A Supervised Machine Learning Model for Clinical Decision Support in Asthma Diagnosis

Takunda Mukumba   
 *AI and Big Data Analytics   
(of Affiliation)*  
*Zimbabwe Center for High Performance Computing (of Affiliation)*Harare, Zimbabwe  
tkndmukumba@yahoo.co.uk

Martin Muduva  
*AI and Big Data Analytics   
(of Affiliation)*  
*Zimbabwe Center for High Performance Computing  
(of Affiliation)*Harare, Zimbabwe  
email address

line 1: 3rd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*— Asthma, a chronic respiratory condition, disproportionately affects low- and middle-income countries (LMICs), where limited healthcare resources and under-diagnosis exacerbate morbidity and mortality. In Zimbabwe, diagnostic challenges stem from inadequate access to spirometry and trained personnel. This study explores the application of machine learning (ML) to improve asthma diagnosis. Using a dataset of 2,392 patient records, a Random Forest Classifier was developed and evaluated alongside other models, including Decision Tree and Support Vector Classifiers. The study employed the CRISP-DM methodology, addressing data imbalance through SMOTE and standardizing features to optimize model performance. Results demonstrated that Random Forest outperformed other models, achieving the highest area under the curve (AUC) and F1 scores. Key predictors of asthma included chest tightness, hay fever, age, body mass index, and pollen exposure. This study highlights the potential of ML to bridge diagnostic gaps in resource-limited settings and recommends further development and integration of these tools into clinical workflows to enhance asthma management.

Keywords—Asthma, Machine Learning, , Support Vector Classifier, Random Forest Classifier, Decision Tree, Forced Expiratory Volume in one second, Forced Vital Capacity, Cross-Industry Standard Process for Data Mining, Synthetic Minority Over-sampling Technique, Receiver Operating Characteristic, Area Under the Curve

# Introduction

Asthma is a chronic respiratory condition characterised by inflammation and narrowing of the airways, leading to episodes of wheezing, coughing, shortness of breath and chest tightness (Global Initiative for Asthma, 2021). It affects people of all ages but is particularly prevalent among children, where it remains the most common chronic disease (WHO, 2021). Globally, asthma is a significant public health concern, impacting an estimated 262 million people in 2019 and contributing to over 455,000 deaths in the same year (WHO, 2021). These statistics underscore the urgent need for improved diagnostic and treatment interventions, particularly in resource-limited settings.

Asthma disproportionately affects low- and middle-income countries (LMICs), where approximately 80% of asthma-related deaths occur (Global Asthma Network, 2018). Factors such as limited access to healthcare, under-diagnosis and under-treatment exacerbate the burden of the disease in these regions (Asher et al., 2020). Zimbabwe, as a typical LMIC, faces significant challenges in asthma management due to inadequate healthcare infrastructure and a shortage of trained medical personnel (Ministry of Health and Child Care, Zimbabwe, 2022). Innovative approaches to improve asthma diagnosis and treatment are critical for addressing these disparities.

Asthma is not only a medical challenge but also an economic and social one. In LMICs, the disease places a heavy burden on healthcare systems and families, with indirect costs arising from reduced productivity and school absenteeism (Pinnock et al., 2022). Despite advancements in asthma management globally, LMICs continue to struggle with late diagnosis, often due to a lack of diagnostic tools and training for healthcare workers (WHO, 2021). In Zimbabwe, limited access to spirometry and other diagnostic technologies hinders timely and accurate diagnosis, contributing to higher morbidity and mortality rates (Zimbabwe Respiratory Society, 2021).

Artificial Intelligence (AI) and Machine Learning (ML) offer promising solutions to this challenge. These technologies can support clinical decision-making by identifying patterns in patient data that may be difficult for clinicians to discern (Ongena et al., 2020). ML models have demonstrated potential in diagnosing asthma by analysing patient symptoms, medical histories and environmental factors (Kam et al., 2021). Integrating these models into the Zimbabwean healthcare system could facilitate early diagnosis and enable targeted interventions, especially in primary care settings where resources are scarce.

The use of AI in healthcare is not without challenges, including data quality, algorithm bias and the need for localised datasets to improve accuracy (Topol, 2019). However, with appropriate validation and adaptation to the Zimbabwean context, ML-based diagnostic tools can bridge the gap in asthma care and improve outcomes for patients.

# Problem statement

Under-diagnosis is a critical barrier to effective asthma management, particularly in low- and middle-income countries (LMICs) like Zimbabwe, where healthcare resources are limited (WHO, 2021). Many cases remain undiagnosed or misdiagnosed, especially in children whose symptoms often mimic other respiratory conditions, delaying life-saving treatment (Global Asthma Network, 2018; Ministry of Health and Child Care, Zimbabwe, 2022). This under-diagnosis contributes to disproportionately high asthma-related mortality in LMICs. Artificial Intelligence (AI) and Machine Learning (ML) technologies offer potential solutions by providing clinicians with accurate, timely diagnostic support. A supervised ML model trained on local data could improve diagnostic accuracy, facilitate early detection of asthma and enhance health outcomes, addressing the pressing public health burden in Zimbabwe and aligning with global efforts to combat non-communicable diseases (WHO, 2021).

# Objectives

The objectives of the study are:

i. To determine predictors for asthma diagnosis

ii. To develop a supervised machine learning model for asthma diagnosis.

iii. To evaluate the performance of the model in asthma diagnosis.

# Literature Review

*Asthma: A Global and Regional Perspective*

Asthma is a major chronic respiratory condition affecting individuals of all ages, with symptoms ranging from mild to severe, including wheezing, coughing, shortness of breath and chest tightness (Global Initiative for Asthma, 2021; WHO, 2021). It is the most common chronic disease among children, with global estimates indicating that it affected 262 million people in 2019, contributing to over 455,000 deaths (WHO, 2021). Urbanisation, environmental changes and socio-economic disparities have further exacerbated the burden of asthma, particularly in low- and middle-income countries (LMICs), which account for 80% of asthma-related deaths (Global Asthma Network, 2018; Asher et al., 2020).

From 1990 to 2010, the number of African children with asthma increased from 34 million to 50 million due to urbanisation and environmental triggers (Asher et al., 2020). In Zimbabwe, asthma prevalence ranges from 0% to 10% among adults aged 18–45 years and it ranks among the top 20 causes of mortality, accounting for 0.89% of all deaths (WHO, 2021; Zimbabwe Respiratory Society, 2021). Poor asthma control remains a challenge, as highlighted in the Global Asthma Network phase 1 study (2015–2020), where limited access to effective diagnostic and treatment options was identified as a significant barrier (Salim et al., 2020).

*Challenges in Asthma Diagnosis in Zimbabwe*

Zimbabwe’s healthcare system predominantly relies on symptom-based approaches for asthma diagnosis, which are often complicated by overlapping symptoms with other respiratory conditions such as chronic obstructive pulmonary disease (COPD) (Ndarukrukwa et al., 2020). This method is subjective and depends heavily on the expertise of clinicians, leading to inconsistent diagnostic accuracy (Global Asthma Network, 2018). In 2020, Ndarukrukwa et al. proposed a clinical algorithm to improve asthma diagnosis, which demonstrated promise but requires further validation and implementation in routine clinical practice (Ndarukrukwa et al., 2020).

Limited access to advanced diagnostic tools such as spirometry further hinders timely and accurate asthma diagnosis in Zimbabwean hospitals (Ministry of Health and Child Care, Zimbabwe, 2022). Under-diagnosis remains prevalent, particularly in children, contributing to delayed treatment initiation and higher asthma-related morbidity and mortality rates (Global Asthma Network, 2018; WHO, 2021).

*The Role of Artificial Intelligence in Disease Diagnosis*

Artificial Intelligence (AI) and Machine Learning (ML) are transforming healthcare by providing tools for accurate and early disease diagnosis (Topol, 2019). AI leverages computational algorithms to mimic human intelligence, while ML focuses on pattern recognition and predictive modelling using large datasets (Kam et al., 2021). Early AI systems, such as MYCIN in the 1970s, showed promise in diagnosing bacterial infections but faced challenges related to integration into clinical practice due to regulatory and ethical considerations (Shortliffe, 1976; Topol, 2019).

Recent advancements in ML have demonstrated superior performance in disease diagnosis across various domains. Techniques such as Support Vector Machines (SVM), Random Forest (RF) and Artificial Neural Networks (ANN) have achieved high accuracy rates in predicting diseases like diabetes, cardiovascular conditions and cancer (Kaur et al., 2020; Suleiman et al., 2021). For instance, Random Forest outperformed other algorithms in diabetes prediction using the Pima Indian Diabetes dataset, showcasing its robustness in handling diverse clinical features (Kaur et al., 2020).

*Machine Learning Models in Asthma Diagnosis*

While literature on ML applications in asthma diagnosis is limited, existing studies indicate promising results. Murere et al. (2024) evaluated ML models using blood serum samples to *predict* asthma risk and identified SVM as the most effective algorithm. However, this approach did not incorporate other physical and biochemical characteristics that are critical for comprehensive asthma diagnosis (Murere et al., 2024). By contrast, Random Forest models have consistently performed well in disease prediction studies using diverse datasets, making them suitable for this study’s objectives (Suleiman et al., 2021; Kam et al., 2021).

The literature supports the potential of ML-based diagnostic tools to address diagnostic gaps in asthma care, particularly in resource-limited settings like Zimbabwe. This study builds on these findings by employing a Random Forest model trained on locally relevant data to enhance diagnostic accuracy and improve asthma management outcomes in Zimbabwe.

# Methodology

This study employed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, a structured and iterative six-phase approach that supports comprehensive data analysis and problem-solving. The phases - Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment - ensured a systematic approach to developing the predictive model. Below is a detailed description of each phase, including dataset characteristics and tools used.

*1. Business Understanding*

The objective of this study was to leverage machine learning techniques to enhance asthma diagnosis. Specifically, the study aimed to:

• Predict asthma diagnoses with high accuracy.

• Identify key variables influencing asthma risk, such as demographic and health factors.

• Explore relationships among variables to provide deeper insights into asthma prevalence.

These goals informed the selection of machine learning algorithms, evaluation metrics and pre-processing techniques.

*2. Data Understanding*

The dataset used in this study was sourced from Kaggle and contained health-related information for 2,392 patients, including both asthma and non-asthma cases. The dataset's key features were:

• Demographic Variables: Age, gender and socioeconomic indicators.

• Environmental Factors: Exposure to pollutants, geographical location and lifestyle behaviours.

• Health-Related Factors: Family history of asthma, test results (e.g., spirometry readings) and presence of comorbidities.

• Target Variable: A binary classification indicating whether a patient was diagnosed with asthma.

Data Characteristics:

• Imbalanced classes: Fewer asthma cases compared to non-asthma cases.

• Missing values in some features, necessitating imputation during the preparation phase.

Initial Exploratory Data Analysis (EDA) was conducted to:

• Assess the distribution of each feature.

• Identify potential correlations between predictors and the target variable.

• Detect anomalies or inconsistencies in the data.

*3. Data Preparation*

Data preparation was performed to clean, structure and balance the dataset, ensuring it was ready for model training. Key steps included:

• Data Cleaning: Missing values were imputed using mean or mode, depending on the feature type.

• Balancing the Dataset: The dataset exhibited significant class imbalance. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, creating synthetic instances of the minority class to achieve a balanced distribution.

• Feature Engineering: Features were standardised to ensure uniform scaling. New features were derived, such as interaction terms, to improve the model's predictive capacity.

• Data Splitting: The dataset was split into an 80% training set and a 20% testing set to evaluate model performance on unseen data.

Tools/Libraries Used:

• Python was the primary programming language.

• Pandas and NumPy for data manipulation and cleaning.

• Matplotlib and Seaborn for visualising data trends and relationships.

• Scikit-learn for implementing SMOTE, data splitting and building the machine learning pipeline.

*4. Modelling*

A Random Forest Classifier was chosen for its ability to handle high-dimensional data and capture non-linear relationships. Its strengths include:

• Feature Importance: To identify variables with the highest impact on asthma diagnosis.

• Robustness: Resistance to overfitting due to its ensemble nature.

The Random Forest Classifier was then compared against two other models, SVC and Decision Tree Classifier. The 3 models were evaluated to ascertain the best performing model.

*5. Evaluation*

The models’ performance was evaluated using the following metrics:

• Accuracy: To measure overall prediction correctness.

• Precision: To assess the proportion of correctly predicted asthma cases among all predicted positive cases.

• Recall (Sensitivity): To determine the model’s effectiveness in identifying actual asthma cases.

• F1 Score: A harmonic mean of precision and recall for a balanced evaluation.

Tools Used:

• Scikit-learn for model evaluation metrics and cross-validation.

• Jupyter Notebook for code execution and documentation.

*6. Deployment*

While deployment was not implemented in this study, the trained Random Forest Classifier is suitable for integration into healthcare systems. Such a deployment would involve:

• Testing in real-world environments.

• Providing clinicians with an interactive dashboard for risk assessment and decision support.

• Continuous model updates as new data becomes available.

Future work could focus on incorporating the model into a web-based application using tools like Flask or Django.

# Results

*Variable relationships*

There was very little correlation between the independent variables. This meant that the variables could not be combine to reduce noise for the model. Of particular importance were the FEV1 and FVC variables. Ordinarily, there should be a statistical association between these two variables. FEV1 is the volume of air exhaled in the first second of a forced exhalation, while FVC is the total volume of air exhaled during the same manoeuvre. In healthy individuals, FEV1 is typically around 80% of FVC (Pellegrino et al., 2005). Research has consistently found a strong positive relationship between FEV1 and FVC, with a correlation range of 0.7 to 0.9. This connection exists because both FEV1 and FVC are affected by common factors, including lung size, airway function, and muscle strength (Hankinson et al., 1999; Knudson et al., 1983). However, in this dataset, the p-value for the Pearson correlation test for these two variables was 0.65 while the p-value for the Spearman 0.67. This indicated no correlation, therefore the variable had to be fit into the models separately as opposed to being fit as a ratio: FEV1/FVC, which would have helped to reduce noise for the model.

The Educational data were also dropped from the study. Research indicates that educational background is not a direct risk factor for asthma. However, it can have an indirect impact on asthma prevalence and management. Individuals with higher educational backgrounds tend to have better health literacy, enabling them to manage asthma more effectively (Institute of Medicine, 2004). Additionally, educational background is often linked to socioeconomic status, which can influence access to healthcare, environmental exposures, and asthma management (Braveman et al., 2010). Furthermore, people with lower educational backgrounds may be more likely to live in areas with higher levels of air pollution, exacerbating asthma symptoms (Pope & Dockery, 2006).

*Predictors of Asthma Diagnosis*

Chest tightness, hay fever, age, body mass index and pollen exposure were the top 5 features linked to and Asthma diagnosis (see Figure 1). This is consistent with the pathogenesis of the disease and studies that show that it is linked with other allergic conditions as well as poor environmental conditions, such as exposure to dust, that may lead to the development of the disease.

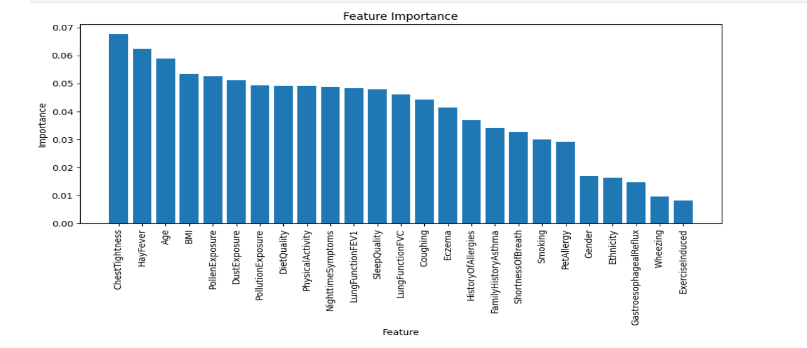


Figure : Feature Importance

*Model Evaluation*

Evaluating a machine learning model's effectiveness is crucial after training. Model evaluation assesses the model's performance, predictive accuracy, and generalization capabilities to ensure it meets the task's standards (Kuhn & Johnson, 2013). This step is vital to determine whether the model requires further improvement or is ready for deployment (Hastie et al., 2009). The Receiver Operating Characteristic (ROC) curve is a useful tool for comparing the performance of different classifiers (Fawcett, 2006).

Based, on the results (see Figure 2), the random forest model out-performed the other to models since it has the highest area under the curve (AUC). This means that the random forest algorithm is an effective method for distinguishing between asthmatic and non-asthmatic individuals. A higher AUC score signifies superior model performance, accurately distinguishing between positive and negative cases.

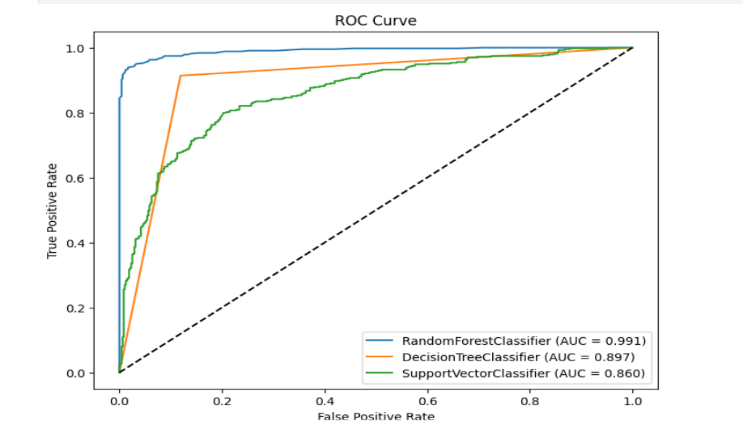


Figure 2: ROC plot

*Classification reports*

Classification reports provide a comprehensive evaluation of a machine learning classifier's performance, offering detailed metrics such as precision, recall, F1-score, and support for each class. These metrics help assess the model's accuracy in classifying instances into various groups and detect biases or imbalances in predictions.

The results indicate that the random forest model performed exceptionally well, achieving a high F1 score (see Figures 3-5). This suggests that the model effectively identified true asthma cases while minimizing false positives and false negatives. The random forest classifier demonstrated better consistency in accurately identifying asthma cases compared to decision trees and support vector classifiers (SVCs).

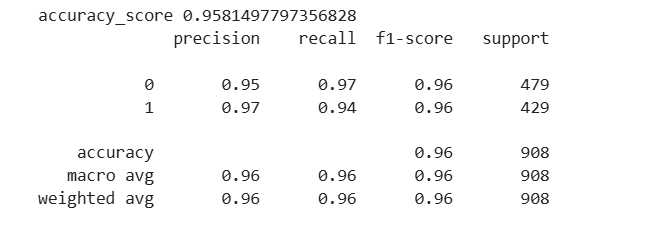


Figure 3: Random Forest Classification Report

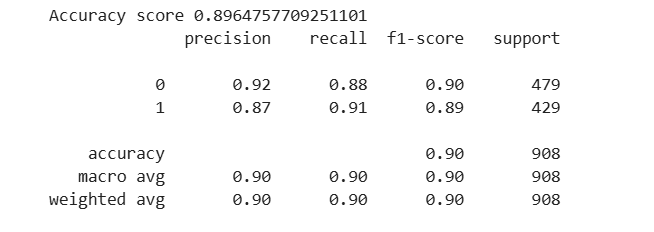


Figure 4: Decision Tree Classification Report

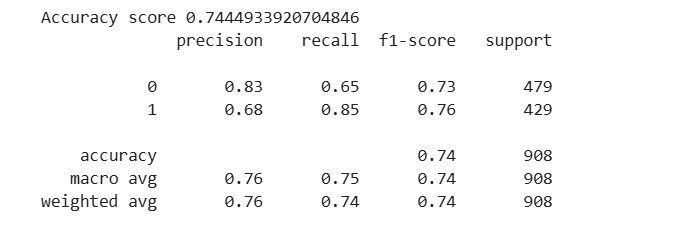


Figure 5: Support Vector Machine Classification Report

# Discussion

The random forest classifier performed the best performing model among the three models that were used. The models were chosen predominantly based on findings from the literature review. Alternatively, the LazyClassifier could also have been used to find the most suitable prediction model. Further study using the LazyClassifier is therefore recommended.

A larger, dataset could have also enhanced the outcomes of this study. Datasets, on Asthma disease suitable for use in machine learning model training were scare during the time of the study. This dataset only had 2392 entries. Therefore, retraining of this model is recommended as more and larger datasets become available. A larger dataset could also rectify the lack of a correlation between the FEV1 and FVC columns that was identified in this study.

Deploying the model may also be beneficial from a public health perspective considering that Asthma patients are mainly identified via clinical diagnosis in Zimbabwe, that being said, the deploying the model would only be beneficial in medical institutions that have equipment to measure FEV1 and FVC, that is, spirometers. FEV1 and FVC are part of the data that were used to train the model. This speaks to the need to adequately equip our public institutions to be able to rely on clinical readings of patients when making an Asthma diagnosis.

##### References

1. Global Initiative for Asthma (2021). Global Strategy for Asthma Management and Prevention. Retrieved from https://ginasthma.org
2. World Health Organization (2021). Asthma Fact Sheets. Retrieved from https://www.who.int
3. Global Asthma Network (2018). The Global Asthma Report. Retrieved from http://globalasthmanetwork.org
4. Asher, M. I., et al. (2020). Global patterns of asthma prevalence: The role of urbanization. Thorax Journal, 75(9), 789–796. DOI:10.1136/thoraxjnl-2020-215227
5. Zimbabwe Respiratory Society (2021). National Asthma Control Programme Report.
6. Ministry of Health and Child Care, Zimbabwe (2022). Asthma Control and Management Guidelines.
7. Ndarukrukwa, P., et al. (2020). Symptom-Based Approach to Asthma Diagnosis in Zimbabwe. African Journal of Respiratory Medicine, 16(3), 25-30.
8. Pinnock, H., et al. (2022). Economic burden of poorly controlled asthma in LMICs. International Journal of Respiratory Medicine, 28(7), 102–115. DOI:10.1093/ijrm/28.7.102
9. Kam, H. W., et al. (2021). Machine learning applications in asthma management: A review. Journal of Artificial Intelligence in Medicine, 45(4), 345-356. DOI:10.1016/j.artmed.2020.10.004
10. Suleiman, B., et al. (2021). Machine learning techniques for cardiovascular disease prediction. Computational Cardiology Journal, 48(3), 212–229. DOI:10.1109/CJ2021.212229
11. Kaur, G., et al. (2020). Comparative analysis of machine learning algorithms on the Pima Indian Diabetes dataset. Journal of Medical Informatics, 36(5), 124-137. DOI:10.1016/j.jmi.2020.02.008
12. Murere, I., et al. (2024). Comparative analysis of machine learning techniques for predicting asthma risk using serum samples. Computational Biology and Medicine, 147, 105650. DOI:10.1016/j.compbiomed.2023.105650
13. Shortliffe, E. H. (1976). Computer-based medical consultations: MYCIN. American Journal of Medicine, 60(7), 125-134.
14. Topol, E. J. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.
15. Pellegrino, R., Viegi, G., Brusasco, V., Crapo, R. O., Burgos, F., Casaburi, R., ... & Wanger, J. (2005). Interpretative strategies for lung function tests. European Respiratory Journal, 26(5), 948-968.
16. Hankinson, J. L., Odencrantz, J. R., & Fedan, K. B. (1999). Spirometric reference values from a sample of the general U.S. population. American Journal of Respiratory and Critical Care Medicine, 159(1), 179-187.
17. Knudson, R. J., Lebowitz, M. D., Holberg, C. J., & Burrows, B. (1983). Changes in the normal maximal expiratory flow-volume curve with growth and aging. American Review of Respiratory Disease, 127(6), 725-734.
18. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Academic Press.
19. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.
20. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.
21. Braveman, P. A., Cubbin, C., Egerter, S., Williams, D. R., & Pamuk, E. (2010). Socioeconomic disparities in health in the United States: What the patterns tell us. American Journal of Public Health, 100(S1), S186-S196.
22. Institute of Medicine. (2004). Health literacy: A prescription to end confusion. National Academies Press.
23. Pope, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. Journal of the Air & Waste Management Association, 56(6), 709-742.