Distribution Agreement

Identifying air pollution mixtures and investigating their associations with pediatric asthma in a time-series framework

By

Katherine Gass, MPH
Doctor of Philosophy
Epidemiology

Matthew Strickland, Ph.D.
Committee Chair

Howard Chang, Ph.D.
Committee Member

W. Dana Flanders, M.D., DSc.
Committee Member

Mitch Klein, Ph.D.
Committee Member

Stefanie Sarnat, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Identifying air pollution mixtures and investigating their associations with pediatric asthma in a time-series framework

By

Katherine Gass
M.P.H., Emory University, 2009
B.A., Oberlin College 2003

Advisor: Matthew Strickland, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Epidemiology

2014

Abstract

Identifying air pollution mixtures and investigating their associations with pediatric asthma in a time-series framework

By Katherine Gass

Every time we take a breath outside we inhale a mixture of different pollutants; however, examining the associations between these pollutant mixtures and health endpoints is challenging. This dissertation uses different methodological approaches to understand the associations between air pollution mixtures and emergency department visits for pediatric asthma. Time series with daily counts of emergency department visits for any diagnosis of asthma or wheeze among pediatric patients were obtained from hospitals in metropolitan Atlanta (1999-2010), Dallas (2006-09) and St. Louis (2001-07). Daily measurements of ambient concentrations of ozone, carbon monoxide, nitrogen dioxide, and particulate matter <2.5 μm in diameter ($PM_{2.5}$) were obtained from monitors in all three metropolitan areas. In addition, daily estimates of $PM_{2.5}$ source concentrations were made available for Atlanta using a Bayesian-based ensemble source apportionment technique.

A modified classification and regression tree algorithm was developed to enable the identification of multipollutant joint effects. This algorithm was then used to determine the multipollutant joint effects associated with pediatric asthma in Atlanta, as well as the common multipollutant joint effects identified in Atlanta, Dallas and St. Louis. These analyses found certain types of days, characterized by their multipollutant profiles, to be associated with a statistically significant increase in asthma emergency department visits. $PM_{2.5}$ appeared to be one of the pollutants driving the formation of these harmful day types and thus further analyses were conducted to determine the associations between pediatric asthma and $PM_{2.5}$ sources. A positive association was observed for the cumulative, seven-day effect a 1 μg increase in biomass burnings (rate ratio: 1.02, 95% confidence interval: 1.01, 1.03), diesel vehicle emissions (rate ratio: 1.05, 95% confidence interval: 1.01, 1.08), and gasoline vehicle emissions (rate ratio: 1.07, 95% confidence interval: 1.03, 1.11). These confidence intervals account for uncertainties in the source apportionment estimates using multiple imputation methods.

This dissertation makes methodological contributions to the field of epidemiology with the development of a classification and regression algorithm that is well-suited for identifying joint effects of exposure mixtures. It also adds to the growing body of literature which suggests a harmful effect of multipollutant exposures on pediatric asthma.

Identifying air pollution mixtures and investigating their associations with pediatric asthma in a time-series framework

By

Katherine Gass

M.P.H., Emory University, 2009

B.A., Oberlin College 2003

Advisor: Matthew Strickland, Ph.D.

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor Of Philosophy

in Epidemiology

2014

Acknowledgements

I would like to express my sincere gratitude to all members of my dissertation committee, all of whom have been incredibly accessible, supportive and insightful during the entire dissertation process. First and foremost I would like to thank Matt Strickland for his support in fostering my intellectual and professional growth by including me in his thought processes and research outside the aims of my dissertation, for his willingness to stop his work and engage in meaningful conversations about my work multiple times a day, and for his friendship and encouragement throughout this process. I thank Mitch Klein for spending hours with me in his office discussing methodological issues related (and not) to my dissertation and for his friendship and mentorship over the course of my entire Rollins career, and for making me a better SAS programmer. I would like to thank Dana Flanders who, despite his brilliance, always treated me like an intellectual peer and was willing to wrestle with methodological issues brought up by this dissertation. I thank Howard Chang his incredibly helpful statistical insights, for driving me to make my graphs in R, and his lightning fast response times. Finally, I thank Stefanie Sarnat for her thorough and detailed reviews of my work, for providing much needed help about what’s really behind all these pollutant measurements and for helping to keep me grounded and sane throughout.

In addition I want to thank all members of the Study of Particles and Health in Atlanta (SOPHIA) for letting me sit in on their meetings and teaching me more than I ever dreamed I’d know about air pollution, and all their wonderful feedback on my work; I’d particularly like to thank Paige Tolbert, Andrea Winquist and Lyndsey Darrow.

Thank you to all members of the Epidemiology department faculty, staff and students for fostering such a great learning environment and supportive group.

Finally I want to thank my family, particularly Heather Tinguely, for encouraging, supporting, and loving me over the wild ride of these past five years!

Financial support: This dissertation was made possible in part by US EPA grant R834799 and NIEHS K01ES019877. This dissertation’s contents are solely the responsibility of the grantee and do not necessarily represent the official views of the US EPA. Further, US EPA does not endorse the purchase of any commercial products or services mentioned in the publication.

Table of Contents

List of Tables and Figures

Chapter 1: Introduction

Overview

Exposure mixtures are all around us. Every day we receive influence from the mixture of genes in our genetic code, ingest a mixture of nutrients at each meal and breathe a mixture of pollutants with each breath. Such mixtures comprise a blend of individual exposures that we experience simultaneously, and yet despite this most fields of research have examined each exposure individually, as if it were an independent phenomenon.

This is particularly true for the field of air pollution. Until this past decade the majority of research was spent understanding the effect of a single pollutant, say nitrogen dioxide, on human health; however, we know that pollutants aren't emitted in isolation, nor do we selectively inhale one pollutant and not another. So why haven't pollution mixtures received more attention by researchers?

For one, studying pollution mixtures (a.k.a. 'multipollutants') is incredibly challenging. It requires changes to the current scientific approaches for air pollution, the development of new modeling methods, and modifications to how we conceptualize health risks [1]. Nonetheless researchers *have* been exploring multipollutant effects for some time, they have just been doing so through single pollutant models, in which a single pollutant effect is believed to act as a surrogate for the air pollution mixture [2-5]. One reason to favor single pollutant models is that the Environmental Protection Agency [6] regulates pollutants individually. Therefore results from single pollutant models have more direct relevance to regulatory decisions.

In 2004, the National Research Council put out a report in which it called for a multipollutant approach to air quality management [7]. This resulted in a paradigm shift for air pollution research, as more scientists began exploring the health effects of multipollutant exposures. However the field of multipollutant research is still in its infancy; there is still no clear consensus as to what a multipollutant approach means nor how it should be executed [8]. A particular challenge is how to model multipollutant exposures for health research.

Motivation

Recently two review articles were published in which the authors describe different statistical approaches for handling multipollutant exposures [9, 10]. One approach described by both articles is classification and regression trees (C&RT). C&RT is a recursive partitioning approach in which the initial dataset is repeatedly split into subsets, such that each resulting subset contains observations that are more similar with regards to the outcome. C&RT results in the formation of ‘terminal nodes’, which are collections of observations that form a complete partition of the initial dataset. An appealing aspect of C&RT is that it results in a dendogram; a visually intuitive and informative tree diagram describing the partitioning process and the resulting classification of terminal nodes. C&RT has been proposed as a useful tool for identifying complex multipollutant interactions [9], though has yet to be used in an epidemiologic study for this purpose.

An alternative way to formulate multipollutant exposures is by the sources from which they are emitted. Source apportionment (SA) is an approach used to describe the components of particulate matter [11] according to the contributing sources. Though PM is often measured, referenced and regulated as if it were a single pollutant, it is in fact a mixture of many different particles that differ in size, chemistry and reactivity. SA is a technique for decomposing the PM mixture and determining the health effects of each source. An added advantage is that it can lead to source-based risk assessment (i.e. what is the effect of reducing emissions from a given source?) and ultimately more targeted regulation [12].

This dissertation takes a multipollutant approach to air pollution epidemiology. In particular two types of air pollution mixtures are examined. Studies 1 and 2 consider pollution mixtures by day and attempt to classify types of days according to their association with the outcome. Study 3 examines the mixtures encompassed in total fine particulate matter mass, by measuring the health effect associated with each different source. Throughout this dissertation novel methodologies are developed or applied. In the first study a modified C&RT algorithm is introduced and its applicability to air pollution mixtures demonstrated. The second study then applies this C&RT algorithm to a three-city examination of multipollutant joint effects and compares the results to more conventional modeling approaches. Finally, the third study is the first epidemiologic application of a novel ensemble-based source apportionment method.

Throughout this dissertation the effects of multipollutants are examined with regards to emergency department visits for acute asthma events among children. Children are particularly vulnerable to effects of air pollution for several reasons including ongoing lung growth and development, incomplete metabolic systems, immature host defenses, high rates of infection with respiratory diseases, and activity patterns that increase their exposure to air pollutants [13]. The association between individual pollutants and childhood asthma has been well-documented in the literature [14-22]. This dissertation seeks to advance the current understanding of the association between ambient air pollution and pediatric asthma by considering the effects air pollution mixtures.

This dissertation has the following aims:

- **Study 1:** To demonstrate how a modified C&RT approach can be used to generate hypotheses about multipollutant joint effects, by investigating the association between ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO) and fine particulate matter <2.5$\mu g/m^3$ ($PM_{2.5}$) and emergency department (ED) visits for pediatric asthma.

- **Study 2:** To extend our understanding of multipollutant joint effects associated with pediatric asthma by comparing those identified via C&RT to those identified through more conventional multipollutant regression modeling approaches.

- **Study 3:** To examine the association between ensemble-based $PM_{2.5}$ source impacts and ED visits for pediatric asthma.

It is the hope that the results presented in this dissertation can help to not only further the understanding of multipollutants' impacts on childhood asthma but also offer new approaches and insight for dealing with mixtures of exposures as a whole.

References

1. Dominici F, Peng RD, Barr CD, Bell ML: **Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach**. *Epidemiology* 2010, **21**(2):187-194.

2. Sarnat JA, Schwartz J, Catalano PJ, Suh HH: **Gaseous pollutants in particulate matter epidemiology: confounders or surrogates?** *Environ Health Perspect* 2001, **109**(10):1053-1061.

3. Janssen NA, Lanki T, Hoek G, Vallius M, de Hartog JJ, Van Grieken R, Pekkanen J, Brunekreef B: **Associations between ambient, personal, and indoor exposure to fine particulate matter constituents in Dutch and Finnish panels of cardiovascular patients**. *Occupational and environmental medicine* 2005, **62**(12):868-877.

4. Laden F, Neas LM, Dockery DW, Schwartz J: **Association of fine particulate matter from different sources with daily mortality in six U.S. cities**. *Environ Health Perspect* 2000, **108**(10):941-947.

5. Sarnat SE, Suh HH, Coull BA, Schwartz J, Stone PH, Gold DR: **Ambient particulate air pollution and cardiac arrhythmia in a panel of older adults in Steubenville, Ohio**. *Occupational and environmental medicine* 2006, **63**(10):700-706.

6. US Department of Health and Human Services, Agency for Healthcare Research and Quality: **2008 National Healthcare Disparities Report**. In. Rockville, MD: US Department of Health and Human Services, Agency for Healthcare Research and Quality; 2008

7. National Research Council: **Air quality management in the United States**. In. Edited by National Academic Press. Washington, D.C.; 2004.

8. Vedal S, Kaufman JD: **What does multi-pollutant air pollution research mean?** *Am J Respir Crit Care Med* 2011, **183**(1):4-6.

9. Billionnet C, Sherrill D, Annesi-Maesano I, Study G: **Estimating the Health Effects of Exposure to Multi-Pollutant Mixture**. *Annals of Epidemiology* 2012, **22**(2):126-141.

10. Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, Batterman SA, Mukherjee B: **Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons**. *Environ Health* 2013, **12**(1):85.

11. Thurston GD, Lippmann M, Scott MB, Fine JM: **Summertime haze air pollution and children with asthma**. *Am J Respir Crit Care Med* 1997, **155**(2):654-660.

12. Hopke PK, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Larson TV *et al*: **PM source apportionment and health effects: 1. Intercomparison of source apportionment results**. *J Expo Sci Environ Epidemiol* 2006, **16**(3):275-286.

13. Samoli E, Nastos PT, Paliatsos AG, Katsouyanni K, Priftis KN: **Acute effects of air pollution on pediatric asthma exacerbation: evidence of association and effect modification**. *Environ Res* 2011, **111**(3):418-424.

14. Auten RL, Foster WM: **Biochemical effects of ozone on asthma during postnatal development**. *Biochimica et biophysica acta* 2011, **1810**(11):1114-1119.

15. Balmes JR: **The role of ozone exposure in the epidemiology of asthma**. *Environ Health Perspect* 1993, **101 Suppl 4**:219-224.

16. Dockery DW, Speizer FE, Stram DO, Ware JH, Spengler JD, Ferris BG, Jr.: **Effects of inhalable particles on respiratory health of children**. *Am Rev Respir Dis* 1989, **139**(3):587-594.

17. Gent JF, Triche EW, Holford TR, Belanger K, Bracken MB, Beckett WS, Leaderer BP: **Association of low-level ozone and fine particles with respiratory symptoms in children with asthma**. *JAMA : the journal of the American Medical Association* 2003, **290**(14):1859-1867.

18. Koren HS: **Associations between criteria air pollutants and asthma**. *Environ Health Perspect* 1995, **103 Suppl 6**:235-242.

19. Peel JL, Tolbert PE, Klein M, Metzger KB, Flanders WD, Todd K, Mulholland JA, Ryan PB, Frumkin H: **Ambient air pollution and respiratory emergency department visits**. *Epidemiology* 2005, **16**(2):164-174.

20. Rossi OV, Kinnula VL, Tienari J, Huhti E: **Association of severe asthma attacks with weather, pollen, and air pollutants**. *Thorax* 1993, **48**(3):244-248.

21. Strickland MJ, Darrow LA, Klein M, Flanders WD, Sarnat JA, Waller LA, Sarnat SE, Mulholland JA, Tolbert PE: **Short-term associations between ambient air pollutants and pediatric asthma emergency department visits**. *Am J Respir Crit Care Med* 2010, **182**(3):307-316.

22. Tolbert PE, Mulholland JA, MacIntosh DL, Xu F, Daniels D, Devine OJ, Carlin BP, Klein M, Dorley J, Butler AJ *et al*: **Air quality and pediatric emergency room visits for asthma in Atlanta, Georgia, USA**. *Am J Epidemiol* 2000, **151**(8):798-810.

Chapter 2: Ambient Air Pollutants and Associated Health Effects

Introduction

Ambient air refers to freely moving air in the outdoor environment. The ambient air found in our atmosphere is a comprised of a natural mix of substances, the most abundant and crucial for life being oxygen (O_2) and nitrogen (N_2). Pollution of the ambient air occurs when gases or particles that are not part of the natural mix are introduced into the air with adverse effects to humans or the environment. Sources of such pollutants can be naturally occurring (e.g. dust storms, volcano eruptions, and forest fires) or caused by human activity (e.g. industrial processes, vehicular transportation, and heating and cooling emissions). Pollutants emitted directly from a source are called primary pollutants, whereas secondary pollutants are formed through chemical reactions involving primary pollutants and some weather conditions including heat and sunlight.

Ambient air pollutants are frequently categorized into gaseous and particulate-phased pollutants. The main gaseous pollutants include carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), volatile organic compounds (VOCs), certain toxic air pollutants and some gaseous forms of metals. Particle pollutants can be either solid or liquid phase and are often characterized according to size as particulate matter <10 micrometers (PM_{10}) and particulate matter <2.5 micrometers ($PM_{2.5}$). A more detailed description of the most common gaseous pollutants and particulate matter follows.

History of Ambient Air Pollution Control

Over the course of the 20th century mounting evidence shed light on the harmful effects of ambient air pollution on public health. Fueled by air pollution disasters, including deadly smogs in Donora, PA and London, England, the U.S. government began a series of procedures to monitor and regulate air pollution, culminating in the 1970 Clean Air Act [1]. The Clean Air Act established regulations on the concentrations of key pollutants for the good of public health. One important result of the Clean Air Act was the development of a pollution control strategy based on National Ambient Air Quality Standards (NAAQS) to be achieved at the state and national level [2]. The Clean Air Act required the Environmental Protection Agency [3] to set NAAQS for six "criteria" pollutants at levels believed to protect the public health with an adequate margin of safety, regardless of economic or technological feasibility of attainment [4]. For a complete list of the criteria pollutants and NAAQS see Table 2.1.

Significant amendments to the Clean Air Act were introduced by Congress in 1990. Highlights of these 1990 amendments include setting attainment deadlines for cities failing to meet the NAAQs, strengthening the power of the EPA and States to enforce standards on individual pollution sources, setting fuel and emission standards for motor vehicles, and reducing acid rain and ozone depletion [5]. The amendment also authorized control of 189 hazardous air pollutants (also known as "toxic air contaminants" or "air toxics") that are known to cause serious health effects including cancer, birth defects, respiratory tract and neurological illness [6, 7].

The success of the Clean Air Act and other pollution control strategies can be seen in the reduction of ambient air pollution levels. Ambient air pollution levels have been declining since 1990 at the same time the motor vehicle use, energy consumption and the U.S. population have been increasing [8]. By 2008, direct $PM_{2.5}$ emissions had declined by over one half, CO and SO_2 dropped by nearly 50%, and PM_{10}, NOx and VOCs had declined by over a third [8]. However, despite these significant declines in pollutant concentrations, epidemiologic studies continue to find health effects from ambient air pollution at levels below the current EPA standards [4]. It is therefore imperative that rigorous studies continue to be conducted on the health effects of low-level exposures and that continuous consideration be given to the revision of the NAAQS to reflect what is best for public health.

Ambient Air Pollution and Health

The detrimental effects of air pollution on human health are well documented. Ambient air pollution has been found to be associated with infectious respiratory diseases, including pneumonia [9, 10], otitis media [11, 12], and bronchiolitis [13, 14] as well as chronic lung diseases with acute manifestations including asthma [15-20] and chronic obstructive pulmonary disease (COPD) [21, 22]. Though the pathways by which air pollution exposure affects the respiratory system may be more explicit, there is also substantial evidence that air pollution is associated with cardiovascular and circulatory diseases. The health effects of ambient air pollution on the cardiovascular and circulatory systems include cardiovascular mortality [23, 24], ischemic heart disease [25], deep vein thrombosis [26] and myocardial infarction [27-29].

The following section describes the main air pollutants in greater detail, highlighting some of the main sources, characteristics and key health effects associated with each pollutant.

Ozone (O_3)

Depending on its location in the atmosphere, ozone can be beneficial or harmful for the planet. Stratospheric ozone, occurring at altitudes above 10km, is naturally created and helps shield the earth from solar radiation. On the contrary, tropospheric ozone occurs in the lower atmosphere and acts as a greenhouse gas, trapping heat from the sun and warming the earth's surface [8]. It is this latter type of ground-level ambient ozone that is considered one of the criteria pollutants and is associated with adverse human health effects.

Ambient ozone is a secondary pollutant, formed through a photochemical process of sunlight acting on nitrogen oxides and hydrocarbons [2], both of which are primarily produced by motor vehicle emissions [7]. Ozone levels tend to highest on warm, sunny windless days with peak levels found in the afternoon [7]. In Atlanta ozone levels are highest in the spring and summer and tend to be homogenous over space [30, 31]. Though the general trend shows O_3 levels decreasing nationally by 10% from 2001 to 2008, in 2008 there were still many areas with O_3 concentrations above the NAAQS [8]. Twenty-three sites around the U.S. showed an increase in ambient O_3 concentration from 2001 to 2008, including Atlanta, Los Angeles and Seattle [8].

High levels of exposure to ambient ozone are known to have detrimental consequence to human health. Ozone is a highly reactive gas that is capable of producing oxidative damage in the lung [32]. It is also a respiratory tract irritant for both adults and children, causing shortness of breath, wheezing, cough and chest pain when inhaled deeply [7]. Study results suggest that O_3 exposure increases asthma morbidity by causing airway inflammation and epithelial permeability [33]. A 5-year follow-up cohort study in southern CA by McConnell et al. found that among children living in high ozone areas, those playing sports were 3.3 times more likely to develop a new case of asthma than those playing no sports [34]. Similarly, greater time spent outside was associated with higher incidence of asthma among children living in high ozone areas [34].

Nitrogen oxides (NO_2, NO and NOx)

Nitrogen dioxide (NO_2) is a pungent gas that varies from yellow to brown in color, depending on its concentration [2]. Nitrogen dioxide is readily produced in a number of atmospheric reactions including direct oxidation of NO and photochemical oxidation of NO with ozone. Though NO_2 is the primary pollutant regulated by the EPA, both NO and NO_2 are commonly referred to together as NOx (nitrogen oxides) because of how readily they convert from one to the other [2]. Nitrogen oxides are formed naturally in the atmosphere by lightning, forest fires and bacterial activity in the soil; however, in North America, anthropogenic sources dominate [35]. Anthropogenic NO_2 is produced by high-temperature combustion, particularly from gasoline and diesel powered engines as well as power plants and industrial combustion [7, 35]. In areas without heavy industrial activity, such as Atlanta, almost all of the ambient NO_2 concentration is from traffic. Because NO_2 is primarily traffic generated, it is sometimes used as an indicator of the proportion of pollutants stemming from a vehicular source [36].

Separating out the effects of NO_2 is often difficult as the pollutant rarely occurs outdoors by itself, but rather as a complex mixture of primary and secondary pollutants [1]. In epidemiologic studies, NO_2 is often used as a surrogate or marker for traffic exposure, and it is often difficult to establish whether there is an independent association of NO_2 due to its high correlation with other ambient pollutants, particularly traffic-derived $PM_{2.5}$ which has the potential to confound associations with NO_2 [7, 37]. Although NO_2 exhibits little seasonality, studies in Atlanta have found NO_2 to exhibit significant spatiotemporal heterogeneity [30, 31].

The most significant health effects of NO_2 result indirectly through its role in the formation of ozone and other secondary pollutants such as $PM_{2.5}$ nitrate [1]. Nonetheless, studies suggest that exposure to NO_2 alone compromises respiratory health. The Chattanooga School Children Study, one of the earliest to examine the respiratory effects of elevated NO_2 exposure, found that lung function of second grade children living near a stationary source of NO_2 was significantly lower than that of children living in the control area [38]. Since then, additional studies have found outdoor exposure to NO_2 to be associated with respiratory morbidity in young children [39, 40]. Findings from a multi-city European study of air pollution and health (APHEA) suggest an independent effect of NO_2 on respiratory and cardiovascular mortality among adults [41]; however it is difficult to rule out the possibility that NO_2 is acting as a surrogate for unmeasured pollutants. A multipollutant study in Atlanta investigating emergency department visits for upper respiratory illness found a significant single-pollutant effect of NO_2 but the effect was attenuated to the null when an additional pollutants were added to the model [18].

Carbon Monoxide (CO)

Carbon monoxide is a colorless, odorless, tasteless gas that is emitted from a variety of natural and anthropogenic sources [2]. Anthropogenic CO is produced by the incomplete combustion of fossil fuels, with the principal sources of ambient CO including on-road and off-road vehicles [35]. In the U.S., highway and non-roadway mobile vehicles contribute up to 80% of CO emissions [8]. Regulations on vehicular emissions in the U.S. have helped lead to a declining trend in CO levels, which decreased by 41% from 2001 to 2008 [8]; however, the opposite trend is being observed in developing countries where CO levels are on the rise [2]. Carbon monoxide typically shows significant spatial and temporal heterogeneity, with the highest concentrations found along roadways during morning and evening rush hours when traffic is at its peak [2, 30]. As with NO_2, in epidemiological studies CO is typically considered a surrogate for traffic-related pollutants; it is difficult to attribute measures of association to CO exposure alone.

Though exposure to extremely high levels of carbon monoxide can lead to acute poisoning and death, this level of exposure is rarely seen in ambient CO concentrations. The primary health concerns associated with low-level exposure to ambient CO are detriments to the cardiovascular system [2]. Once it is inhaled, CO is readily absorbed into the bloodstream where it may have direct and indirect effects on the cardiovascular system [2]. Carbon monoxide interferes with oxygen transport through the formation of carboxyhemoglobin [7], thereby reducing the amount of oxygen flowing to the bodily organs and tissue [8]. CO exposure during pregnancy has been associated with preterm birth [42] and a 23 gram reduction in birth weight per 1 part per million increase in CO [43]. Preexisting conditions such as chronic respiratory disease, diabetes and CVD may make people more susceptible to harmful effects from CO exposure; however, study results are contradictory. While some studies have reported an increased risk in hospitalization for acute cardiovascular events among subjects with preexisting cardiopulmonary conditions [28, 44], Peel et al. found that patients with congestive heart failure had a decreased risk of emergency department visits for ischemic heart disease [45].

Particulate Matter (PM)

Particulate matter (PM) refers to solid and liquid-phase particles that come from natural and anthropogenic sources [2]. Particles can be classified as primary or secondary in origin and vary in composition, size and density. Primary particles are emitted directly from a source (e.g. a combustion engine) while secondary particles are formed from chemical reactions involving gaseous precursors, including sulfur oxides and NOx [1]. The majority of PM compounds can be grouped into five categories: sulfates, nitrates, elemental (black) carbon, organic carbon and crustal metals [8].

The term particulate matter refers to a complex mix of pollutant particles, the components of which are likely to exhibit substantial spatiotemporal variability. As a result, it is sometimes difficult to isolate the effects of PM from other gaseous pollutants. Results from several studies have indicated that confounding due to SO_2 and O_3 can largely be dismissed for studies of particulate matter effects [36, 46, 47]. However these results may apply only to PM_{10}, as more recent studies have found that $PM_{2.5}$ sulfate can confound associations with ozone [48]. Studies on confounding of PM by NO_2 are less conclusive, with some U.S. studies finding little evidence of confounding [47] while other European and U.S. studies have found significant confounding [36, 37]. One explanation for this may be the difference in source contribution between the U.S. and Europe, particularly emissions from diesel vehicles that tend to be greater in Europe, which will affect the components of the PM mix [36]. Another factor could be the prevailing size of the particulate matter, as smaller particles have been found to be highly correlated with NO_2 due their common motor vehicle source [37].

In order for particles to have an effect on human health, they must first have a pathway into the body, namely through the respiratory system. Studies on the deposition and clearance of particles in the body defined inhalable particles as those less than 10 μm in aerodynamic diameter, commonly referred to as PM_{10} [49]. Later evidence in the 1990's suggested that even smaller particles, those with an aerodynamic diameter of 2.5 μm were able to penetrate the alveolar gas-exchange region of the lungs and may present a greater health risk to humans [49]. In lieu of these findings, in the context of health effect studies, particulate matter is frequently classified by its aerodynamic diameter as PM_{10} or $PM_{2.5}$.

$PM_{2.5}$

Fine particulate matter is defined as all biogenic and anthropogenic aerodynamic particles measuring less than 2.5 micrometers in diameter. The U.S. EPA did not begin monitoring $PM_{2.5}$ until 1999 [50]; prior to that point any health effects of $PM_{2.5}$ would have been attributed to the larger PM_{10} classification (all particles <10 micrometers). Despite the recent focus on $PM_{2.5}$, annual and 24-hour $PM_{2.5}$ concentrations declined by 17 and 19 percent from 2001 to 2008, respectively [8].

Primary $PM_{2.5}$ is emitted from combustion processes, particularly diesel-powered engines, power generation and wood burning whereas secondary $PM_{2.5}$ is formed from atmospheric processes [7]. Because of their low settling velocity, fine particles can be transported relatively long distances downwind from their sources [2]. In Atlanta, fine particulate matter has been found to be relatively homogenous over space [30, 51]. The composition of $PM_{2.5}$ has been shown to vary spatially and temporally. One multi-city study in the U.S. found that the association of $PM_{2.5}$ and cardiovascular and respiratory outcomes varied by season, with the strongest effects found in the spring [50].

One of the first studies to look at the health effects of $PM_{2.5}$ found it to be associated with lung cancer and cardiopulmonary mortality [52]. As the body of research on $PM_{2.5}$ grows, the more health effects previously attributed to all PM exposure are now being attributed to fine particulate matter. A reanalysis of the Harvard Six City study found that fine, and not coarse, particulate matter was associated with acute respiratory tract effects in children [53]. A recent update to the American Heart Association's statement on air pollution concluded that acute $PM_{2.5}$ exposure can lead to CVD mortality and non-fatal events [54].

$PM_{2.5}$ Components

Recently there has been increasing interest in understanding the $PM_{2.5}$ components that pose the greatest health risk. Particulate matter is, by definition, a mix of different chemical species that varies by emission source, seasonality, geographic location, and meteorological conditions. The major components that comprise $PM_{2.5}$ are sulfates, nitrates, organic carbon (OC), elemental carbon (EC) and crustal material [8]. According to data from the U.S. EPA, in the eastern U.S. sulfate is the largest component by mass, due primarily to electricity production and industrial boilers. OC is the primary component of $PM_{2.5}$ on the west coast, with woodstoves and fireplaces comprising the primary sources. Additional sources of OC include highway vehicles, waste burning and wildfires. Nitrate, another important component on $PM_{2.5}$, is produced by highway vehicles and non-road mobile machinery. EC emitted from the incomplete production of fossil fuels is typically one of the smallest $PM_{2.5}$ components by mass [8]. Crustal material is comprised of minerals from the Earth's crust that become airborne through soil erosion, weathering and dust storms.

The composition of $PM_{2.5}$ mass has been shown to significantly modify the association of $PM_{2.5}$ and hospital admissions [50]. Studies examining the chemical composition of $PM_{2.5}$ have found nickel, EC and vanadium [55]; EC, OC, and ammonium [56]; and nickel, arsenic, chromium, bromine and OC [50] to be associated with hospital admissions for cardiovascular and respiratory disease. Data from Atlanta and other southeastern monitoring sites indicate that organic material and sulfates comprise 60% of $PM_{2.5}$ [57]. Though sulfate has been found to be relatively spatially homogenous in Atlanta [30], it does exhibit significant seasonality with maximum concentrations occurring during the warm months [57].

The health risk posed by $PM_{2.5}$ varies by component. In Atlanta, sulfate-rich secondary $PM_{2.5}$ and $PM_{2.5}$ stemming from mobile sources, primarily EC and OC, was found to be associated with asthma [20] and all respiratory ED visits [58]. Emergency department visits for all CVD outcomes were significantly associated with same-day $PM_{2.5}$ concentrations of OC-related sources, including diesel and gasoline as well as biomass burning [58].

Table 2.1 National Ambient Air Quality Standards

Pollutant	NAAQS[1]	Primary/ Secondary	Health Effects[2,3]
CO	8hr – 9ppm 1hr – 35ppm	Primary	Decreased exercise capacity, reduction in oxygen to bodily organs, aggravation of heart disease, chest pain, premature mortality
NO_2	1hr – 100ppb	Primary	Decreased lung function, increased airway reactivity, increased susceptibility to lung infection and respiratory symptoms
	Annual – 53ppb	Primary and Secondary	
O_3	8hr – 0.075ppm	Primary and Secondary	Decreased lung function, causes respiratory symptoms (i.e. coughing and shortness of breath), lung inflammation, decreased exercise capacity, aggravation of asthma and lung disease, leads to hospital admissions and premature mortality
SO_2	1hr – 75ppb	Primary	Decreased lung function, increased respiratory symptoms including aggravated asthma and wheezing, and respiratory mortality
	3hr – 0.5ppm	Secondary	
PM_{10}	24hr - 150 µg/m^3	Primary and Secondary	Decreased lung function, increased respiratory symptoms and illness, increase asthma exacerbations, hospital admissions for cardiovascular and respiratory diseases, and premature mortality,
$PM_{2.5}$	24hr - 35 µg/m^3 Annual - 15µg/m^3	Primary and Secondary	Increased hospitalizations for cardiovascular disease, myocardial infarction

[1] U.S. Environmental Protection Agency. "National Ambient Air Quality Standards (NAAQS)." Retrieved 11/29/2011, from http://www.epa.gov/air/criteria.html.

[2] U.S. Environmental Protection Agency (2010). Our Nation's Air: Status and trends through 2008. Research Park Triangle, NC, Office of Air Quality Planning and Standards

[3] Bernard, S. M., J. M. Samet, et al. (2001). "The potential impacts of climate variability and change on air pollution-related health effects in the United States." Environ Health Perspect **109 Suppl 2**: 199-209.

References

1. Bernard SM, Samet JM, Grambsch A, Ebi KL, Romieu I: **The potential impacts of climate variability and change on air pollution-related health effects in the United States**. *Environ Health Perspect* 2001, **109 Suppl 2**:199-209.

2. Godish T: **Air Quality**, 4th Edition edn. Boca Raton: Lewis Publishers; 2004.

3. US Department of Health and Human Services, Agency for Healthcare Research and Quality: **2008 National Healthcare Disparities Report**. In. Rockville, MD: US Department of Health and Human Services, Agency for Healthcare Research and Quality; 2008

4. American Thoracic Society Committee of the Environmental and Occupational Health Assembly: **What constitutes an adverse health effect of air pollution?** *Am J Respir Crit Care Med* 2000, **161**:665-673.

5. U.S. Environmental Protection Agency: **Highlights of the 1990 Clean Air Act Amendments**. *EPA Journal* 1991, **January/February**.

6. **History of the Clean Air Act** [http://www.epa.gov/oar/caa/caa_history.html#caa90]

7. Kim JJ: **Ambient air pollution: health hazards to children**. *Pediatrics* 2004, **114**(6):1699-1707.

8. U.S. Environmental Protection Agency: **Our Nation's Air: Status and trends through 2008**. In. Research Park Triangle, NC: Office of Air Quality Planning and Standards; 2010.

9. Cheng MF, Tsai SS, Wu TN, Chen PS, Yang CY: **Air pollution and hospital admissions for pneumonia in a tropical city: Kaohsiung, Taiwan**. *J Toxicol Environ Health A* 2007, **70**(24):2021-2026.

10. Kappos AD, Bruckmann P, Eikmann T, Englert N, Heinrich U, Hoppe P, Koch E, Krause GH, Kreyling WG, Rauchfuss K *et al*: **Health effects of particles in ambient air**. *Int J Hyg Environ Health* 2004, **207**(4):399-407.

11. Brauer M, Gehring U, Brunekreef B, de Jongste J, Gerritsen J, Rovers M, Wichmann HE, Wijga A, Heinrich J: **Traffic-related air pollution and otitis media**. *Environ Health Perspect* 2006, **114**(9):1414-1418.

12. MacIntyre EA, Karr CJ, Koehoorn M, Demers PA, Tamburic L, Lencar C, Brauer M: **Residential air pollution and otitis media during the first two years of life**. *Epidemiology* 2011, **22**(1):81-89.

13. Karr C, Lumley T, Schreuder A, Davis R, Larson T, Ritz B, Kaufman J: **Effects of subchronic and chronic exposure to ambient air pollutants on infant bronchiolitis**. *Am J Epidemiol* 2007, **165**(5):553-560.

14. Karr CJ, Demers PA, Koehoorn MW, Lencar CC, Tamburic L, Brauer M: **Influence of ambient air pollutant sources on clinical encounters for infant bronchiolitis**. *Am J Respir Crit Care Med* 2009, **180**(10):995-1001.

15. Friedman MS, Powell KE, Hutwagner L, Graham LM, Teague WG: **Impact of changes in transportation and commuting behaviors during the 1996 Summer Olympic Games in Atlanta on air quality and childhood asthma**. *JAMA : the journal of the American Medical Association* 2001, **285**(7):897-905.

16. Lin S, Liu X, Le LH, Hwang SA: **Chronic exposure to ambient ozone and asthma hospital admissions among children**. *Environ Health Perspect* 2008, **116**(12):1725-1730.

17. McConnell R, Islam T, Shankardass K, Jerrett M, Lurmann F, Gilliland F, Gauderman J, Avol E, Kunzli N, Yao L *et al*: **Childhood incident asthma and traffic-related air pollution at home and school**. *Environ Health Perspect* 2010, **118**(7):1021-1026.

18. Peel JL, Tolbert PE, Klein M, Metzger KB, Flanders WD, Todd K, Mulholland JA, Ryan PB, Frumkin H: **Ambient air pollution and respiratory emergency department visits**. *Epidemiology* 2005, **16**(2):164-174.

19. Samoli E, Nastos PT, Paliatsos AG, Katsouyanni K, Priftis KN: **Acute effects of air pollution on pediatric asthma exacerbation: evidence of association and effect modification**. *Environ Res* 2011, **111**(3):418-424.

20. Strickland MJ, Darrow LA, Klein M, Flanders WD, Sarnat JA, Waller LA, Sarnat SE, Mulholland JA, Tolbert PE: **Short-term associations between ambient air pollutants and pediatric asthma emergency department visits**. *Am J Respir Crit Care Med* 2010, **182**(3):307-316.

21. Brown DW, Croft JB, Greenlund KJ, Giles WH: **Trends in hospitalization with chronic obstructive pulmonary disease-United States, 1990-2005**. *COPD* 2010, **7**(1):59-62.

22. Liang WM, Liu WP, Kuo HW: **Diurnal temperature range and emergency room admissions for chronic obstructive pulmonary disease in Taiwan**. *Int J Biometeorol* 2009, **53**(1):17-23.

23. Analitis A, Katsouyanni K, Dimakopoulou K, Samoli E, Nikoloulopoulos AK, Petasakis Y, Touloumi G, Schwartz J, Anderson HR, Cambra K *et al*: **Short-term effects of ambient particles on cardiovascular and respiratory mortality**. *Epidemiology* 2006, **17**(2):230-233.

24. Chen R, Li Y, Ma Y, Pan G, Zeng G, Xu X, Chen B, Kan H: **Coarse particles and mortality in three Chinese cities: the China Air Pollution and Health Effects Study (CAPES)**. *Sci Total Environ* 2011, **409**(23):4934-4938.

25. Pope CA, 3rd, Burnett RT, Thurston GD, Thun MJ, Calle EE, Krewski D, Godleski JJ: **Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease**. *Circulation* 2004, **109**(1):71-77.

26. Baccarelli A, Martinelli I, Pegoraro V, Melly S, Grillo P, Zanobetti A, Hou L, Bertazzi PA, Mannucci PM, Schwartz J: **Living near major traffic roads and risk of deep vein thrombosis**. *Circulation* 2009, **119**(24):3118-3124.

27. Dockery DW: **Epidemiologic evidence of cardiovascular effects of particulate air pollution**. *Environ Health Perspect* 2001, **109 Suppl 4**:483-486.

28. Nuvolone D, Balzi D, Chini M, Scala D, Giovannini F, Barchielli A: **Short-term association between ambient air pollution and risk of hospitalization for acute myocardial infarction: results of the cardiovascular risk and air pollution in Tuscany (RISCAT) study**. *Am J Epidemiol* 2011, **174**(1):63-71.

29. Zanobetti A, Schwartz J: **Air pollution and emergency admissions in Boston, MA**. *J Epidemiol Community Health* 2006, **60**(10):890-895.

30. Sarnat SE, Klein M, Sarnat JA, Flanders WD, Waller LA, Mulholland JA, Russell AG, Tolbert PE: **An examination of exposure measurement error from air pollutant spatial variability in time-series studies**. *J Expo Sci Environ Epidemiol* 2010, **20**(2):135-146.

31. Tolbert PE, Klein M, Metzger KB, Peel J, Flanders WD, Todd K, Mulholland JA, Ryan PB, Frumkin H: **Interim results of the study of particulates and health in Atlanta (SOPHIA)**. *J Expo Anal Environ Epidemiol* 2000, **10**(5):446-460.

32. Schwartz J: **Air pollution and children's health**. *Pediatrics* 2004, **113**(4 Suppl):1037-1043.

33. D'Amato G, Cecchi L: **Effects of climate change on environmental factors in respiratory allergic diseases**. *Clin Exp Allergy* 2008, **38**(8):1264-1274.

34. McConnell R, Berhane K, Gilliland F, London SJ, Islam T, Gauderman WJ, Avol E, Margolis HG, Peters JM: **Asthma in exercising children exposed to ozone: a cohort study**. *Lancet* 2002, **359**(9304):386-391.

35. Holman C: **Sources of air pollution**. In: *Air Pollution and Health.* Edited by Holgate S, Samet JM, Koren HS, Maynard RL. San Diego, CA: Academic Press; 1999: 115-148.

36. Katsouyanni K, Touloumi G, Samoli E, Gryparis A, Le Tertre A, Monopolis Y, Rossi G, Zmirou D, Ballester F, Boumghar A *et al*: **Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 project**. *Epidemiology* 2001, **12**(5):521-531.

37. Kim D, Sass-Kortsak A, Purdham JT, Dales RE, Brook JR: **Associations between personal exposures and fixed-site ambient measurements of fine particulate matter, nitrogen dioxide, and carbon monoxide in Toronto, Canada**. *J Expo Sci Environ Epidemiol* 2006, **16**(2):172-183.

38. Shy CM, Creason JP, Pearlman ME, McClain KE, Benson FB, Young MM: **The Chattanooga school children study: effects of community exposure to nitrogen dioxide. 1. Methods, description of pollutant exposure, and results of ventilatory function testing**. *J Air Pollut Control Assoc* 1970, **20**(8):539-545.

39. Dales RE, Cakmak S, Doiron MS: **Gaseous air pollutants and hospitalization for respiratory disease in the neonatal period**. *Environ Health Perspect* 2006, **114**(11):1751-1754.

40. Esplugues A, Ballester F, Estarlich M, Llop S, Fuentes-Leonarte V, Mantilla E, Vioque J, Iniguez C: **Outdoor, but not indoor, nitrogen dioxide exposure is associated with persistent cough during the first year of life**. *Sci Total Environ* 2011, **409**(22):4667-4673.

41. Samoli E, Aga E, Touloumi G, Nisiotis K, Forsberg B, Lefranc A, Pekkanen J, Wojtyniak B, Schindler C, Niciu E *et al*: **Short-term effects of nitrogen dioxide on mortality: an analysis within the APHEA project**. *Eur Respir J* 2006, **27**(6):1129-1138.

42. Ritz B, Wilhelm M, Hoggatt KJ, Ghosh JK: **Ambient air pollution and preterm birth in the environment and pregnancy outcomes study at the University of California, Los Angeles**. *Am J Epidemiol* 2007, **166**(9):1045-1052.

43. Gouveia N, Bremner SA, Novaes HM: **Association between ambient air pollution and birth weight in Sao Paulo, Brazil**. *J Epidemiol Community Health* 2004, **58**(1):11-17.

44. Mann JK, Tager IB, Lurmann F, Segal M, Quesenberry CP, Jr., Lugg MM, Shan J, Van Den Eeden SK: **Air pollution and hospital admissions for ischemic heart disease in persons with congestive heart failure or arrhythmia**. *Environ Health Perspect* 2002, **110**(12):1247-1252.

45. Peel JL, Metzger KB, Klein M, Flanders WD, Mulholland JA, Tolbert PE: **Ambient air pollution and cardiovascular emergency department visits in potentially sensitive groups**. *Am J Epidemiol* 2007, **165**(6):625-633.

46. Schwartz J, Spix C, Touloumi G, Bacharova L, Barumamdzadeh T, le Tertre A, Piekarksi T, Ponce de Leon A, Ponka A, Rossi G *et al*: **Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions**. *J Epidemiol Community Health* 1996, **50 Suppl 1**:S3-11.

47. Samet JM, Zeger SL, Dominici F, Dockery DW, J. S: **The National Morbidity, Mortality, and Air Pollution Study (NMMAPS): Methods and Methodological Issues.** In. Boston: Health Effects Institute; 1999.

48. Franklin M, Schwartz J: **The impact of secondary particles on the association between ambient ozone and mortality**. *Environ Health Perspect* 2008, **116**(4):453-458.

49. Dockery D: **Health effects of particulate air pollution**. *Annals of Epidemiology* 2009, **19**(4):257-263.

50. Zanobetti A, Franklin M, Koutrakis P, Schwartz J: **Fine particulate air pollution and its components in association with cause-specific emergency admissions**. *Environ Health* 2009, **8**:58.

51. Darrow LA, Klein M, Sarnat JA, Mulholland JA, Strickland MJ, Sarnat SE, Russell AG, Tolbert PE: **The use of alternative pollutant metrics in time-series studies of ambient air pollution and respiratory emergency department visits**. *J Expo Sci Environ Epidemiol* 2011, **21**(1):10-19.

52. Pope CA, 3rd, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD: **Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution**. *JAMA : the journal of the American Medical Association* 2002, **287**(9):1132-1141.

53. Schwartz J, Neas LM: **Fine particles are more strongly associated than coarse particles with acute respiratory health effects in schoolchildren**. *Epidemiology* 2000, **11**(1):6-10.

54. Brook RD, Rajagopalan S, Pope CA, 3rd, Brook JR, Bhatnagar A, Diez-Roux AV, Holguin F, Hong Y, Luepker RV, Mittleman MA *et al*: **Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association**. *Circulation* 2010, **121**(21):2331-2378.

55. Bell ML, Ebisu K, Peng RD, Samet JM, Dominici F: **Hospital admissions and chemical composition of fine particle air pollution**. *Am J Respir Crit Care Med* 2009, **179**(12):1115-1120.

56. Peng RD, Bell ML, Geyh AS, McDermott A, Zeger SL, Samet JM, Dominici F: **Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution**. *Environ Health Perspect* 2009, **117**(6):957-963.

57. Edgerton ES, Hartsell BE, Saylor RD, Jansen JJ, Hansen DA, Hidy GM: **The Southeastern Aerosol Research and Characterization Study: Part II. Filter-based measurements of fine and coarse particulate matter mass and composition**. *J Air Waste Manag Assoc* 2005, **55**(10):1527-1542.

58. Sarnat JA, Marmur A, Klein M, Kim E, Russell AG, Sarnat SE, Mulholland JA, Hopke PK, Tolbert PE: **Fine particle sources and cardiorespiratory morbidity: an application of chemical mass balance and factor analytical source-apportionment methods**. *Environ Health Perspect* 2008, **116**(4):459-466.

Chapter 3: Asthma and Air Pollution

Asthma is a chronic lung disease with 235 million sufferers worldwide and the most common lung disease among children [1]. In the United States, approximately 7.8% of the population had been diagnosed with asthma in 2008, while the rate among children was 9.3% [2]. These rates represent a significant increase in asthma prevalence over the past decade, which is seen across all age groups [3]. Although the probable cause for this increase is unclear, many genetic, biological and environmental risk factors have been found to be associated with the onset of asthma. It is important to point out that though there is a statistically significant trend of increasing prevalence from 1980 – 2009, the prevalence of asthma attacks has remained level since 1997 [4].

The economic costs associated with asthma exacerbations are significant. From 2002 to 2007 the annual economic cost of asthma in the U.S. was $56 billion, with the bulk of this burden stemming from direct health care costs [5]. Each year there are approximately 15 million outpatient visits, 2 million emergency room visits, and 500,000 hospitalizations for acute asthma management [6]. In addition to direct healthcare costs, asthma-related morbidity is responsible for significant productivity loss. Asthma is the leading cause of activity limitation in the U.S. In 2008, asthma accounted for 14.2 million lost work days in adults and 14.4 million lost school days in children [7].

Asthma is a chronic inflammatory lung disease characterized by reversible airway obstruction and airway hyperresponsiveness [8]. Asthma diagnosis is typically based on the presence of characteristic symptoms including episodic breathlessness, wheezing, cough and chest tightness [9]. Measurements of lung function variability, limitation and reversibility of obstruction are used to confirm asthma diagnoses [9].

Among asthmatics, the factors known to precipitate an acute response are heterogeneous; different triggers may be more or less important for susceptible individuals [6]. The numerous precipitants of acute asthma exacerbations include smoke, mold, pet dander, infectious diseases, stress, exercise, and air pollution. While respiratory tract viruses are the most common causes leading to acute asthma exacerbations, cigarette smoke is one of the most important modifiable risk factors for asthma exacerbations, along with exposure to allergens from cats and dogs [6]. Hospital asthma mortality is highest in the winter months and may be reflective of higher rates of influenza infection [6].

Severe asthma exacerbations have been described as “events that require urgent action on the part of the patient and physician to prevent a serious outcome, such as hospitalization or death” [10]. Poorly controlled asthma can lead to severe exacerbations. Preventative asthma medications are the primary strategy for managing persistent asthma in children; however, recent data suggest that these preventative medications are underutilized [11]. A national sample indicated that less than one-third of children and adolescents with diagnosed asthma used preventative medications [12].

Populations Susceptible to Asthma

Disparities in asthma prevalence are seen across ethnic, racial, socio-economic, age and gender strata. In the U.S., the population-based prevalence, emergency department visits, and hospitalization rates for asthma were higher among blacks than whites, higher among females than males, higher in poor than non-poor, and higher in children than adults [2]. The socio-demographic group with the highest rate of asthma is poor people of Hispanic Puerto Rican descent, with approximately 22.4% of the population currently diagnosed with asthma [2].

Asthma prevalence data suggest some form of interaction exists between age and gender. Although overall women experience higher rates of asthma than males, among children 0-17 years the prevalence is highest in boys [6]. The prevalence of asthma in children under 15 years old has been reported to be 25-70% greater in boys than girls, whereas overall the prevalence is approximately 30% greater in females [13, 14]. This finding suggests that young boys may be more susceptible to mechanisms that trigger asthma attacks [15]. One reason commonly put forth for this increased susceptibility in childhood is that boys have proportionally smaller airways in relation to their lung size than girls [13]. In adulthood, the shift in asthma burden from males to females may represent the role of sex hormones in the complex pathways of asthma [6, 15].

Another sub-population with increased susceptibility to asthma is people born preterm. Preterm birth is defined as birth <37 weeks gestation, with very preterm 26-33 weeks and late-preterm 34-36 weeks [16]. The rate of preterm births in developed countries showed a dramatic rise over the past couple decades; however, the past 5 years have seen a steady decline. In 2009, 12.2% of all U.S. births were preterm [17]. The rise in preterm births seen during the 1990’s and 2000’s may have been due in part to changes in how labor and delivery were managed, namely more frequent induction of labor and cesarean delivery prior to 39 weeks [17], as well as increased incidence of multiple births [16]. Any increase in preterm birth is not to be taken lightly, as preterm birth is a leading cause of neonatal morbidity and mortality and has been shown to have long-term health consequences persisting into adulthood [18]. In particular, preterm birth is associated with chronic lung disease and the development of asthma-like symptoms in adolescence and adulthood [19].

Studies on the risk of adverse birth outcomes and their role on the development of asthma are inconclusive. A systematic meta-analysis found that infants born preterm have a 7-36% increased risk of developing asthma compared to term, with a trend of increasing risk as gestational age decreases [20]. In a case-control study of young adults, preterm birth and low birth weight were associated with a reduced risk of developing allergen responses later in life [21]. A national cohort study in Sweden found that adults born extremely preterm (23-27 weeks gestation) were 2.4 times more likely to develop asthma, while no additional risk of asthma was seen in those born preterm and late-preterm (28-36 weeks) [22]. Conversely, a retrospective cohort study found children born late-preterm (34-36 weeks) had a 68% increased odds of developing asthma by 18 months [23]. Part of the discrepancy in these findings with late-preterm birth may be due to the age at which asthma was assessed, as the effect of preterm birth on asthma appears to be strongest in young age and decreases into adulthood [20].

The proposed mechanisms for how preterm birth may affect the risk of asthma development include genetic, environmental, and perinatal factors. Several studies have suggested that the increased risk of developing asthma may be due to stunted fetal development and bacterial infections, such as chorioamnionitis [20, 23, 24]. Indeed, premature birth results in a deficit in lung development which may leave lungs more vulnerable to insult from environmental factors such as cigarette smoke [19]. Another hypothesis for the observed associations is that preterm birth and asthma share a common genetic determinant [22].

Air Pollution and Childhood Respiratory Health

Exposure to ambient air pollution is a risk factor for asthma and other respiratory diseases that merits special attention. It is one of the few risk factors to which all people are exposed, albeit to varying degrees. Children are particularly vulnerable to effects of air pollution for several reasons including ongoing lung growth and development, incomplete metabolic systems, immature host defenses, high rates of infection with respiratory diseases, and activity patterns that increase their exposure to air pollutants [15].

The reasons for the increased vulnerability of children to adverse health effects from air pollution exposure are both physiological and behavioral. The lungs are not well formed at birth. Eighty percent of the alveoli in the lungs are formed postnatally [25] and full lung functionality not achieved until 6 years of age [26]. During the lung development period, the immature epithelium allow greater permeability of inhaled toxicants leading to damage of the airways in young children [26]. This suggests that early childhood exposures to air pollutants during this critical development process could have lasting detrimental effects on lung function. A Southern CA study found that children living <500 meters from a freeway had substantial deficits in 8-year growth of lung function compared with children living >1500 meters from a freeway, which was independent of regional air quality [27].

Relative to their size, children have much greater exposure to inhaled air pollutants than adults. Children have larger lung surface area per kilogram of body weight than adults and breathe 50% more air per kilogram of body weight than adults [26]. Children also have narrower airways than adults, thus irritants which may elicit a minor response in adults could lead to potentially significant airway obstruction in young children [28].

In addition to the respiratory system, exposure to air pollution in early childhood might have adverse effects on the immune system, which is not fully formed at birth [26]. For example exposure to ambient air pollution, particularly traffic-related pollutants, may increase susceptibility to RSV infection by mediating the immune response and increasing increased inflammation [29]. Children also spend more time playing outdoors than adults, particularly in the afternoon when many pollutant levels are at their highest [26], and therefore may have more direct exposure to ambient concentrations of air pollutants than adults. One study found that children living in communities with high levels of air pollution had decreased lung function growth, with larger deficits seen in the children that spent more time outdoors [25].

The next chapter will discuss approaches for characterizing air pollution exposure and methods for modeling the health effects.

References

1. **Asthma** [http://www.who.int/mediacentre/factsheets/fs307/en/index.html]

2. Moorman JE, Zahran H, Truman BI, Molla MT: **Current asthma prevalence - United States, 2006-2008**. In: *Morbidity and Mortality Weekly Report.* vol. 60. Atlanta, GA: Centers for Disease Control and Prevention; 2011: 84.

3. Centers for Disease Control: **Vital Signs: Asthma Prevalence, Disease Characteristics, and Self-Management Education --- United States, 2001--2009**. *Morbidity and Mortality Weekly Report* 2011, **60**(17):547-552.

4. Akinbami LJ, Moorman JE, Liu X: **Asthma prevalence, health care use, and mortality: United States, 2005-2009**. *Natl Health Stat Report* 2011(32):1-14.

5. **U.S. asthma rates continue to rise** [http://www.cdc.gov/media/releases/2011/p0503_vitalsigns.html]

6. Dougherty RH, Fahy JV: **Acute exacerbations of asthma: epidemiology, biology and the exacerbation-prone phenotype**. *Clin Exp Allergy* 2009, **39**(2):193-202.

7. **Trends in asthma morbidity and mortality** [http://www.lungusa.org/finding-cures/our-research/trend-reports/asthma-trend-report.pdf]

8. Centers for Disease Control and Prevention: **National surveillance for asthma--1980 - 2004**. *Morbidity and Mortality Weekly Report* 2007, **56**(S 108):1-14; 18-54.

9. Global Initiantive for Asthma (GINA): **Global Strategy for Asthma Management and Prevention**. In.; 2010.

10. Forno E, Celedon JC: **Predicting asthma exacerbations in children**. *Curr Opin Pulm Med* 2012, **18**(1):63-69.

11. Kit BK, Simon AE, Ogden CL, Akinbami LJ: **Trends in preventive asthma medication use among children and adolescents, 1988-2008**. *Pediatrics* 2012, **129**(1):62-69.

12. US Department of Health and Human Services, Agency for Healthcare Research and Quality: **2008 National Healthcare Disparities Report**. In. Rockville, MD: US Department of Health and Human Services, Agency for Healthcare Research and Quality; 2008

13. Schatz M, Clark S, Emond JA, Schreiber D, Camargo CA: **Sex differences among children 2-13 years of age presenting at the emergency department with acute asthma**. *Pediatric Pulmonology* 2004, **37**:523-529.

14. Akinbami LJ, Moorman JE, Liu X: **Asthma prevalence, health care use and mortality: United States, 2005-2009**. *National Health Statistics Reports* 2011, **32**.

15. Samoli E, Nastos PT, Paliatsos AG, Katsouyanni K, Priftis KN: **Acute effects of air pollution on pediatric asthma exacerbation: evidence of association and effect modification**. *Environ Res* 2011, **111**(3):418-424.

16. McCormick MC, Litt JS, Smith VC, Zupancic JA: **Prematurity: an overview and public health implications**. *Annu Rev Public Health* 2011, **32**:367-379.

17. Martin JA, Hamilton BE, Ventura SJ, Osterman MJ, Kirmeyer S, Mathews T, Wilson E: **Births: Final Data for 2009.** In: *National Vital Statistics Reports.* vol. 60. Hyattsville, MD: National Center for Health Statistics; 2011.

18. Beck S, Wojdyla D, Say L, Betran AP, Merialdi M, Requejo JH, Rubens C, Menon R, Van Look PF: **The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity**. *Bulletin of the World Health Organization* 2010, **88**:31-38.

19. Baraldi E, Filippone M: **Chronic lung disease after premature birth**. *N Engl J Med* 2007, **357**(19):1946-1955.

20. Jaakkola JJ, Ahmed P, Ieromnimon A, Goepfert P, Laiou E, Quansah R, Jaakkola MS: **Preterm delivery and asthma: a systematic review and meta-analysis**. *J Allergy Clin Immunol* 2006, **118**(4):823-830.

21. Siltanen M, Wehkalampi K, Hovi P, Eriksson JG, Strang-Karlsson S, Jarvenpaa AL, Andersson S, Kajantie E: **Preterm birth reduces the incidence of atopy in adulthood**. *J Allergy Clin Immunol* 2011, **127**(4):935-942.

22. Crump C, Winkleby MA, Sundquist J, Sundquist K: **Risk of asthma in young adults who were born preterm: a Swedish national cohort study**. *Pediatrics* 2011, **127**(4):e913-920.

23. Goyal NK, Fiks AG, Lorch SA: **Association of late-preterm birth with asthma in young children: practice-based study**. *Pediatrics* 2011, **128**(4):e830-838.

24. Dombkowski KJ, Leung SW, Gurney JG: **Prematurity as a predictor of childhood asthma among low-income children**. *Ann Epidemiol* 2008, **18**(4):290-297.

25. Kim JJ: **Ambient air pollution: health hazards to children**. *Pediatrics* 2004, **114**(6):1699-1707.

26. Schwartz J: **Air pollution and children's health**. *Pediatrics* 2004, **113**(4 Suppl):1037-1043.

27. Gauderman WJ, Vora H, McConnell R, Berhane K, Gilliland F, Thomas D, Lurmann F, Avol E, Kunzli N, Jerrett M *et al*: **Effect of exposure to traffic on lung development from 10 to 18 years of age: a cohort study**. *Lancet* 2007, **369**(9561):571-577.

28. Salvi S: **Health effects of ambient air pollution in children**. *Paediatr Respir Rev* 2007, **8**(4):275-280.

29. Harrod KS, Jaramillo RJ, Rosenberger CL, Wang SZ, Berger JA, McDonald JD, Reed MD: **Increased susceptibility to RSV infection by exposure to inhaled diesel engine emissions**. *Am J Respir Cell Mol Biol* 2003, **28**(4):451-463.

Chapter 4: Epidemiologic Methods for Multipollutants

The past decade has seen a shift away from studying single-pollutant effects towards multipollutant effects, and in 2004 the National Research Council recommended a multipollutant approach to air quality management be adopted [1]. According to the EPA, a "multipollutant approach takes into account that humans and ecosystems are exposed to many air pollutants at the same time" [2]. Humans breathe in a mixture of pollutants so it seems logical that we would want to study the health effects of these multipollutant mixtures; however, how these mixtures are studied poses a challenge. Firstly it is important to clarify the multipollutant research question being put forth and secondly to understand the different statistical tools available for answering the question.

There is not yet consensus about what the term "multipollutant mixture" means with regards to air pollution and health studies. Conceptual issues that underlie much of the multipollutant health discourse include pollutant covariation, joint effects, pollutant interaction, novel exposure definitions, and disentangling effects. Pollutants frequently covary, which may be due to a common source, weather conditions, geographical features or photochemistry. In multipollutant studies such covariation can cause confounding of the health effect and must be considered in the analysis and interpretation. Often of interest in multipollutant research is the joint effect of two or more pollutants, whether it is from the perspective of source apportionment (e.g. what is the health effect of reducing emissions from a given source?) or from an interest in interaction (e.g. when combined do pollutants A and B act synergistically or antagonistically on a given outcome?). In this context, pollutant interaction is meant to encompass statistical interaction, occurring from having more than one pollutant in then model, in which a departure from additivity or multiplicity of effect is observed; biological interaction, in which two pollutants co-participate in a causal mechanism (e.g. black carbon may act as a carrier mechanism transporting O_3 into the distal areas of the lung, leading to inflammation [3]); or chemical interaction, in which the presence of two substances precipitates a reaction and the formation of a new substance (e.g. sulfuric acid aerosols being neutralized by exogenous ambient ammonia or endogenously derived ammonia [4]). Pollutant mixtures are sometimes studied in order to characterize the health risk posed by a set of exposures. A prime example of this is source apportionment, in which statistical techniques are used to assign measured pollutants to a single source in order to characterize the health risk presented by the source. Finally, Klein et al. suggests that multipollutant mixtures may be studied in order to disentangle the single effects of a certain pollutant, chemical, component or species from the joint effects. The methodological and statistical techniques used to study multipollutant mixtures will depend upon which of these conceptual issues is being addressed.

If the primary interest is the joint effect of two or more pollutants, studying this multipollutant effect is not as simple as putting all the pollutants into a generalized linear model and examining the resulting coefficients. Unlike most multiple exposure studies, pollutant exposures are not independent and placing more than one in the same model often leads to parameter estimates that are unstable due to their high correlation [5]. In fact, pollutants are so intrinsically intertwined with one another that it is nearly impossible to disentangle the effect of a single pollutant. It is commonly understood that even in single-pollutant studies, the health burden attributed to a pollutant is more likely caused by multiple pollutants [6]. Issues to consider when putting multiple pollutants into a single model include: differential measurement error, colinearity of pollutants, whether or not one pollutant is an intermediate of the other (i.e. in the causal pathway), interactions between pollutants as well as meteorological variables, and that the pollutants may be surrogates for the same underlying mixture (i.e. may represent a traffic source) [7]. Research groups and individual scientists have put forth a myriad of different suggestions for how multipollutants can be assessed. What follows is a description of the different study design and modeling approaches used in this dissertation for identifying pollution-related health effects.

Study Design

Before diving into the specifics of statistical approaches it is important to understand the way in which air pollution is measured, as this dictates, to some degree, the types of analyses that can be performed. Ambient air pollution, that is the freely moving air in the outdoor environment, is typically measured by ground station monitors that may be calibrated to measure the concentration of one or several pollutants. The simplest way to represent a population's exposure to ambient air pollution is to use air pollution data from a single central monitor [8]. Additional approaches include nearest monitor, spatial averaging, kriging, and population-weighted averaging. Which method is best depends on the spatiotemporal variability of the pollutant(s) of interest, the population to which inference of exposure is being made, and the amount of measurement error one is willing to tolerate. In general primary pollutants tend to be more spatially heterogeneous, while secondary pollutants are spatially homogenous [9].

Time Series

When both the exposure and outcome make use of aggregate data, in which the unit of observation is a group of people, the study is referred to as ecological [10]. Ecological studies are common in air pollution epidemiology, where the exposure is typically a series of measurements from a single monitor and the outcome is daily counts of disease or death, and are frequently analyzed using time series analysis.

In a time series study, a population is followed through time with exposure and confounders measured at the population level and the outcome frequently recorded as counts of binary events. This aggregation results in a loss of information about the relationship between the exposure and the outcome due to reduced variation in the exposure; however, this loss in power due is overwhelmed by the gains in power that result from the size of the population that can feasibly be studied in time series designs [11]. Although time series studies do not offer precise information about the relationship between air pollution and health for a specific individual, they do provide great insight into the health impact of air pollution, as measured by central monitoring sites, on a population [12]. In most time series studies the unit of observation is the day. Though we might expect the underlying risk of an individual to vary based on factors such as age, smoking, and SES, because time series studies are concerned with the aggregate risk of the population these factors will not affect the measure of overall risk. That is to say the distributions of age, smoking and SES in the population do not vary from day to day with the ambient exposure and thus do not need to be considered for confounding in time series analysis; an important advantage of time series studies [13]. Only factors that co-vary with ambient pollution levels (e.g. meteorology, season, week-day, etc.) have the potential to confound a time series study and therefore must be adequately controlled for in the model.

Time series studies typically model the expected number of events in a given day, which is suggestive of a Poisson distribution. The probability of observing y events on a given day, assuming a mean of λ is:

$$P(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

In a classic Poisson model the variance is equal to λ; however, it is common in count studies for the variance to exceed the mean, which is referred to as overdispersion. In a well-specified model, overdispersion may occur when the underlying population is not homogenous in their risk of morbidity [12]. When the model is not well-specified, excess variation may be a sign of unmeasured predictors, or that the variable does not follow a Poisson distribution. Overdispersion in the Poisson variance can be accounted for by scaling the variance proportional to λ [13].

In Poisson regression, also referred to as log-linear regression, the log of the expected daily count is modeled as a linear function of predictors. In air pollution studies the model typically takes the form: $log(E(Y_t)) = \sum_i \beta_i X_{ti}$ where $E(Y_t)$ represents the expected number of events on a given day and X_{ti} represents the predictor variables, including air pollution exposure. Poisson generalized linear models (GLM), which allow for the control of temporal trends with splines and can account for overdispersion, are commonly used in air pollution and health studies [14-17].

Case-crossover

Case-crossover studies offer an alternative to the time series approach in which the control of individual level confounders is more explicit. In the case-crossover model, the study population consists solely of cases, who serve as their own controls in the analysis. Rather than choose an external comparison group, the referent period is chosen from the case's history or future [18]. Much like a matched case-control design, the day on which a case observed an event (the "index" period) is matched with similar non-event days (the "referent" period). The selection of this referent period is based on days within a short time period of exposure to minimize time-varying factors. One common choice in the selection of referent periods is to match on month and day of week [19], while others have chosen to match on month and maximum temperature [20].

A major advantage of the case-crossover design is that individuals serve as their own controls. Potential confounders that do not vary with ambient air pollution levels will not confound a case-crossover study. This includes personal smoking status, additional indoor pollution sources, age, and gender.

The case-crossover framework has been shown to be a special case of the time series approach when exposure is common to the cohort at each time, as in air pollution studies [21]. An important distinction between the two approaches is how the measures of association are modeled. Conditional logistic regression (CLR) is typically used to get estimates of the odds ratio in case-crossover studies, whereas time series studies utilize a log-linear regression approach to express the expected number of counts each day [21]. When the number of time intervals and case groupings is large CLR approach is computationally inefficient compared to log-linear regression [21]. Another disadvantage of case-crossover studies is that conditional logistic regression cannot account for overdispersion, as is typically done in a time series approach.

This dissertation will employ both time series and case-crossover designs to examine air pollution and health associations.

Confounding Control

As previously mentioned, when using a time series or case-crossover design, factors that co-vary with ambient pollution levels have the potential to confound the study and therefore must be adequately controlled for in the model. Of particular concern for air pollution and health time series studies are long-term trends in morbidity and mortality, meteorology, seasonality, and infectious diseases. Many of these variables show systematic variation over the course of the year, which will induce a correlation that is not necessarily causal. To reduce these spurious associations and focus on what is potentially causal it is necessary to remove these time-varying patterns in the data.

Splines are often used to capture and control for trends in data that do not follow a simple parametric (i.e. linear) form. In spline models the variable exhibiting the temporal trend is categorized and the boundaries between these categories are called the “knots” of the spline [10]. The degree of “smoothness” of the spline is dependent upon the number of knots, which corresponds to the degrees of freedom [22]. How closely the splines control for trends is determined by the number of knots. Over-controlling runs the risk of missing the exposure effect (bias towards the null), while under-controlling may inappropriately attribute temporal trends to the exposure effect (bias away from null). Cubic splines fit a cubic polynomial within each interval and the data are required to join smoothly at the interval. In Chapter 7 we employ splines in a time series study to model the associations of air pollution sources and pediatric asthma. We run sensitivity analyses, varying the number of knots in our day-of-year spline, to consider how the results would change under differing degrees of smoothness in the spline.

In addition to controlling for smooth temporal trends, it is important to consider meteorological control. Air pollution and weather are intrinsically intertwined (e.g. sunlight is necessary in the formation of ozone) and the association between weather and health has been well documented, making confounding a practical concern. As a result, it is crucial that models contain adequate control for the relevant meteorological covariates. Extensive analyses and investigations conducted by members of the Study of Particles of Health in Atlanta (SOPHIA) have suggested that for the assessment of asthma emergency department visits, cubic terms for maximum temperature and dew point typically provide adequate control.

When the effect of temperature or air pollution is delayed by one or may days then a lag model may be best. Several studies looking and the effects of extreme temperature on mortality have found the heat effects to be more immediate [23, 24], while the effects of cold can be delayed by up to 4 weeks [23, 25-27]. In general, the more flexible the lag model is, the better it will control for confounding (i.e. by temperature in air pollution studies). The trade-off, however, is in interpretability of the lag model.

There are several different approaches for dealing with lag effects in time series studies. The simplest is to include each lag term in the model with its own coefficient, sometimes called a “unconstrained distributed lag model” [28]. A disadvantage is that if the number of lags is large, this approach will use up many degrees of freedom [29]. The coefficients may also become unstable due to multicollinearity, as pollution on a given day is highly correlated with pollution on previous days [29]. An alternative is to use a constrained distributed lag model, which requires the coefficients for all of the lag parameters to be equal. One common approach is to constrain the shape of the lag coefficients to follow a polynomial function. For example, a third degree polynomial distributed lag function, p(x), can be stated as:

$$p(x) = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \alpha_3 X^3$$

where X represents the lagged values. The simplest constrained lag approach is to use a zero degree polynomial lag structure, which is equivalent to the moving average approach [29]. Air pollution and health studies have often found that using a multi-day moving average of air pollution exposure (2- or 3-day averages) has a better fit than single day pollution or longer moving averages, suggesting that the effect of an increase in pollution on a single day is distributed across several days [28]. Ultimately, a balance must be struck between too much constraint, which risks producing a distorted shape, and too little constraint, which can result in estimates that are too noisy to be informative [30].

Analytic Approaches

A simplistic way to characterize a pollutant mixture is to use a single pollutant as a surrogate for an underlying mixture or emission source. This has historically been the most common approach for characterizing pollution mixtures [31-34]. For example SO_2 is often thought of as a surrogate for power plant emissions, while NO_2 and CO may represent vehicular traffic [7]. Single pollutant models are popular because they are simpler to conduct, easy to interpret, and can be directly applied to air quality regulation, which is conducted independently for each pollutant [35].

Nonetheless, there is something dissatisfying about single pollutant models because we know the air is a complex mixture of pollutants that are inhaled simultaneously. Furthermore, there are important limitations to single pollutant models, including confounding by correlated pollutants, which may lead to an overestimation of the main effects. In the spirit of mixtures and an effort to better characterize exposures as they occur, multipollutant models have increasingly been employed in air pollution epidemiology [36-39]. In multipollutant models, two or more pollutants are included in the same model to get the combined effect. While these models have the potential to better capture exposure, they also bring new complexities to the model. For example, it has been shown that if the pollutants included in the model have different measurement error, there is the potential for the effect of the more poorly measured pollutant to be transferred to the better measured pollutant, resulting in bias of the point estimates [40]. When two pollutants are both independent risk factors and correlated with one another, then including both in the model will control for confounding and lead to more accurate measures of association. Multipollutant models also offer the possibility to evaluate pollutant-pollutant interactions, which some studies have suggested is a real concern [37]; however, as the number of interaction terms increase the power to detect any significant association will be reduced. Additionally, when two or more highly correlated pollutants are included in the same model along with their interaction terms, such as PM10 and NO_2, the results can become unstable [35].

Source Apportionment

All ambient pollution originates from a source. Even secondary pollutants such as ozone, sulfates and nitrates require primary source emissions to form. Another way of characterizing multipollutants is to characterize the main source emissions. This is done by apportioning the particles found in the air to their respective sources. Chemical species that are characteristic of a given source profile may serve as adequate tracers of that source when present in samples above their limit of detection. For example, silicon may be used as a tracer for soil components of $PM_{2.5}$ [41], levoglucosan as a tracer for biomass burnings [42], and hopanes and elemental carbon for vehicle emissions [43].

When the sources in the region are known, a method known as chemical mass balance [44] can be used to apportion the mass from total $PM_{2.5}$ to the relevant sources based on the presence of different tracer species [45]. When the sources are unknown a latent variable analysis is needed, approaches of which include principle components analysis, UNMIX [46, 47] and positive matrix factorization [48].

A benefit of source apportionment is that it provides a more detailed description of particulate exposure. Particulate matter is itself a mixture of many different chemical components and there is evidence to support that these components have varying degrees of toxicity [41, 49-51]. Current EPA practice regulates all $PM_{2.5}$ mass equally; it does not distinguish between particles shown to be more or less toxic. As the literature for source apportionment epidemiology grows and more evidence is available to identify the most harmful sources, there will be increasing pressure on the EPA to change their regulatory practices and start regulation according to source emissions, rather than total $PM_{2.5}$ mass.

Dimension Reduction

A practical challenge of multipollutant studies is how to handle the myriad of measured pollutant variables, which have the potential to exceed the number of observations. Some type of dimension reduction is often needed in order to reduce the data to the key set of predictors and remove variables that do not have any explanatory power [35]. One commonly used method for reducing dataset dimensionality into a core set of latent factors is principal component analysis (PCA) [7, 52, 53]. One drawback of PCA is that the principal components are constructed solely on the pollutant covariates and without consideration for how these covariates are related to the health outcome [5]. This lack of consideration for the outcome or response variable is typically referred to as an "unsupervised" approach. Supervised PCA (SPCA) was developed by Bair et al. to account for the response variable in the formation of the components [54] and has since been modified by Roberts et al. to fit the multipollutant context [5]. An alternative to SPCA, regression shrinkage techniques, such as ridge regression and LASSO, help to identify a subset of key variables that are highly predictive of the response while shrinking to zero the coefficients of non-predictor variables [55].

Classification and Regression Trees

Classification and regression trees (C&RT) offer an alternative to traditional regression models. C&RT uses recursive partitioning to split data into groups that contain similar responses for the outcome variable. No assumptions of a parametric relationship or monotonic relationship with the outcome are required in C&RT. C&RT techniques have been used in clinical medicine to develop diagnostic algorithms for predicting disease presence and severity [56, 57]. C&RT can also be used to identify high-order interactions in the data, even when the main effects may be relatively weak, and does not require the interaction terms be pre-specified. Many research groups, particularly in the field of genetics have used recursive partitioning to identify interaction among many predictor variables [58-60]; however to our knowledge only one study has used C&RT to identify interaction effects of temperature and ambient air pollution on total mortality [61].

For studies of air pollution the outcome is frequently a continuous or count variable (i.e. emergency department visits per day) and thus a regression tree approach is most appropriate. A regression tree always starts with a root node that contains the sample of data from which the tree will be grown. The data are then split into two child nodes based on the value of a predictor variables such that the branching results in child nodes that are "more pure" than the parent node. At each node, the variable that is most strongly associated with the response variable (i.e. produces the most impurity improvement) will be used for the split. The commonly used approach for selecting the best split is a least squares deviation criterion, based on within-node variance:

Impurity at node $\tau = i(\tau) \ = \ \sum(Y_i - \bar{Y}(\tau))^2$ [62](*section 10.3)*

This process continues until the sample space is partitioned by a sequence of binary splits into *n* terminal nodes, such that all observations in a given terminal node have the same predicted value for the outcome variable. The user can set *a priori* the minimum number of observations to be included at each terminal node.

A full tree is formed once there are no further splits that result in a reduction of node impurity *or* the minimum number of observations per terminal node has been reached. Full trees must then be "pruned". Pruning is done to cope with overfitting and to find the tree that is most predictive of the outcome while being the least vulnerable to noise in the data. Partitioning and pruning can be thought of as synonymous to forward and backward model selection procedures [62]. Similarly, the number of nodes in a tree is analogous to the model degrees of freedom [63].

The binary splits in the final pruned regression tree represent points at which the data are stratified based on the response variable. The mean responses at the terminal nodes can be compared to gain a better understanding of how the sequential stratifications (i.e. interactions) of the predictor variables modify the response. Terminal nodes with significantly different response variables suggest the binary splits may have identified an important interaction in the data.

Although several statistical packages are capable of running C&RT, including the 'rpart' and 'tree' packages in R and S-plus, CART® by Salford Systems, SYSTAT, and DTREG, they are limited to varying degrees in their applicability to epidemiologic research. One key feature lacking from all of these packages is the ability to control for confounding at the time of tree partitioning. When the outcome is continuous the researcher can avoid this pitfall by using residuals from the model that controls for confounding in tree partitioning; however, there is no straight forward way to do this when the outcome is Poisson. Additionally, most existing packages choose the best split according to node impurity. While this is ideal for prediction models, when the goal is to identify statistically significant associations splitting on node impurity will not always result in an optimal tree.

The next chapter provides a description of a modified classification and regression tree algorithm that is better-suited for epidemiologic research and the study of air pollution mixtures in particular.

References

1. National Research Council: **Air quality management in the United States**. In. Edited by National Academic Press. Washington, D.C.; 2004.

2. Costa D: **One atmosphere: the intersection of air quality and climate change**. In. Edited by Office of Research and Development: U.S. Environmental Protection Agency; 2010.

3. Jakab GJ, Hemenway DR: **Concomitant exposure to carbon black particulates enhances ozone-induced lung inflammation and suppression of alveolar macrophage phagocytosis**. *J Toxicol Environ Health* 1994, **41**(2):221-231.

4. Brown JS, Graham JA, Chen LC, Postlethwait EM, Ghio AJ, Foster WM, Gordon T: **Panel discussion review: session four--assessing biological plausibility of epidemiological findings in air pollution research**. *J Expo Sci Environ Epidemiol* 2007, **17 Suppl 2**:S97-105.

5. Roberts S, Martin MA: **Using supervised principal components analysis to assess multiple pollutant effects**. *Environ Health Perspect* 2006, **114**(12):1877-1882.

6. Mauderly JL, Burnett RT, Castillejos M, Ozkaynak H, Samet JM, Stieb DM, Vedal S, Wyzga RE: **Is the air pollution health research community prepared to support a multipollutant air quality management framework?** *Inhal Toxicol* 2010, **22 Suppl 1**:1-19.

7. Kim JY, Burnett RT, Neas L, Thurston GD, Schwartz J, Tolbert PE, Brunekreef B, Goldberg MS, Romieu I: **Panel discussion review: session two--interpretation of observed associations between multiple ambient air pollutants and health effects in epidemiologic analyses**. *J Expo Sci Environ Epidemiol* 2007, **17 Suppl 2**:S83-89.

8. Ivy D, Mulholland JA, Russell AG: **Development of Ambient Air Quality Population-Weighted Metrics for Use in Time-Series Health Studies**. *Journal of the Air & Waste Management Association* 2008, **58**(5):711-720.

9. Wade KS, Mulholland JA, Marmur A, Russell AG, Hartsell B, Edgerton E, Klein M, Waller L, Peel JL, Tolbert PE: **Effects of instrument precision and spatial variability on the assessment of the temporal variation of ambient air pollution in Atlanta, Georgia**. *J Air Waste Manag Assoc* 2006, **56**(6):876-888.

10. Rothman KJ, Greenland S, Lash TL: **Modern Epidemiology, 3rd Edition**, Third Edition edn. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

11. Sheppard L: **Acute air pollution effects: consequences of exposure distribution and measurements**. *Journal of Toxicology and Environmental Health, Part A* 2005, **68**:1127-1135.

12. Koken PJ, Piver WT, Ye F, Elixhauser A, Olsen LM, Portier CJ: **Temperature, air pollution, and hospitalization for cardiovascular diseases among elderly people in Denver**. *Environ Health Perspect* 2003, **111**(10):1312-1317.

13. Schwartz J, Spix C, Touloumi G, Bacharova L, Barumamdzadeh T, le Tertre A, Piekarksi T, Ponce de Leon A, Ponka A, Rossi G *et al*: **Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions**. *J Epidemiol Community Health* 1996, **50 Suppl 1**:S3-11.

14. Peel JL, Tolbert PE, Klein M, Metzger KB, Flanders WD, Todd K, Mulholland JA, Ryan PB, Frumkin H: **Ambient air pollution and respiratory emergency department visits**. *Epidemiology* 2005, **16**(2):164-174.

15. Strickland MJ, Darrow LA, Klein M, Flanders WD, Sarnat JA, Waller LA, Sarnat SE, Mulholland JA, Tolbert PE: **Short-term associations between ambient air pollutants and pediatric asthma emergency department visits**. *Am J Respir Crit Care Med* 2010, **182**(3):307-316.

16. Metzger KB, Tolbert PE, Klein M, Peel JL, Flanders WD, Todd K, Mulholland JA, Ryan PB, Frumkin H: **Ambient air pollution and cardiovascular emergency department visits**. *Epidemiology* 2004, **15**(1):46-56.

17. Ito K, Thurston GD, Silverman RA: **Characterization of PM2.5, gaseous pollutants, and meteorological interactions in the context of time-series health effects models**. *J Expo Sci Environ Epidemiol* 2007, **17 Suppl 2**:S45-60.

18. Lu Y, Symons JM, Geyh AS, Zeger SL: **An approach to checking case-crossover analyses based on equivalence with time-series methods**. *Epidemiology* 2008, **19**(2):169-175.

19. Levy D, Sheppard L, Checkoway H, Kaufman J, Lumley T, Koenig J, Siscovick D: **A case-crossover analysis of particulate matter air pollution and out-of-hospital primary cardiac arrest**. *Epidemiology* 2001, **12**(2):193-199.

20. Darrow LA, Klein M, Strickland MJ, Mulholland JA, Tolbert PE: **Ambient air pollution and birth weight in full-term infants in Atlanta, 1994-2004**. *Environ Health Perspect* 2011, **119**(5):731-737.

21. Lu Y, Zeger SL: **On the equivalence of case-crossover and time series methods in environmental epidemiology**. *Biostatistics* 2007, **8**(2):337-344.

22. Armstrong B: **Models for the relationship between ambient temperature and daily mortality**. *Epidemiology* 2006, **17**(6):624-631.

23. Anderson BG, Bell ML: **Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States**. *Epidemiology* 2009, **20**(2):205-213.

24. Medina-Ramon M, Zanobetti A, Cavanagh DP, Schwartz J: **Extreme temperatures and mortality: assessing effect modification by personal characteristics and specific cause of death in a multi-city case-only analysis**. *Environ Health Perspect* 2006, **114**(9):1331-1336.

25. Carder M, McNamee R, Beverland I, Elton R, Cohen GR, Boyd J, Agius RM: **The lagged effect of cold temperature and wind chill on cardiorespiratory mortality in Scotland**. *Occupational and environmental medicine* 2005, **62**(10):702-710.

26. Diaz J, Garcia R, Lopez C, Linares C, Tobias A, Prieto L: **Mortality impact of extreme winter temperatures**. *Int J Biometeorol* 2005, **49**(3):179-183.

27. Keatinge WR, Donaldson GC: **Mortality related to cold and air pollution in London after allowance for effects of associated weather patterns**. *Environ Res* 2001, **86**(3):209-216.

28. Schwartz J: **The distributed lag between air pollution and daily deaths**. *Epidemiology* 2000, **11**(3):320-326.

29. Pope CA, 3rd, Schwartz J: **Time series for the analysis of pulmonary health data**. *Am J Respir Crit Care Med* 1996, **154**(6 Pt 2):S229-233.

30. Zanobetti A, Schwartz J, Samoli E, Gryparis A, Touloumi G, Atkinson R, Le Tertre A, Bobros J, Celko M, Goren A *et al*: **The temporal pattern of mortality responses to air pollution: a multicity assessment of mortality displacement**. *Epidemiology* 2002, **13**(1):87-93.

31. Sarnat JA, Schwartz J, Catalano PJ, Suh HH: **Gaseous pollutants in particulate matter epidemiology: confounders or surrogates?** *Environ Health Perspect* 2001, **109**(10):1053-1061.

32. Janssen NA, Lanki T, Hoek G, Vallius M, de Hartog JJ, Van Grieken R, Pekkanen J, Brunekreef B: **Associations between ambient, personal, and indoor exposure to fine particulate matter constituents in Dutch and Finnish panels of cardiovascular patients**. *Occupational and environmental medicine* 2005, **62**(12):868-877.

33. Laden F, Neas LM, Dockery DW, Schwartz J: **Association of fine particulate matter from different sources with daily mortality in six U.S. cities**. *Environ Health Perspect* 2000, **108**(10):941-947.

34. Sarnat SE, Suh HH, Coull BA, Schwartz J, Stone PH, Gold DR: **Ambient particulate air pollution and cardiac arrhythmia in a panel of older adults in Steubenville, Ohio**. *Occupational and environmental medicine* 2006, **63**(10):700-706.

35. Dominici F, Peng RD, Barr CD, Bell ML: **Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach**. *Epidemiology* 2010, **21**(2):187-194.

36. Katsouyanni K, Touloumi G, Samoli E, Gryparis A, Le Tertre A, Monopolis Y, Rossi G, Zmirou D, Ballester F, Boumghar A *et al*: **Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 project**. *Epidemiology* 2001, **12**(5):521-531.

37. Mauderly JL, Samet JM: **Is there evidence for synergy among air pollutants in causing health effects?** *Environ Health Perspect* 2009, **117**(1):1-6.

38. Tolbert PE, Klein M, Peel JL, Sarnat SE, Sarnat JA: **Multipollutant modeling issues in a study of ambient air quality and emergency department visits in Atlanta**. *J Expo Sci Environ Epidemiol* 2007, **17 Suppl 2**:S29-35.

39. Jerrett M, Burnett RT, Beckerman BS, Turner MC, Krewski D, Thurston G, Martin RV, van Donkelaar A, Hughes E, Shi Y *et al*: **Spatial analysis of air pollution and mortality in California**. *Am J Respir Crit Care Med* 2013, **188**(5):593-599.

40. Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery D, Cohen A: **Exposure measurement error in time-series studies of air pollution: concepts and consequences**. *Environ Health Perspect* 2000, **108**(5):419-426.

41. Sarnat JA, Marmur A, Klein M, Kim E, Russell AG, Sarnat SE, Mulholland JA, Hopke PK, Tolbert PE: **Fine particle sources and cardiorespiratory morbidity: an application of chemical mass balance and factor analytical source-apportionment methods**. *Environ Health Perspect* 2008, **116**(4):459-466.

42. Simoneit BR, Elias VO: **Detecting organic tracers from biomass burning in the atmosphere**. *Marine pollution bulletin* 2001, **42**(10):805-810.

43. Birch ME, Cary RA: **Elemental carbon-based method for occupational monitoring of particulate diesel exhaust: methodology and exposure issues**. *The Analyst* 1996, **121**(9):1183-1190.

44. Watson JG, Cooper JA, Huntzicker JJ: **The effective variance weighting for least squares calculations applied to the mass balance receptor model**. *Atmospheric Environment* 1984, **18**(7):1347-1355.

45. Hopke PK: **The use of source apportionment for air quality management and health assessments**. *J Toxicol Environ Health A* 2008, **71**(9-10):555-563.

46. Henry RC, Kim BM: **Extension of self-modeling curve resolution to mixtures of more than three components. Part 1: finding the basic feasible region.** . *Chemom Intell Lab Systems* 1990, **8**:205–216.

47. Kim B.M., Henry RC: **Extension of self-modeling curve resolution to mixtures of more than three components. Part 2: finding the complete solution**. *Chemom Intell Lab Systems* 1999, **49**:67-77.

48. Paatero P, Tapper U: **Positive matrix factorization--A nonnegative factor model with optimal utilization of error-estimates of data values**. *Environmetrics* 1994, **5**(2):111-126.

49. Andersen ZJ, Wahlin P, Raaschou-Nielsen O, Scheike T, Loft S: **Ambient particle source apportionment and daily hospital admissions among children and elderly in Copenhagen**. *J Expo Sci Environ Epidemiol* 2007, **17**(7):625-636.

50. Bell ML, Ebisu K, Leaderer BP, Gent JF, Lee HJ, Koutrakis P, Wang Y, Dominici F, Peng RD: **Associations of PM Constituents and Sources with Hospital Admissions: Analysis of Four Counties in Connecticut and Massachusetts (USA) for Persons >/= 65 Years of Age**. *Environ Health Perspect* 2013.

51. Gent JF, Koutrakis P, Belanger K, Triche E, Holford TR, Bracken MB, Leaderer BP: **Symptoms and medication use in children with asthma and traffic-related sources of fine particle pollution**. *Environ Health Perspect* 2009, **117**(7):1168-1174.

52. Yu TY, Chang IC: **Spatiotemporal features of severe air pollution in northern Taiwan**. *Environ Sci Pollut Res Int* 2006, **13**(4):268-275.

53. Wilhelm M, Ghosh JK, Su J, Cockburn M, Jerrett M, Ritz B: **Traffic-related air toxics and preterm birth: a population-based case-control study in Los Angeles County, California**. *Environ Health* 2011, **10**:89.

54. Bair E, Hastie T, Paul D, Tibshirani R: **Prediction by supervised principal components**. *Journal of the American Statistical Association* 2006, **101**:119-137.

55. Tibshirani R: **Regression shrinkage and selection via LASSO**. *Journal of the Royal Statistical Society* 1996, **Series B (Methodological) 58**(1):267-288.

56. Samanta B, Bird GL, Kuijpers M, Zimmerman RA, Jarvik GP, Wernovsky G, Clancy RR, Licht DJ, Gaynor JW, Nataraj C: **Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms**. *Artif Intell Med* 2009, **46**(3):201-215.

57. Potts JA, Gibbons RV, Rothman AL, Srikiatkhachorn A, Thomas SJ, Supradish PO, Lemon SC, Libraty DH, Green S, Kalayanarooj S: **Prediction of dengue disease severity among pediatric Thai patients using early clinical laboratory indicators**. *PLoS Negl Trop Dis* 2010, **4**(8):e769.

58. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P: **Identifying SNPs predictive of phenotype using random forests**. *Genet Epidemiol* 2005, **28**(2):171-182.

59. Garcia-Magarinos M, Lopez-de-Ullibarri I, Cao R, Salas A: **Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction**. *Ann Hum Genet* 2009, **73**(Pt 3):360-369.

60. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P: **Screening large-scale association study data: exploiting interactions using random forests**. *BMC Genet* 2004, **5**:32.

61. Hu W, Mengersen K, McMichael A, Tong S: **Temperature, air pollution and total mortality during summers in Sydney, 1994-2004**. *Int J Biometeorol* 2008, **52**(7):689-696.

62. Zhang H, Singer BH: **Recursive Partitioning and Applications**, Second Edition edn. New York: Springer; 2010.

63. Therneau TM, Atkinson EJ: **An Introduction to Recursive Partitioning Using the RPART Routines**. In: *Technical Report #61*. Mayo Foundation; 1997.

Chapter 5: Classification and Regression Trees for Epidemiologic Research (Study 1)

Introduction

Every day we breathe a blend of air pollutants, ingest an assortment of nutrients, and are influenced by a unique combination of genes. Throughout the course of a day and lifetime our total exposure can be conceptualized as a complex mixture of different individual exposures. Advances in science have improved our ability to measure these exposures; a major challenge is how best to characterize and relate these mixtures to health endpoints.

Characterization of mixtures for epidemiologic research depends upon both the data that can be obtained as well as the research question of interest. For some research questions interest may center on estimating the combined “joint effects” of two or more individual exposures on a given outcome. Encompassed in this issue of joint effects is the concept of interaction. While some joint effects may be indicative of interaction, it is not always the case. For example, given an additive or multiplicative scale, exposures A and B may combine synergistically, antagonistically, or without interaction to promote disease, and our conceptualization of joint effects encompasses all of these. Here we refer to “interaction” as statistical interaction or effect measure modification, that is a deviation from the expected independent joint effect of two or more risk factors[1].

Statistical interaction is often assessed by including the product of two or more risk factors (exposures) in a regression model and using statistical tests to determine whether the resulting coefficient differs significantly from zero. As the number of exposures increases, the number of possible third-, fourth-, fifth-, and higher-order interactions becomes too large to include in any one model and these are rarely considered in conventional analyses. Testing only a specific sub-set of these interaction terms requires substantial *a priori* knowledge about complex interactions. As model complexity grows so does the challenge of interpretation[2]. In addition, parameter estimates may become unstable as the number of interaction terms increases.

In this paper we describe how classification and regression trees (C&RT) can be used as an alternative method for identifying complex joint effects, including interactions, for multiple exposures. The proposed approach expands the applicability of C&RT to epidemiologic research by demonstrating how it can be used for risk estimation. We view this method as a means to generate hypotheses about joint effects that may merit further investigation. We illustrate this approach with an investigation of the effect of outdoor air pollutant concentrations on emergency department visits for pediatric asthma.

Methods

Data

The data we use to demonstrate our C&RT approach are from the Study of Particles and Health in Atlanta (SOPHIA) [3]. The 3-day moving average population-weighted concentrations of ambient carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), and particulate matter less than 2.5 microns in diameter ($PM_{2.5}$) were calculated using measurements from stationary monitors from January 1st, 1999 – December 31st, 2009 [4]. During the same period, daily counts of hospital emergency department (ED) visits for asthma in children 2-18 years old were collected from all hospitals in the area. We defined emergency department visits for asthma as all visits with an International Classification of Disease, 9th edition code for asthma (493.0-493.9) or wheeze (786.07). For a greater description of this dataset see Strickland et al, 2010 [5].

Conceptual example

We illustrate our method assuming the goal is examining the joint health risks of CO, NO_2, O_3 and $PM_{2.5}$ on ED visits. To simplify this example and aid comprehension, we have chosen to reduce the set of all possible joint effects by classifying the daily concentrations of each pollutant into quartiles. This simplification yields 4^4 or 256 different types of days, each of which can be viewed as a unique mixture. To study the association of health with these types of mixtures we could calculate a risk ratio for every type of day, choosing the days when all pollutants are in their lowest quartile as the referent group. This would result in 255 risk ratios.

This approach quickly becomes cumbersome as the number of pollutants (or quantiles) increases. Furthermore, it is unlikely that the joint effects for every pollutant-quantile combination are of interest. Some of these mixtures may never occur due to pollutant covariation, while statistical power will be lacking for rarely occurring mixtures. In addition, as the number of quantiles used to classify the pollutant concentrations increases, the differences in the joint effects between two adjacent quantiles of the same pollutant may be trivial. In this situation statistical efficiency would be improved if similar days were grouped. But how should days be grouped? C&RT methods address this issue by taking all possible joint effects and collapsing them into groups that have similar predicted values for the outcome through a recursive partitioning process.

Statistical methods

C&RT is a non-parametric regression approach. It represents a supervised form of hierarchical clustering in which the data are sequentially split into dichotomous groups, such that each resulting group contains increasingly similar responses for the outcome [6, 7]. The end product of a typical C&RT analysis is a dendogram illustrating the paths of dichotomous splits. Every tree starts with a “root node” that contains the observations from which the tree will be grown. The observations are then partitioned into two “child nodes” based on the value of an independent predictor variable. The resulting child nodes each contain a subset of the original observations. Each child node may be further partitioned, again based on the value of an independent predictor variable. This process continues until a set of partitioning criteria are no longer met, resulting in terminal nodes. Terminal nodes, by definition, cannot have offspring. The collection of terminal nodes forms a complete partition of the observations in the root node.

When the identification of joint effects is of interest, the C&RT approach offers some potential advantages over traditional parametric modeling approaches. C&RT makes no assumption of a monotonic or parametric relationship with the outcome, is able to identify complex interactions among the predictor variables without *a priori* specification of the interaction terms, and can handle datasets where the number of predictors is high relative to the number of observations. C&RT is a supervised learning approach, meaning it creates partitions based on an outcome variable. This is in contrast to unsupervised learning approaches, such as principal components analysis [8], k-means [9], and self-organizing maps [10], which do not consider the outcome.

Although several statistical packages are capable of running C&RT, including the ‘rpart’ and ‘tree’ packages in R and S-plus, CART® by Salford Systems, SYSTAT, and DTREG, they are limited to varying degrees in their applicability to epidemiologic research. In the health sciences, C&RT is most commonly used as a prediction tool [2]; however, for epidemiologic research, we are more often interested in estimating effects than prediction. In the next section we describe a modified C&RT approach that we believe is more appropriate for effect estimation.

Modified C&RT approach

As a first step, before performing any partitions of the observations, a referent group of days is selected from all study days and held aside; this referent group is not used in tree construction. The purpose of excluding a referent group is to enable statistical comparisons (i.e. risk ratio) between risk associated with days in the terminal nodes and those in the referent group. For our example, we chose as a referent group the days in which all four pollutants were in their lowest quartile. This is analogous to our referent group selection in the conceptual example.

When attempting to estimate causal effects, it is necessary to have a well-specified epidemiologic regression model that controls for confounding. For this example we chose a Poisson generalized linear model using a framework equivalent to the conditional logistic case-crossover model [11], with time trends controlled by matching on weekday, month and year, and meteorology controlled with cubic terms for the three-day moving average: maximum temperature, maximum temperature interacted with an indicator for season, and dew point. A spline for day-of-year with two knots was included to provide additional control for seasonal trends. At this first step the model should not include any of the exposure variables of primary interest (i.e. the pollutant variables). Indicator variables are created representing all possible ways to split days into two groups, using each of the individual exposures in the analysis. The number of indicators needed for each exposure will be one less than the number of distinct levels of the exposure. For example, three indicator variables were created for ozone: one indicator comparing quartiles 1 vs. 2-4, a second comparing quartiles 1 and 2 vs. 3 and 4, and a third comparing quartiles 1-3 vs. 4. This was done for all four pollutants, resulting in 12 indicator variables for the 12 possible splitting points. If one prefers to keep the pollutant variable continuous, indicator variables could be created for every possible comparison. For example if the pollutant contained 80 levels, 79 indicators would be defined. Power may be limited with this approach, due to some joint effects having low representation; however, this is not a limitation of the method but rather a consequence of exploring joint effects that occur infrequently.

Each indicator is then included one at a time in the regression model with control for confounding using all the observations (save for those held out in the referent group). After each run of the model the null hypothesis of independence between the outcome and each of the exposure indicators, conditional on the confounding control, is tested and the *P*-value saved. The *P*-values for all possible exposure indicators from the model runs are compared and the smallest *P*-value below a pre-specified alpha level is selected as the first splitting variable. The observations (excluding the referent group) are then partitioned into two child subsets or nodes, each containing the subset of the original observations according to the indicator variable that produced the optimal split. The process repeats itself for each child node, with the regression model being run separately on the two subsets of data and the best splitting point chosen from among the remaining indicators to further partition the child nodes.

Partitioning stops if a minimum child node size is not met, the null hypothesis cannot be rejected for any of the eligible exposure indicators at a pre-specified alpha level, or no further partitions remain. When any of the stopping criteria are met the node becomes a terminal node. The investigator must specify the significance level (alpha) and minimum node size, though this latter criteria is optional; how conservative the stopping criteria are will partly determine the size of the tree. Consequently there is a trade-off between growing a tree large enough to identify potentially important joint effects and running the risk of over-fitting the tree. In our example we specified a two-sided alpha of 0.15 and a minimum node size of 60 observations. The joint effects for the terminal nodes were calculated by including indicator variables for each terminal node simultaneously in the previously described case-crossover model, with the held out data when all pollutants were in the first quartile as the referent, to get adjusted risk ratios for each terminal node. Analytic code was created in SAS® v9.3 (Statistical Analysis System; North Carolina).

Results

A total of 4,010 days, out of 4,018, with no missing data on air pollution levels and hospital emergency department visits for pediatric asthma were analyzed. There were 131 days with the concentrations of CO, NO_2, O_3 and $PM_{2.5}$ all in their lowest quartiles, which were held aside to serve as the referent group, leaving 3,879 days in the dataset to be partitioned.

The C&RT algorithm produced a tree with 13 terminal nodes, based on an alpha of 0.15 (Figure 5.1). Each terminal node represents a subset of days with a specific pattern of pollutants that the algorithm could not split further, conditional on the confounders included in the model. Referring back to the 256 types of days conceptual example, the terminal nodes will form a partition of the 255 joint effects in the tree. For example, terminal node T1, which represents the subset of days where $PM_{2.5}$ is in the highest quartile and NO_2 is in the 1st or 2nd quartiles, is equivalent to grouping 32 unique types of days – all the combinations of CO and O_3, characterized by quartiles, holding $PM_{2.5}$ constant at the 4th quartile and NO_2 at *either* the 1st or 2nd quartile.

The tree depicted in Figure 5.1 is configured such that the right-hand branch of each split always corresponds to the higher concentration. As a result, the mean concentration of the pollutants in the terminal nodes generally increases from left to right in the tree. Table 5.1 contains the mean and standard deviations of the pollutant concentrations at each terminal node. The right-most terminal node, T4, has the highest mean concentrations. The referent group contains the lowest concentrations for all four pollutants, by design.

The dotted lines in Figure 5.1 show how the tree size, and hence the number of terminal nodes, would change if a more conservative alpha of 0.1 or 0.05 were selected. Note that the dotted line for alpha = 0.05 does not mean that the *P*-values for all subsequent splits are greater than 0.05; it only indicates that the *P*-values for the splits occurring at internal (non-terminal) nodes 4, 5 and 7 were greater than 0.05. The *P*-values for the selected splits at each internal node as well as the subset of data to which the splits apply are presented in Table 5.2. This information can be used to see how the tree size would differ under alternative choices of alpha.

A simultaneous Wald test for the inclusion of all 13 terminal nodes in the model was significant, with a chi-square statistic of 34.3 (p=0.001, with 13 degrees of freedom), a result that was not unexpected, given that the terminal nodes were created through binary splits determined via hypothesis tests. The joint risk associated with days in each terminal node in comparison with risk associated within the held-out referent group are presented as adjusted risk ratios, estimated in a time series analysis using the same case-crossover model and confounding covariates (Table 5.3). The largest risk ratio was for terminal node T1 (RR: 1.10, 95% CI: 1.05, 1.16) and corresponds to days where concentrations of $PM_{2.5}$ are in the highest quartile and NO_2 are in the lowest two quartiles. Terminal nodes T2 (RR: 1.08, 95% CI: 1.03, 1.14) and T7 (RR: 1.08, 95% CI 1.01, 1.15) had the next largest risk ratios compared to the referent.

Discussion

Many research groups, particularly in genetics, have used recursive partitioning to identify interactions among large numbers of predictor variables [12-14]; however, for the purposes of epidemiologic research we have found the standard C&RT packages to be lacking, to varying degrees. In this paper we present a new C&RT algorithm that is better-suited to epidemiologic research when generating hypotheses about complex joint effects is of interest.

Perhaps the most important way in which the proposed algorithm differs from available C&RT programs is in its control for confounding. Rarely in observational epidemiologic research are we immune to the hazards of confounding. Nonetheless, because most C&RT programs were developed for the purposes of prediction and classification, and not causal inference, they do not directly account for confounding. The typical C&RT approach is to consider all covariates one-at-a-time in the search for the optimal split [7]; however, this one-at-a-time approach ignores confounding. One approach for handling confounding is to first remove the association with the confounders and then fit a regression tree to the residuals [15]; unfortunately, this approach is appropriate only for Gaussian outcomes and cannot be easily applied to the residuals from generalized linear models (e.g. binomial or Poisson data) [16]. Conditional inference trees, first proposed by Hothorn et al in 2006, offer a framework for recursive partitioning in which the best split is chosen conditional on all possible model splits [17]; however, this approach requires that all covariates in the conditional model be eligible for partitioning. The C&RT algorithm we propose differentiates exposure covariates from control covariates, i.e., it allows for user-defined *a priori* control of confounding while restricting the selection of the optimal splits to the exposure covariates, thereby making this approach better aligned to epidemiologic research when effect estimation is of interest. Bertolet et al identified many of the same limitations to the existing C&RT approaches and go on to present a similar method for using classification and regression trees that control for confounding with Cox proportional hazards models and survival data [18].

A cited drawback to existing C&RT algorithms is their inability to quantify exposure effect estimates [19]. The C&RT algorithm we have proposed enables effect estimation through the withholding of a common referent group of days during tree construction. This allows for estimation of joint effects across terminal nodes in relation to the pre-specified reference group. Selecting the referent group *a priori* ensures that it does not depend on the analysis (i.e. how the algorithm groups the data); otherwise each analysis might yield a different referent group and hinder comparisons across studies. Additionally, such *a priori* selection allows the researcher to define a meaningful referent group.

C&RT does not provide a single statistic that summarizes all the joint effects, nor is it possible to look at the tree and assess whether the algorithm "worked;" C&RT merely identifies the joint effects present in the data. Therefore, we suggest using C&RT as an intermediary step to generate hypotheses about joint effects that exist in the data in order to inform future analyses and studies. For example, terminal node T4 has higher mean concentrations for all pollutants relative to T1 and yet the RR for T1 is greater than T4 (RR 1.10 vs 1.07). While this difference could be due to the relatively small sample sizes or random error, one hypothesis is that splitting NO_2 at its 50th percentile (quartiles 1 & 2 vs. 3 & 4), which resulted in terminal node T1, may represent a particularly harmful type of $PM_{2.5}$ mixture with regards to pediatric asthma. Alternatively there may be certain meteorological factors that promote this specific pollutant covariation and influence personal exposure levels, such as relative humidity. These hypotheses lead to several researchable questions. For example, do days in T1 appear to be dominated by a single source? Is there evidence that this joint effect is associated with increased risk in other datasets? Does residual confounding or effect measure modification by meteorological factors further explain the relative risks associated with each terminal node?

CO appears only once in the final tree, as a split at internal node 9, which results in terminal nodes 5 and 6 (Figure 5.1). This suggests that in Atlanta CO may be less associated with pediatric asthma visits than O_3, NO_2, and $PM_{2.5}$. The minimal role of CO in the final tree is not entirely surprising since ambient concentrations of CO in isolation pose no appreciable health risk to the general population [20]; we chose to include CO in our model to act as a potential surrogate for other pollutants emitted from combustion sources that were not included in the model. Removing CO from the analysis – assuming no change to the referent group – would only affect the final tree by collapsing terminal nodes 5 and 6 into a single terminal node.

The RRs in Table 5.3 do not appear to be dominated by any single pollutant. Instead they suggest that higher levels of pollution are generally more harmful, with the RRs appearing relatively robust to the components of the mixture. Terminal nodes T1, T3, T4, T6, T7 and T8 all have high overall mean concentrations, but from Table 5.1 it is clear that the distribution of pollutants in these terminal nodes is different. For example, T8 is driven by high NO_2, CO and $PM_{2.5}$; T4 by high O_3 and $PM_{2.5}$; and T6 by high O_3, and yet all three terminal nodes are associated with a similarly elevated risk relative to the referent group. These results are consistent with a recent multipollutant study by Winquist et al, which found that the joint effects of an inter-quartile range increase pollutant combinations (oxidants, secondary pollutants, indicators of traffic, power plants, and five criteria pollutants) resulted in statistically significant health effects but that the point estimates for the different pollutant combinations were not appreciably different from each other [21]. From the perspective of multipollutant risk assessment, the C&RT approach of classifying day types may offer valuable insight by identifying specific pollution mixtures that are detrimental to health, which could lead to simultaneous regulation of pollutants or identification of harmful sources. In addition, by calculating a single joint effect for each terminal node, this approach helps to avoid over estimation of the RRs that could occur from joint effect calculations based on single pollutant models in which the single pollutant associations may be capturing the effects of correlated pollutants.

The confidence intervals presented in Table 5.3 should be viewed in the framework of hypothesis generation and not as a tested result. Multiple significance tests were conducted to identify the terminal nodes. Ideally the joint effects for the terminal nodes would be estimated using independent observations; however, because another independent study was not available at the time of analysis, confidence intervals should not be interpreted at their nominal level. Instead, in the spirit of hypothesis generating, the confidence intervals should be used to motivate future analyses, which may lead to substantive results.

Each terminal node can be interpreted as representing a specific mixture or a collection of mixtures that has a similar association with the outcome. Although the path of exposure indicator terms leading to each terminal node in a C&RT tree may indicate interaction, this is not always the case. For example, suppose a tree splits first after the third quartile of $PM_{2.5}$ and then both branches go on to split between the second and third quartiles of NO_2 (similar to the tree in Figure 5.1). If the risk ratios comparing the higher

NO_2 terminal nodes with the respective lower NO_2 terminal nodes within levels of $PM_{2.5}$ are the same (a subjective decision) then interaction is not present. In this scenario, the effect of NO_2 on the outcome does not depend on the level of $PM_{2.5}$, and therefore the tree is not suggestive of interaction. The tree in Figure 5.1, however, does not meet this criterion. Instead we conclude that interaction between NO_2 and $PM_{2.5}$ *is* present in our data because a calculation of the relative risks comparing days in internal nodes 5 vs. 4 and internal nodes 7 vs. 6 suggests a different direction of effect of NO_2 at low vs. high levels of $PM_{2.5}$ (RR: 1.03 and RR: 0.96 respectively).

In our air pollution example, the fact that internal nodes 4 and 5 split on different exposures (O_3 and $PM_{2.5}$, respectively) suggests that there is something different about the association of the pollution mixtures on pediatric asthma visits on moderate $PM_{2.5}$ days when NO_2 is below vs. above its median level. By looking at the C&RT tree we cannot determine whether this is due to some chemical or physiological interaction between $PM_{2.5}$ and the other pollutants, a difference in the particles that comprise $PM_{2.5}$ on high days as compared to low or normal days, the covariation of $PM_{2.5}$ with other pollutants, random error, or some other factor. Instead, we can use the C&RT tree to generate such hypotheses regarding relationships that exist in the data, which can then be investigated in subsequent analyses. For example, an interesting follow-up analysis would be to perform C&RT on just the $PM_{2.5}$ constituents to identify the components that appear to be driving the health association.

While most C&RT packages utilize measures of node impurity, including the Gini index for classification trees and least squares for regression trees [7], to guide the splitting decisions there are situations in which other criteria may be justifiable. One approach is to base the best split on statistical significance, as was done in this paper and has been favored by others [17, 18]. Selecting splits based on the smallest *P*-value (or largest Chi-square statistic) illustrates how recursive partitioning can be used to capture the strongest association present in the data.

The selection of α in the proposed algorithm is analogous to pruning in the traditional C&RT programs, with larger values of α generating larger trees and smaller values generating nested sub-trees. A frequently cited downside of C&RT is the instability of the tree, leading many investigators to favor random forests instead, which is an approach that incorporates information from an ensemble of trees [6]. Although random forests offer a solution to tree stability, because there is no summary tree created, identification of joint effects is difficult. In the example we have presented, tree size and stability will be affected by the cut-points selected to categorize the exposures. Because the purpose of the proposed approach is hypothesis generation, and not prediction or classification, the stability of any individual tree may be of less concern. Once C&RT has been used to identify potentially harmful joint effects, further refinement of these effects, including investigating a dose-response relationship or finding specific cut-point values, can be conducted using other statistical approaches. Furthermore, knowledge of the *P*-values at each splitting point in the tree, including the most significant and several runners-up, may offer a guide for the stability of any given branch.

C&RT is sometimes criticized for displaying a selection bias towards predictor variables with more splits [17, 22]. We tried to address this by assessing equal numbers of potential splits (e.g. quartiles) for each predictor. In our example we chose to create quartile indicators to bridge the C&RT results with the conceptual example; however, the proposed algorithm places no restrictions on how the splits are created. One could create finer splitting points (e.g. deciles or centiles) to better approximate the continuous nature of the exposures; however, if statistical significance is used to determine the best split, the aforementioned tendency to select more balanced splits could become more pronounced as the number of potential splits increases. Allowing the exposures to remain continuous is currently infeasible with this modified C&RT algorithm, due to computational challenges posed by the quantity of GLM models needed, and this direction warrants future methodological development. Alternatively, splits could be based on substantive knowledge (e.g., the U.S. National Ambient Air Quality Standards). An advantage of this approach is that it would allow for greater generalizability, as the splitting points would not be data-based.

A particular challenge in mixtures research is how to deal with highly correlated exposures; while not unique to C&RT, it is important to consider how it may influence the regression tree results. If two exposures are highly correlated, and one is causally associated with the outcome while the other is merely a surrogate for the former, the algorithm will not necessarily split on the causal exposure. This is of particular concern if the two exposures have differential measurement error, as the exposure with the least amount of measurement error can have the estimated greatest effect, even if not causal [3]. Pollutant correlation also affects the frequency at which specific mixtures occur. Of the 256 possible day types referred to in the conceptual example, 37 never occurred during the 11-year period, and another 58 occurred less than 0.1% of the time. This happens because the three-day moving averages of O_3 and $PM_{2.5}$ are strongly correlated (ρ=0.61), as are CO and NO_2 (ρ=0.59). The regression tree will have limited power to identify whether rarely occurring exposures are harmful. As a result, the terminal nodes in the resulting tree can be considered as *either* indicative of homogeneity of effect *or* as lacking sufficient power to split further.

C&RT is one of many statistical tools that can be used to address the challenge of multipollutant exposures. Among the more frequently cited approaches are single pollutant regression models [23], two-pollutant regression models [23-25], source apportionment [26], clustering [8-10], recursive partitioning [7, 27], dimension reduction [28-30], and Bayesian model averaging [31]. Two recent reviews offer a detailed overview of the advantages and disadvantages of these and other approaches for multipollutant research [19, 32]. Recursive partitioning approaches, including C&RT, are attractive because unlike traditional regression models they require no distributional assumptions and can easily handle large numbers of predictors. While C&RT is frequently utilized for its ability to identify complex interactions [33-35], we feel that this should be broadened to "complex joint effects." Such a broadening of scope would not only help to caution against the misinterpretation of interaction in C&RT trees, a problem that others have documented in the literature [6], but it would also expand the utility of C&RT. Identifying joint effects associated with the outcome may be sufficient if one is interested in describing health associations in terms of covarying exposures where interaction may not exist, as in the case with air pollution. C&RT has the additional advantage over other mixture approaches of producing output that is both visually intuitive and informative.

In air pollution epidemiology, while there is currently interest in moving from a single pollutant to a multipollutant framework, the term "multipollutant" is often used broadly and may encompass many different conceptual issues [23, 24, 36]. When the multipollutant interest involves the joint effects of several pollutants, we feel that C&RT, particularly with the modifications mentioned in this paper, is a very appropriate tool. We suggest that C&RT be used as an intermediary step for identifying and refining potentially harmful multipollutant joint effects for further investigation. A good example of the benefits of incorporating C&RT into the modeling strategy is demonstrated in Sun et al, who show how a two-step multipollutant modeling strategy involving C&RT and dimension reduction techniques can offer substantial improvements on variable selection [19].

For illustrative purposes we have shown how C&RT can be used to address challenges in the field of air pollution; however, there are many other fields in which exposure mixtures are of interest that may benefit from this C&RT approach. As previously mentioned, researchers in genetics have been using C&RT to identify gene-gene joint effects. The proposed C&RT approach would enable these researchers to expand their current approach to include simultaneous control for biological and environmental factors that may confound the gene-gene associations. Other fields that may benefit from C&RT include nutrition, where understanding the joint effects of nutrient mixtures is of interest, and infectious disease research, where advancements in multiplex assays allow scientists to measure an individual’s exposure to many different antibodies.

Conclusions

With advances in science and technology, high dimensional datasets are increasingly common, leading many researchers to question how best to characterize and analyze these mixtures of exposures. Many issues arise when dealing with mixtures, including exposure covariation, physiological and chemical interaction, joint effects, and novel exposure metrics. Classification and regression trees offer an alternative to traditional regression approaches and may be well-suited for identifying complex patterns of joint effects in the data. While recursive partitioning approaches such as C&RT are not new, they are seldom used in epidemiologic research. We believe that the aforementioned modifications to the C&RT algorithm, namely the differentiation of exposure and control covariates to account for confounding and the withholding of a referent group, can aid researchers interested in generating hypotheses about exposure mixtures.

References

1. Rothman KJ, Greenland S, Lash TL: **Modern Epidemiology, 3rd Edition**, Third Edition edn. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

2. Zhang H, Singer BH: **Recursive Partitioning and Applications**, Second Edition edn. New York: Springer; 2010.

3. Tolbert PE, Klein M, Peel JL, Sarnat SE, Sarnat JA: **Multipollutant modeling issues in a study of ambient air quality and emergency department visits in Atlanta**. *J Expo Sci Environ Epidemiol* 2007, **17 Suppl 2**:S29-35.

4. Ivy D, Mulholland JA, Russell AG: **Development of Ambient Air Quality Population-Weighted Metrics for Use in Time-Series Health Studies**. *Journal of the Air & Waste Management Association* 2008, **58**(5):711-720.

5. Strickland MJ, Darrow LA, Klein M, Flanders WD, Sarnat JA, Waller LA, Sarnat SE, Mulholland JA, Tolbert PE: **Short-term associations between ambient air pollutants and pediatric asthma emergency department visits**. *Am J Respir Crit Care Med* 2010, **182**(3):307-316.

6. Strobl C, Malley J, Tutz G: **An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging, and Random Forests**. *Psychological Methods* 2009, **14**(4):323-348.

7. Breiman L, Friedman JH, Olshen RA, Stone CJ: **Classification and Regression Trees**. Belmont: Wadsworth International Group; 1984.

8. Chakraborty G, Chakraborty B: **A novel normalization technique for unsupervised learning in ANN**. *IEEE Trans Neural Netw* 2000, **11**(1):253-257.

9. Hartigan JA, Wong MA: **A K-means clustering algorithm**. *Applied Statistics* 1979, **28**:100-108.

10. Kohonen T: **Self-Organizing Maps**. Berlin: Springer; 1995.

11. Lu Y, Zeger SL: **On the equivalence of case-crossover and time series methods in environmental epidemiology**. *Biostatistics* 2007, **8**(2):337-344.

12. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P: **Identifying SNPs predictive of phenotype using random forests**. *Genet Epidemiol* 2005, **28**(2):171-182.

13. Garcia-Magarinos M, Lopez-de-Ullibarri I, Cao R, Salas A: **Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction**. *Ann Hum Genet* 2009, **73**(Pt 3):360-369.

14. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P: **Screening large-scale association study data: exploiting interactions using random forests**. *BMC Genet* 2004, **5**:32.

15. Hastie T, Tibshirani R: **Generalized Additive Models**. In. London: Chapman & Hall; 1990.

16. O'hara Hines R, Carter E: **Improved added variable and partial residual plots for the detection of influential observations in generalized linear models**. *Applied Statistics* 1993, **42**(1):3-20.

17. Hothorn T, Hornik K, Zeileis A: **Unbiased recursive partitioning: a conditional inference framework**. *Journal of Computational and Graphical Statistics* 2006, **15**(3):651-674.

18. Bertolet M, Brooks MM, Bittner V: **Tree-based identification of subgroups for time-varying covariate survival data**. *Stat Methods Med Res* 2012.

19. Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, Batterman SA, Mukherjee B: **Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons**. *Environ Health* 2013, **12**(1):85.

20. Kuller LH, Radford EP: **Epidemiological bases for the current ambient carbon monoxide standards**. *Environ Health Perspect* 1983, **52**:131-139.

21. Winquist A, Kirrane E, Klein M, Strickland MJ, Darrow LA, Sarnat SE, Gass KM, Mulholland JA, Russell AG, Tolbert PE: **Joint Effects of Ambient Air Pollutants on Pediatric Asthma Emergency Department Visits in Atlanta, 1998-2004**. In: *Abstracts of the 2013 Conference of the International Society of Environmental Epidemiology (ISEE).* Basel, Switzerland: Environ Health Persp; 2013.

22. Shih YS: **A note on split selection bias in classification trees**. *Computational Statistics & Data Analysis* 2004, **45**(3):457-466.

23. Mauderly JL, Burnett RT, Castillejos M, Ozkaynak H, Samet JM, Stieb DM, Vedal S, Wyzga RE: **Is the air pollution health research community prepared to support a multipollutant air quality management framework?** *Inhal Toxicol* 2010, **22 Suppl 1**:1-19.

24. Dominici F, Peng RD, Barr CD, Bell ML: **Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach**. *Epidemiology* 2010, **21**(2):187-194.

25. Johns DO, Stanek LW, Walker K, Benromdhane S, Hubbell B, Ross M, Devlin RB, Costa DL, Greenbaum DS: **Practical advancement of multipollutant scientific and risk assessment approaches for ambient air pollution**. *Environ Health Perspect* 2012, **120**(9):1238-1242.

26. Grahame T, Hidy GM: **Pinnacles and pitfalls for source apportionment of potential health effects from airborne particle exposure**. *Inhal Toxicol* 2007, **19**(9):727-744.

27. Breiman L: **Random Forests**. *Machine Learning* 2001, **45**:5-32.

28. Roberts S, Martin MA: **Using supervised principal components analysis to assess multiple pollutant effects**. *Environ Health Perspect* 2006, **114**(12):1877-1882.

29. Roberts S, Martin MA: **A critical assessment of shrinkage-based regression approaches for estimating the adverse health effects of multiple air pollutants**. *Atmospheric Environment* 2005, **39**(33):6223–6230.

30. Tibshirani R: **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society Series B* 1996, **58**(1):267-288.

31. Thomas DC, Jerrett M, Kuenzli N, Louis TA, Dominici F, Zeger S, Schwarz J, Burnett RT, Krewski D, Bates D: **Bayesian model averaging in time-series studies of air pollution and mortality**. *J Toxicol Environ Health A* 2007, **70**(3-4):311-315.

32. Billionnet C, Sherrill D, Annesi-Maesano I, Study G: **Estimating the Health Effects of Exposure to Multi-Pollutant Mixture**. *Annals of Epidemiology* 2012, **22**(2):126-141.

33. Zhang H, Bonney G: **Use of classification trees for association studies**. *Genet Epidemiol* 2000, **19**(4):323-332.

34. Camp NJ, Slattery ML: **Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States)**. *Cancer Causes Control* 2002, **13**(9):813-823.

35. Roetker NS, Yonker JA, Lee C, Chang V, Basson JJ, Roan CL, Hauser TS, Hauser RM, Atwood CS: **Multigene interactions and the prediction of depression in the Wisconsin Longitudinal Study**. *BMJ Open* 2012, **2**(4).

36. Vedal S, Kaufman JD: **What does multi-pollutant air pollution research mean?** *Am J Respir Crit Care Med* 2011, **183**(1):4-6.

TABLE 5.3. Mean and Standard Deviation for Pollutant Concentrations in Each Terminal Node, Atlanta, Georgia, 1999 - 2009

Terminal Node	N	CO	NO_2	O_3	$PM_{2.5}$
		Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Overall	4010	0.57 (0.3)	21.07 (7)	43.76 (17.45)	14.06 (5.78)
Referent Group	131	0.27 (0.04)	11.9 (2.74)	24.68 (3.88)	6.81 (1.4)
T1	316	0.5 (0.18)	16.68 (2.91)	57.09 (14.05)	21.13 (3.76)
T2	279	0.53 (0.21)	24.13 (2.74)	35.77 (9.84)	8.47 (1.05)
T3	1039	0.68 (0.31)	26.05 (4.31)	41.02 (14.89)	13.26 (1.98)
T4	441	0.67 (0.28)	27.3 (5.31)	72.19 (12.72)	23.66 (4.74)
T5	91	0.33 (0.07)	17.03 (2.42)	60.25 (5.05)	13.77 (2.51)
T6	68	0.66 (0.09)	16.54 (3.17)	62.49 (5.81)	13.94 (2.36)
T7	76	0.71 (0.37)	22.63 (1.29)	41.33 (13.27)	19.26 (2.07)
T8	168	1.13 (0.42)	33.31 (7.08)	38.65 (11.11)	21.18 (4.36)
T9	458	0.44 (0.21)	12.69 (2.58)	30.83 (7.76)	9.83 (2.67)
T10	263	0.5 (0.24)	18.46 (1.22)	23.34 (4.87)	10.18 (2.67)
T11	435	0.44 (0.17)	18.27 (1.28)	42.35 (7.06)	11.46 (2.95)
T12	160	0.37 (0.15)	12.74 (2.3)	47.08 (3.42)	10.1 (1.9)
T13	85	0.45 (0.17)	13.65 (1.98)	49.26 (3.99)	14.69 (0.98)

Table 5.2. Quartile Contrasts at Each Internal (Non-Terminal) Node.

Internal Node No.[a]	N	Quartile Contrast[b]	Wald *P*-value[c]	Subset of pollutant quartiles to which contrast applies[d]			
				CO	NO_2	O_3	$PM_{2.5}$
1	3879	$PM_{2.5}$: 4 vs. 1-3	0.000	All	All	All	All
2	2878	NO_2: 3-4 vs. 1-2	0.003	All	All	All	1-3
3	1001	NO_2: 3-4 vs. 1-2	0.019	All	All	All	4
4	1560	O_3: 4 vs. 1-3	0.096	All	1,2	All	1-3
5	1318	$PM_{2.5}$: 2-3 vs. 1	0.123	All	3,4	All	1-3
7	685	O_3: 4 vs. 1-3	0.128	All	3,4	All	4
8	1401	NO_2: 2 vs. 1	0.086	All	1,2	1-3	1-3
9	159	CO: 3-4 vs. 1-2	0.043	All	1,2	4	1-3
14	244	NO_2: 4 vs. 3	0.096	All	3,4	1-3	4
16	703	O_3: 3 vs. 1-2	0.140	All	1	1-3	1-3
17	698	O_3: 2-3 vs. 1	0.062	All	2	1-3	1-3
33	309	$PM_{2.5}$: 3 vs. 1-2	0.033	All	1	3	1-3

[a] The node numbers correspond to the numbering in Figure 5.1 (where each node, *n,* produces two child nodes numbered *2n* and *2n+1).*

[b] Based on the indicator variable chosen for the best split.

[c] *P*-value based on a Wald test that the beta coefficient for the quartile contrast indicator is zero.

[d] Each subset of pollutant concentration levels represents an effect modifier of the quartile contrast and relates directly to the branching of the tree in Figure 5.1. Note that in the first split of the tree there is no effect modification by any of the pollutants because the entire dataset is used.

Table 5.3. Risk Ratios of Emergency Department Visits for Pediatric Asthma for Days in the Terminal Nodes as Compared to the Referent Group,[a] Atlanta, Georgia, 1999-2009.

Terminal Node[b]	N[c]	Risk Ratio	95% Confidence Interval	Type of Days (pollutant quartiles)			
				CO	NO_2	O_3	$PM_{2.5}$
Referent	131	1.00		1	1	1	1
T1	316	1.10	(1.05, 1.16)	1-4	1,2	1-4	4
T2	279	1.08	(1.03, 1.14)	1-4	3,4	1-4	1
T3	1039	1.05	(1.01, 1.1)	1-4	3,4	1-4	2,3
T4	441	1.07	(1.02, 1.13)	1-4	3,4	4	4
T5	91	1.03	(0.97, 1.1)	1,2	1,2	4	1-3
T6	68	1.07	(0.98, 1.17)	3,4	1,2	4	1-3
T7	76	1.08	(1.01, 1.15)	1-4	3	1-3	4
T8	168	1.07	(1.01, 1.14)	1-4	4	1-3	4
T9	458	1.01	(0.97, 1.05)	1-4	1	1,2	1-3
T10	263	1.04	(0.99, 1.09)	1-4	2	1	1-3
T11	435	1.03	(0.98, 1.07)	1-4	2	2,3	1-3
T12	160	1.02	(0.97, 1.08)	1-4	1	3	1,2
T13	85	1.04	(0.97, 1.11)	1-4	1	3	3

[a] Days when all pollutants are in the lowest quartile.

[b] Terminal nodes represent different types of days that can be described in terms of the pollutant quartiles.

[c] Each day is in one and only one terminal node; the column sums to 4010

[d] *P*-values are associated with the null hypothesis that the risk ratio for the pollutant indicator is 1.0.

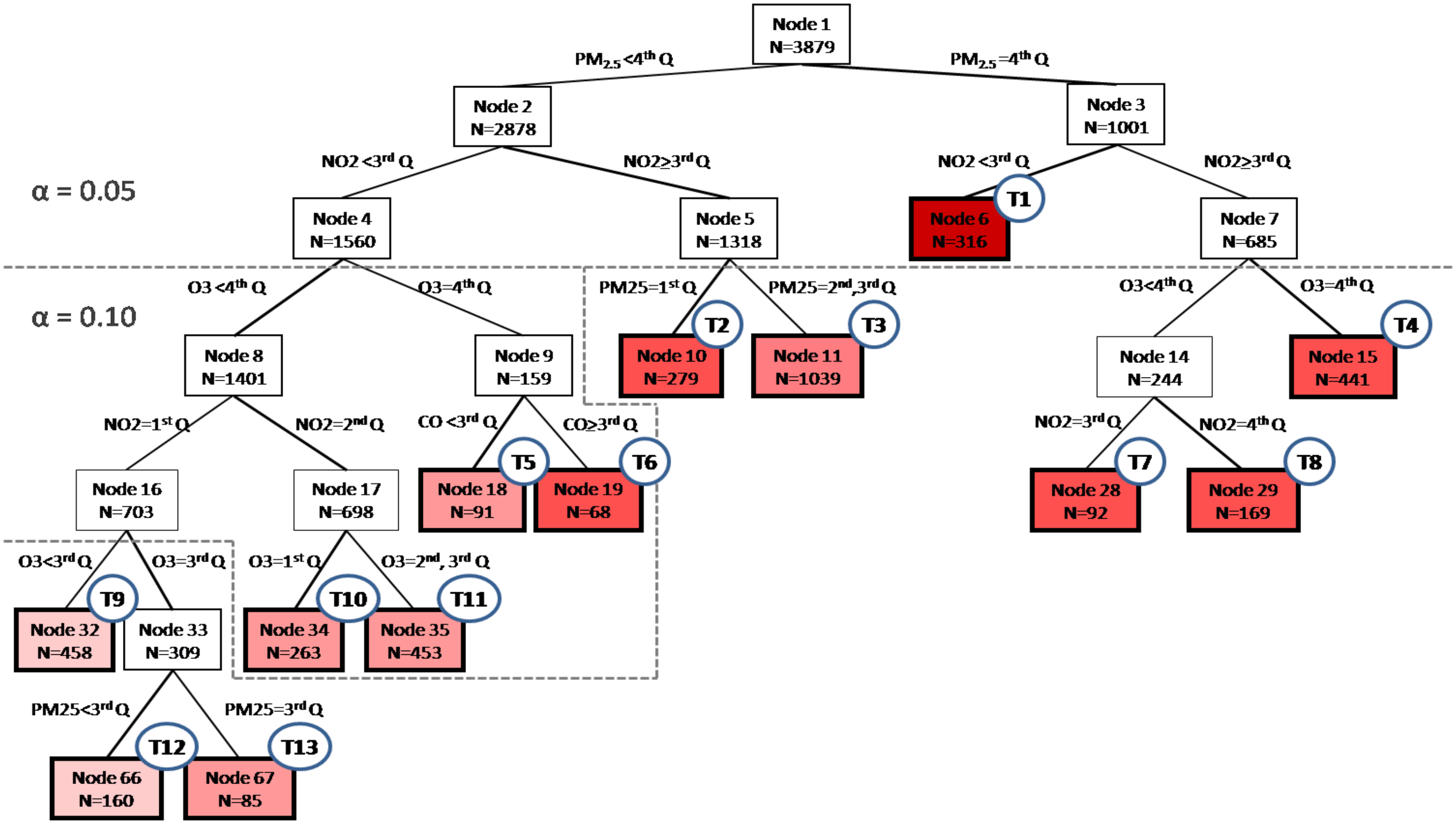


Figure 5.1. Tree resulting from C&RT analysis illustrating the joint effects of CO, NO_2, O_3, and $PM_{2.5}$, treated as ordinal variables by quartile, for pediatric asthma ED visits in Atlanta from 1/1/1999 – 12/31/2009. The tree was grown using an alpha of 0.15 and a minimum node size of 60 observations. Nodes are numbered such that each node, n, produces two child nodes numbered 2n and 2n+1. Nodes with a bold border are terminal nodes for two-sided $\alpha = 0.15$, labeled T1 – T13, as indicated by the circle in the upper right-hand corner and are colored according to the strength of association; redder colors indicate a more harmful association. The dotted lines indicate how the tree would appear under different levels of α. For each split of the tree the branch with the more harmful association is bolded.

Chapter 6: A Three-City Analysis of Multipollutant Joint Effects: a comparison of classification and regression trees with conventional multipollutant models (Study 2)

Introduction

Everyday humans breathe a mixture of different air pollutants. Characterizing these multipollutant mixtures for health effects has been tackled by air pollution research groups for decades. Historically, the most common epidemiological approach for addressing mixtures has been through single pollutant models, in which a single pollutant effect is believed to act as a surrogate for the air pollution mixture [1-4]. Multipollutant models, in which two or more pollutants are included in the same model to get the combined effect, have been the predominant approach for examining mixtures [5-8]. While these conventional approaches have made significant advances to our understanding of air pollution epidemiology, it is important to consider alternative approaches as the information we can learn is limited by the constraints and assumptions of the model [9, 10].

Two reviews of alternative statistical approaches for multipollutant research have recently been published [11, 12]. In both reviews, classification and regression trees (C&RT), a recursive partitioning approach, was cited as a method for handling multipollutant exposures; however, there have been few applications of C&RT in assessing the health effects of ambient pollution exposure [13, 14]. In a recent paper we showed how C&RT can be adapted for epidemiologic research and become a useful tool for generating hypotheses about multipollutant joint effects [15].

C&RT classifies days according to their multipollutant profiles, and this can be conceptually appealing, as it adds a discrete summary of what is observed and can also help to elucidate patterns of meteorology, seasonality, and emission sources that cause certain pollutants to covary more strongly than others. From a health perspective, classification of days can enable identification of day types that are more harmful to human health and help to improve risk prediction systems. From a regulatory perspective, identifying the most harmful multipollutant joint effects can lead to more targeted regulation.

In this analysis, we sought to extend our exploration of multipollutant joint effects associated with pediatric asthma emergency department (ED) visits by comparing those identified via C&RT to those identified through more conventional multipollutant regression modeling approaches. To increase the power and generalizability of our findings, we take a three-city approach, utilizing data from Atlanta, Dallas and St. Louis to investigate multipollutant joint effects. The results of our three-city C&RT approach are compared to those of more conventional multipollutant models and in doing so we discuss the relative merits, limitations and assumptions tied to each modeling approach and make suggestions for ways forward.

Methods

Data

Our analysis utilized emergency department (ED) visit and criteria pollutant data available for three metropolitan areas: 20-county Atlanta, GA; 12-county Dallas-Fort Worth, TX; and 16-county St. Louis, MO-IL.

Emergency Department Visit Data

Computerized billing records for ED visits to acute care hospitals in each city were obtained as follows: for Atlanta, from individual hospitals and the Georgia Hospital Association for an 11-year study period (January 1, 1999 through December 31, 2009); for Dallas, from the Dallas-Fort Worth Hospital Council Foundation for a 3.5-year study period (January 1, 2006 through August 31, 2009); and for St. Louis, from the Missouri Hospital Association for a 6.5-year study period (January 1, 2001 through June 27, 2007). Relevant data elements included a patient identifier, admission date, patient age, primary and secondary International Classification of Diseases 9th Revision (ICD-9) diagnosis codes, and ZIP code of patient residence. We used these data in accordance with our data use agreement with the individual hospitals and/or hospital associations; this study was also approved by the Emory University Institutional Review Board. Visits by patients living in ZIP codes outside of the city-specific study areas were excluded.

The individual-level data were restricted to visits by pediatric patients (i.e. ages 2-18 years), and aggregated to daily counts of asthma and/or wheeze, identified as any ICD-9 code of 493 and/or 786.07. Visits by the same patient for the same condition on the same day were counted as a single visit.

Air Pollution Data

Our analysis focused on three criteria pollutants, ozone (O_3), nitrogen dioxide (NO_2), and particulate matter less than 2.5 microns in diameter ($PM_{2.5}$), shown to have strong associations with asthma/wheeze in previous analyses using these data (Strickland et al., 2009;[16]) and by others. The 3-day moving average population-weighted concentrations of ambient O_3 (8-hr max), NO_2 (1-hr max) and $PM_{2.5}$ (24-hr average) were calculated using measurements from stationary monitors in each of the three cities (Ivy et al., 2008).

Statistical Methods

Base Model

The base model used to analyze each of the city-specific time series was a Poisson generalized linear model using a framework equivalent to the conditional logistic case-crossover model (Lu and Zeger., 2007). Time trends were controlled by matching on weekday, month and year, and meteorology was controlled with cubic terms for the three-day moving average of: maximum temperature, maximum temperature interacted with an indicator for season, and dew point. A spline for day-of-year with two knots was included to provide additional control for seasonal trends. Both the C&RT and conventional multipollutant approaches utilized this same base model for estimating effects, with the only difference being how the air pollution exposures were modeled.

Classification and Regression Trees

This analysis utilized classification and regression trees (C&RT) to estimatemultipollutant joint effects. C&RT is a non-parametric regression approach that represents a supervised form of hierarchical clustering in which the observations are sequentially split into dichotomous groups, such that each resulting group contains increasingly similar responses for the outcome [17, 18]. Every tree starts with a “root node” that contains the observations from which the tree will be grown. The observations are then partitioned into two “child nodes” based on the value of an independent predictor variable. The resulting child nodes each contain a subset of the original observations. Each child node may be further partitioned, again based on the value of an independent predictor variable. This process continues until a set of partitioning criteria are no longer met, resulting in terminal nodes. Terminal nodes, by definition, cannot have offspring. The collection of terminal nodes forms a complete partition of the observations in the root node. Each terminal node can be viewed as a unique mixture, defined by the path of partitions, or splits, leading from the root node to that particular terminal node.

While several statistical packages for running automated C&RT analyses exist, none allow for simultaneous control for confounding. As a result of this and other constraints we developed a modified C&RT approach aimed at estimating joint effects. The following description contains a summary of the modified C&RT approach, which has been described in detail in Gass et al[15].

Three-city C&RT Approach

To generate hypotheses about multipollutant joint effects across three cities, we first sought to identify joint effects *within* each city by growing city-specific regression trees. We then compared the resulting trees and identified types of pollution mixtures that were common across all three cities. We hypothesized that exposure to one or more of these pollution mixtures might be a cause of pediatric asthma and tested this in a three-city regression analysis.

The first step in our three-city C&RT comparison was to determine a set of standard concentration cut-offs based on the daily levels of O_3, NO_2, and $PM_{2.5}$ in each city. While every city-day might have a unique multipollutant profile based on continuous pollutant distributions, these profiles are conceptually difficult to manage and inhibit comparison across cities. To better enable comparison, we used the standard cut-offs to classify every day into a “Day-Type” according to its concentration of O_3, NO_2, and $PM_{2.5}$. Using these cut-offs, each pollutant was divided into four levels, resulting in 4^3 or 64 Day-Types. Throughout the paper, the nomenclature used to describe these Day-Types is O_3-level/ NO_2-level/ $PM_{2.5}$-level. For example, Day-Type “2/2/4” refers to days where both O_3 and NO_2 concentrations are in the 2nd level and the $PM_{2.5}$ concentration is in the 4th level (with levels 1 – 4 ranging from lowest to highest concentration).

A referent group was defined as days when all three pollutants were in the lowest level (i.e. days designated as Day-Type 1/1/1). It was decided *a priori* that the referent group should include at least 100 days in each city. In order to reach this minimum, the referent group was defined as: $O_3 \leq 35$ppb, $NO_2 \leq 21$ppb, $PM_{2.5} \leq 11$ug, which corresponds to roughly the 40th percentile of the overall distribution for each pollutant from all 3 cities combined. The remaining level cut-offs were defined at approximately the 60th and 80th percentiles. Figure 6.1 shows the concentration cut-offs for each pollutant level as well as the frequency of days in each level across the three cities.

For each city, the referent days were withheld to serve as a comparison group while the remaining days formed the root node in the C&RT algorithm. We considered the 9 possible ways these days can be split based on the three pollutant concentrations. This was accomplished by creating three mutually exclusive indicator variables for each pollutant representing the different comparisons: level 1 vs. 2-4, levels 1 and 2 vs. 3 and 4, and levels 1-3 vs. 4. These 9 indicator variables represented the 9 possible partitions of the days in the C&RT procedure.

Each of the 9 indicators was considered one-at-a-time in the base model described above. The indicator resulting in the smallest *P*-value for the null hypothesis that the beta for that indicator was 0, in the model containing just the indictor and confounders as independent variableswas selected. The days were partitioned accordingly. Partitioning continued until one of three stopping criteria were met: there were no more remaining ways to partition the days, the remaining splits were not significant at a pre-specified level of alpha (α=0.15), or the minimum number of days for each node (n=60) was not met. In this case partitioning stopped and the node becomes a terminal node.

This C&RT approach was used to grow three *separate* trees using the same algorithm and splitting-indicator definitions for Atlanta, Dallas, and St. Louis. For the purposes of this study we were not interested in comparing the shapes of the trees or the ordering of the splits, but rather the mixture of air pollutants encompassed in the terminal nodes. For each of the three cities, the C&RT algorithm partitioned the 63 Day-Types (all but the withheld referent group) into terminal nodes according to their association with the outcome. As such, we were interested in identifying Day-Types that were grouped together in a terminal node in all three cities, as these may indicate homogeneity of effect of Day-Types *within* each city, across all three cities. For example, suppose Day-Types 2/3/4 and 3/3/4 are together in a terminal node in all three cities, then we say this pair of Day-Types constitutes a “Group” of Day-Types in which we are potentially interested. We further examined the effects of these Day-Type groups directly in our

Poisson GLM via indicator variables. This was done individually for each city, as well as a single three-city model with city-specific effects for all covariates in the model.

Conventional Approach

Finally, we compared our C&RT findings to those obtained from conventional multipollutant regression modeling approaches. Specifically we considered the following three modeling approaches, described in detail below: linear effects, linear effects with all first-order interactions, and linear effects with quadratic effects. For each modeling approach we approximated the comparable joint effect estimates to those obtained for the groups of Day-Types from our C&RT model.

Conventional model with linear effects: Our conventional multipollutant model consisted of the base model described above with the inclusion of a single linear term (β*pollutant) for each of the three pollutants, modeled using continuous 3-day moving average concentrations. A single linear effects model was run for the three cities combined, with city-specific effects for all covariates. To estimate the joint effects comparable to the C&RT model, we used the results of the three-city linear effects model to estimate the joint effect for a change in the mean concentration of each Group relative to the referent group mean. This was done by multiplying the coefficient for the linear effect of each pollutant by the difference in Group mean versus referent mean, for each pollutant. We them summed these products across the pollutants in the combination, and exponentiated the sum to get the rate ratio for the joint effect. Standard errors for the joint effects were calculated using the variance-covariance matrices for the individual-pollutant effect estimates. As a sensitivity analysis, we also calculated the joint effects using the change in the median concentration of each Group relative to the referent group median.

Conventional model with linear effects and first-order interaction: The same conventional linear effects multipollutant model was run with the addition of all first-order multiplicative interaction terms (O_3* NO_2, O_3* $PM_{2.5}$, and NO_2* $PM_{2.5}$). We also considered the same model with the second-order interaction term (O_3* NO_2* $PM_{2.5}$). The same approach described above was used to calculate the comparable joint effects.

Conventional model with linear and quadratic effects: Finally we ran the same linear effects model with the addition of a quadratic term for each pollutant. Again, the same approach described above was used to calculate the comparable joint effects.

Results

After excluding days with missing air pollution levels or hospital ED visits, 4,012 observations remained for analysis for Atlanta (1999-2009), 2,354 days for St. Louis (2001-6/2007) and 1337 days for Dallas (2006-8/2009). The referent group, identified as days where all pollutants were in the lowest level, contained 606 days (15%) for Atlanta, 115 days (5%) for St. Louis, and 121 days (9%) for Dallas (Figure 6.1). A greater description of the referent group is provided in Table 6.3, including the monthly distribution as well as the percent of days with any precipitation and average wind speed. Table 6.4 contains the frequency of each of the 64 Day-Types (that is the joint distribution of the three pollutants parameterized as ordinal variables) in each city. All

Day-Types occurred at least once in Atlanta, in St. Louis there was 1 Day-Type that never occurred and in Dallas there were 5 Day-Types that never occurred.

The C&RT algorithm was run separately for each city, generating three regression trees with seven, six, and seven terminal nodes in Atlanta, Dallas, and St. Louis, respectively (Figures 6.2-6.4). Comparing terminal nodes across the three cities, there were 17 Groups of two or more Day-Types that occurred together in the same terminal node in all three cities (Table 6.1). While the numeric labeling of the Groups is arbitrary, we decided to order the Groups according to the level of O_3, followed by NO_2. Of the 7709 days from the three cities combined, 842 were in the referent group and 5446 were in one of the 17 Groups. Table 6.1 contains the number of days in each Group by city, as well as the mean concentrations of O_3, NO_2, and $PM_{2.5}$. There were 10 Day-Types, corresponding to 1421 days from Atlanta, St. Louis and Dallas combined, that did not appear in any of the 17 Groups; these Day-Types did not appear together in terminal nodes with other day types consistently in all cities.

The rate ratios (RR) for the 17 Groups, included as indicator variables in the base case-crossover model using the three-city dataset (Atlanta, St. Louis, and Dallas combined), are shown in Figure 6.5. Nearly every RR, with the exception of the RR for Groups 2 and 10, is suggestive of a harmful association with pediatric asthma. Groups 11 and 14 had the two strongest RRs (RR: 1.07, 95% CI: 1.03, 1.12; and RR: 1.06, 95% CI: 1.02, 1.09).

The RRs for an IQR increase in the 3-day moving average concentration of O_3, NO_2, and $PM_{2.5}$ from the conventional linear effects model are plotted for the single city and three-city models in Figure 6.6. In all three cities the association with O_3 is the strongest, followed by NO_2; $PM_{2.5}$ appears to have a null, or in the case of Dallas protective, association in these models.

The C&RT Group RR results are presented side by side with the three conventional multipollutant modeling approaches to provide more direct comparison between the results (Table 6.2). The RRs shown for the conventional models are for a concentration change from the referent mean to the Group mean. A Wald test of significance for the exposure terms was significant for the C&RT and all three conventional models (Table 6.2). A test of significance for the exposure terms beyond the linear effects was non-significant for the quadratic model (p=0.63), suggesting that the multipollutant effect is not quadratic. However this test was significant for the first-order interaction terms (p=0.007). The RR results for the linear effects and interaction effects models are quite different, with the latter suggesting approximately double the increase in risk for each of the joint effects, albeit with a loss of precision. Results from the second-order interaction model are not included because the term was non-significant (p=0.21) and the resulting confidence intervals were too large to be informative.

The RRs from the conventional linear effects model have relatively good agreement with the C&RT Group results for the lower mean concentration Groups. At higher pollution levels both the linear effects and interaction models suggest increasing risk with concentration, while the C&RT Group results suggest risk plateaus at the highest concentrations. For example, the RR for Group 18, which contains the Day-Type when all pollutants are at their highest level (4/4/4), is 1.05 in the C&RT model (95% CI: 1.01, 1.09) vs. an RR of 1.10 in the conventional model with linear effects (95% CI: 1.06, 1.14) and 1.17 (95% CI: 1.1, 1.25) in the conventional model with interaction. The lower

95% confidence levels for Groups 2, 10, 15, 16 and 18 for the conventional models are all *greater* than the point estimate from the corresponding groups in the C&RT model.

The sensitivity analysis using the Group median (as opposed to mean) to calculate the joint effects for the conventional models yielded similar estimates and thus are not shown.

Discussion

In this paper we utilized classification and regression trees, a non-parametric recursive partitioning approach, to identify multipollutant joint effects associated with pediatric asthma in Atlanta, Dallas and St. Louis. It is difficult to identify complex interactions of two, three or four pollutants using conventional regression models due to power limitations [19, 20]. A known advantage of C&RT is that can be used to detect complex and multiple interactions between covariates [11, 12]. We have previously shown that with few modifications, C&RT can be used to detect interactions between pollutant concentrations while simultaneously controlling for temporal and meteorological confounding [15].

One key finding of this analysis is that the C&RT approach yielded different results than would have been generated under more conventional regression approaches. All three of the conventional multipollutant models suggest that increasing pollution leads to increasing rate ratios; however the C&RT Group results suggest a non-linear relationship with RRs plateauing when all pollutants are high (Table 6.2). While this lack of a synergistic --or even multiplicative-- response is surprising, it is not unprecedented. In a review of the literature, Mauderly and Samet found that 22 out of 36 laboratory studies failed to demonstrate a synergistic response [6]. It is plausible that the true biological response is less than multiplicative and that this is masked by the constraints placed upon regression models when pollution is modeled linearly.

An alternative hypothesis that might support of the C&RT findings is that on the highest pollution days asthmatic children change their behavior and limit exposure. A cross-sectional study by Wen et al lends some support to this theory, which found that asthmatic adults had a greater odds of modifying their outdoor activity compared with non-asthmatics on days with media alerts due to a high air quality index [21].

Examination of the differences between the C&RT and conventional model results suggests that the role of individual pollutants was different. In all three conventional models the joint effects are driven by O_3. This is demonstrated by the increasing RRs in columns 4-6 of Table 6.3, an artifact of assigning arbitrary labels to the Groups based on increasing O_3 concentration. Conversely, if one were to sort the RRs for the C&RT Groups according to the mean concentrations of any single pollutant (Table 6.3 column 3), no pollutant would appear to drive the results. This difference in pollutant-specific association is most striking when looking at $PM_{2.5}$. The linear effects models for each city imply that $PM_{2.5}$ has a null association (Figure 6.6), while the C&RT results suggest $PM_{2.5}$ plays an important role in determining the Group joint effect; Groups 11, 3, 18, 17, and 14 have the five greatest mean $PM_{2.5}$ concentrations and also five of the highest RRs from the C&RT model.

Though the linear effects and C&RT models are not measuring the same thing (e.g., the C&RT results model exposure categorically while the conventional results model it continuously) the differences implied by the results is striking and merits further attention. It is possible that C&RT is able to identify joint effects driven by $PM_{2.5}$

constituents. For example, the conventional model with linear effects treats all $PM_{2.5}$ concentrations equally; it could not distinguish between a high $PM_{2.5}$ day that is primarily elemental carbon vs. a high day that is primarily sulfates. Conversely the C&RT model has the potential to distinguish $PM_{2.5}$ mixtures through the interactions generated by subsequent partitioning. By partitioning on a pollutant (e.g., NO_2) that is correlated with certain $PM_{2.5}$ components, C&RT has the ability differentiate $PM_{2.5}$ mixtures through their correlation with other independent variables in the model. While the model with first-order interaction terms can discriminate between $PM_{2.5}$ mixtures, its discriminatory power is limited to a linear effect for each of the interaction terms. As such it could not, for example, identify the same complex interactions as seen in the St. Louis C&RT tree through nodes 1-2-4-9 (Figure 6.4).

By binning days the C&RT model may be able to account for unmeasured confounding that is non-smooth (i.e. that varies with terminal node classification, not pollution). For example, people may modify their behavior under certain types of days in a way that affects ED visits for asthma. As a result, it is possible that the point estimates for the Group results are measuring not only the multipollutant effect but also the effects of other factors that are correlated with those Day-Types. While this could be a disadvantage if one intends to use the point estimates to conduct risk assessment, it could be beneficial if the interest is in identifying types of days that are most harmful for a particular health outcome. Knowing the harmful types of days could lead to a more targeted warning system to alert vulnerable populations.

When interpreting these results it is important to consider the modeling assumptions and how they are likely to affect estimates. C&RT models are nonparametric with no assumptions of monotonicity; each terminal node has its own estimated multipollutant joint effect for the outcome, relative to the referent group. By contrast the conventional linear effects model imposes a monotonic relationship with the outcome. While this may be desirable for a single pollutant model where it is well-known that the dose-response is strictly increasing, it may be too restrictive for multipollutant models where so much remains unknown about the joint effects. Likewise, the interaction model used only looks for deviations from a multiplicative joint effect, rather than an additive one, which again may be desirable but one should be purposeful in their decision. The results from the quadratic model suggest that the quadratic effects do not add any explanatory power (p=0.633).

Furthermore, it is important to remember that when two or more Day-Types appear together in the same terminal node, it is either indicative of homogeneity of effect or lack of power to detect any further effect. It is likely that one or more of the 17 Groups were formed as a result of insufficient power to further partition the terminal node in a given city. This would result in an “artificial group”, that is a Group in which the Day-Types do not have a similar association with the outcome.

One downside to presenting the combined Group RRs in this analysis is that any heterogeneity across the cities will be masked. Some between city heterogeneity is to be expected due to tangible and intangible city differences, including socio-economic status, air conditioning use, climate acclimation and behavior patterns that are likely to modify the health associations found. Nonetheless, it seems likely that there exist some ambient pollution mixtures that are universally harmful, despite city-specific differences which may accentuate or attenuate the underlying true association.

Conclusion

As we have shown, C&RT can be used to investigate multipollutant joint effects and may lead to different conclusions than more conventional models. In particular, the results from this study suggest C&RT and conventional models lead to different joint effects of O_3, NO_2 and $PM_{2.5}$ when concentrations are high. It is possible that the monotonicity assumptions of conventional models are leading to an overestimation of risk on high pollution days. Furthermore we have shown how C&RT models can be beneficial for identifying types of days that are particularly harmful to health, which can help to improve warning systems and lead to more targeted regulation. Understanding the potential risk air pollution mixtures pose to human health is a complex and challenging undertaking that has only just begun. Exploring alternative models with different sets of assumptions can be a useful way to generate new ideas and perhaps gain greater insight into air pollution mixtures.

References

1. Sarnat JA, Schwartz J, Catalano PJ, Suh HH: **Gaseous pollutants in particulate matter epidemiology: confounders or surrogates?** *Environ Health Perspect* 2001, **109**(10):1053-1061.

2. Janssen NA, Lanki T, Hoek G, Vallius M, de Hartog JJ, Van Grieken R, Pekkanen J, Brunekreef B: **Associations between ambient, personal, and indoor exposure to fine particulate matter constituents in Dutch and Finnish panels of cardiovascular patients**. *Occupational and environmental medicine* 2005, **62**(12):868-877.

3. Laden F, Neas LM, Dockery DW, Schwartz J: **Association of fine particulate matter from different sources with daily mortality in six U.S. cities**. *Environ Health Perspect* 2000, **108**(10):941-947.

4. Sarnat SE, Suh HH, Coull BA, Schwartz J, Stone PH, Gold DR: **Ambient particulate air pollution and cardiac arrhythmia in a panel of older adults in**

Steubenville, Ohio. *Occupational and environmental medicine* 2006, **63**(10):700-706.

5. Katsouyanni K, Touloumi G, Samoli E, Gryparis A, Le Tertre A, Monopolis Y, Rossi G, Zmirou D, Ballester F, Boumghar A *et al*: **Confounding and effect modification in the short-term effects of ambient particles on total mortality: results from 29 European cities within the APHEA2 project**. *Epidemiology* 2001, **12**(5):521-531.

6. Mauderly JL, Samet JM: **Is there evidence for synergy among air pollutants in causing health effects?** *Environ Health Perspect* 2009, **117**(1):1-6.

7. Tolbert PE, Klein M, Peel JL, Sarnat SE, Sarnat JA: **Multipollutant modeling issues in a study of ambient air quality and emergency department visits in Atlanta**. *J Expo Sci Environ Epidemiol* 2007, **17 Suppl 2**:S29-35.

8. Jerrett M, Burnett RT, Beckerman BS, Turner MC, Krewski D, Thurston G, Martin RV, van Donkelaar A, Hughes E, Shi Y *et al*: **Spatial analysis of air pollution and mortality in California**. *Am J Respir Crit Care Med* 2013, **188**(5):593-599.

9. Dominici F, Peng RD, Barr CD, Bell ML: **Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach**. *Epidemiology* 2010, **21**(2):187-194.

10. Mauderly JL, Burnett RT, Castillejos M, Ozkaynak H, Samet JM, Stieb DM, Vedal S, Wyzga RE: **Is the air pollution health research community prepared to support a multipollutant air quality management framework?** *Inhal Toxicol* 2010, **22 Suppl 1**:1-19.

11. Billionnet C, Sherrill D, Annesi-Maesano I, Study G: **Estimating the Health Effects of Exposure to Multi-Pollutant Mixture**. *Annals of Epidemiology* 2012, **22**(2):126-141.

12. Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, Batterman SA, Mukherjee B: **Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons**. *Environ Health* 2013, **12**(1):85.

13. Oiamo TH, Luginaah IN: **Extricating sex and gender in air pollution research: a community-based study on cardinal symptoms of exposure**. *International journal of environmental research and public health* 2013, **10**(9):3801-3817.

14. Hu W, Mengersen K, McMichael A, Tong S: **Temperature, air pollution and total mortality during summers in Sydney, 1994-2004**. *Int J Biometeorol* 2008, **52**(7):689-696.

15. Gass K, Klein M, Chang HH, Flanders WD, Strickland MJ: **Classification and regression trees for epidemiologic research: an air pollution example**. *Environmental Health* 2014, **13**(17).

16. Winquist A, Klein M, Tolbert P, Flanders WD, Hess J, Sarnat SE: **Comparison of emergency department and hospital admissions data for air pollution time-series studies**. *Environ Health* 2012, **11**:70.

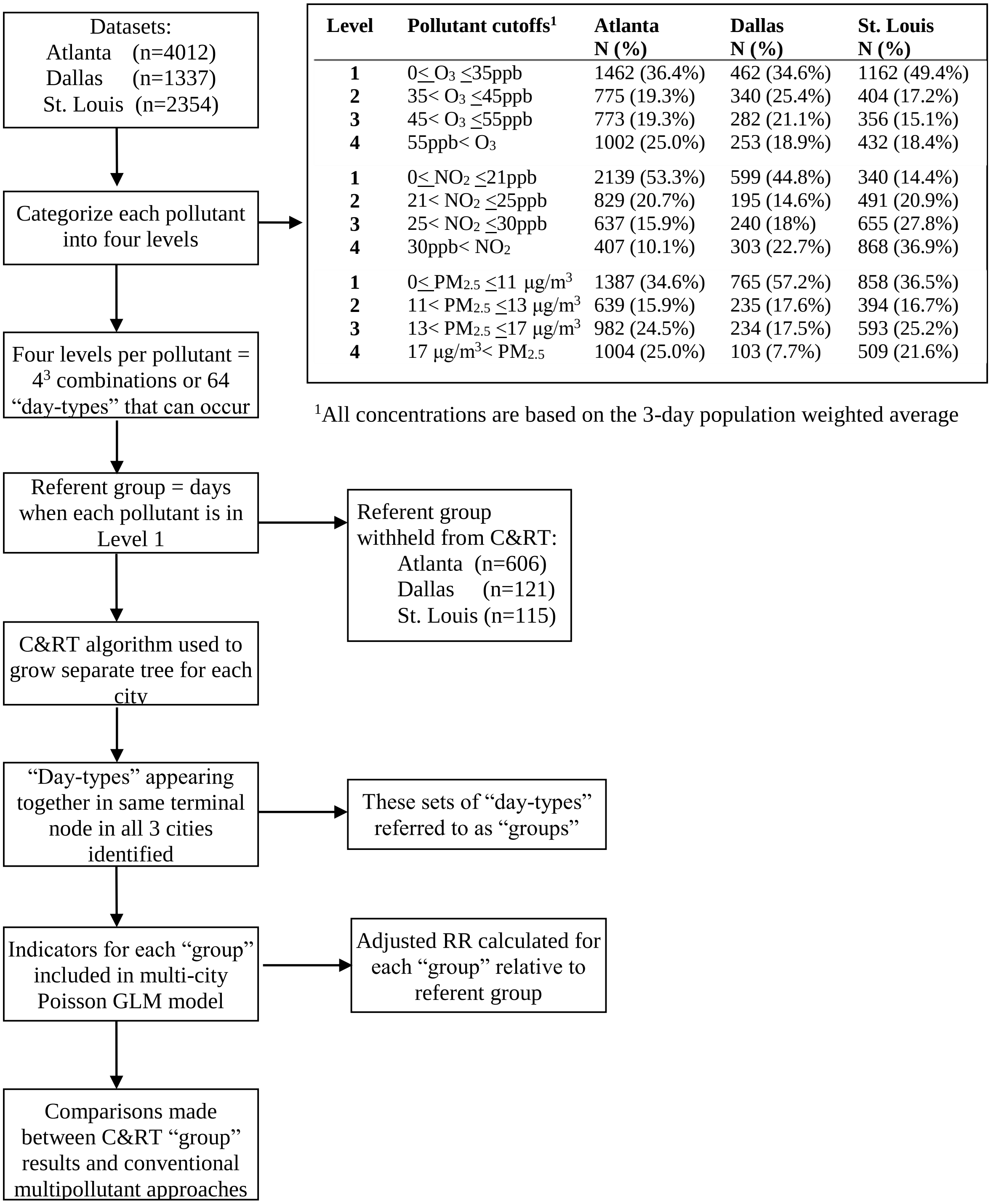
17. Strobl C, Boulesteix AL, Zeileis A, Hothorn T: **Bias in random forest variable importance measures: illustrations, sources and a solution**. *BMC Bioinformatics* 2007, **8**:25.

18. Breiman L, Friedman JH, Olshen RA, Stone CJ: **Classification and Regression Trees**. Belmont: Wadsworth International Group; 1984.

19. Greenland S: **Tests for interaction in epidemiologic studies: a review and a study of power**. *Statistics in medicine* 1983, **2**(2):243-251.

20. Rothman KJ, Greenland S, Lash TL: **Modern Epidemiology, 3rd Edition**, Third Edition edn. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

21. Wen XJ, Balluz L, Mokdad A: **Association between media alerts of air quality index and change of outdoor activity among adult asthma in six states, BRFSS, 2005**. *Journal of community health* 2009, **34**(1):40-46.


Datasets:
Atlanta (n=4012)
Dallas (n=1337)
St. Louis (n=2354)
Categorize each pollutant into four levels
Four levels per pollutant = 4^3 combinations or 64 “day-types” that can occur
Referent group = days when each pollutant is in Level 1
Referent group withheld from C&RT:
Atlanta (n=606)
Dallas (n=121)
St. Louis (n=115)
C&RT algorithm used to grow separate tree for each city
“Day-types” appearing together in same terminal node in all 3 cities identified
These sets of “day-types” referred to as “groups”
Indicators for each “group” included in multi-city Poisson GLM model
Adjusted RR calculated for each “group” relative to referent group
Comparisons made between C&RT “group” results and conventional multipollutant approaches

Level	Pollutant cutoffs[1]	Atlanta N (%)	Dallas N (%)	St. Louis N (%)
1	$0\leq O_3 \leq 35$ppb	1462 (36.4%)	462 (34.6%)	1162 (49.4%)
2	$35< O_3 \leq 45$ppb	775 (19.3%)	340 (25.4%)	404 (17.2%)
3	$45< O_3 \leq 55$ppb	773 (19.3%)	282 (21.1%)	356 (15.1%)
4	55ppb< O_3	1002 (25.0%)	253 (18.9%)	432 (18.4%)
1	$0\leq NO_2 \leq 21$ppb	2139 (53.3%)	599 (44.8%)	340 (14.4%)
2	$21< NO_2 \leq 25$ppb	829 (20.7%)	195 (14.6%)	491 (20.9%)
3	$25< NO_2 \leq 30$ppb	637 (15.9%)	240 (18%)	655 (27.8%)
4	30ppb< NO_2	407 (10.1%)	303 (22.7%)	868 (36.9%)
1	$0\leq PM_{2.5} \leq 11$ µg/m^3	1387 (34.6%)	765 (57.2%)	858 (36.5%)
2	$11< PM_{2.5} \leq 13$ µg/m^3	639 (15.9%)	235 (17.6%)	394 (16.7%)
3	$13< PM_{2.5} \leq 17$ µg/m^3	982 (24.5%)	234 (17.5%)	593 (25.2%)
4	17 µg/m^3< $PM_{2.5}$	1004 (25.0%)	103 (7.7%)	509 (21.6%)

[1]All concentrations are based on the 3-day population weighted average

Figure 6.1. Flow diagram outlining three-city C&RT approach. The table in the upper right contains the concentration cutoffs used to categorize each pollutant into four levels and the frequency at which each level occurred by city.

Table 6.1. Groups described by Day-Type, number of days and mean concentration. Group labels are arbitrary and ordered according to level of O_3, followed by NO_2. Day-Types are represented by the level of O_3, NO_2 and $PM_{2.5}$ respectively. The number of days in each group is presented for all three cities combined, as well as by city. Mean concentrations and standard deviations are presented for the 3-day population weighted average of O_3 (ppb), NO_2 (ppb) and $PM_{2.5}$ (μg/m^3).

Label	Day-Type $O_3/NO_2/PM_{2.5}$	All N	Atlanta N	Dallas N	St. Louis N	O_3 Mean(SD)	NO_2 Mean(SD)	$PM_{2.5}$ Mean(SD)
Group1 (Referent)	1/1/1	842	606	121	115	25.9 (6.3)	15.4 (3.8)	8.2 (1.7)
Group2	1/1/2, 1/1/3, 2/1/2, 2/1/3	652	431	134	87	32.3 (9.4)	16.4 (3.4)	13.2 (1.6)
Group3	1/1/4, 2/1/4	99	49	25	25	33.1 (10.2)	16.5 (3.5)	19.8 (3)
Group4	1/2/2, 1/2/3	248	126	13	109	23 (7.2)	22.9 (1.2)	13.3 (1.5)
Group5	1/3/1, 1/3/2, 1/3/3, 1/3/4	642	210	85	347	24.8 (7)	27.5 (1.4)	12 (3.9)
Group6	1/4/1, 1/4/2, 1/4/3, 1/4/4, 2/4/1, 2/4/2, 2/4/3, 2/4/4	855	173	196	486	30 (8.9)	34.7 (4.1)	13.3 (5.2)
Group7	2/2/2, 2/2/3	107	63	15	29	40.3 (2.8)	23 (1.1)	13.1 (1.6)
Group8	2/3/1, 2/3/2, 2/3/3, 2/3/4	257	110	38	109	40.1 (2.8)	27.3 (1.4)	11.3 (3.5)
Group9	3/1/1, 3/2/1	267	156	76	35	49.3 (2.8)	18 (4.2)	9.2 (1.4)
Group10	3/1/2, 3/1/3	333	242	64	27	50 (3)	16.3 (3)	13.7 (1.6)
Group11	3/1/4, 3/2/4	161	119	16	26	50.9 (2.8)	18.7 (3.9)	19.9 (2.8)
Group12	3/2/2, 3/2/3	131	68	27	36	49.4 (2.8)	22.9 (1.2)	13.7 (1.6)
Group13	3/3/1, 3/4/1	163	25	63	75	49.7 (2.7)	31 (4.3)	9.1 (1.4)
Group14	3/3/2, 3/3/3, 3/3/4, 3/4/2, 3/4/3, 3/4/4	356	163	36	157	50 (2.9)	31.7 (5.4)	15.6 (4.2)
Group15	4/1/2, 4/1/3,	205	149	42	14	61.4 (5.6)	17.3 (2.8)	14.3 (1.7)
Group16	4/2/2, 4/2/3	128	77	36	15	63.1 (6.1)	23.1 (1.1)	14.3 (1.7)
Group17	4/3/1, 4/3/2, 4/3/3, 4/3/4	392	207	72	113	67.2 (10.7)	27.4 (1.4)	19.1 (6.6)
Group18	4/4/1, 4/4/2, 4/4/3, 4/4/4	450	156	58	236	68.5 (11.2)	35.7 (4.9)	19.7 (6.3)

Table 6.2. Rate Ratios (RR) for the multipollutant joint effects from the C&RT and conventional models. Rate ratios for the conventional models are calculated for the effect of an increase equal to the Group mean minus the referent mean for all three pollutants. AIC offers information about model fit (lower values indicate better fit) while the Wald test provides information on the significance of the exposure covariates in each model.

	C&RT results with indicators for each Group		**Group Mean - Referent Mean**			**Conventional Model: linear effects**		**Conventional Model: linear effects + first-order interactions**		**Conventional Model: linear effects+ quadratic effects**	
Label	**RR**	**95% CI**	**O_3**	**NO_2**	**$PM_{2.5}$**	**RR**	**95% CI**	**RR**	**95% CI**	**RR**	**95% CI**
Group 2	0.99	(0.97, 1.02)	6.35	1.01	5.04	1.01	(1.00, 1.02)	1.05	(1.02, 1.08)	1.01	(1.00, 1.03)
Group 3	1.04	(0.99, 1.09)	7.2	1.18	11.64	1.01	(1.00, 1.03)	1.07	(1.03, 1.11)	1.02	(1.00, 1.04)
Group 4	1.01	(0.98, 1.04)	-2.89	7.56	5.18	1.01	(1.00, 1.02)	1.04	(1.02, 1.06)	1.01	(1.00, 1.03)
Group 5	1.03	(1.01, 1.06)	-1.05	12.11	3.83	1.02	(1.00, 1.03)	1.04	(1.02, 1.07)	1.02	(1.01, 1.04)
Group 6	1.04	(1.01, 1.07)	4.06	19.34	5.14	1.04	(1.02, 1.06)	1.07	(1.04, 1.10)	1.04	(1.02, 1.06)
Group 7	1.03	(0.99, 1.08)	14.37	7.63	4.97	1.03	(1.02, 1.05)	1.08	(1.05, 1.12)	1.04	(1.02, 1.06)
Group 8	1.02	(0.99, 1.06)	14.24	11.91	3.12	1.04	(1.02, 1.06)	1.09	(1.05, 1.13)	1.05	(1.02, 1.07)
Group 9	1.04	(1.01, 1.08)	23.43	2.68	0.98	1.04	(1.02, 1.06)	1.10	(1.05, 1.15)	1.04	(1.02, 1.07)
Group 10	1.00	(0.97, 1.03)	24.15	0.94	5.56	1.04	(1.02, 1.06)	1.10	(1.05, 1.15)	1.04	(1.02, 1.07)
Group 11	1.07	(1.03, 1.12)	24.96	3.35	11.74	1.04	(1.02, 1.07)	1.11	(1.06, 1.16)	1.05	(1.02, 1.08)
Group 12	1.04	(0.99, 1.08)	23.48	7.53	5.52	1.05	(1.03, 1.07)	1.11	(1.06, 1.16)	1.05	(1.03, 1.08)
Group 13	1.04	(0.99, 1.08)	23.84	15.68	0.92	1.06	(1.04, 1.09)	1.13	(1.08, 1.19)	1.06	(1.03, 1.09)
Group 14	1.06	(1.02, 1.09)	24.15	16.29	7.46	1.06	(1.04, 1.09)	1.13	(1.08, 1.18)	1.07	(1.04, 1.10)
Group 15	1.01	(0.97, 1.05)	35.53	1.98	6.09	1.06	(1.03, 1.09)	1.13	(1.07, 1.20)	1.06	(1.03, 1.10)
Group 16	1.03	(0.98, 1.08)	37.21	7.71	6.16	1.07	(1.04, 1.10)	1.15	(1.08, 1.22)	1.08	(1.04, 1.11)
Group 17	1.05	(1.01, 1.09)	41.33	12.04	10.94	1.08	(1.05, 1.12)	1.16	(1.09, 1.23)	1.09	(1.05, 1.13)
Group 18	1.05	(1.01, 1.09)	42.56	20.37	11.55	1.10	(1.06, 1.14)	1.17	(1.10, 1.25)	1.10	(1.06, 1.15)
AIC	62674					62667		62645		62669	
Wald Test*	$\chi^2_{df=18}$=36.6, p=0.006					$\chi^2_{df=3}$=27.06, p<0.0001		$\chi^2_{df=6}$=39.14, p<0.0001		$\chi^2_{df=6}$=28.77, p<0.0001	
Wald Test**	*NA*					*NA*		$\chi^2_{df=3}$=12.00, p=0.007		$\chi^2_{df=3}$=1.72, p=0.633	

*Simultaneous Wald test for all exposure parameters in the model.

**Simultaneous Wald test for additional exposure parameters beyond the linear effects

Table 6.3. Distribution of referent and non-referent (i.e. used to generate C&RT trees) days by city.

	Atlanta				**Dallas**				**St. Louis**			
	Referent (n=606)		Non-referent (n=3406)		Referent (n=121)		Non-referent (n=1261)		Referent (n=115)		Non-referent (n=2239)	
Month	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
January	106	17.49	229	6.72	15	12.4	107	8.8	25	21.74	190	8.49
February	41	6.77	270	7.93	6	4.96	107	8.8	5	4.35	192	8.58
March	26	4.29	315	9.25	14	11.57	110	9.05	1	0.87	203	9.07
April	12	1.98	318	9.34	3	2.48	117	9.62	5	4.35	205	9.16
May	24	3.96	317	9.31	10	8.26	114	9.38	4	3.48	213	9.51
June	15	2.48	315	9.25	13	10.74	107	8.8	0	--	207	9.25
July	9	1.49	332	9.75	8	6.61	116	9.54	0	--	186	8.31
August	18	2.97	323	9.48	2	1.65	122	10.03	4	3.48	182	8.13
September	54	8.91	276	8.1	11	9.09	79	6.5	9	7.83	171	7.64
October	79	13.04	262	7.69	4	3.31	89	7.32	25	21.74	161	7.19
November	96	15.84	234	6.87	11	9.09	79	6.5	21	18.26	159	7.1
December	126	20.79	215	6.31	24	19.83	69	5.67	16	13.91	170	7.59
Precipitation	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Days with precipitation	246	40.59	968	28.45	49	40.5	213	17.52	39	33.91	663	29.61
Wind	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Wind speed	9.36	3.54	7.62	3.16	12.49	4.62	10.68	4.31	10.03	3.36	8.71	3.21

Table 6.4. Frequency at which each Day-Type (n=64) occurred by city, as well as the terminal node designation from the city-specific trees.

Day-Type	Pollutant Level			Atlanta			Dallas			St. Louis		
	O_3	NO_2	$PM_{2.5}$	N	%	Terminal Node[1]	N	%	Terminal Node[1]	N	%	Terminal Node[1]
1/1/1	1	1	1	606	(15.1%)	Referent	121	(9.1%)	Referent	115	(4.9%)	Referent
1/1/2	1	1	2	137	(3.4%)	1A	29	(2.2%)	1D	30	(1.3%)	6S
1/1/3	1	1	3	98	(2.4%)	1A	30	(2.2%)	1D	25	(1.1%)	6S
1/1/4	1	1	4	24	(0.6%)	1A	6	(0.4%)	1D	11	(0.5%)	1S
1/2/1	1	2	1	137	(3.4%)	5A	46	(3.4%)	1D	152	(6.5%)	3S
1/2/2	1	2	2	56	(1.4%)	5A	9	(0.7%)	1D	46	(2%)	7S
1/2/3	1	2	3	70	(1.7%)	5A	4	(0.3%)	1D	63	(2.7%)	7S
1/2/4	1	2	4	21	(0.5%)	5A	1	(0.1%)	1D	26	(1.1%)	1S
1/3/1	1	3	1	61	(1.5%)	5A	70	(5.2%)	1D	171	(7.3%)	2S
1/3/2	1	3	2	46	(1.1%)	5A	9	(0.7%)	1D	47	(2%)	2S
1/3/3	1	3	3	70	(1.7%)	5A	5	(0.4%)	1D	86	(3.7%)	2S
1/3/4	1	3	4	33	(0.8%)	5A	0	(0%)	1D	43	(1.8%)	2S
1/4/1	1	4	1	11	(0.3%)	4A	114	(8.5%)	1D	84	(3.6%)	2S
1/4/2	1	4	2	23	(0.6%)	4A	18	(1.3%)	1D	63	(2.7%)	2S
1/4/3	1	4	3	41	(1%)	4A	0	(0%)	1D	99	(4.2%)	2S
1/4/4	1	4	4	28	(0.7%)	4A	0	(0%)	1D	101	(4.3%)	2S
2/1/1	2	1	1	221	(5.5%)	1A	116	(8.7%)	1D	30	(1.3%)	3S
2/1/2	2	1	2	92	(2.3%)	1A	34	(2.5%)	1D	13	(0.6%)	6S
2/1/3	2	1	3	104	(2.6%)	1A	41	(3.1%)	1D	19	(0.8%)	6S
2/1/4	2	1	4	25	(0.6%)	1A	19	(1.4%)	1D	14	(0.6%)	1S
2/2/1	2	2	1	80	(2%)	6A	16	(1.2%)	1D	47	(2%)	3S
2/2/2	2	2	2	32	(0.8%)	6A	10	(0.7%)	1D	12	(0.5%)	7S
2/2/3	2	2	3	31	(0.8%)	6A	5	(0.4%)	1D	17	(0.7%)	7S
2/2/4	2	2	4	10	(0.2%)	6A	1	(0.1%)	1D	4	(0.2%)	1S

2/3/1	2	3	1	42	(1%)	7A	28	(2.1%)	1D	69	(2.9%)	2S
2/3/2	2	3	2	21	(0.5%)	7A	7	(0.5%)	1D	18	(0.8%)	2S
2/3/3	2	3	3	31	(0.8%)	7A	2	(0.1%)	1D	19	(0.8%)	2S
2/3/4	2	3	4	16	(0.4%)	7A	0	(0%)	1D	3	(0.1%)	2S
2/4/1	2	4	1	7	(0.2%)	4A	54	(4%)	1D	58	(2.5%)	2S
2/4/2	2	4	2	11	(0.3%)	4A	6	(0.4%)	1D	31	(1.3%)	2S
2/4/3	2	4	3	26	(0.6%)	4A	0	(0%)	1D	33	(1.4%)	2S
2/4/4	2	4	4	26	(0.6%)	4A	1	(0.1%)	1D	17	(0.7%)	2S
3/1/1	3	1	1	123	(3.1%)	2A	59	(4.4%)	5D	11	(0.5%)	3S
3/1/2	3	1	2	88	(2.2%)	2A	34	(2.5%)	2D	10	(0.4%)	6S
3/1/3	3	1	3	154	(3.8%)	2A	30	(2.2%)	2D	17	(0.7%)	6S
3/1/4	3	1	4	80	(2%)	2A	12	(0.9%)	2D	12	(0.5%)	1S
3/2/1	3	2	1	33	(0.8%)	2A	17	(1.3%)	5D	24	(1%)	3S
3/2/2	3	2	2	23	(0.6%)	2A	11	(0.8%)	2D	12	(0.5%)	7S
3/2/3	3	2	3	45	(1.1%)	2A	16	(1.2%)	2D	24	(1%)	7S
3/2/4	3	2	4	39	(1%)	2A	4	(0.3%)	2D	14	(0.6%)	1S
3/3/1	3	3	1	21	(0.5%)	2A	28	(2.1%)	6D	31	(1.3%)	2S
3/3/2	3	3	2	26	(0.6%)	2A	11	(0.8%)	2D	18	(0.8%)	2S
3/3/3	3	3	3	37	(0.9%)	2A	7	(0.5%)	2D	23	(1%)	2S
3/3/4	3	3	4	26	(0.6%)	2A	1	(0.1%)	2D	14	(0.6%)	2S
3/4/1	3	4	1	4	(0.1%)	2A	35	(2.6%)	6D	44	(1.9%)	2S
3/4/2	3	4	2	9	(0.2%)	2A	7	(0.5%)	2D	43	(1.8%)	2S
3/4/3	3	4	3	23	(0.6%)	2A	6	(0.4%)	2D	45	(1.9%)	2S
3/4/4	3	4	4	42	(1%)	2A	4	(0.3%)	2D	14	(0.6%)	2S
4/1/1	4	1	1	22	(0.5%)	3A	15	(1.1%)	3D	0	(0%)	3S
4/1/2	4	1	2	38	(0.9%)	3A	17	(1.3%)	3D	7	(0.3%)	6S
4/1/3	4	1	3	111	(2.8%)	3A	25	(1.9%)	3D	7	(0.3%)	6S
4/1/4	4	1	4	216	(5.4%)	3A	11	(0.8%)	3D	19	(0.8%)	1S
4/2/1	4	2	1	11	(0.3%)	3A	14	(1%)	4D	8	(0.3%)	3S

4/2/2	4	2	2	17	(0.4%)	3A	10	(0.7%)	4D	5	(0.2%)	7S
4/2/3	4	2	3	60	(1.5%)	3A	26	(1.9%)	4D	10	(0.4%)	7S
4/2/4	4	2	4	164	(4.1%)	3A	5	(0.4%)	4D	27	(1.1%)	1S
4/3/1	4	3	1	7	(0.2%)	3A	15	(1.1%)	4D	6	(0.3%)	4S
4/3/2	4	3	2	19	(0.5%)	3A	15	(1.1%)	4D	11	(0.5%)	4S
4/3/3	4	3	3	43	(1.1%)	3A	24	(1.8%)	4D	38	(1.6%)	4S
4/3/4	4	3	4	138	(3.4%)	3A	18	(1.3%)	4D	58	(2.5%)	4S
4/4/1	4	4	1	1	(0%)	3A	17	(1.3%)	4D	8	(0.3%)	5S
4/4/2	4	4	2	1	(0%)	3A	8	(0.6%)	4D	28	(1.2%)	5S
4/4/3	4	4	3	38	(0.9%)	3A	13	(1%)	4D	68	(2.9%)	5S
4/4/4	4	4	4	116	(2.9%)	3A	20	(1.5%)	4D	132	(5.6%)	5S

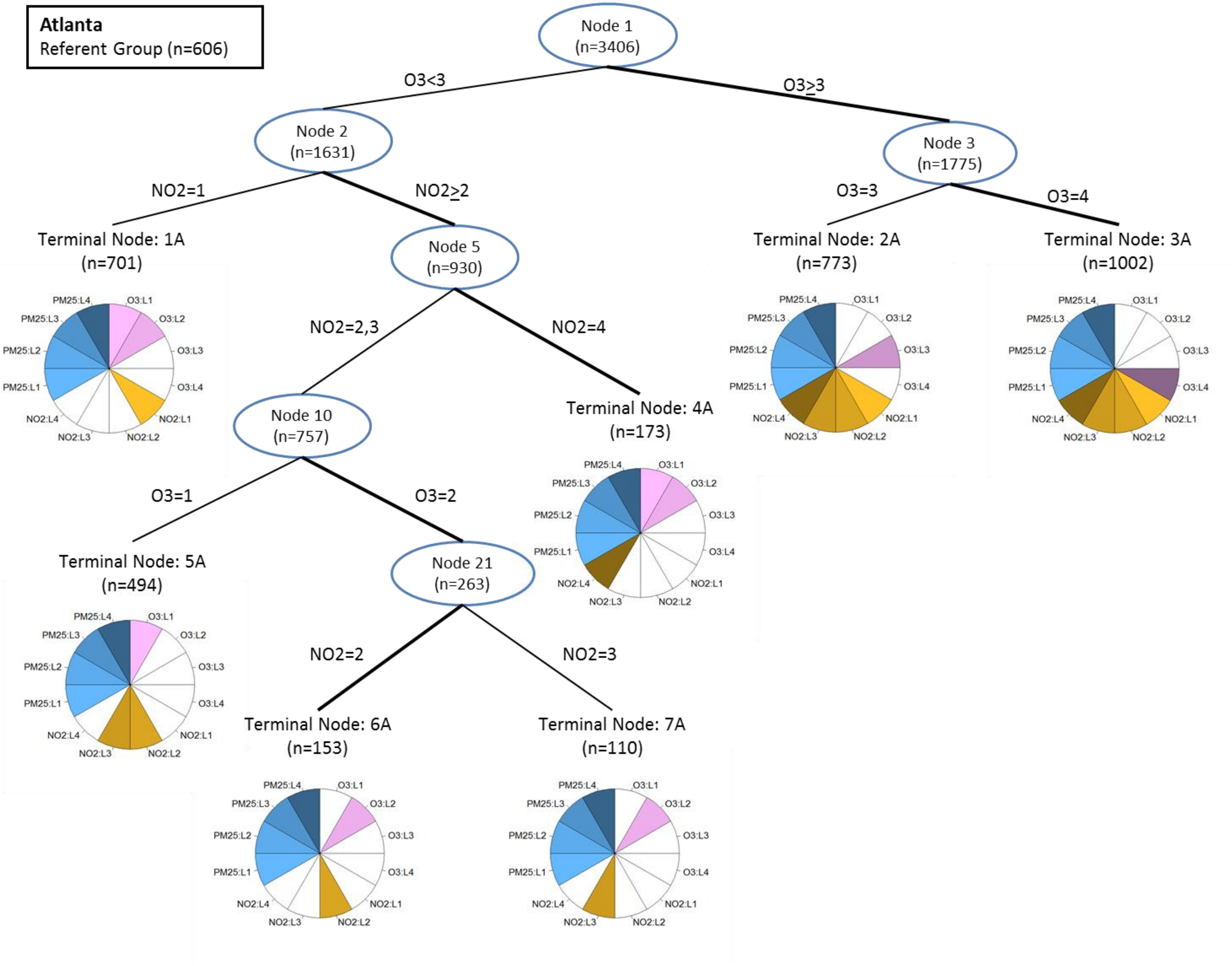

Atlanta
Referent Group (n=606)
Node 1
(n=3406)
O3<3
O3≥3
Node 2
(n=1631)
Node 3
(n=1775)
NO2=1
NO2≥2
O3=3
O3=4
Terminal Node: 1A
(n=701)
Node 5
(n=930)
Terminal Node: 2A
(n=773)
Terminal Node: 3A
(n=1002)
NO2=2,3
NO2=4
Node 10
(n=757)
Terminal Node: 4A
(n=173)
O3=1
O3=2
Terminal Node: 5A
(n=494)
Node 21
(n=263)
NO2=2
NO2=3
Terminal Node: 6A
(n=153)
Terminal Node: 7A
(n=110)
PM25:L4
PM25:L3
PM25:L2
PM25:L1
O3:L1
O3:L2
O3:L3
O3:L4
NO2:L1
NO2:L2
NO2:L3
NO2:L4

Figure 6.2. Classification tree illustrating the joint effects of O_3, NO_2 and **$PM_{2.5}$ associated with emergency department visits for pediatric asthma in Atlanta (1999 – 2009).** Internal nodes are designated with an oval and numbered such that each node, n, produces two child nodes numbers 2n and 2n+1. The branches of the tree are labeled according to the level of the pollutant used to partition the tree. For each partition, the branch with the more harmful association is bolded. Terminal nodes are numbered 1A-7A (A for Atlanta). The pie graphs at each terminal node are a graphical representation of the Day-Types that fall into each terminal node. Each pie graph has 12 wedges, four representing each level (L1-L4) of O_3 (shades of purple), four representing each level of NO_2 (shades of gold), and four representing each level of $PM_{2.5}$ (shades of blue). Pie wedges are colored if a pollutant level is classified into that terminal node and left white if the pollutant level is absent from the terminal node. Day-Types present in the terminal node can be identified by finding every combination of one O_3 wedge (purple), one NO_2 wedge (gold) and one $PM_{2.5}$ wedge (blue). For example terminal node 7A contains 4 Day-Types: O_3 level 2, NO_2 level 3 and $PM_{2.5}$ levels 1-4 (2/3/1, 2/3/2, 2/3/3, 2/3/4).

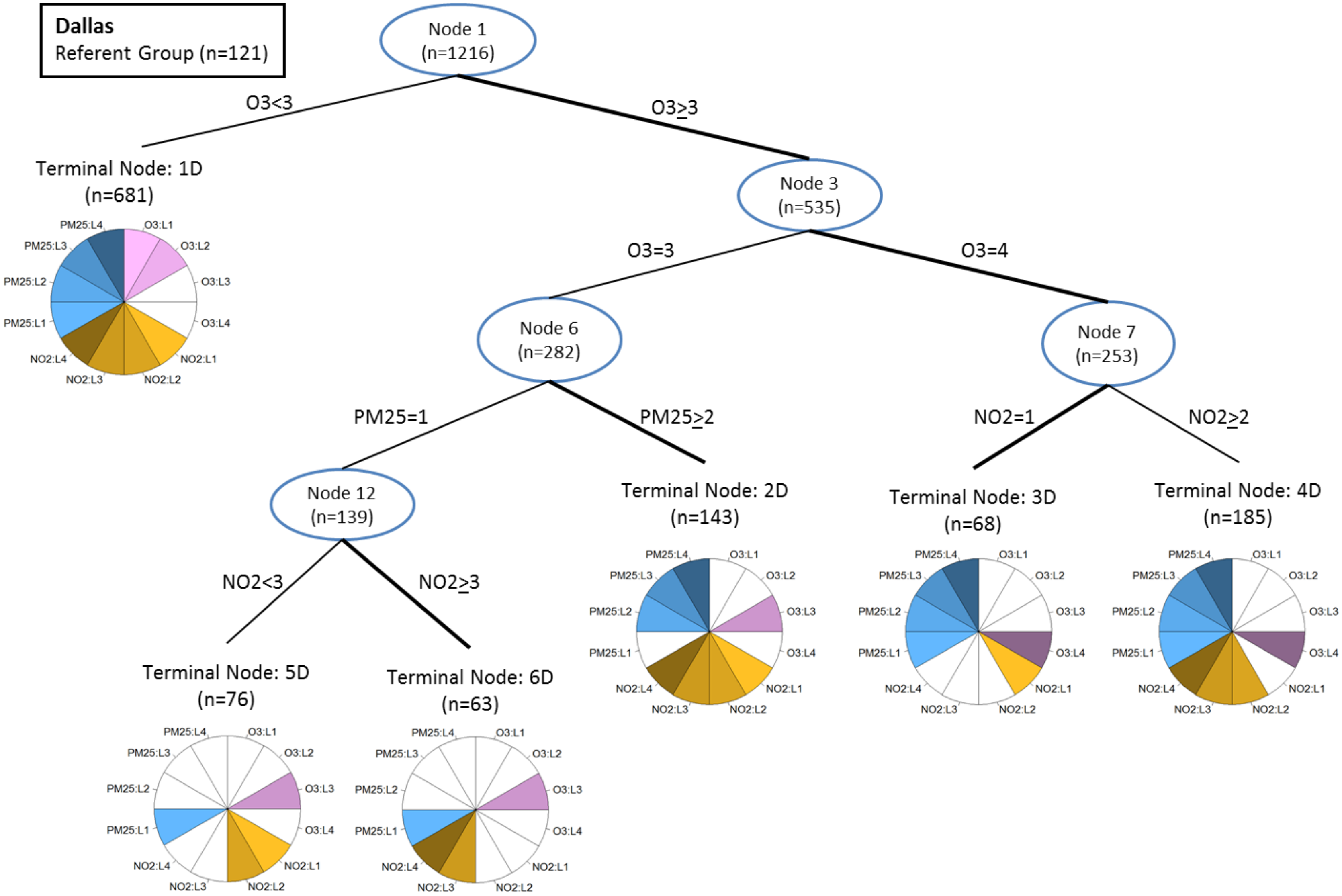

Dallas
Referent Group (n=121)
Node 1
(n=1216)
O3<3
O3≥3
Terminal Node: 1D
(n=681)
PM25:L4
O3:L1
PM25:L3
O3:L2
PM25:L2
O3:L3
PM25:L1
O3:L4
NO2:L4
NO2:L1
NO2:L3
NO2:L2
Node 3
(n=535)
O3=3
O3=4
Node 6
(n=282)
Node 7
(n=253)
PM25=1
PM25≥2
NO2=1
NO2≥2
Node 12
(n=139)
Terminal Node: 2D
(n=143)
Terminal Node: 3D
(n=68)
Terminal Node: 4D
(n=185)
NO2<3
NO2≥3
Terminal Node: 5D
(n=76)
Terminal Node: 6D
(n=63)

Figure 6.3. Classification tree illustrating the joint effects of O_3, NO_2 and $PM_{2.5}$ associated with emergency department visits for pediatric asthma in Dallas (2006 –2009). Terminal nodes are numbered 1D-6D (D for Dallas). For additional description see Figure 6.2.

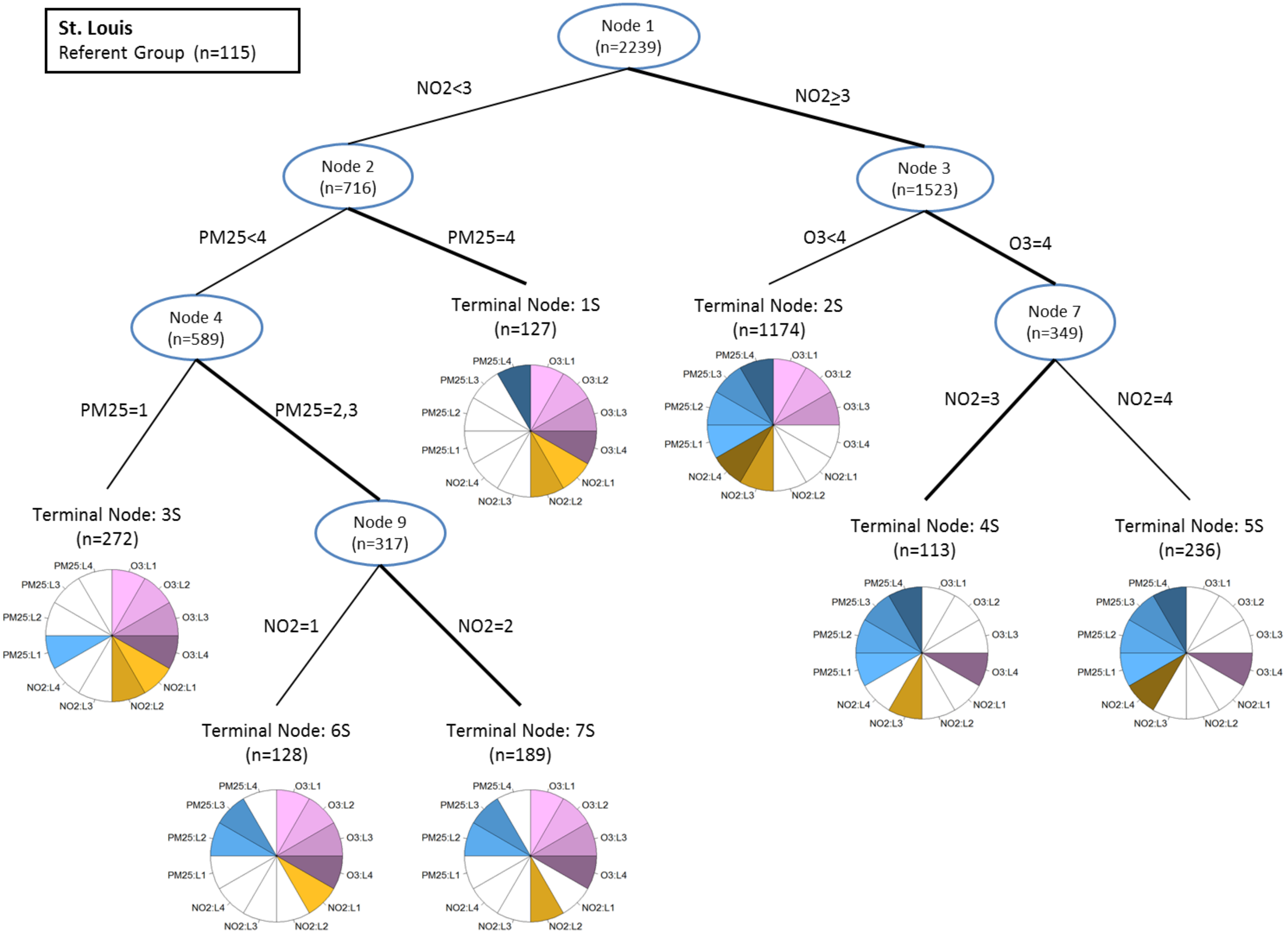

St. Louis
Referent Group (n=115)
Node 1
(n=2239)
NO2<3
NO2≥3
Node 2
(n=716)
Node 3
(n=1523)
PM25<4
PM25=4
O3<4
O3=4
Node 4
(n=589)
Terminal Node: 1S
(n=127)
Terminal Node: 2S
(n=1174)
Node 7
(n=349)
PM25=1
PM25=2,3
NO2=3
NO2=4
Terminal Node: 3S
(n=272)
Node 9
(n=317)
Terminal Node: 4S
(n=113)
Terminal Node: 5S
(n=236)
NO2=1
NO2=2
Terminal Node: 6S
(n=128)
Terminal Node: 7S
(n=189)
PM25:L4
PM25:L3
PM25:L2
PM25:L1
O3:L1
O3:L2
O3:L3
O3:L4
NO2:L1
NO2:L2
NO2:L3
NO2:L4

Figure 6.4. Classification tree illustrating the joint effects of O_3, NO_2 and $PM_{2.5}$ associated with emergency department visits for pediatric asthma in St. Louis (2001 –2007). Terminal nodes are numbered 1S-7S (S for St. Louis). For additional description see Figure 6.2.

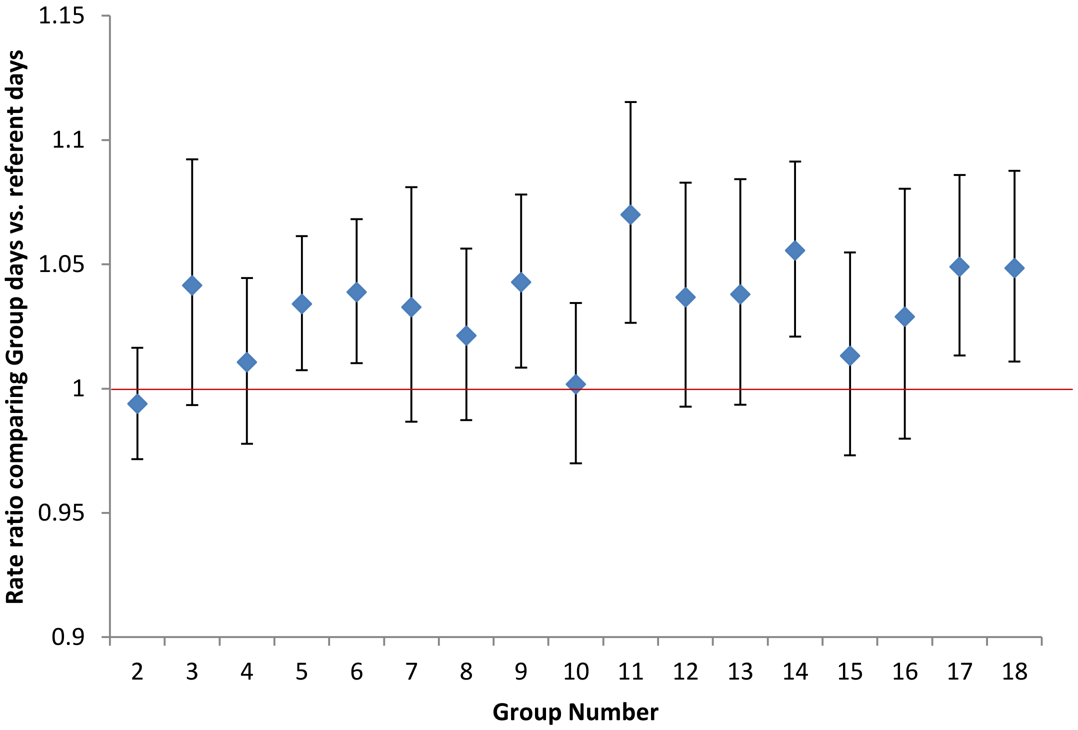


Figure 6.5. Rate ratios and 95% confidence intervals presented for the 17 C&RT Groups relative to the referent group (all three pollutants in the lowest level) for the three-city model. Rate ratios were calculated using indicator variables for each Group in a Poisson GLM model with city-specific control for confounding.

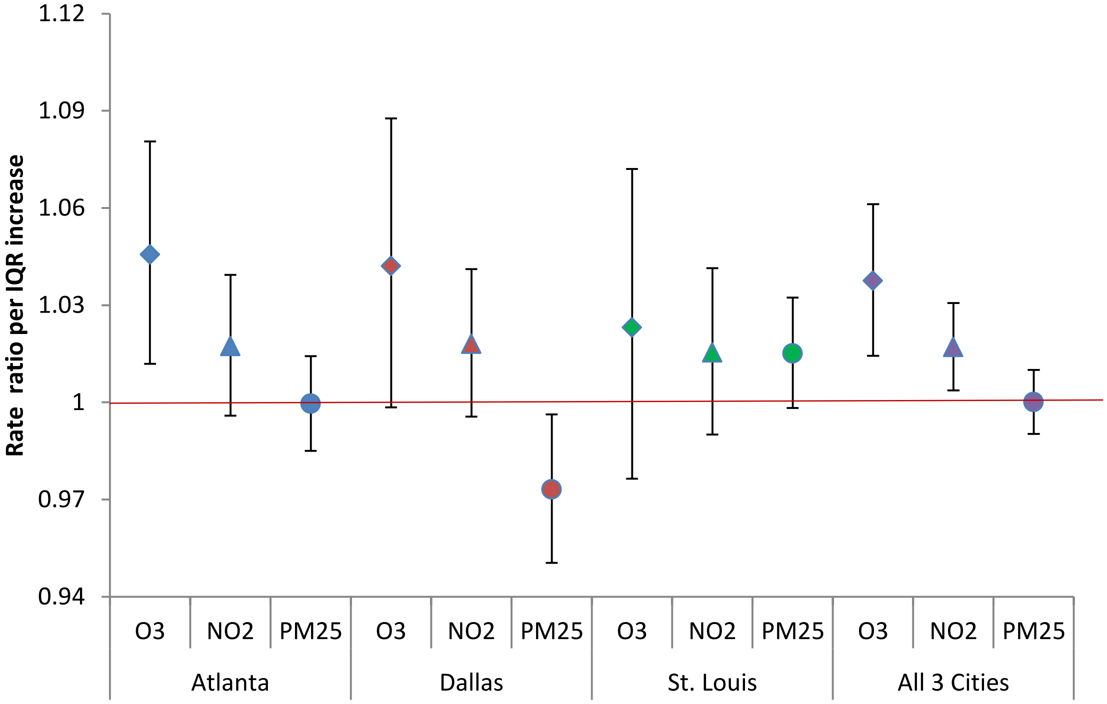


Figure 6.6. Rate ratios per IQR increase from the conventional multipollutant model with linear effects; results are presented separately for each city and all three-cities combined.

Chapter 7: Ensemble-Based Source Apportionment of Fine Particulate Matter and Emergency Department Visits for Pediatric Asthma (Study 3)

Introduction

There is increasing epidemiologic evidence that fine particulate matter with aerodynamic diameter <2.5 microns ($PM_{2.5}$) has a harmful effect on childhood asthma [1-7]. However $PM_{2.5}$ is a heterogeneous mixture of particles and there is a growing body of evidence that some particles are more detrimental to health than others [8-16]. Despite this, current regulatory strategy treats all particles that contribute to PM mass equally. Conducting epidemiological studies on source-apportioned PM, rather than total PM, may help identify the causal agents that precipitate acute asthmatic events and ultimately lead to more effective source-based regulation.

Apportioning total fine particulate matter mass into its contributing sources is traditionally done with receptor-based modeling [17]. Common techniques include chemical mass balance (CMB) modeling [18] and factor analytic approaches such as principle components analysis [19], UNMIX [20, 21] and positive matrix factorization [22]. More recently chemical transport models, such as the Community Multiscale Air Quality (CMAQ), which utilize emission inventories and meteorology, have been used to estimate the impact of $PM_{2.5}$ sources [23, 24]. While source apportionment (SA) techniques have proven to be a powerful tool in epidemiologic studies, each approach has its own set of limitations when included in health studies [25]. Indeed, a challenge with SA is that an accepted “gold standard” does not exist [26]. Without a gold standard approach for measuring source impacts, it is difficult to quantify the uncertainty associated with each of the techniques, which may lead to biased health estimates and underestimated standard errors.

One approach to overcome the limitations of individual SA models is to use an ensemble of source-apportioned results. Lee et al (2009) showed that an ensemble average of five different SA models led to a higher predicted-to-observed $PM_{2.5}$ ratio, fewer zero-impact days and a reduction in the day-to-day variability of the source impact estimates compared to single SA models [25]. Balachandran et al (2012) builds on this approach by propagating the uncertainties of each SA approach in the ensemble average [27] and then using Bayesian techniques to obtain multiple realizations of the source profiles to further capture the day-to-day uncertainties in the source apportionment techniques [28]. This Bayesian-based ensemble approach has been shown to provide results that are more consistent with independent observations and known emission sources compared with other single SA methods [28].

This is the first study to apply results from this novel ensemble-based source apportionment technique in an epidemiologic analysis. In this study we examine the association between ensemble-based $PM_{2.5}$ source impacts and emergency department visits for childhood asthma. By utilizing data from an 8.5-year time series study (1/1/2002 - 6/30/2010) in metropolitan Atlanta, with daily exposure and outcome observations, we also have the ability to examine lagged associations.

Methods

Exposure Data: Bayesian-based ensemble source apportionment

The Bayesian-based ensemble approach combines four independent SA methods, three of which were receptor-based: CMB using molecular markers [29] and CMB using gas-based constraints [30] and positive matrix factorization [31], as well as one chemical transport model [32]. All SA methods were conducted using concentrations from the Jefferson Street monitoring site in downtown Atlanta. Ensemble averaging was conducted iteratively. In the first step estimates of the source impacts and the uncertainties for each of the four SA methods were averaged with equal weighting. Next the root mean square error (RMSE) was calculated between each method with respect to the average for each source category. The inverse of the source-specific and method-specific RMSE was then used as weights to re-estimate the ensemble average source impacts and uncertainties. A more detailed description of these first two steps is provided in Balachandran et al (2012) [27]. The above algorithm results in each day having the same estimated source impact uncertainty because the RMSE was constant across days. To further account for the uncertainties in the ensemble weights, for each day, the source-specific and method-specific RMSEs were sampled independently from their posterior distributions and used as weights to recalculate the ensemble-averaged source impact. Finally, each posterior sample of daily source impact time series was used to derive CMB source profiles. [28].

This Bayesian ensemble method was applied to estimate two seasonal source profiles (July 2001 and January 2002), which in turn were used to estimate daily source impacts for the 8.5 year time series (1/1/2002 – 6/30/2010). Each day 10 source profiles were sampled from the seasonal source distribution and used in a CMB equation to get the daily concentration of each source. This resulted in 10 separate time series with daily

SA concentrations. A total of nine categories of sources were classified from the individual four SA methods: five primary sources and four secondary sources [25]. Primary sources included biomass burning (BURN), primary PM from coal combustion (COAL), construction and road dust (DUST), diesel vehicles and non-road engines (DV), and gasoline-fuel vehicles and engine sources (GV). Secondary sources included ammonium bisulfate, ammonium sulfate ammonium nitrate and other organic carbon, which was grouped as secondary organic carbon (SOC). Because the CMAQ simulations that contributed to the chemical transport model were biased high for the sulfates and nitrates [33] these three secondary sources were dropped from the analysis, leaving SOC the only secondary source.

Health Data

Daily emergency department (ED) visit data were collected from all hospitals in metropolitan Atlanta for the 8.5 year time series. Individual visits were restricted to pediatric patients ($\leq$18 years) living in zip codes within 5-county metropolitan Atlanta. We defined emergency department visits for asthma as all visits with an International Classification of Disease, 9th edition code for asthma (493.0-493.9) or wheeze (786.07).

Statistical Methods

We performed time-series analyses to estimate associations between the $PM_{2.5}$ sources and ED visits for pediatric asthma. We had an a priori interest in the association with lags 0-2 (same day and previous two days' exposure); but also an interest in exploring longer lag periods given previous published findings suggesting the effect of pollutants on asthma may be prolonged over many days [1, 2, 34]. As a result, we considered the following two initial models: 1) a single-source model with same day (lag 0) and previous week (lag 1-7) source impacts using an unconstrained distributed lag, that is, with an individual term for each lag in the same model; and 2) the same single-source model with exposure considered for only lags 0-2, controlling for source concentrations on lag days 3-7, again with individual terms for each lag in the same model. All associations were calculated as the cumulative rate ratio for a 1 $\mu g/m^3$ increase in the source concentration for each exposure day.

Long-term temporal trends were controlled using a cubic spline with 8 knots per year. Separate cubic terms were included for average maximum temperature lag 0-2 and lag 3-7. Similarly, two cubic terms were included for average dew point lag 0-2 and lag 3-7. Indicator terms were included for season, day-of-week, federal holidays, and the days after Thanksgiving and Christmas. To further control for temporal and meteorological trends we included interaction terms for season and day-of-week, as well as season and the maximum temperature cubic terms (for lag 0-2 and lag 3-7).

The single-source model analysis was conducted separately for each of the six sources (BURN, COAL, DUST, DV, GV, and SOC) as well as for total $PM_{2.5}$. Because previous epidemiologic analyses in Atlanta have found a strong association between O_3 and pediatric asthma [2, 3, 35, 36]), we also ran the same models controlling for O_3, using the same unconstrained 8-day distributed lag structure for O_3, as for the main exposure. To account for potential confounding by sources not included in the model, we ran the single-source model with simultaneous control for the 8-day moving average of the other five sources. Additional sensitivity analyses were conducted to compare the following lag structures: lag 0-7 constrained with a cubic polynomial, the 8-day moving average of lags 0-7, and lags 0 through 7 considered separately (undistributed).

To account for the additional uncertainty from the ensemble-averaged SA concentrations, each analysis was run 10 times, once for each of the 10 separate ensemble time series (aka “runs”). Multiple imputation methods were used to arrive at a combined point estimate and variance for each analysis as follows. The summary regression coefficient, Q was obtained by taking the mean of the regression coefficients from each run, where *m*=10.

$$\overline{Q} = \frac{1}{m}\sum\nolimits_{i=1}^{m}\hat{Q}_i$$

Imputation-corrected variances were calculated according to the method described by Rubin, 1987 [37]. The first step required calculating the average variance from the ensemble runs (within imputation variance) (W) and the variance of the ensemble run coefficients (between imputation variance) (B).

$$\overline{W} = \frac{1}{m}\sum\nolimits_{i=1}^{m}\hat{W}_i$$

$$B = \frac{1}{m-1}\sum\nolimits_{i=1}^{m}(\overline{Q} - \hat{Q}_i)^2$$

With these two quantities the total imputation-corrected variance (T) can be calculated as:

$$T = \overline{W} + (1 + {}^{1}\!/_{m})B$$

and follows a t-distribution with degrees of freedom (ν) equal to:

$$\nu_m = (m-1)\left[1 + \frac{\overline{W}}{(1 + {}^{1}\!/_{m})B}\right]^2 \text{ [37].}$$

All analyses were conducted using SAS® v9.3 (Statistical Analysis System; North Carolina).

Results

There were 2,170 days with SA observations for lags 0 through 7 for all sources and for $PM_{2.5}$. Table 7.1 summarizes the pollutant, meteorological and hospitalization data included in the model. On average, 71 ED visits for acute asthma or wheeze occurred each day among children 0-18 years. Mean concentrations of fine particulate matter and O_3 during the study period were 14.51 $\mu g/m^3$ and 40.61ppb, respectively.

Summary statistics of the daily source concentrations from each of the 10 ensemble runs are presented in Table 7.2. When averaged across all ensemble runs, BURN had the highest mean concentration (2.81 $\mu g/m^3$) and greatest standard deviation (2.59), while COAL had the lowest (0.12 $\mu g/m^3$, 0.12). The mean concentration of DV was greater than GV (1.01 vs. 0.88 $\mu g/m^3$) with DV showing greater average standard deviation within each ensemble run (0.95 vs. 0.69). The last column in Table 7.2 shows the average correlation between each of the 10 ensemble runs by source. DUST had the highest correlation between the runs (r=0.98), while the other five sources had similarly moderate average correlations between the runs, ranging from r=0.66 to r=0.74.

Table 7.3 contains the correlations between each of the sources and total $PM_{2.5}$ and O_3. The strongest correlations were observed between O_3 and DUST, as well as $PM_{2.5}$ and DV (both with r=0.48). Among the sources, SOC and BURN exhibited the strongest negative correlation (r=-0.46) while SOC and DV exhibited the strongest positive correlation (r=0.44).

The results from the primary health analyses for each of the sources are shown in Figure 7.1. The results for each source are shown for three separate models: the single-source model (with exposure modeled using an unconstrained distributed lag), the single-source model with the addition of O_3 control and the single-source model with simultaneous control for the other sources. Results for exposure to total $PM_{2.5}$ are presented for the single-source model and single-source model with O_3 control in Figure 7.2. For each model in Figures 7.1 and 7.2, two separate exposures were considered: cumulative exposure to lag 0-2, controlling for lag 3-7 separately, and cumulative exposure to lag 0-7. It is important to note that the y-axes in Figures 7.1 and 7.2 differ for each source, a consequence of reporting the associations in terms of a 1 $\mu g/m^3$ increase, as opposed to an IQR increase.

The RR for lag 0-7 was larger than lag 0-2 for all sources and $PM_{2.5}$, with the exception of DUST and SOC, in all three models (Figures 7.1 and 7.2). The single-source model resulted in significant associations for BURN lag 0-7 (RR: 1.02, 95% CI: 1.01, 1.03), DV lag 0-7 (RR: 1.05, 95% CI: 1.01, 1.08), GV lag 0-2 (RR: 1.03, 95% CI: 1.01, 1.05) and lag 0-7 (RR: 1.07, 95% CI: 1.03, 1.11) and $PM_{2.5}$ lag 0-2 (RR: 1.00, 95% CI: 1.00, 1.01) and lag 0-7 (RR: 1.01, 95% CI: 1.00, 1.01). When O_3 was added to the model, the associations with DV lag 0-7, GV lag 0-2, and $PM_{2.5}$ lag 0-2 were null, while the lag 0-7 RRs for BURN, GV and $PM_{2.5}$ remained significant. Controlling for all other sources in the same model resulted in a decrease in the point estimates and only DV lag 0-7 remained significant (RR: 1.06, 95% CI: 1.00, 1.12). The RRs for the O_3 associations, from the single-source models with O_3 control, are presented in Figure 7.5. For comparison we also present the results from an O_3-only model for the association of lags 0-7, modeled with an unconstrained distributed structure. There is no appreciable difference in the point estimates in Figure 7.5 between the O_3-only model and the association with O_3 from the single-source models with O_3 control. When $PM_{2.5}$ and O_3 are in the same model the O_3 association is non-significant (RR: 1.06, 95% CI: 0.99, 1.13).

Table 7.4 uses the results from the single-source model measuring cumulative exposure to lag 0-7 to demonstrate how the 10 ensemble runs contributed to the summary point estimates and standard errors. The standard error of the point estimates was nearly 10 times greater for COAL compared with the other sources. The ratio of the imputation-corrected standard error to the average standard error shows the degree to which the confidence intervals were inflated due to the propagation of error from the ensemble runs. The degree of inflation ranged from 3% (DUST) to 20% (DV).

Results from the sensitivity analyses considering alternative lag specifications, specifically two constrained lag structures (a cubic polynomial and 8-day moving average) and lags 0 through 7 modeled separately (i.e. undistributed), resulted in nearly identical point estimates for all sources and are therefore not shown. The individual unconstrained distributed lag results from the single-source model are shown for the sources and total $PM_{2.5}$ in Figures 7.3 and 4, respectively. The graphs for BURN, DV, GV and $PM_{2.5}$ suggest an immediate same-day and day-after association (lag 0 and 1) followed by a delayed association occurring between lags 4-7. For both BURN and DV the greatest association was seen on lag 7. Throughout all analysis the associations for primary coal combustion, dust, and secondary organic carbons were consistent with the null.

Discussion

These analyses using source apportioned-exposure estimates suggest that some, but not all, sources of fine particulate matter are hazardous to a child's respiratory health. In particular, traffic-related sources (diesel and gasoline vehicles), as well as biomass burning, were associated pediatric asthma ED visits when considering the cumulative eight-day exposure. The exposures to gasoline vehicles and biomass burning was statistically significant after controlling for O_3 in the model. When all sources were included in the same model, only the eight-day exposure to diesel vehicles was statistically significant.

The finding that diesel and gasoline vehicle sources were associated with childhood asthma is well supported in the literature. Studies have found that residential proximity to roadways is associated with both incident asthma [38] and asthma exacerbation [39]. A study looking at the associations between source apportioned $PM_{2.5}$ and asthmatic children found traffic-related exposures to be most harmful, leading to a statistically significant increase in asthmatic symptoms [40]. In particular, previous studies have found indicators of diesel exhaust to be associated with hospital admissions for asthma [34] and airway inflammation in asthmatics [41]. The results of these studies are further corroborated by our model with all sources, which found that diesel vehicles (DV) exhibited the strongest association with asthma ED visits.

Our decision to control for ozone was driven by a concern for confounding, given that previous studies in Atlanta that have found O_3 to be strongly associated with pediatric asthma [2, 3, 35, 36, 42]. Nonetheless, there is the potential for O_3 to be on one of the causal pathways if source emissions lead to O_3 formation, which in-turn leads to an increase in asthma ED visits. If this were the case then controlling for O_3 would result in a bias towards the null. Looking at the correlations between O_3 and the sources in Table 7.3, this appears to be of minimal concern; only DUST and SOC exhibited moderate correlations with O_3 (r=0.48 and r=0.43, respectively). The slight attenuation of rate ratios in Figure 7.1 for the single-source models with O_3 control, as well as the similarity between the O_3-only and single-source with O_3 model results in Figure 7.5, provide further evidence that confounding by O_3 is unlikely to be a major concern for the sources. There does, however, appear to be some confounding by O_3 of the $PM_{2.5}$ association for lags 0-2 (Figure 7.2).

Similarly we chose to incorporate all sources in the third model to account for potential between-source confounding. The results in Figure 7.1 suggest that there may be some confounding present in the single-source model results, particularly in the associations of BURN and GV, which were significant in the single-source model but non-significant in the model with control for all sources.

The availability of daily source concentrations over an 8.5 year period enabled us to examine different extended lag structures. Many past studies have been limited in their ability to examine the lag structure of $PM_{2.5}$ sources and constituents because much of the USA Environmental Protection Agency [43] monitoring data are only available on every third or sixth day [11, 14, 44, 45]. Studies that have looked at the temporal patterns of $PM_{2.5}$ exposure and acute asthma exacerbations have consistently found evidence of a lagged effect [1, 2, 34]. In particular, the lag pattern we observed for total $PM_{2.5}$ (Figure 7.4) is consistent with an earlier Atlanta-based study by Peel et al (2005) that analyzed the association between $PM_{2.5}$ and asthma ED visits from 1998 – 2000, with an unconstrained distributed lag structure, and found lags 0 and 6 to have the strongest association with asthma ED visits [2]. A source apportionment study by Halonen et al (2008) found the strongest associations between traffic sources and pediatric asthma occurred between lags 3-5 [1], similar to the results in Figure 7.3 which suggest the greatest diesel associations occurred between lags 4-7. There is biological plausibility behind these findings of a delayed effect, as ultrafine particles have been shown to penetrate deep into the lung, particularly in persons suffering from asthma [46] which could lead to inflammation in the alveolar region of the lungs [47].

Further evidence of the lagged effect can be seen when the RR for the cumulative association of lag 0-2 is compared with that of lag 0-7 in Figures 7.1 and 7.2. With the exception of DUST and SOC, both of which are consistent with the null, all sources and total $PM_{2.5}$ showed greater associations for lag 0-7, compared with lag 0-2. We chose to include lag 0-2 because it is commonly reported in both the asthma and source apportionment literature [3, 9, 40, 48-50]; however we did so *controlling* for lags 3-7. Our individual lag results from both the distributed and undistributed models suggested a strong association with lags 3-7 and thus excluding these lags may result in confounding of the lag 0-2 association. As a result our lag 0-2 associations are likely to be smaller than those reported from other studies, which may be biased upwards due to confounding by longer lags not included in the model.

In our exploration of different lag structures, as part of the sensitivity analysis, we found that lag structure had very little effect on our results. The point estimates for the cumulative lag 0-2 and 0-7 exposures were nearly identical between the unconstrained, cubic constrained, and moving average models, suggesting that the overall temporal associations captured in the results are robust. Similarly, the individual lag results from the distributed and undistributed models displayed comparable patterns and point estimates for each lag across all sources. There appears to be minimal temporal correlation between the lags for most sources; the Spearman correlation coefficient between today and yesterday's concentrations was less than r=0.5 for all sources, with the exception of DUST which was r=0.63.

While we typically conceptualize lagged exposures presented in this, and other, analyses as the cumulative effect of an increase of 1 μg/m^3 in each of the lagged days examined, for some sources such an exposure may be unlikely to occur. For example, the cumulative effect of 1 μg/m^3 increase in biomass burning over eight days may not be realistic, given that most burn events occur sporadically and over short time intervals. An alternative and equivalent way to conceptualize the RR is the effect of a single-day 1 μg/m^3 increase in biomass burning *sustained* over 8 days. For many source exposures this interpretation of the RR may be more realistic. This interpretation may also have a more direct correspondence to regulation (e.g. what will be the sustained effect of shutting down a power plant for one day?).

An important contribution of this paper is the use of ensemble-based source apportionment data in a health analysis. By using ensemble-averaged results from four different source apportionment methods we were better able to account for the uncertainties of each approach, while alleviating some of the concerns regarding inter-method variability. In order to propagate the uncertainty, all reported model results for the sources are the combination of 10 separate ensemble runs, with the net result of inflating the summary confidence intervals by 3-20% (Table 7.4). The relatively small increase in the CIs for DUST (3%) can be attributed to the strong correlation between the ensemble runs (r=0.98, Table 7.2), which in turn is a reflection of the relative agreement between SA methods. Conversely, the results suggest that BURN has the greatest between-ensemble variability in source concentration, as is evident from the larger standard errors and weakest between-run correlation in Table 7.2. The greatest increase in confidence interval width occurred with DV, and is a function of the ratio of the average standard error for the ensemble runs (within-run SE) vs. the standard error of the point estimates (between-run SE). In the case of DV this ratio was relatively small, meaning that there was greater variability in the RR results between the ensemble runs.

Indeed as with any study that uses ambient concentrations to represent population exposure, measurement error caused by spatial misalignment is a concern. In our study, SA data from a single monitor was used to represent exposure for a 5-county area. While extrapolating SA data from a single monitoring site to a greater metropolitan area is common [26, 51, 52], one must do so with caution. Some sources of PM are likely to be more spatially homogenous than others (e.g. secondary organic carbons will be more homogenous than local vehicle emissions) and these differences in spatial variation will lead to differing degrees of spatial misalignment [53]. For the more heterogeneous sources where exposure misclassification is expected to be the greatest (e.g. BURN, COAL, DV, & GV) the observed point estimates are likely to be biased to the null [54].

Conclusion

In this study we found that fine particulate matter generated from diesel and gasoline vehicle sources, as well as biomass burnings, was associated with a significant increase in emergency department visits for acute asthma-related events among children 0-18 years. Our results, which corroborate previous findings, suggest that for children, the harmful effects from a single-day's exposure to these sources are sustained for the following week, with some of the largest effects seen between lags 4-7. This study takes advantage of a novel ensemble-based source apportionment technique, which helps to minimize the potential for bias from relying on any single SA method and provides a means for inflating the confidence intervals around the point estimates to account for the uncertainty in SA methods. As a result of this latter feature, our results may be more conservative than those from single SA studies. Nonetheless, we found some sources to have significantly harmful associations, lending credence to the belief that sources have varying toxicity and providing further incentive for source-based regulation.

References

1. Halonen JI, Lanki T, Yli-Tuomi T, Kulmala M, Tiittanen P, J. P: **Urban air pollution, and asthma and COPD hospital emergency room visits**. *Thorax* 2008, **63**(7):635-641.

2. Peel JL, Tolbert PE, Klein M, Metzger KB, Flanders WD, Todd K, Mulholland JA, Ryan PB, Frumkin H: **Ambient air pollution and respiratory emergency department visits**. *Epidemiology* 2005, **16**(2):164-174.

3. Strickland MJ, Darrow LA, Klein M, Flanders WD, Sarnat JA, Waller LA, Sarnat SE, Mulholland JA, Tolbert PE: **Short-term associations between ambient air pollutants and pediatric asthma emergency department visits**. *Am J Respir Crit Care Med* 2010, **182**(3):307-316.

4. Mar TF, Larson TV, Stier RA, Claiborn C, Koenig JQ: **An analysis of the association between respiratory symptoms in subjects with asthma and daily air pollution in Spokane, Washington**. *Inhal Toxicol* 2004, **16**(13):809-815.

5. Rabinovitch N, Strand M, Gelfand EW: **Particulate levels are associated with early asthma worsening in children with persistent disease**. *Am J Respir Crit Care Med* 2006, **173**(10):1098-1105.

6. Ranzi A, Gambini M, Spattini A, Galassi C, Sesti D, Bedeschi M, Messori A, Baroni A, Cavagni G, Lauriola P: **Air pollution and respiratory status in asthmatic children: hints for a locally based preventive strategy. AIRE study**. *European journal of epidemiology* 2004, **19**(6):567-576.

7. Slaughter JC, Lumley T, Sheppard L, Koenig JQ, Shapiro GG: **Effects of ambient air pollution on symptom severity and medication use in children with asthma**. *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology* 2003, **91**(4):346-353.

8. Andersen ZJ, Wahlin P, Raaschou-Nielsen O, Scheike T, Loft S: **Ambient particle source apportionment and daily hospital admissions among children and elderly in Copenhagen**. *J Expo Sci Environ Epidemiol* 2007, **17**(7):625-636.

9. Bell ML, Ebisu K, Leaderer BP, Gent JF, Lee HJ, Koutrakis P, Wang Y, Dominici F, Peng RD: **Associations of PM Constituents and Sources with Hospital Admissions: Analysis of Four Counties in Connecticut and Massachusetts (USA) for Persons >/= 65 Years of Age**. *Environ Health Perspect* 2013.

10. Bell ML, Ebisu K, Peng RD, Samet JM, Dominici F: **Hospital admissions and chemical composition of fine particle air pollution**. *Am J Respir Crit Care Med* 2009, **179**(12):1115-1120.

11. Ito K, Mathes R, Ross Z, Nadas A, Thurston G, Matte T: **Fine particulate matter constituents associated with cardiovascular hospitalizations and mortality in New York City**. *Environ Health Perspect* 2011, **119**(4):467-473.

12. Mar TF, Norris GA, Koenig JQ, Larson TV: **Associations between air pollution and mortality in Phoenix, 1995-1997**. *Environ Health Perspect* 2000, **108**(4):347-353.

13. Ostro B, Feng WY, Broadwin R, Green S, Lipsett M: **The effects of components of fine particulate air pollution on mortality in california: results from CALFINE**. *Environ Health Perspect* 2007, **115**(1):13-19.

14. Peng RD, Bell ML, Geyh AS, McDermott A, Zeger SL, Samet JM, Dominici F: **Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution**. *Environ Health Perspect* 2009, **117**(6):957-963.

15. Zanobetti A, Franklin M, Koutrakis P, Schwartz J: **Fine particulate air pollution and its components in association with cause-specific emergency admissions**. *Environ Health* 2009, **8**:58.

16. Ozkaynak H, Thurston GD: **Associations between 1980 U.S. mortality rates and alternative measures of airborne particle concentration**. *Risk analysis : an official publication of the Society for Risk Analysis* 1987, **7**(4):449-461.

17. Hopke PK, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Larson TV *et al*: **PM source apportionment and health effects: 1. Intercomparison of source apportionment results**. *J Expo Sci Environ Epidemiol* 2006, **16**(3):275-286.

18. Watson JG, Cooper JA, Huntzicker JJ: **The effective variance weighting for least squares calculations applied to the mass balance receptor model**. *Atmospheric Environment* 1984, **18**(7):1347-1355.

19. Blifford JIH, Meaker GO: **A factor analysis model of large scale pollution**. *Atmospheric Environment* 1967, **1**:147-157.

20. Henry RC, Kim BM: **Extension of self-modeling curve resolution to mixtures of more than three components. Part 1: finding the basic feasible region.** . *Chemom Intell Lab Systems* 1990, **8**:205–216.

21. Kim B.M., Henry RC: **Extension of self-modeling curve resolution to mixtures of more than three components. Part 2: finding the complete solution**. *Chemom Intell Lab Systems* 1999, **49**:67-77.

22. Paatero P: **Least squares formulation of Robust, non-negative factor analysis**. *Chemom Intell Lab Systems* 1997, **37**:23–35.

23. Koo B, Wilson GM, Morris RE, Dunker AM, Yarwood G: **Comparison of source apportionment and sensitivity analysis in a particulate matter air duality model**. *Environmental science & technology* 2009, **43**(17):6669-6675.

24. Marmur A, Park SK, Mulholland JA, Tolbert PE, Russell AG: **Source apportionment of PM2.5 in the southeastern United States using receptor and emissions-based models: conceptual differences and implications for timeseries health studies**. *Atmospheric Environment* 2006, **40**:2533-2551.

25. Lee D, Balachandran S, Pachon J, Shankaran R, Lee S, Mulholland JA, Russell AG: **Ensemble-trained PM2.5 source apportionment approach for health studies**. *Environmental science & technology* 2009, **43**(18):7023-7031.

26. Thurston GD, Ito K, Mar T, Christensen WF, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Larson TV *et al*: **Workgroup report: workshop on source apportionment of particulate matter health effects--intercomparison of results and implications**. *Environ Health Perspect* 2005, **113**(12):1768-1774.

27. Balachandran S, Pachon J, Hu Y, Lee D, Mulholland JA, Russell AG, : **Ensemble-trained source apportionment of fine particulate matter and method uncertainty analysis**. *Atmospheric Environment* 2012, **61**:387-394.

28. Balachandran S, Chang HH, Pachon JE, Holmes HA, Mulholland JA, Russell AG: **Bayesian-based ensemble source apportionment of PM2.5**. *Environmental science & technology* 2013, **47**(23):13511-13518.

29. Zheng M, Cass GR, Schauer JJ, Edgerton ES: **Source apportionment of PM2.5 in the Southeastern United States using solvent-extractable organic compounds as tracers**. *Environmental science & technology* 2002, **36**(11):2361-2371.

30. Marmur A, Unal A, Mulholland JA, Russell AG: **Optimization-based source apportionment of PM2.5 incorporating gas-to-particle ratios**. *Environmental science & technology* 2005, **39**(9):3245-3254.

31. Paatero P, Tapper U: **Positive matrix factorization--A nonnegative factor model with optimal utilization of error-estimates of data values**. *Environmetrics* 1994, **5**(2):111-126.

32. Byun D, Schere KL: **Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeline system**. *Appl Mech Rev* 2006, **59**(1-6):51-77.

33. Dennis R, Fox T, Fuentes M, Gilliland A, Hanna S, Hogrefe C, Irwin J, Rao ST, Scheffe R, Schere K *et al*: **A framework for evaluating regional-scale numerical photochemical modeling systems**. *Environmental fluid mechanics (Dordrecht, Netherlands : 2001)* 2010, **10**(4):471-489.

34. Kim SY, Peel JL, Hannigan MP, Dutton SJ, Sheppard L, Clark ML, Vedal S: **The temporal lag structure of short-term associations of fine particulate matter chemical constituents and cardiovascular and respiratory hospitalizations**. *Environ Health Perspect* 2012, **120**(8):1094-1099.

35. Tolbert PE, Mulholland JA, MacIntosh DL, Xu F, Daniels D, Devine OJ, Carlin BP, Klein M, Dorley J, Butler AJ *et al*: **Air quality and pediatric emergency room visits for asthma in Atlanta, Georgia, USA**. *Am J Epidemiol* 2000, **151**(8):798-810.

36. Friedman MS, Powell KE, Hutwagner L, Graham LM, Teague WG: **Impact of changes in transportation and commuting behaviors during the 1996 Summer Olympic Games in Atlanta on air quality and childhood asthma**. *JAMA : the journal of the American Medical Association* 2001, **285**(7):897-905.

37. Rubin DB: **Multiple imputation for nonresponse in surveys**. New York: J. Wiley & Sons; 1987.

38. McConnell R, Islam T, Shankardass K, Jerrett M, Lurmann F, Gilliland F, Gauderman J, Avol E, Kunzli N, Yao L *et al*: **Childhood incident asthma and traffic-related air pollution at home and school**. *Environ Health Perspect* 2010, **118**(7):1021-1026.

39. Boehmer TK, Foster SL, Henry JR, Woghiren-Akinnifesi EL, Yip FY: **Residential proximity to major highways - United States, 2010**. *Morbidity and mortality weekly report Surveillance summaries (Washington, DC : 2002)* 2013, **62 Suppl 3**:46-50.

40. Gent JF, Koutrakis P, Belanger K, Triche E, Holford TR, Bracken MB, Leaderer BP: **Symptoms and medication use in children with asthma and traffic-related sources of fine particle pollution**. *Environ Health Perspect* 2009, **117**(7):1168-1174.

41. Patel MM, Chillrud SN, Deepti KC, Ross JM, Kinney PL: **Traffic-related air pollutants and exhaled markers of airway inflammation and oxidative stress in New York City adolescents**. *Environ Res* 2013, **121**:71-78.

42. White MC, Etzel RA, Wilcox WD, Lloyd C: **Exacerbations of childhood asthma and ozone pollution in Atlanta**. *Environ Res* 1994, **65**(1):56-68.

43. US Department of Health and Human Services, Agency for Healthcare Research and Quality: **2008 National Healthcare Disparities Report**. In. Rockville, MD: US Department of Health and Human Services, Agency for Healthcare Research and Quality; 2008

44. Ito K, Christensen WF, Eatough DJ, Henry RC, Kim E, Laden F, Lall R, Larson TV, Neas L, Hopke PK *et al*: **PM source apportionment and health effects: 2. An investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in Washington, DC**. *J Expo Sci Environ Epidemiol* 2006, **16**(4):300-310.

45. Ostro B, Roth L, Malig B, Marty M: **The effects of fine particle components on respiratory hospital admissions in children**. *Environ Health Perspect* 2009, **117**(3):475-480.

46. Anderson PJ, Wilson JD, Hiller FC: **Respiratory tract deposition of ultrafine particles in subjects with obstructive or restrictive lung disease**. *Chest* 1990, **97**:1115-1150.

47. Peters A, Wichmann HE, Tuch T, Hienrich J, Heyder J: **Respiratory effects are associated with the number of ultrafine particles**. *Am J Respir Crit Care Med* 1997, **155**:1376–1383.

48. Sacks JD, Rappold AG, Davis Jr JA, Richardson DB, Waller AE, Luben TJ: **Influence of Urbanicity and County Characteristics on the Association**

between Ozone and Asthma Emergency Department Visits in North Carolina. *Environ Health Perspect* 2014.

49. Darrow LA, Hess J, Rogers CA, Tolbert PE, Klein M, Sarnat SE: **Ambient pollen concentrations and emergency department visits for asthma and wheeze**. *J Allergy Clin Immunol* 2012, **130**(3):630-638.e634.

50. Delfino RJ, Zeiger RS, Seltzer JM, Street DH, McLaren CE: **Association of asthma symptoms with peak particulate air pollution and effect modification by anti-inflammatory medication use**. *Environ Health Perspect* 2002, **110**(10):A607-617.

51. Sarnat JA, Marmur A, Klein M, Kim E, Russell AG, Sarnat SE, Mulholland JA, Hopke PK, Tolbert PE: **Fine particle sources and cardiorespiratory morbidity: an application of chemical mass balance and factor analytical source-apportionment methods**. *Environ Health Perspect* 2008, **116**(4):459-466.

52. Lall R, Ito K, Thurston GD: **Distributed lag analyses of daily hospital admissions and source-apportioned fine particle air pollution**. *Environ Health Perspect* 2011, **119**(4):455-460.

53. Bell ML, Ebisu K, Peng RD: **Community-level spatial heterogeneity of chemical constituent levels of fine particulates and implications for epidemiological research**. *J Expo Sci Environ Epidemiol* 2011, **21**(4):372-384.

54. Strickland MJ, Gass KM, Goldman GT, Mulholland JA: **Effects of ambient air pollution measurement error on health effect estimates in time-series studies: a simulation-based analysis**. *J Expo Sci Environ Epidemiol* 2013.

Table 7.5 Summary statistics for fine particulate matter (concentrations in µg/m^3), ozone (ppb), meteorology, and emergency department visits for pediatric asthma in 5-county Atlanta (2002 – June, 2010).

Variable	Number of days[a]	Median	Mean (SD)	Minimum	Maximum	IQR[b]
Pollutant						
Fine particulate matter ($PM_{2.5}$ µg/m^3)	2170	13.18	14.51 (7.33)	1.06	72.56	9.16
Ozone (O_3 ppb)	2090	39.34	40.61 (19.16)	0.52	116.37	28.09
Meteorology						
Maximum temperature (°C)	2170	23	21.96 (8.37)	-1	40	13
Dew point (°C)	2163	11	9.64 (9.34)	-20	24	16
Health Outcome						
Asthma/Wheeze ED visits	2170	68	70.68 (28.75)	13	220	39

[a] The analysis was restricted to days when all sources and fine particulate matter were non-missing for the 8-day lag.

[b] Inter-quartile range

Table 7.7. Summary statistics for the source impacts, averaged across 10 ensemble runs. The standard deviation across runs is given in parentheses. All results are reported in μg/m^3.

Source	**Minimum**	**Median**	**Mean**	**Maximum**	**Standard Deviation**	**Inter-Quartile Range**	**Correlation[a] Between Ensemble Runs**
	Mean (SD)	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*
Biomass burning (BURN)	0 (0)	2.05 (0.048)	2.81 (0.029)	27.68 (6.04)	2.59 (0.081)	2.72 (0.063)	0.66 (0.012)
Primary coal combustion (COAL)	0 (0)	0.09 (0.001)	0.12 (0.001)	1.08 (0.196)	0.12 (0.002)	0.14 (0.004)	0.70 (0.012)
Dust/resuspended soil (DUST)	0 (0.002)	0.25 (0.001)	0.38 (0.001)	7.75 (1.485)	0.46 (0.012)	0.27 (0.002)	0.98 (0.001)
Diesel vehicles (DV)	0 (0)	0.79 (0.016)	1.01 (0.012)	9.52 (0.646)	0.95 (0.017)	0.99 (0.016)	0.74 (0.009)
Gasoline vehicles (GV)	0.02 (0.01)	0.66 (0.012)	0.81 (0.008)	7.12 (0.702)	0.69 (0.008)	0.66 (0.013)	0.74 (0.010)
Secondary organic carbon (SOC)	0 (0)	1.32 (0.029)	1.6 (0.013)	27.26 (1.305)	1.65 (0.019)	2.03 (0.040)	0.67 (0.019)

[a] Mean Spearman correlation calculated from all pairwise runs

Table 7.9. Spearman correlation between sources (averaged across all 10 ensemble runs), fine particulate matter, and ozone.

	BURN	COAL	DUST	DV	GV	SOC	$PM_{2.5}$	O_3
Biomass burning (BURN)	1.00							
Primary coal combustion (COAL)	0.21	1.00						
Dust/resuspended soil (DUST)	0.02	0.13	1.00					
Diesel vehicles (DV)	0.00	0.20	0.25	1.00				
Gasoline vehicles (GV)	0.40	0.09	0.14	0.22	1.00			
Secondary organic carbon (SOC)	-0.46	0.03	0.27	0.44	-0.12	1.00		
Fine particulate matter ($PM_{2.5}$)	0.20	0.19	0.41	0.48	0.30	0.46	1.00	
Ozone (O_3)	-0.23	0.02	0.48	0.12	-0.11	0.43	0.47	1.00

Table 7.11. Summary statistics for the standard error of the point estimate from the ensemble runs, where each measure is computed from the mean of 10 ensemble runs using the single-source model to calculate the rate ratio of a combined increase of 1µg/m^3 over lags 0-7.

Source	Standard Error of Point Estimates Across Ensembles	Average Standard Error for Ensemble Runs	Imputation-Corrected Standard Error	Ratio of Imputation-Corrected SE / Average SE
BURN	0.0027	0.0049	0.0057	1.1539
COAL	0.0675	0.1161	0.1359	1.1714
DUST	0.0064	0.0269	0.0277	1.0309
DV	0.0085	0.0134	0.0161	1.1992
GV	0.0077	0.0181	0.0198	1.0939
SOC	0.0049	0.0083	0.0097	1.1780

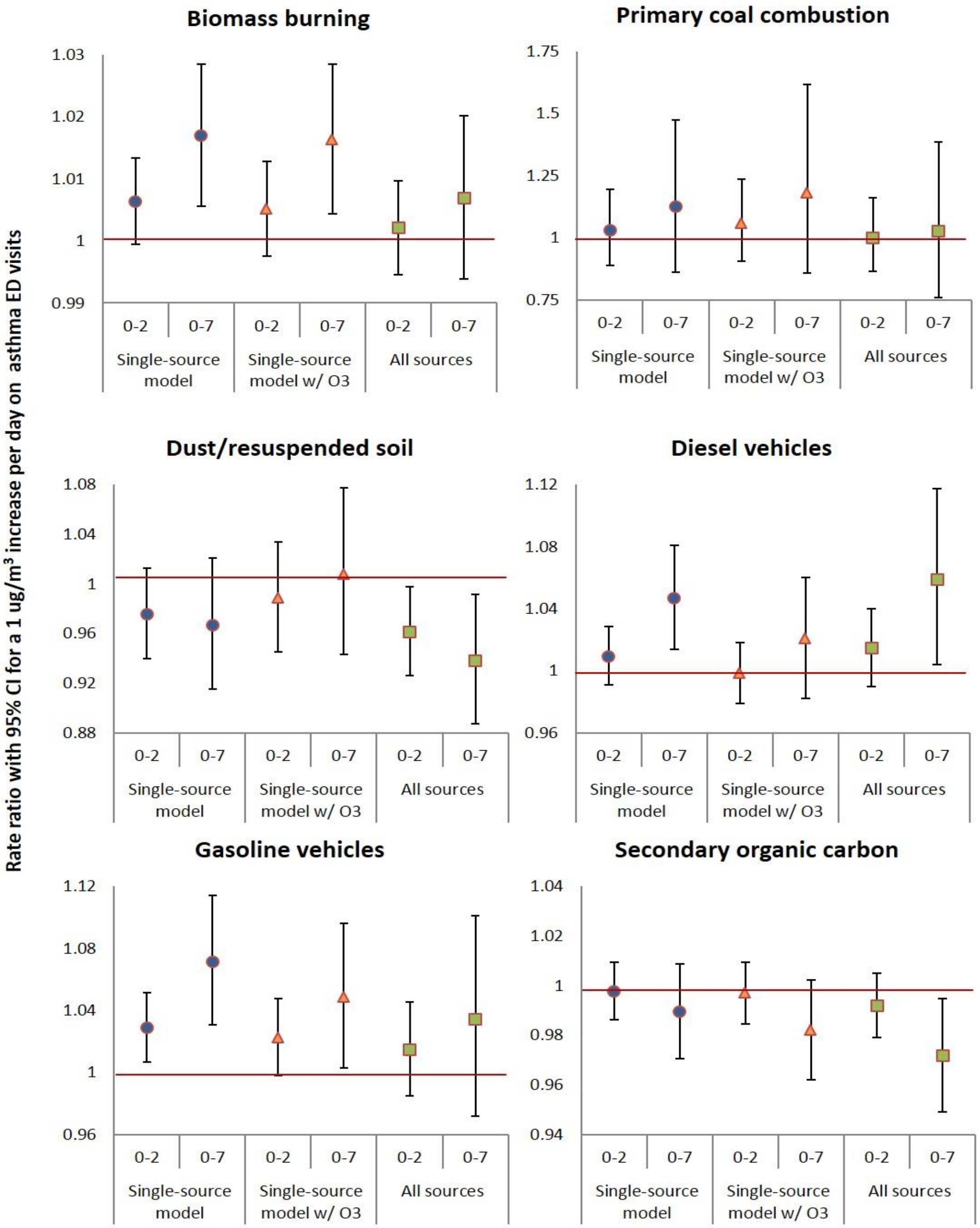


Figure 7.2. Rate ratios and 95% confidence intervals for the effect of a 1µg/m^3 increase in source concentration on pediatric asthma ED visits, presented for lags 0-2 (controlling for lags 3-7) and lags 0-7. Blue circles, orange triangles and green squares represent results from the single-source model with an unconstrained distributed lag structure, the same model with O_3 control, and the same model controlling for all sources simultaneously, respectively.

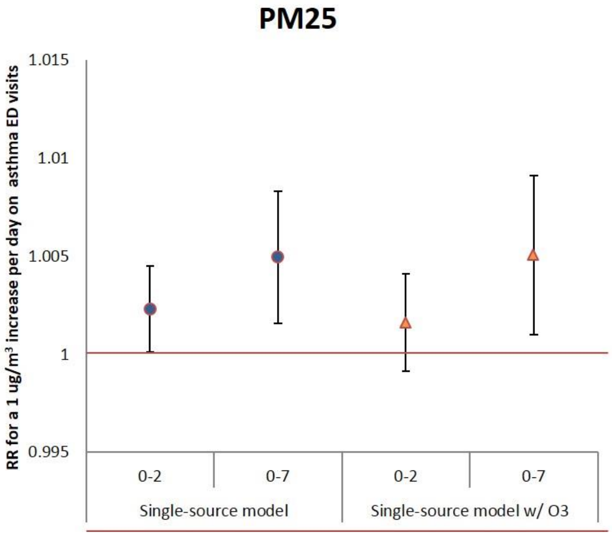


Figure 7.4 Rate ratios and 95% confidence intervals for the effect of a 1µg/m^3 increase in total PM$_{2.5}$ concentration on pediatric asthma ED visits, presented for lags 0-2 (controlling for lags 3-7) and lags 0-7. Blue circles represent results from the single-source model with an unconstrained distributed lag structure, while orange triangles represent the same model with O$_3$ control.

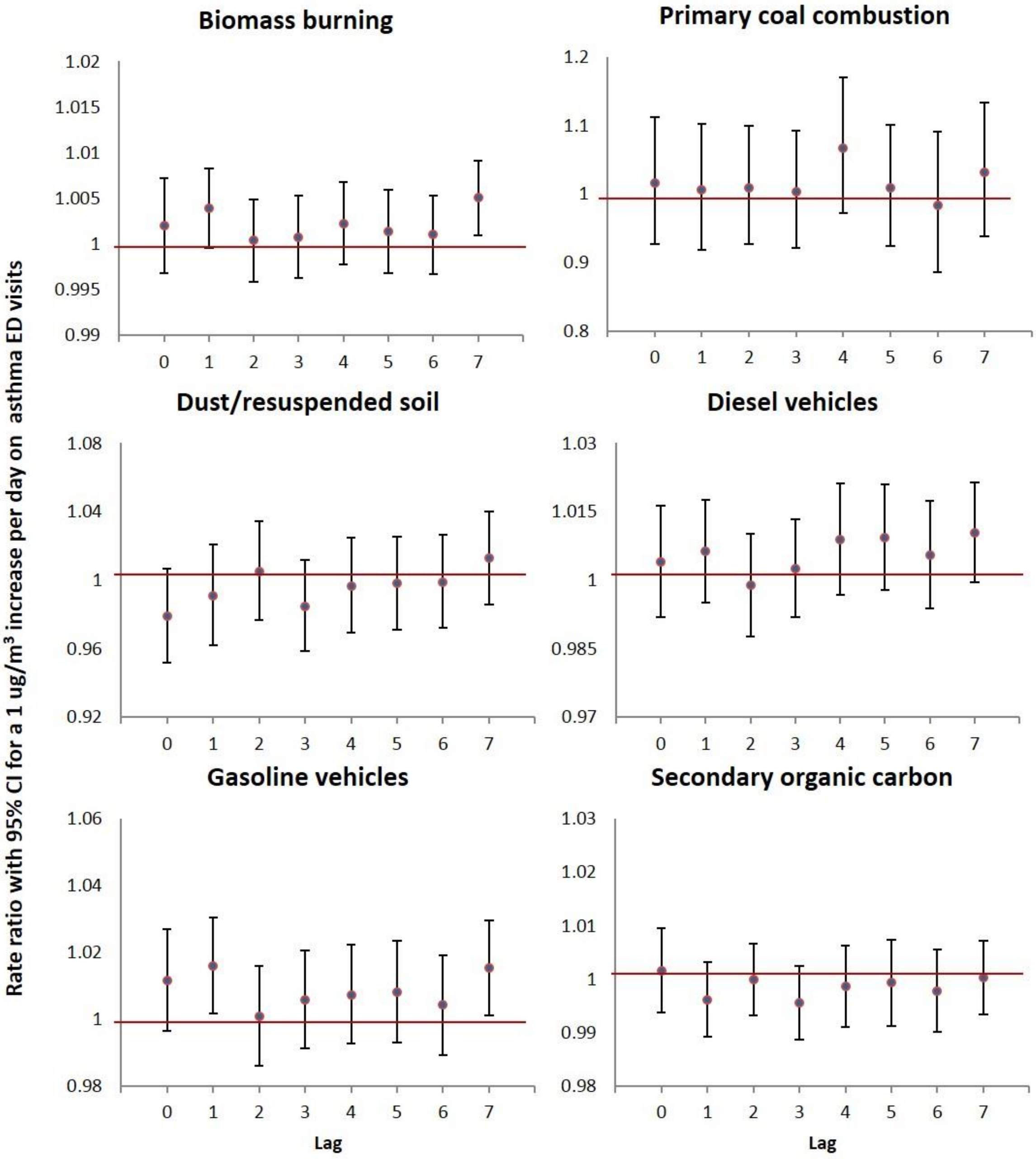


Figure 7.6. Rate ratios and 95% confidence intervals for the single-day effect of a 1µg/m^3 increase in source concentration on pediatric asthma ED visits presented for lags 0 through 7. Results are generated from the single-source model with an unconstrained distributed lag structure.

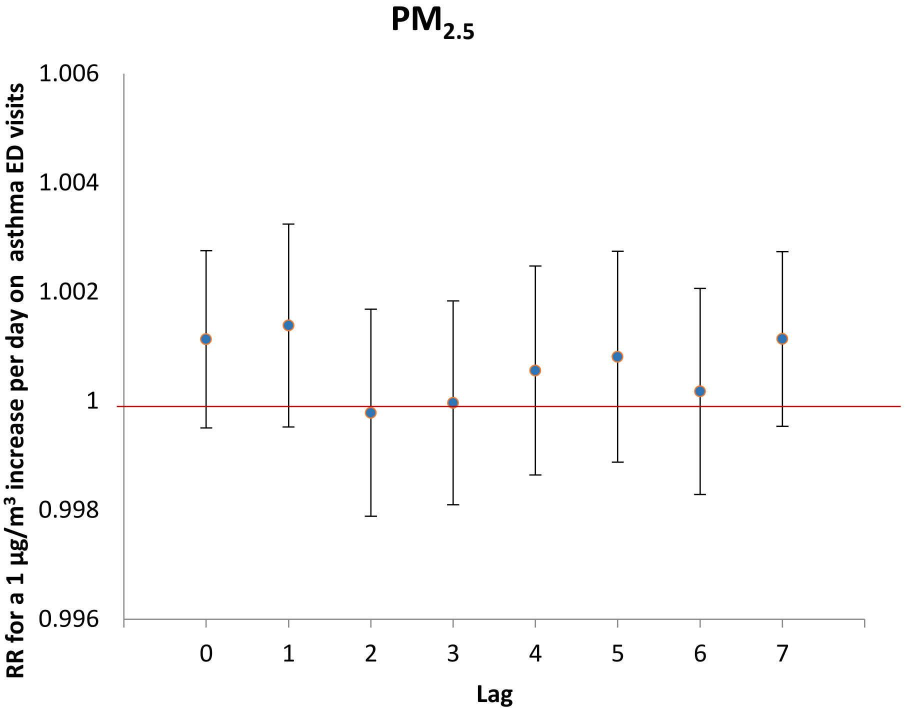


Figure 7.8. Rate ratios and 95% confidence intervals for the single-day effect of a 1µg/m^3 increase in total PM$_{2.5}$ concentration on pediatric asthma ED visits presented for lags 0 through 7. Results are generated from the single-source model with an unconstrained distributed lag structure.

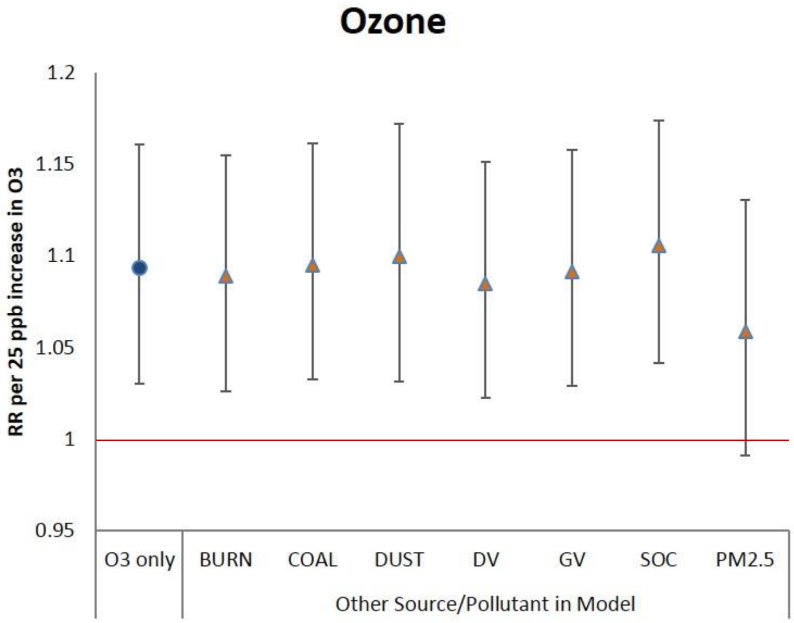


Figure 7.5. Rate ratios and 95% confidence intervals for the cumulative effect of a 25 ppb increase in O_3 for lags 0-7 generated from the single-source models with O_3 control. For comparison the results of an 'O_3 only' model (i.e. with no sources or $PM_{2.5}$) are shown.

Chapter 8: Conclusion

This dissertation makes both methodological as well as substantive contributions to the field of air pollution epidemiology. Existing tools for conducting C&RT analyses have all been lacking, to varying degrees, in their applicability to epidemiologic research. The modified C&RT algorithm presented in Study 1 offers an approach that addresses several of these limitations. In particular, the modified algorithm offers a way to control for confounding that is separate from tree construction; has more direct correspondence with statistical inference by choosing the best split based on statistical significance; and enables estimation of joint effects through the withholding of a common referent group prior to tree construction. In Study 2, we highlight how a modeling approach using C&RT can lead to conclusions that are different from those generated using conventional modeling approaches. In particular, the C&RT results suggest that the assumptions of monotonicity and a log-linear relationship inherent in many conventional modeling approaches may lead to an overestimation of risk on high pollution days. Though we cannot know which, if any, of these modeling approaches is correct, we would argue that the incorporation of alternative models with different sets of assumptions, such as C&RT, can be a useful way to generate new ideas and perhaps gain greater insight into air pollution mixtures.

Throughout this dissertation we find that the association between $PM_{2.5}$ and ED visits for pediatric asthma is both important and complex. For example, in Study 1 the first split in Figure 4.1 is for $PM_{2.5}$ between the 3rd and 4th quartiles, suggesting that when pollutants are categorized by quartiles, $PM_{2.5}$ (and not O_3) is most significantly associated with the outcome. In Study 2, the results from the C&RT Groups suggest that the days with the most harmful exposures for pediatric asthma are also the days with the highest $PM_{2.5}$ concentrations. This finding is a stark departure from the conventional models, which suggest that O_3 is the biggest driver of the multipollutant effect and that $PM_{2.5}$ has a null association (Figure 6.6). These results from Studies 1 and 2 together suggest that the $PM_{2.5}$ effect identified by C&RT may be due to the varying mixtures encompassed in total $PM_{2.5}$ mass.

We examine this possibility in Study 3 by looking at the association according to the different sources of $PM_{2.5}$ and find that $PM_{2.5}$ sources do vary in their toxicity, with traffic and biomass burning sources found to be the most harmful. Results from Study 3 also add to the growing body of literature that suggests the effect of $PM_{2.5}$ on pediatric asthma may be extended over several days. In our study, the effects of a single-day's exposure to diesel and gasoline vehicles, as well as biomass burning, were sustained for the following week, with some of the largest effects seen between lag days 4-7 (Figures 7.1 and 7.3).

Understanding the health associations related to air pollution mixtures poses many challenges, including what statistical approaches to use and how to characterize pollution mixtures. This dissertation tackles some of these challenges while examining the relationship between multipollutant exposures and emergency department (ED) visits for pediatric asthma in Atlanta.