dependent variable and predictors included the levels of the two toxics, their product as a representative of their interaction, and covariates. For combinations with two air toxics, this regression model was inferred from the whole cohort, while for combinations with three air toxics, analyses were conducted for the last two toxics on the sample meeting the threshold for the first toxic level in the combination. A significant

490 interaction was identified if the p-value of the interaction term in the model was lower than
491 0.05.

492 The DEEP framework is implemented in the Python programming language (71). The
493 XGBoost, model evaluation (AUC score calculation) and regression components are
494 implemented using the xgboost (33), scikit-learn (72) and statsmodels (73) Python packages
495 respectively.

496

497 **Study Approval**

498 The study was approved by the Mount Sinai Institutional Review Board, New York, USA.
499 Parents of participants provided written informed consent.

500

501 **Author contributions**

502 HLH, PC, GP and SB conceived the study. GP and SB supervised the work. YCL, HLH, PC, ZA,
503 and YC managed and analyzed the data. YCL and HLH drafted the manuscript. All the
504 authors reviewed, edited, and approved the manuscript. Order among co-first authors was
505 determined based on contribution to results generation.

506

507 **Acknowledgements**

508 This work was supported by a pilot grant from the Department of Genetics and Genomic
509 Sciences at Mount Sinai and NIH grants R01 HL147328, R01 AI118833, UG3 OD023337,
510 R01HG011407-01A1, and P30 ES23515. It was also supported in part through the
511 computational resources provided by Scientific Computing at the Icahn School of Medicine
512 at Mount Sinai. We thank Alfin Vicencio of the Mount Sinai Hospital System for his
513 assistance with cohort recruitment and Jeanette Stingone of Columbia University for her
514 technical advice.

515

516 **Conflicts of interest**

517 The authors have declared that no conflict of interest exists.

References

1. United States Environmental Protection Agency. About Urban Air Toxics https://www.epa.gov/urban-air-toxics/about-urban-air-toxics. Accessed 8/12/2021.
2. Walters GI, Robertson AS, Moore VC, and Burge PS. Occupational asthma caused by acrylic compounds from SHIELD surveillance (1989-2014). *Occup Med (Lond).* 2017;67(4):282-9.
3. North ML, Takaro TK, Diamond ML, and Ellis AK. Effects of phthalates on the development and expression of allergic disease and asthma. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology.* 2014;112(6):496-502.
4. Dumas O, Despreaux T, Perros F, Lau E, Andujar P, Humbert M, et al. Respiratory effects of trichloroethylene. *Respiratory medicine.* 2018;134:47-53.
5. Schenker MB, and Jacobs JA. Respiratory effects of organic solvent exposure. *Tubercle and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease.* 1996;77(1):4-18.
6. Nurmatov UB, Tagiyeva N, Semple S, Devereux G, and Sheikh A. Volatile organic compounds and risk of asthma and allergy: a systematic review. *European respiratory review : an official journal of the European Respiratory Society.* 2015;24(135):92-101.
7. Rumchev K, Spickett J, Bulsara M, Phillips M, and Stick S. Association of domestic exposure to volatile organic compounds with asthma in young children. *Thorax.* 2004;59(9):746-51.
8. Weisel CP. Assessing exposure to air toxics relative to asthma. *Environ Health Perspect.* 2002;110 Suppl 4:527-37.
9. Ye D, Klein M, Chang HH, Sarnat JA, Mulholland JA, Edgerton ES, et al. Estimating Acute Cardiorespiratory Effects of Ambient Volatile Organic Compounds. *Epidemiology.* 2017;28(2):197-206.
10. Rosa MJ, Jung KH, Perzanowski MS, Kelvin EA, Darling KW, Camann DE, et al. Prenatal exposure to polycyclic aromatic hydrocarbons, environmental tobacco smoke and asthma. *Respiratory medicine.* 2011;105(6):869-76.
11. Ran J, Kioumourtzoglou MA, Sun S, Han L, Zhao S, Zhu W, et al. Source-Specific Volatile Organic Compounds and Emergency Hospital Admissions for Cardiorespiratory Diseases. *International journal of environmental research and public health.* 2020;17(17).
12. Delfino RJ. Epidemiologic evidence for asthma and exposure to air toxics: linkages between occupational, indoor, and community air pollution research. *Environmental health perspectives.* 2002;110 Suppl 4:573-89.

13. Bably M, Arif AA, and Post A. Prenatal use of cleaning and scented products and its association with childhood asthma, asthma symptoms, and mental health and developmental comorbidities. *The Journal of asthma : official journal of the Association for the Care of Asthma.* 2021;58(1):46-51.
14. Adgent MA, Carroll KN, Hazlehurst MF, Loftus CT, Szpiro AA, Karr CJ, et al. A combined cohort analysis of prenatal exposure to phthalate mixtures and childhood asthma. *Environment international.* 2020;143:105970.
15. Stingone JA, Pandey OP, Claudio L, and Pandey G. Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among U.S. children. *Environ Pollut.* 2017;230:730-40.
16. Patel CJ. Analytic Complexity and Challenges in Identifying Mixtures of Exposures Associated with Phenotypes in the Exposome Era. *Curr Epidemiol Rep.* 2017;4(1):22-30.
17. Choi H, Schmidbauer N, Sundell J, Hasselgren M, Spengler J, and Bornehag CG. Common household chemicals and the allergy risks in pre-school age children. *PloS one.* 2010;5(10):e13423.
18. Oakes M, Baxter L, and Long TC. Evaluating the application of multipollutant exposure metrics in air pollution health studies. *Environ Int.* 2014;69:90-9.
19. Khreis H, Kelly C, Tate J, Parslow R, Lucas K, and Nieuwenhuijsen M. Exposure to traffic-related air pollution and risk of development of childhood asthma: A systematic review and meta-analysis. *Environ Int.* 2017;100:1-31.
20. Esposito S, Tenconi R, Lelii M, Preti V, Nazzari E, Consolo S, et al. Possible molecular mechanisms linking air pollution and asthma in children. *BMC Pulm Med.* 2014;14:31.
21. Orellano P, Quaranta N, Reynoso J, Balbi B, and Vasquez J. Effect of outdoor air pollution on asthma exacerbations in children and adults: Systematic review and multilevel meta-analysis. *PLoS One.* 2017;12(3):e0174050.
22. Edwards SC, Jedrychowski W, Butscher M, Camann D, Kieltyka A, Mroz E, et al. Prenatal exposure to airborne polycyclic aromatic hydrocarbons and children's intelligence at 5 years of age in a prospective cohort study in Poland. *Environ Health Perspect.* 2010;118(9):1326-31.
23. Gehring U, Wijga AH, Hoek G, Bellander T, Berdel D, Bruske I, et al. Exposure to air pollution and development of asthma and rhinoconjunctivitis throughout childhood and adolescence: a population-based birth cohort study. *Lancet Respir Med.* 2015;3(12):933-42.
24. Chang HH, Pan A, Lary DJ, Waller LA, Zhang L, Brackin BT, et al. Time-series analysis of satellite-derived fine particulate matter pollution and asthma morbidity in Jackson, MS. *Environ Monit Assess.* 2019;191(Suppl 2):280.

25. Brokamp C, Jandarov R, Rao MB, LeMasters G, and Ryan P. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmos Environ (1994).* 2017;151:1-11.
26. Coker E, Liverani S, Ghosh JK, Jerrett M, Beckerman B, Li A, et al. Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County. *Environ Int.* 2016;91:1-13.
27. Deng H, Urman R, Gilliland FD, and Eckel SP. Understanding the importance of key risk factors in predicting chronic bronchitic symptoms using a machine learning approach. *BMC Med Res Methodol.* 2019;19(1):70.
28. Centers for Disease Control and Prevention. Measures to Identify and Track Racial Disparities in Childhood Asthma: Prevalence and Outcome Measures. July 14, 2016. Available at https://www.cdc.gov/asthma/asthma_disparities/outcome_measures.htm#anchor_1532268838154, accessed 8/10/2021.
29. Gliklich RE, Castro M, Leavy MB, Press VG, Barochia A, Carroll CL, et al. Harmonized outcome measures for use in asthma patient registries and clinical practice. *The Journal of allergy and clinical immunology.* 2019;144(3):671-81 e1.
30. Covar RA, Fuhlbrigge AL, Williams P, and Kelly HW. The Childhood Asthma Management Program (CAMP): Contributions to the Understanding of Therapy and the Natural History of Childhood Asthma. *Current respiratory care reports.* 2012;1(4):243-50.
31. United States Environmental Protection Agency. National-Scale Air Toxics Assessment. Washington, DC: U.S. Environmental Protection Agency. 2014. Available at https://wwwepagov/national-air-toxics-assessment, accessed 12/10/2020.
32. Do AN, Chun Y, Grishina G, Grishin A, Rogers AJ, Raby BA, et al. Network study of nasal transcriptome profiles reveals master regulator genes of asthma. *J Allergy Clin Immunol.* 2021;147(3):879-93.
33. Tianqi Chen CG. XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Association for Computing Machinery, New York, NY, USA.* 2016:785–94. .
34. Nathan RA, Sorkness CA, Kosinski M, Schatz M, Li JT, Marcus P, et al. Development of the asthma control test: a survey for assessing asthma control. *The Journal of allergy and clinical immunology.* 2004;113(1):59-65.
35. Lever J, Krzywinski M, and Altman N. Classification evaluation. *Nature Methods.* 2016;13(8):603-4.

36. Benjamini Y, and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological).* 1995;57(1):289-300.
37. Chin JY, Godwin C, Parker E, Robins T, Lewis T, Harbin P, et al. Levels and sources of volatile organic compounds in homes of children with asthma. *Indoor air.* 2014;24(4):403-15.
38. Gahleitner F, Guallar-Hoyas C, Beardsmore CS, Pandya HC, and Thomas CP. Metabolomics pilot study to identify volatile organic compound markers of childhood asthma in exhaled breath. *Bioanalysis.* 2013;5(18):2239-47.
39. Delfino RJ, Gong H, Linn WS, Hu Y, and Pellizzari ED. Respiratory symptoms and peak expiratory flow in children with asthma in relation to volatile organic compounds in exhaled breath and ambient air. *J Expo Anal Environ Epidemiol.* 2003;13(5):348-63.
40. Ferm RJ. The Chemistry Of Carbonyl Sulfide. *Chemical Reviews.* 1957;57(4):621-40.
41. Moghtaderi M, Ashraf MA, Moghtaderi T, Teshnizi SH, and Nabavizadeh SH. Heavy metal concentration in classroom dust samples and its relationship with childhood asthma: a study from Islamic Republic of Iran. *Eastern Mediterranean health journal = La revue de sante de la Mediterranee orientale = al-Majallah al-sihhiyah li-sharq al-mutawassit.* 2020;26(5):594-601.
42. Carder M, Seed MJ, Money A, Agius RM, and van Tongeren M. Occupational and work-related respiratory disease attributed to cleaning products. *Occup Environ Med.* 2019;76(8):530-6.
43. Hsieh CY, Jung CR, Lin CY, and Hwang BF. Combined exposure to heavy metals in PM2.5 and pediatric asthma. *J Allergy Clin Immunol.* 2020.
44. Andersson M, Backman H, Nordberg G, Hagenbjork A, Hedman L, Eriksson K, et al. Early life swimming pool exposure and asthma onset in children - a case-control study. *Environ Health.* 2018;17(1):34.
45. New Jersey Department of Health. Hazardous Substance Fact Sheet: Triethylamine. 2010. Available at https://nj.gov/health/eoh/rtkweb/documents/fs/1907.pdf, accessed 8/16/2021.
46. Mukherjee M, Cingolani E, Pritchard DI, and Bosquillon C. Enhanced expression of Organic Cation Transporters in bronchial epithelial cell layers following insults associated with asthma - Impact on salbutamol transport. *Eur J Pharm Sci.* 2017;106:62-70.
47. Laborde-Casterot H, Rosenberg N, Dupont P, and Garnier R. Is the incidence of aliphatic amine-induced occupational rhinitis and asthma underestimated? *Am J Ind Med.* 2014;57(12):1303-10.

48. Ohara T, Sato T, Shimizu N, Prescher G, Schwind H, Weiberg O, et al. *Ullmann's Encyclopedia of Industrial Chemistry*. 2020:1-21.
49. Peng C-H, Fryd M, and Wayland BB. Organocobalt Mediated Radical Polymerization of Acrylic Acid in Water. *Macromolecules.* 2007;40(19):6814-9.
50. Wang X, Schmidt F, Hanaor D, Kamm PH, Li S, and Gurlo A. Additive Manufacturing of Ceramics from Preceramic Polymers: A Versatile Stereolithographic Approach Assisted by Thiol-Ene Click Chemistry. *Additive Manufacturing.* 2019;27.
51. Pozdeeva NN, and Denisov ET. Mechanism of hydroquinone-inhibited oxidation of acrylic acid and methyl methacrylate.
52. Rossberg M, Lendle W, Pfleiderer G, Tögel A, Dreher E-L, Langer E, et al. *Ullmann's Encyclopedia of Industrial Chemistry*. 2006.
53. Agency for Toxic Substances and Disease Registry. Toxicological Profile for 1,1-Dichloroethane. U.S. Department of Health and Human Services, Atlanta, GA.; 1990.
54. Fiserova-Bergerova V, Tichy M, and Di Carlo FJ. Effects of biosolubility on pulmonary uptake and disposition of gases and vapors of lipophilic chemicals. *Drug Metab Rev.* 1984;15(5-6):1033-70.
55. Holder JW. Physical and physicochemical factors effecting transport of chlorohydrocarbon gases from lung alveolar air to blood as measured by the causation of narcosis. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev.* 2012;30(1):42-80.
56. Becher R, Hongslo JK, Jantunen MJ, and Dybing E. Environmental chemicals relevant for respiratory hypersensitivity: the indoor environment. *Toxicol Lett.* 1996;86(2-3):155-62.
57. Sivertsen B, and Clench-Aas J. Exposure to environmental chemicals relevant for respiratory hypersensitivity: European aspects. *Toxicol Lett.* 1996;86(2-3):143-53.
58. Schwela D. Exposure to environmental chemicals relevant for respiratory hypersensitivity: global aspects. *Toxicol Lett.* 1996;86(2-3):131-42.
59. Triggiani M, Loffredo S, Granata F, Staiano RI, and Marone G. Modulation of mast cell and basophil functions by benzene metabolites. *Curr Pharm Des.* 2011;17(34):3830-5.
60. Olumide YM, Akinkugbe AO, Altraide D, Mohammed T, Ahamefule N, Ayanlowo S, et al. Complications of chronic use of skin lightening cosmetics. *International Journal of Dermatology.* 2008;47(4):344-53.
61. World Health Organization, United Nations Environment Programme on Chemical Safety, International Labour Organisation. Hydroquinone: Health and Safety Guide. Geneva: World Health Organization, 1996.

62. George B, Schultz B, Palma T, Vette A, Whitaker D, and Williams RW. An evaluation of EPA s National-Scale Air Toxics Assessment (NATA): Comparison with benzene measurements in Detroit, Michigan. *Atmospheric Environment.* 2011;45:3301-8.
63. Xue Z and Jia C. A Model-to-Monitor Evaluation of 2011 National-Scale Air Toxics Assessment (NATA). *Toxics.* 2019;7(1).
64. Garcia E, Hurley S, Nelson DO, Gunier RB, Hertz A, and Reynolds P. Evaluation of the agreement between modeled and monitored ambient hazardous air pollutants in California. *International journal of environmental health research.* 2014;24(4):363-77.
65. United States Department of Health and Human Services National Institute of Health, National Heart Lung and Blood Institute. National Asthma Education and Prevention Program Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma. NIH Publication Number 08-5846. National Heart Lung and Blood Institute (US), ed. Clinical Practice Guidelines. Bethesda (MD). 2007.
66. Bettiol A, Gelain E, Milanesio E, Asta F, and Rusconi F. The first 1000 days of life: traffic-related air pollution and development of wheezing and asthma in childhood. A systematic review of birth cohort studies. *Environmental health : a global access science source.* 2021;20(1):46.
67. Gehring U, Wijga AH, Koppelman GH, Vonk JM, Smit HA, and Brunekreef B. Air pollution and the development of asthma from birth until young adulthood. *Eur Respir J.* 2020;56(1).
68. Hsu HH, Chiu YH, Coull BA, Kloog I, Schwartz J, Lee A, et al. Prenatal Particulate Air Pollution and Asthma Onset in Urban Children. Identifying Sensitive Windows and Sex Differences. *Am J Respir Crit Care Med.* 2015;192(9):1052-9.
69. Weitekamp Chelsea A, Lein M, Strum M, Morris M, Palma T, Smith D, et al. An Examination of National Cancer Risk Based on Monitored Hazardous Air Pollutants. *Environmental health perspectives.* 2021;129(3):037008.
70. United States Census Bureau. United States Census 2016. ZIP Codes Business Patterns 2016 Data. Available at https://www2.census.gov/programs-surveys/cbp/datasets/2016/: Accessed 5/28/2020.
71. Oliphant TE. Python for Scientific Computing. *Computing in Science & Engineering.* 2007;9(3):10-20.
72. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011;12:2825-30.
73. Seabold S, and Perktold J. Proceedings of the 9th Python in Science Conference. *Austin, TX.* 2010:61.

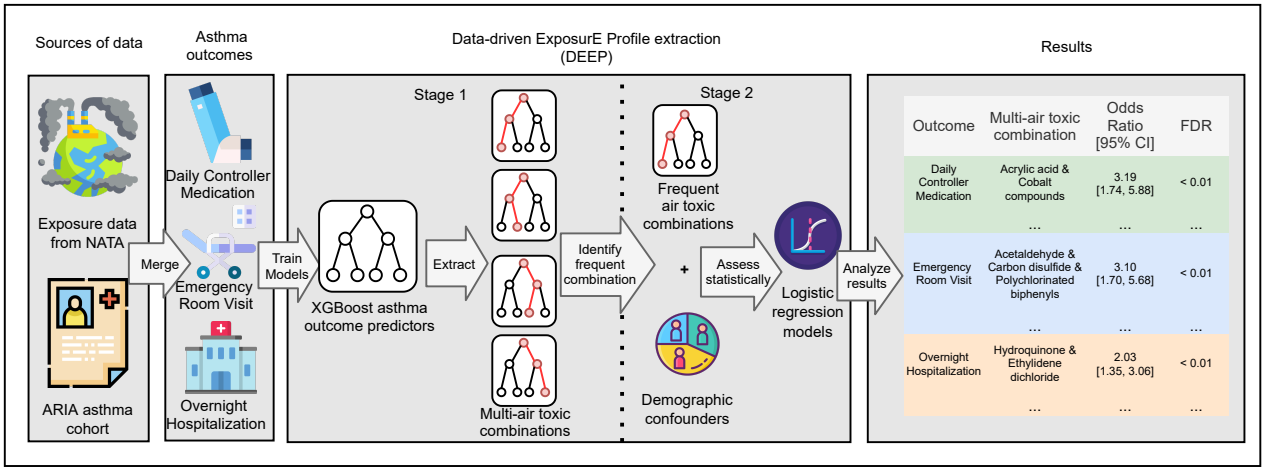


Figure 1: **Study overview.** Exposure data for over a hundred air toxics from the US Environmental Protection Agency's National Air Toxic Assessment (NATA) database were geocoded to Airway in Asthma (ARIA) cohort participants with mild to severe persistent asthma (n=151), based on participants' residential ZIP code. The Data-driven ExposurE Profile extraction (DEEP) method developed in this study was then applied to the air toxic data to identify multi-air toxic combinations associated with three childhood asthma outcomes: use of prescribed daily asthma controller medication, lifetime emergency department visit for asthma, and lifetime overnight hospitalization for asthma. In the first stage of DEEP, multi-air toxic combinations were identified via XGBoost models consisting of decision trees. In the second stage, multivariable logistic regression models were used to identify air toxic combinations significantly associated with childhood asthma outcomes after adjustment for age, gender, race/ethnicity, and family income. (Some images in this figure were obtained from www.flaticon.com and were made by Wanicon, Freepik and Flat Icons.)

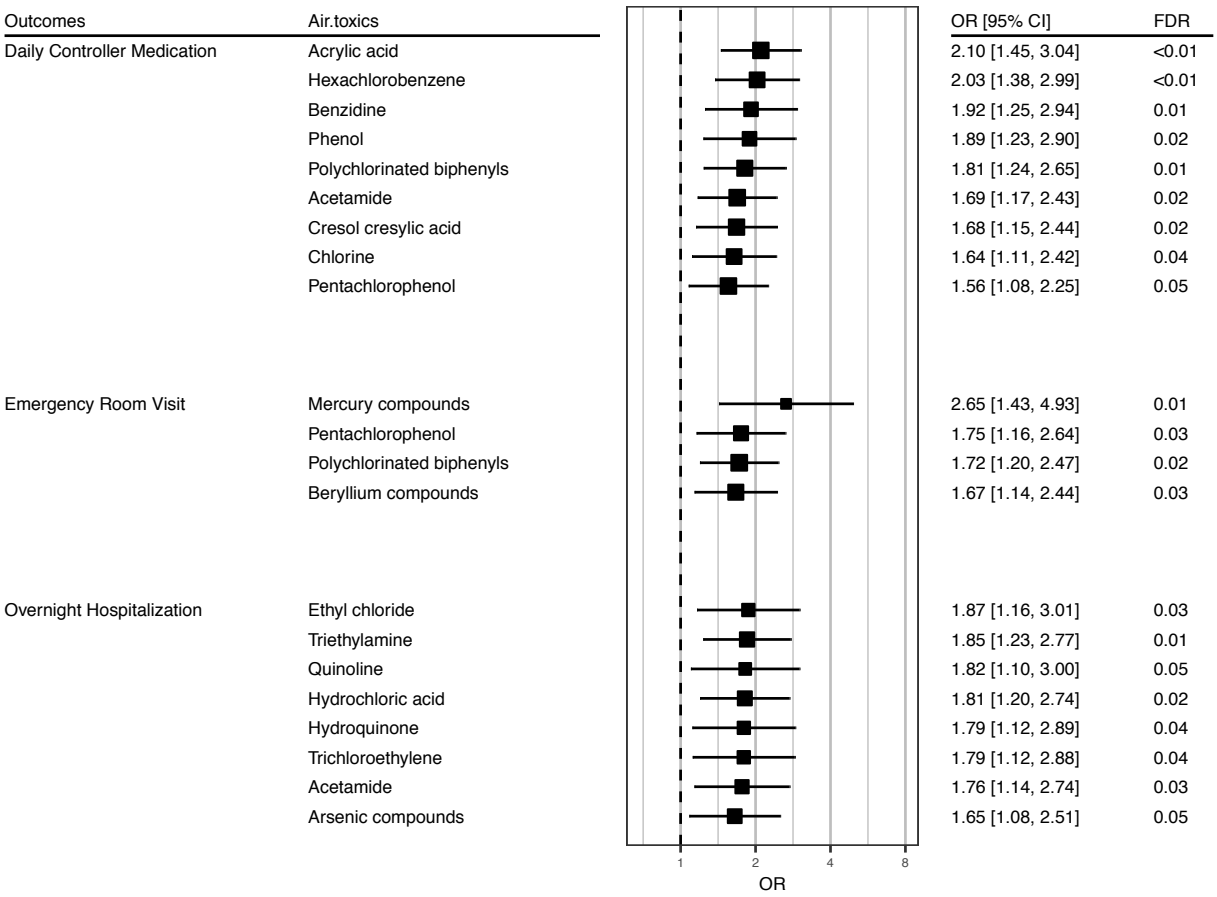


Figure 2: **Air toxics individually associated with childhood asthma outcomes after adjustment for age, gender, race/ethnicity, and family income in ARIA cohort participants with persistent asthma (n=151).** For each outcome and air toxic, the strength of the association is shown in terms of its odds ratio (OR), 95% confidence interval (CI), and false discovery rate (FDR). P-values for individual air toxics were obtained from multivariable logistic regression models and then adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure, yielding FDR values.

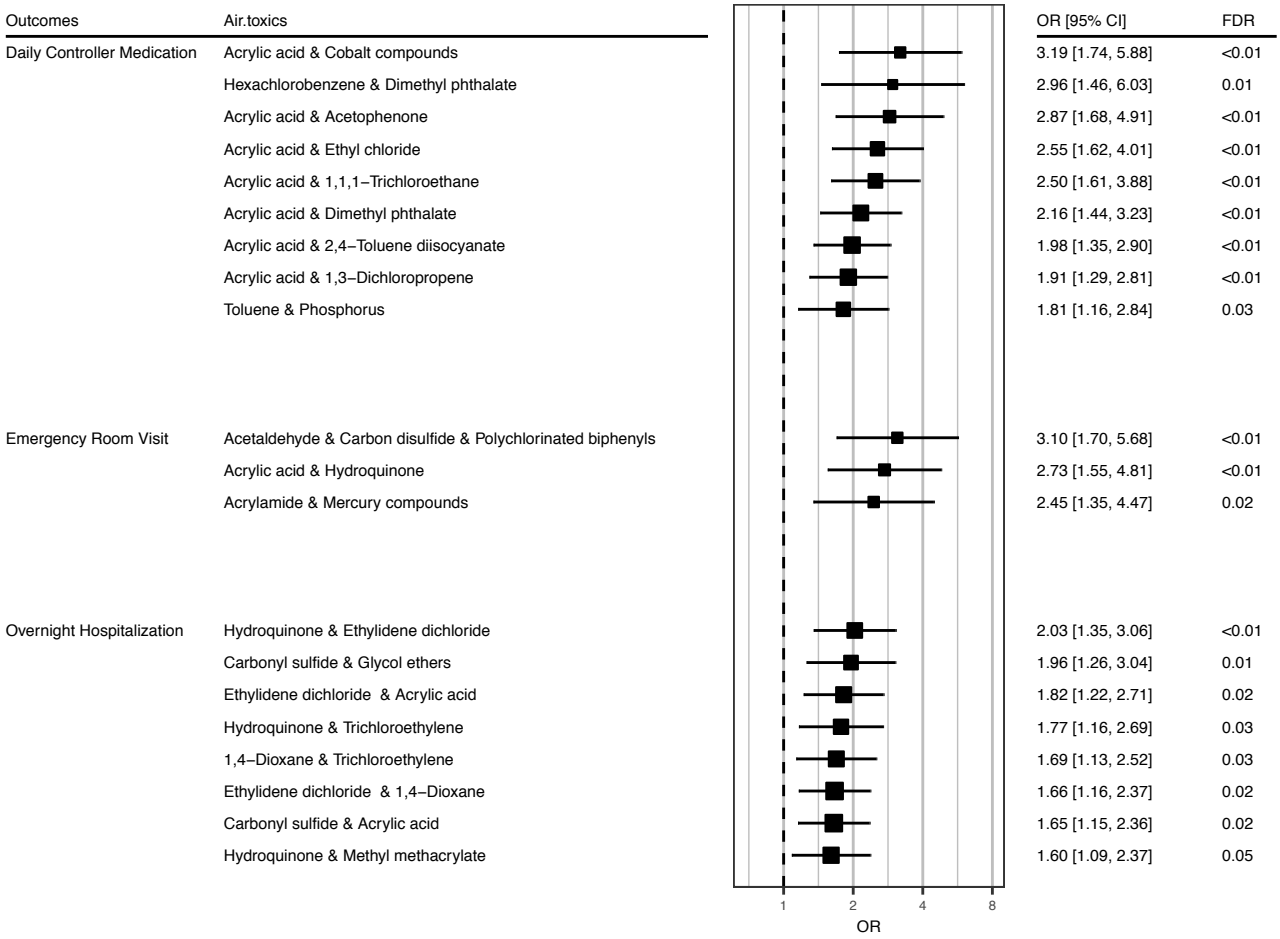


Figure 3: **Multi-air toxic combinations associated with childhood asthma outcomes after adjustment for age, gender, race/ethnicity, and family income in ARIA cohort participants with persistent asthma (n=151).** For each outcome and combination, the strength of the association is shown in terms of its odds ratio (OR), 95% confidence interval (CI), and false discovery rate (FDR). The P-values for multi-air toxic combinations were obtained from multivariable logistic regression models and then adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure, yielding FDR values.

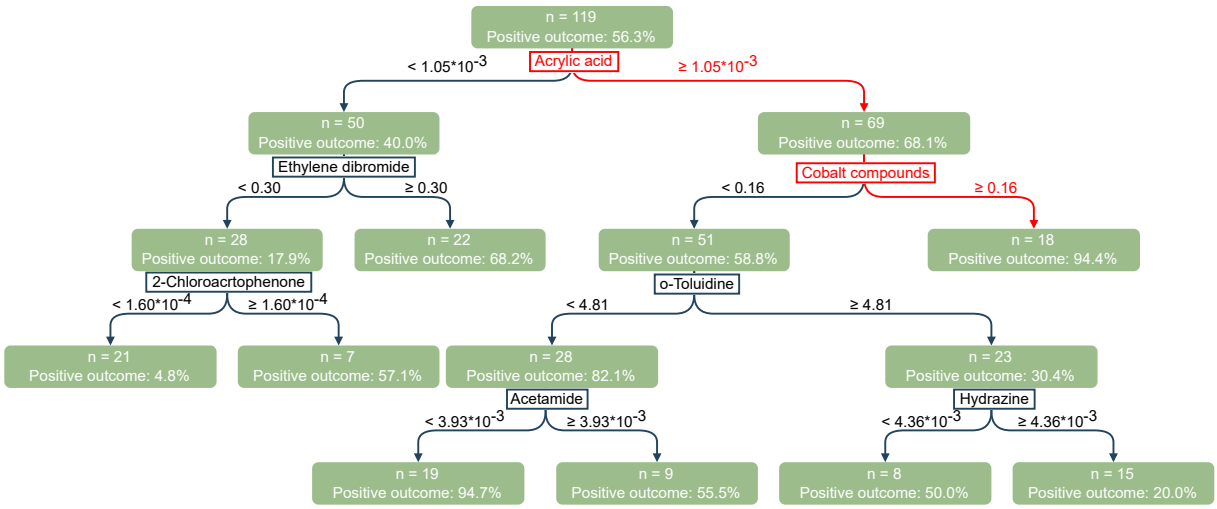


Figure 4: **A sample decision tree learned by DEEP to predict daily asthma controller medication using NATA-derived air toxic data geocoded to subjects (n=149).** Each node in the tree indicates the number of participants satisfying the air toxic decision path until that point and the percentage of participants with that outcome. The sample corresponding to each node is stratified into two subpopulations based on the air toxic and its threshold associated with the node. The multi-air toxic combination acrylic acid & cobalt compounds, which was the most significantly associated with this outcome, is highlighted in red.

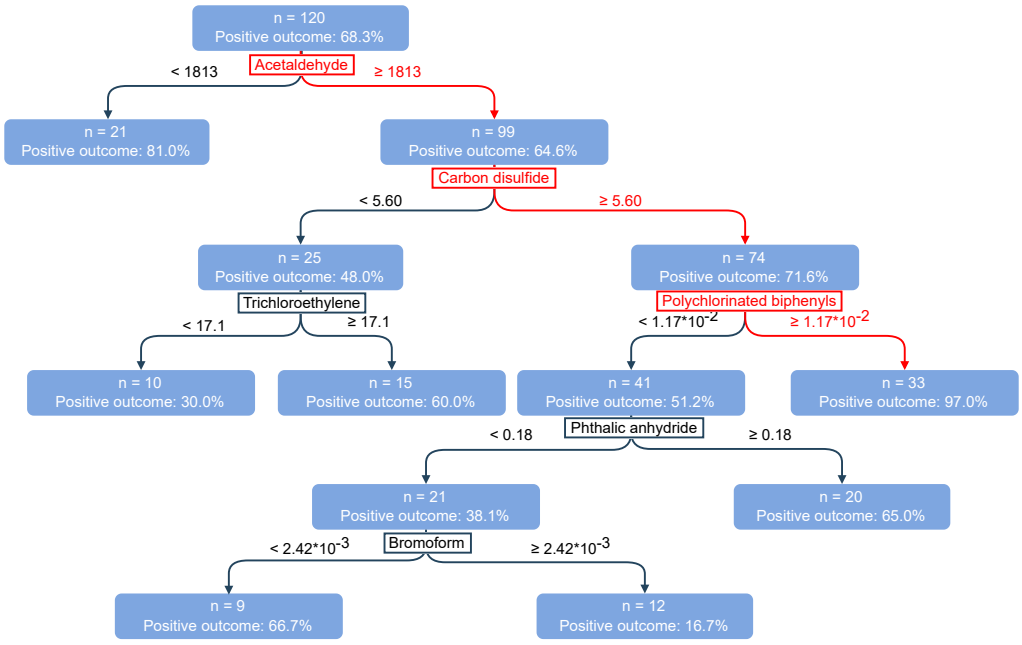


Figure 5: A sample decision tree learned by DEEP to predict lifetime emergency room visit for asthma from NATA-derived air toxic exposure data geocoded to each subject (n=151). Each node in the tree indicates the number of participants satisfying the air toxic decision path until that point and the percentage of participants with that outcome. The sample corresponding to each node is stratified into two subpopulations based on the air toxic and its threshold associated with the node. The multi-air toxic combination acetaldehyde & carbon disulfide & polychlorinated biphenyls, which was the most significantly associated with this outcome, is highlighted in red.

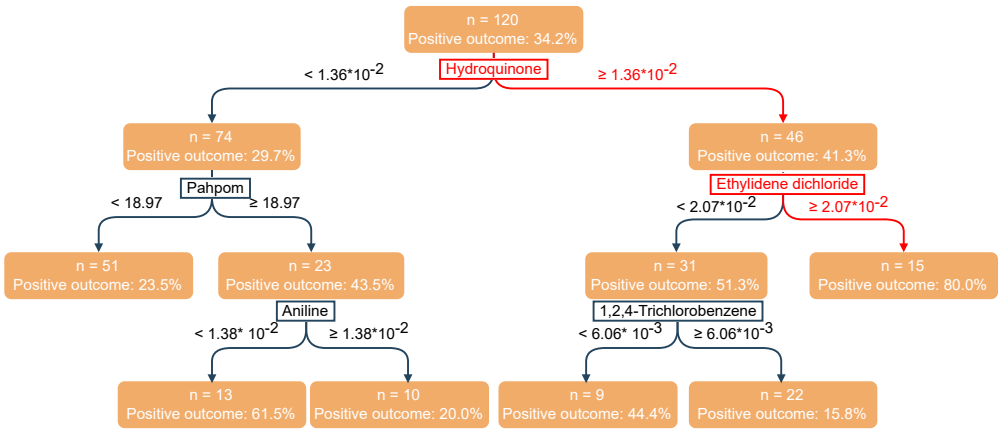


Figure 6: A sample decision tree learned by DEEP to predict lifetime overnight hospitalization for asthma from NATA-derived air toxic data geocoded to each participant (n=151). Each node in the tree indicates the number of participants satisfying the air toxic decision path until that point and the percentage of subjects with that outcome. The sample corresponding to each node is stratified into two subpopulations based on the air toxic and its threshold associated with the node. The multi-air toxic combination hydroquinone & ethylidene dichloride, which was the most significantly associated with this outcome, is highlighted in red.

Table 1: **Characteristics of the ARIA cohort participants included in this study.** Mean (standard deviation) and number (%) are shown for continuous and categorical variables respectively. Characteristics stratified by asthma outcome are also shown. P-values comparing subjects with and without the outcome are shown from two-sided student t-test for continuous variables and a chi-squared test for categorical variables. *Two participants did not provide information about their daily controller medication use.

	All	Daily Controller Medication*			Emergency Room Visit			Overnight Hospitalization		
	(n=151)	Yes (n=84)	No (n=65)	p-value	Yes (n=103)	No (n=48)	p-value	Yes (n=51)	No (n=100)	p-value
Age, years	12.0 (3.2)	11.5 (3.2)	12.6 (3.1)	0.048	11.7 (3.3)	12.4 (2.9)	0.23	12.1 (3.2)	11.9 (3.2)	0.69
Gender, female	62 (41.1%)	33 (39.3%)	29 (44.6%)	0.63	41 (39.8%)	21 (43.8%)	0.78	19 (37.3%)	43 (43%)	0.61
Race/ethnicity				0.50			0.03			0.07
Asian	6 (4.0%)	4 (4.8%)	2 (3.1%)		3 (2.9%)	3 (6.3%)		1 (2.0%)	5 (5.0%)	
Black	26 (17.2%)	13 (15.5%)	13 (20%)		20 (19.4%)	6 (12.5%)		12 (23.5%)	14 (14.0%)	
White	57 (37.7%)	33 (39.3%)	24 (36.9%)		32 (31.1%)	25 (52.1%)		12 (23.5%)	45 (45.0%)	
Latino	51 (33.8%)	25 (29.8%)	24 (36.9%)		42 (40.8%)	9 (18.8%)		23 (45.1%)	28 (28.0%)	
Mixed	10 (6.6%)	8 (9.5)	2 (3.1%)		5 (4.9%)	5 (10.4%)		3 (5.9%)	7 (7.0%)	
Not Reported	1 (0.7%)	1 (1.2%)	0 (0%)		1 (1.0%)	0 (0.0%)		0 (0%)	1 (1.0%)	
Income, US$	49349.3 (18725.8)	48954.8 (18701.7)	49961.7 (19017.3)	0.75	48870.9 (19414.5)	50375.8 (17109.8)	0.63	45576.7 (14822.9)	51273.3 (20161.2)	0.053
Asthma Control Test (ACT) score	16.8 (3.9)	16.4 (3.9)	17.3 (4.0)	0.16	16.2 (4.0)	18.0 (3.5)	$9.54*10^{-3}$	15.9 (4.3)	17.3 (3.7)	0.05
FEV1%	87.1 (17.5)	85.5 (17.7)	88.6 (17.3)	0.32	86.2 (18.0)	88.6 (16.6)	0.45	82.1 (19.7)	89.4 (15.9)	0.04
FEV1/FVC	79.6 (10.4)	78.3 (10.5)	80.8 (10.2)	0.17	79.4 (10.9)	80.0 (9.5)	0.77	77.7 (10.5)	80.4 (10.3)	0.17
Regular use of asthma medicine										
Beta agonist	145 (96.0%)	83 (98.8%)	60 (92.3%)	0.11	102 (99.0%)	43 (89.6%)	0.02	51 (100%)	94 (94.0%)	0.18
Inhaled corticosteroid (ICS)	39 (25.8%)	35 (41.7%)	3 (4.6%)	$7.19*10^{-7}$	27 (26.2%)	12 (25.0%)	1	10 (19.6%)	29 (29.0%)	0.29
Combined ICS/LABA	31 (20.5%)	30 (35.7%)	0 (0%)	$2.16*10^{-7}$	28 (27.2%)	3 (6.3%)	$5.97*10^{-3}$	23 (45.1%)	8 (8.0%)	$2.98*10^{-7}$
Leukotriene receptor antagonist	30 (19.9%)	29 (34.5%)	1 (1.5%)	$1.81*10^{-6}$	23 (22.3%)	7 (14.6%)	0.37	19 (37.3%)	11 (11.0%)	$3.08*10^{-4}$
Omalizumab	2 (1.3%)	1 (1.2%)	1 (1.5%)	1	2 (1.9%)	0 (0.0%)	0.84	0 (0.0%)	2 (2.0%)	0.79

Table 2: **Air toxics identified by DEEP as significantly associated with at least one of the three asthma outcomes, either individually or in combination with other air toxics.** D: Air toxic associated with daily asthma controller medication. E: Air toxic associated with lifetime emergency room visit for asthma. O: Air toxic associated with lifetime overnight hospitalization for asthma.

Air toxic	Level at participant's residential ZIP code Median (IQR) values (in ng/m^3)	Individually significant	Significant in combination(s)
1,1,1-trichloroethane	667.7 (120.5, 1452.9)		D
1,3-dichloropropene	129 (1.75, 405.18)		D
1,4-dioxane	$3.16x10^{-3}$ ($1.3x10^{-3}$, 0.18)		O
2,4-toluene diisocyanate	0.08 ($7.2x10^{-3}$, 0.15)		D
Acetaldehyde	2420.6 (1967.0, 2768.3)		E
Acetamide	$2.61x10^{-4}$ ($1.7x10^{-4}$, $9.1x10^{-4}$)	D, O	
Acetophenone	0.1 (0.02, 0.17)		D
Acrylamide	$4.24x10^{-4}$ ($2.5x10^{-6}$, $6.6x10^{-3}$)		E
Acrylic acid	$1.27x10^{-3}$ ($4.98x10^{-4}$, 0.06)	D	D, E, O
Arsenic compounds	0.73 (0.49, 1.13)	O	
Benzidine	$1.44x10^{-5}$ ($1.2x10^{-5}$, $2.5x10^{-5}$)	D	
Beryllium compounds	0.16 (0.09, 0.27)	E	
Carbon disulfide	5.66 (5.59, 9.57)		E
Carbonyl sulfide	0.03 ($1.77x10^{-3}$, 0.25)		O
Chlorine	0.91 (0.39, 1.28)	D	
Cobalt compounds	0.15 (0.10, 0.19)		D
Cresol cresylic acid	20.05 (15.1, 24.9)	D	
Dimethyl phthalate	0.07 (0.04, 0.08)		D
Ethyl chloride	8.21 ($7.8x10^{-3}$, 20.26)	O	D
Ethylidene dichloride	0.12 (0.02, 0.22)		O
Glycol ethers	12.67 (2.58, 117.36)		O
Hexachlorobenzene	$1.2x10^{-6}$ ($2.47x10^{-7}$, $3.91x10^{-4}$)	D	D
Hydrochloric acid	208.8 (146.8, 301.7)	O	
Hydroquinone	$7.3x10^{-3}$ ($1.4x10^{-3}$, 0.03)	O	E,O
Methyl methacrylate	1.62 (1.06, 2.79)		O
Mercury compounds	0.4 (0.22, 1.87)	E	E
Pentachlorophenol	$4.6x10^{-7}$ (0, $1.4x10^{-6}$)	D, E	
Phenol	3.34 (2.03, 22.35)	D	
Phosphorus	0.21 (0.11, 0.31)		D
Polychlorinated biphenyls	$1.1x10^{-4}$ ($1.2x10^{-5}$, $8.8x10^{-3}$)	D,E	E
Quinoline	0.14 ($5.23x10^{-4}$, 0.17)	O	
Toluene	9392 (4984.58, 21249.3)		D
Trichloroethylene	106.4 (18.1, 132.6)	O	O
Triethylamine	2.81 (1.82, 5.77)	O	

Table 3: Air toxic combinations associated with asthma outcomes with statistically significant interactions between combination members. The p-values are for the interaction term of multivariable logistic regression models where asthma outcome was the dependent variable and independent variables included member air toxic levels, interaction term, and covariates (age, gender, race/ethnicity, family income).

Outcome	Air Toxic Combination	Interactions	P-value
Daily Controller Medication	Acrylic acid & Dimethyl phthalate	Acrylic acid & Dimethyl phthalate	0.02
	Acrylic acid & Cobalt compounds	Acrylic acid & Cobalt compounds	0.02
Emergency Room Visit	Acrylic acid & Hydroquinone	Acrylic acid & Hydroquinone	< 0.004

Table 4: Demographic characteristics of children exposed and not exposed to the acrylic acid & cobalt compounds combination, which was associated with daily asthma controller medication. Mean (standard deviation) and number (%) are shown for continuous and categorical variables, respectively. Also shown are p-values for the differences between the two groups of children, calculated using a two-sided two sample student t-test for continuous variables and a chi-square test for categorical variables.

	Multi-air toxic combination: Acrylic acid & Cobalt compounds			
Variable	**Total (n=149)**	**Children exposed to combination (n=23)**	**Children not exposed to combination (n=126)**	**p-value**
Age, years	12.0 (3.2)	13.7 (3.6)	11.7 (3.0)	0.02
Gender, female	87 (58.4)	14 (60.9)	73 (57.9)	0.97
Race/ethnicity				0.089
Asian	6 (4.0)	0 (0.0)	6 (4.8)	
Black	26 (17.4)	5 (21.7)	21 (16.7)	
White	57 (38.3)	6 (26.1)	51 (40.5)	
Latino	49 (32.9)	8 (34.8)	41 (32.5)	
Mixed	10 (6.7)	3 (13.0)	7 (5.6)	
Not Reported	1 (0.7)	1 (4.3)	0 (0.0)	
Income, US$	49394.0 (18846.6)	46339.2 (15671.1)	49951.7 (19318.1)	0.34

Table 5: Demographic characteristics of children exposed and not exposed to the acetaldehyde & carbon disulfide & polychlorinated biphenyls combination, which was associated with lifetime emergency room visit for asthma. Mean (standard deviation) and number (%) are shown for continuous and categorical variables, respectively. Also shown are p-values for the differences between the two groups of children, calculated using a two-sided student t-test for continuous variables and a chi-square test for categorical variables.

	Multi-air toxic combination: Acetaldehyde & Carbon disulfide & Polychlorinated biphenyls			
Variable	**Total (n=151)**	**Children exposed to combination (n=48)**	**Children not exposed to combination (n=103)**	**p-value**
Age, years	12.0 (3.2)	10.0 (2.6)	12.9 (3.0)	5.34×10^{-8}
Gender, female	89 (58.9)	28 (58.3)	61 (59.2)	0.94
Race/ethnicity				0.0046
Asian	6 (4.0)	3 (6.2)	3 (2.9)	
Black	26 (17.2)	14 (29.2)	12 (11.7)	
White	57 (37.7)	9 (18.8)	48 (46.6)	
Latino	51 (33.8)	19 (39.6)	32 (31.1)	
Mixed	10 (6.6)	2 (4.2)	8 (7.8)	
Not Reported	1 (0.7)	1 (2.1)	0 (0.0)	
Income, US$	49349.3 (18725.8)	44708.3 (14015.1)	51512.0 (20194.5)	0.019

Table 6: Demographic characteristics of children exposed and not exposed to the hydroquinone & ethylidene dichloride combination, which was associated with lifetime overnight hospitalization for asthma. Mean (standard deviation) and number (%) are shown for the continuous and categorical variables, respectively. Also shown are p-values for the differences between the two groups of children, calculated using a two-sided student t-test for continuous variables and a chi-squared test for categorical variables.

	Multi-air toxic combination: Hydroquinone & Ethylidene dichloride			
Variable	**Total (n=151)**	**Children exposed to combination (n=17)**	**Children not exposed to combination (n=134)**	**p-value**
Age, years	12.0 (3.2)	9.9 (2.5)	12.2 (3.2)	0.0022
Gender, female	89 (58.9)	12 (70.6)	77 (57.5)	0.44
Race/ethnicity				0.32
Asian	6 (4.0)	0 (0.0)	6 (4.5)	
Black	26 (17.2)	2 (11.8)	24 (17.9)	
White	57 (37.7)	4 (23.5)	53 (39.6)	
Latino	51 (33.8)	10 (58.8)	41 (30.6)	
Mixed	10 (6.6)	1 (5.9)	9 (6.7)	
Not Reported	1 (0.7)	0 (0.0)	1 (0.7)	
Income, US$	49349.3 (18725.8)	40226.8 (6929.4)	50506.6 (19420.4)	8.26×10^{-5}