Fall 12-2022

Novel Techniques for Quantifying Secondhand Smoke Diffusion into Children's Bedroom

Sunil Ramchandani
Chapman University, sramchandani@chapman.edu

Recommended Citation

S. Ramchandani, "Novel techniques for quantifying secondhand smoke diffusion into children's bedroom," Ph.D. dissertation, Chapman University, Orange, CA, 2022. https://doi.org/10.36837/chapman.000411

Novel techniques for quantifying secondhand smoke diffusion into children's bedroom

A Dissertation by

Sunil Ramchandani

Chapman University

Orange, CA

Schmid College of Science and Technology

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computational and Data Sciences

December 2022

Committee in charge:

Dr. Vincent Berardi Ph.D., Committee Chair

Dr. Melbourne Hovell Ph.D.

Dr. Cyril Rakowski Ph.D.

CHAPMAN UNIVERSITY
SCHMID COLLEGE OF SCIENCE AND TECHNOLOGY
Computational and Data Sciences

The dissertation of Sunil Ramchandani is approved.

Vincent Berardi, Ph.D., Chair

Melbourne Hovell, Ph.D.

Cyril Rakovski

Cyril Rakowski, Ph.D.

August 2022

Novel techniques for quantifying secondhand smoke diffusion into children's bedroom

by Sunil Ramchandani

ACKNOWLEDGEMENTS

I would like to thank Dr. Vincent Berardi, Dr. Hesham El-Askary, Dr. Melbourne Hovell, Dr. Cyril Rakovski, everyone who helped me along the way. I would not have succeeded without you. I also wanted to express additional appreciation for everyone in the Chapman University Computational and Data Science Program for providing the support necessary for me to accomplish everything included in this dissertation.

ABSTRACT

Novel techniques for quantifying secondhand smoke diffusion into children's bedroom

by Sunil Ramchandani

The impact of secondhand smoking to health of general population, and specifically children, is a well-known phenomenon that researchers have studied for years. In this dissertation, I extend work done within a secondhand smoke intervention to understand and quantity the impact of intervention on flow of smoke air particle concentration from the main room where smoking generates air particle contamination to a room where a child living in the home sleeps. The paper also explores potential modelling techniques to proactively identify and the impact of the smoke air particles with the intent to discourage adults from smoking in the home and thus potentially minimizing the impact to children's health. The data was analyzed using hierarchical linear models to quantify the impact of intervention. The analysis finds that smoke air particles attenuated, on average, by 31.6% from the main room to the child's room. Using hierarchical linear models, I also quantified the effect of intervention where the relationship between the main room and child's room concentrations decreased once the intervention became active (-0.146 to -0.034 based on random slope versus random intercept). I also developed an LSTM model that can proactively identify whether a smoking event would be an impact children's health. The results of the model are very encouraging, with an accuracy of approximately 80% when using less than 4 minutes of main room data. The two key outcomes from this study are 1) I can quantify the impact of intervention on the flow of air particle concentration between the

main room and child’s room and 2) I am able to develop a modelling approach that can proactively identify the potential impact of SHS to health of the child. The study open doors for several possibilities including use of the findings by practitioners in counselling sessions to provide metrics to smoking adults and advice on the potential impact of smoking to the health of the child. The modelling approach also lays a foundation for future research to implement proactive, real time monitoring and notification in smart homes.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

1 Background

1.1 Introduction

Smoking combustible tobacco products causes cardiovascular disease (heart disease and stroke), more frequent and severe asthma attacks, fatal diseases such as cancer and chronic obstructive pulmonary disease (COPD)[3]. Besides having negative outcomes for smokers, smoking also produces second-hand smoke (SHS), a combination of mainstream smoke exhaled from smokers and smoke emitted from smoldering tobacco.[9–13] There is substantial evidence that demonstrates the harm that short and long-term exposure to SHS represents to the respiratory and cardiovascular health of adults and children[4]. Studies show that there is no such thing as a "risk-free" level of SHS exposure[5.] The smoke may expose bystanders to harmful constituents such as nicotine, heavy metals, ultrafine particulates, volatile organic compounds, and other toxicants.[7,8]

Children have some of the highest SHS exposure rates in society, with 40% of children aged between 3-11 being exposed to tobacco SHS.[5] SHS is particularly unsafe for children due to biological characteristics (higher breathing rates, immature lungs, and underdeveloped immune systems), making it challenging to filter toxins.[4,5] Furthermore, children inhale a larger volume of air per body mass than adults, which results in higher relative doses of inhalation-related exposure. Studies suggest the health outcomes associated with children's exposure to SHS include sudden infant death syndrome, acute respiratory infections, and increased asthma severity.[1,38]

The home is the primary location where children are exposed to tobacco SHS, especially when their parents smoke.[38,47–51] People are indoors for an average of 80% to 90% of their day,[46] with younger children spending most of that time in their home. Children are at risk of SHS exposure even if they are not in the same room as the smoking parent. Bedtimes for children are typically in the early evening, with one study indicating that grade K-4 children, on average, go to bed at 8:27 pm.[52] In contrast, American adult's average in-bed time is 11:12 pm,[5] which leaves nearly three hours for smoking, on average, while children sleep. Previous studies have reported the presence of a "daily dip-evening incline" class of smokers, with an elevated frequency of smoking later in the evening, possibly due to increased nicotine dependence.[53] If children are sleeping during a late evening smoking event, there is the potential for SHS to infiltrate into their bedrooms, potentially without parents' knowledge. In addition to late evening cigarettes likely impacting children's sleeping environments, smokers often engage in night smoking after they have gone to bed, but before waking up to begin their day. For example, over four weeks, 41% of smokers attempting to quit reported a night smoking event, with these individuals' night smoking on 26% of nights.[54] Beyond sleeping times, children likely spend a significant proportion of remaining in-home time in their bedroom while playing, studying, or engaged in other activities. During these periods, SHS infiltrating into the bedroom from in-home smoking locations could pose a health risk.

Research indicates that parents reduce SHS exposure when they know that it is impacting their children's health.[55] For example, an increase in pulmonary functionality was reported after surgery for children with a history of SHS exposure, but not for children who were not exposed to SHS.[56-58] Typically, in everyday scenarios without severe indicators of vulnerability such as asthma or surgery, parents are less able to discriminate accurately and subsequently mitigate the risks of SHS.[59–61] There have been several studies over time that have called for the transmission of detailed microenvironment information concerning the impact associated with SHS exposure to help parents understand the true scope of risks and develop appropriate strategies to protect their children.[11,63] However, this intervention approach is hampered by an insufficient understanding of SHS diffusion dynamics, with much of our knowledge gained via experiments in controlled environment[64] and computational models.[65,70]

This dissertation aims to provide a more comprehensive assessment of children's potential for in-home SHS exposure, more specifically assessing how SHS diffuses from smoking areas into sensitive, latent environments. SHS has typically been assessed via retrospective questionnaires or objective measures, such as air particle data, collected over short periods in a single location for a limited number of participants. Such designs do not allow SHS diffusion from smoking areas into other locales to be assessed. In contrast, this work will be performed in the context of a secondary analysis (i.e., collecting the data that has been part of an earlier study) of air particle data from a previously conducted SHS reduction trial. It focuses on

a.) Quantifying the relationship between smoking occurring within a home and subsequent impact on the children's bedrooms.

b.) Assessing changes in this relationship as a function of initiating the feedback delivered in the intervention.

c.) Identifying household characteristics (e.g., home size or presence of cannabis smokers) that affect if/how tobacco smoke infiltrates children’s environments.

d) A mechanism to proactively identify the potential impact of SHS on health of child based on presence of smoke air particles within a home.

There are several factors that can impact the health of child. The broader impact of this dissertation will be to help reduce SHS exposure to improve children's health. The aim is for the knowledge gained from this study to be used to design early notification systems will alter the parents' smoking behavior and potential reduce the impact the protect children’s health.

2 Data Description & Pre-Processing

2.1 Introduction to data

The data used for the dissertation was generated by Project Fresh Air (PFA),[27,28] a multiple baseline/randomized control trial aimed at reducing SHS in the households of smokers from a low-socioeconomic status (SES) population. The study enrolled 298 homes with at least one adult who generated indoor SHS (typically via cigarette smoking) and at least one SHS-exposed child living in the home (See Table 1 for sample demographics). Two Dylos DC1700 air particle monitors were installed inside each home to monitor air quality and calibrated to detect particles with sizes consistent with SHS. One monitor was installed in the main smoking room (MNR), and another was placed in the child's bedroom (CHD), with both locations self-reported. To construct a site plan (Figure 1), project personnel used laser distance measurements to record each room's dimensions and the home's physical characteristics, including the distance between the monitors. The concentration of fine air particulate matter (PM 2.5)[110] was measured every ten seconds by both monitors. In half of the homes, designated as the experimental condition, the monitors were fitted with programmed devices to deliver aversive visual and auditory feedback (yellow/red lights and tones) when air particle concentrations exceeded a threshold. For these homes, the trial was stratified into two phases: 1) baseline, a period during which monitor feedback was not active, and 2) treatment, the period during which the feedback was activated. The other half of the homes were enrolled in a control condition, where the monitors passively measured indoor air throughout the study.

The trial lasted approximately three months for each household, with the baseline phase representing, on average, the first two weeks of enrollment. Additionally, after both the first and last week in the study, trained staff administered a comprehensive computer-assisted face-to-face interview to gather data about smoking-related behaviors. It included questions asked about each household member's smoking habits, other potential particle-generating behaviors (e.g., burning candles), and mitigation activities (e.g., opening windows) during the prior seven days. Parents also kept air diaries during this period.

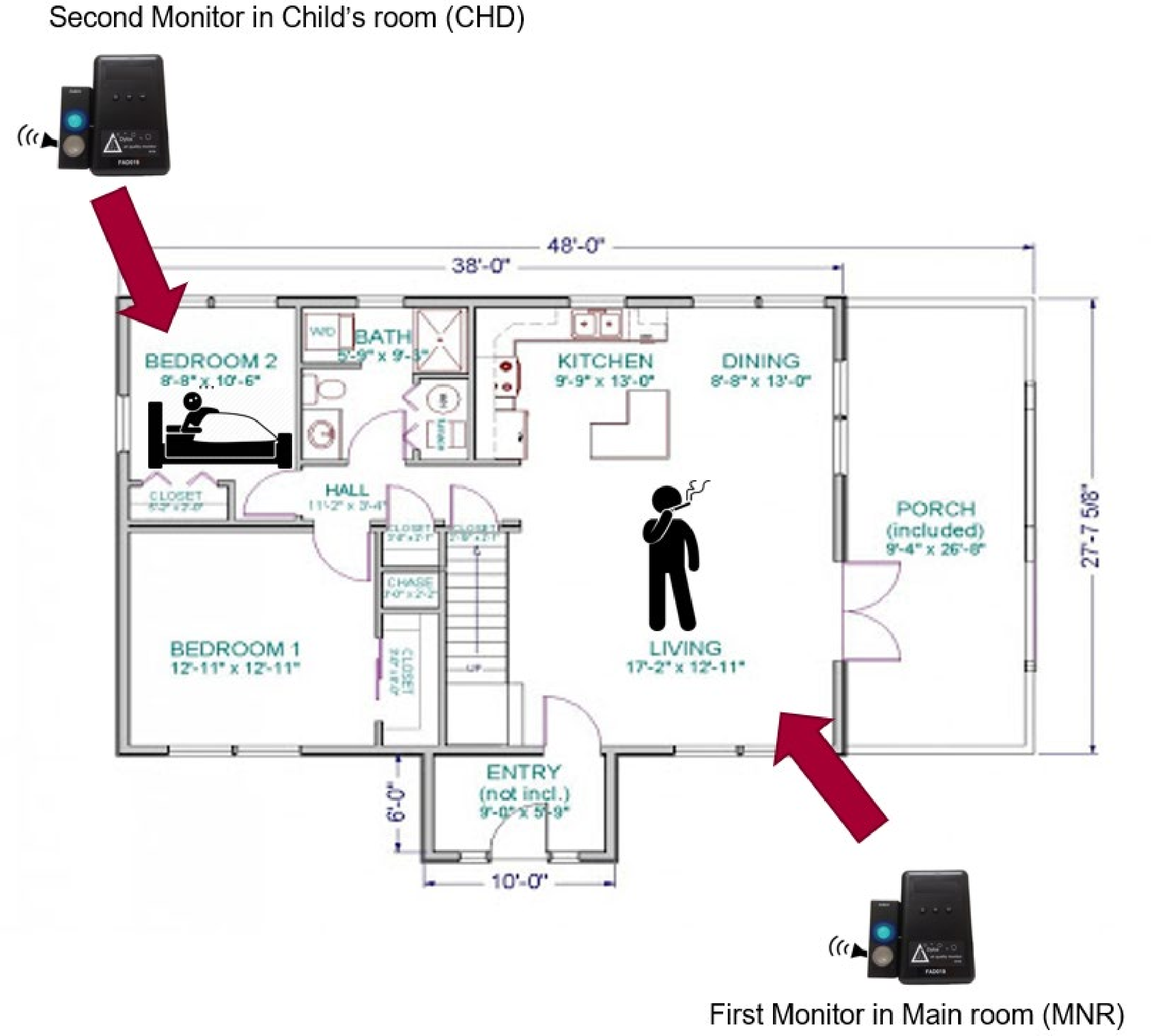


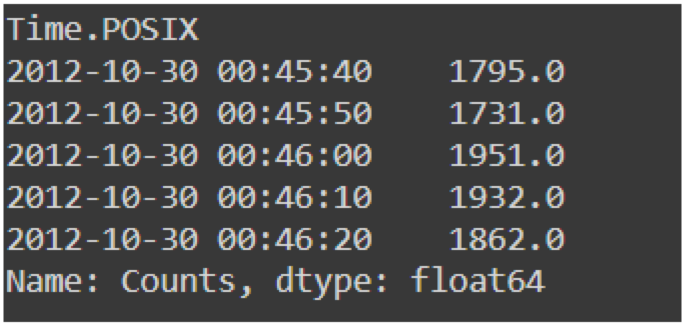
Figure 1 : Generic layout of a sample home with monitors in main smoking room (MNR) and child's room (CHD).

Race/Ethnicity	**% or Mean (Standard Deviation)**
Hispanic	34.20%
Other	23.70%
White	22.40%
Black	19.70%
Marital Status	
Married	36.00%
Single/Never Married	28.40%
Divorced/Separated	17.20%
Not married but living with a partner	17.20%
Widowed	1.20%
Single Parent	
Yes	46.84%
No	53.15%
Marijuana Consumption	
Yes	15.30%
No	84.70%
Home Type	
Apartment/Condo	52.19%
Detached House	33.56%
Townhouse	8.89%
Duplex	4.06%
Trailer/Mobile Home	1.26%
# Adults	2.56 (1.09)
# Children	2.16 (1.23)
Average distance between rooms in a home (feet)	12.02 (10.68)
# Rooms	5.90 (2.33)

Table 1 : Demographic Information

2.2 Preliminary data cleaning

The air particle data for the 298 homes are stored in a separate file for each home. The particle count, along with the capture time (time stamp) is reported. (Figure 2) illustrates a small sample of the data for a single home.



```
Time.POSIX
2012-10-30 00:45:40    1795.0
2012-10-30 00:45:50    1731.0
2012-10-30 00:46:00    1951.0
2012-10-30 00:46:10    1932.0
2012-10-30 00:46:20    1862.0
Name: Counts, dtype: float64
```

Figure 2: Particle count format for each home

Before proceeding with analysis, two data issues were required to be addressed. The first is the presence of outliers in air particle monitor measures. There are 45 homes where select air particle measurements are greater than 3 standard deviations away from the home-specific mean particle count. Previous experience with the devices indicates that these measures are likely due to errors with the air particle devices, so I eliminated them from consideration in subsequent analyses. The second issue was incorrect timestamps. Seven homes had the CHD measurements with a timestamp in 2013, but the MNR measures with a timestamp in 2000. Each file was inspected, and the timestamp issue fixed manually.

2.3 Main room (MNR) peak extraction

A large proportion of the air particle data collected by the monitors consists of low-level, ambient concentrations associated with air particles' background levels. I am interested in the diffusion of particle concentration of diffusion of SHS primarily caused by cigarette smoking between MNR and CHD locations and therefore are most interested in time periods with elevated measures (i.e., peaks) that are reflective of poor air quality and are most likely representative of smoking events[28]. Therefore, the first task in quantifying the relationship between the air particle quality in the two rooms was to identify the time intervals associated with peaks within each home's time series to extract the relevant data. Because most smoking is expected to be initiated in the MNR location, I chose to extract peaks from the data in this location and then to extract data from corresponding times in the CHD location.

Peak extraction methodologies can be based on various criteria, such as level thresholds[106], variance[105], and Fourier coefficients[104]. The level threshold is an adaptive technique where I apply a certain threshold to peaks selected by the algorithm and ignore spurious peaks. The variance technique is based on the principle of statistical dispersion, where a peak is a data point that is 'x' standard deviations away from the moving mean. The Fourier coefficient technique removes noise from the signal using an adaptive short-time discrete Fourier transform. The peaks with the highest intensity among the peak clusters are then recorded based on the signal-to-noise ratio. Since I am specifically interested in peaks that are generally consistent with smoking events, I decided to use a threshold technique as it was the simplest, and the results aligned with the suspected smoking events when I visually inspected the results.

To implement the threshold algorithm, I began by smoothing the air particle data to reduce noise in the signal. The exponential weighted moving average was applied to smooth the signal with a span of 5 minutes. Figure 3 shows an example for a sample signal. A simple mean approach would apply uniform weights to the peaks, but I was more interested in the value of the signal values near the peak maximum. So, I selected the exponential weighted moving average to smooth the signal.

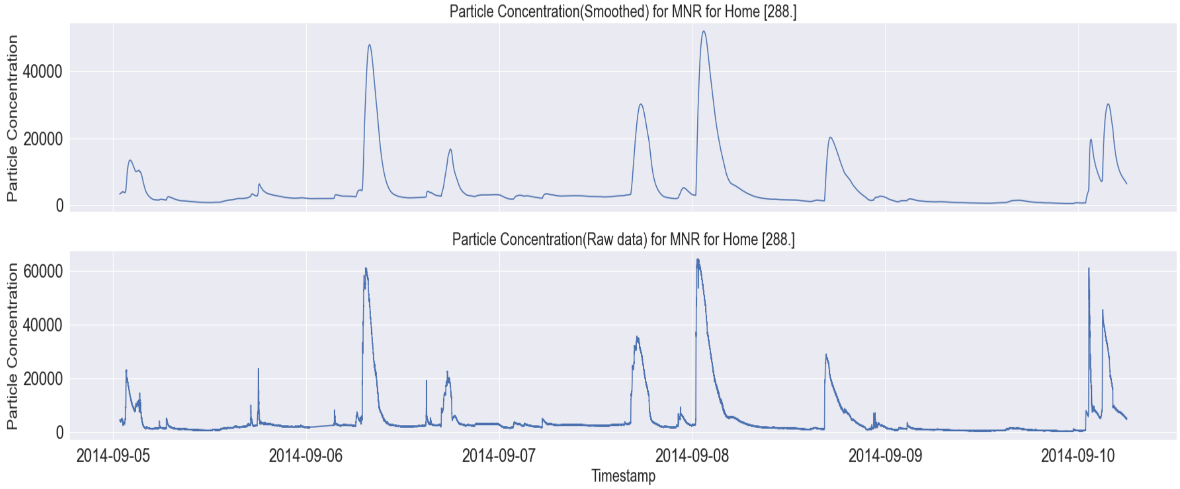


Figure 3 : The effect of exponential weighted moving average (top half) to a raw signal (bottom half)

The *find_peaks*[101] method from the *Scipy* package in Python was then used to identify the smoking events. This approach finds all local maxima by a simple comparison of neighboring values. A peak or local maximum is defined as any sample whose two direct neighbors on either side have a smaller amplitude. For flat peaks (more than one sample of equal amplitude wide), the middle sample index is returned (rounded down if the number of samples is even). Peak identification using this algorithm is an iterative process with three fundamental parameter values that need to be adjusted to identify the peaks.

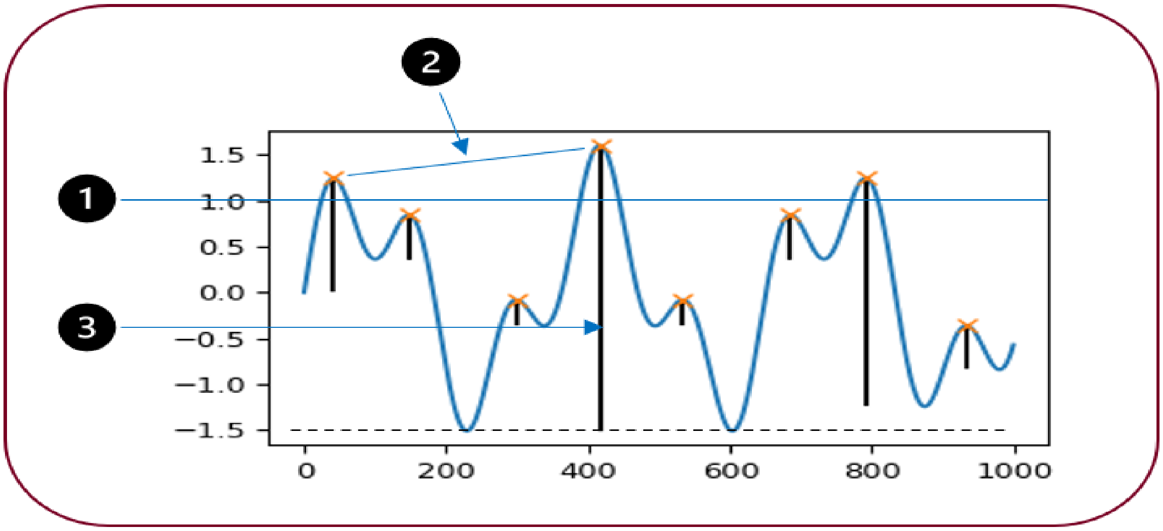


Figure 4 : Key parameters to identify peaks. (1. Threshold Value, 2. The horizontal distance, 3. Prominence)

As shown in Figure 4, these parameters are:

- Threshold value: minimum required height of the peak
- Minimum horizontal distance: measured in samples between neighboring peaks.
- Prominence: how well a peak stands out from the signal's surrounding baseline. It is the vertical distance between the peak and its lowest contour line. The contour line for a peak is identified by finding the lowest point of the adjacent peaks.

I visually compared the peaks identified with various sets of parameters for randomly selected peaks from homes with both few and many peaks. This analysis indicated that the following values: a threshold value of 10,000 counts; minimal distance of 5,000; and prominence of 15 optimally identified peaks with a shape and duration roughly consistent with tobacco smoking. This was determined by the domain expertise of PFA personnel. Since I am interested in peaks that represent smoking events, I used the above-outlined parameters.

Once the peak locations were identified, I then moved to extracting the peak's start and end time, which is vital for obtaining the dynamics of infiltration of SHS from the MNR to the CHD location. To complete this task, I leveraged Python's SciPy package and its *find_widths* function to identify the peak's start and end times. The algorithm to calculate a peak's width is as follows:

1. Calculate the evaluation height h_{eval} with the formula $h_{eval} = h_{peak} - P * R$ where h_{peak} is the height of the peak itself, P is the peak's prominence and R a positive ratio specified with the argument relative height
2. Draw a horizontal line starting with peak evaluation height h_{eval} and extending it in both directions until it intersects the signal.
3. Calculate the width as the horizontal distance between the chosen endpoints on both sides. As a result of this, each peak's maximal possible width is the horizontal distance between its bases.

I passed the location of each peak, along with the relative height (i.e., R) of 0.8, which preliminary visual analysis indicated was an appropriate value to use. I then labeled each of the peaks for a home from $1 \ldots N$, where N is the total number of peaks in the home. The labeling of the peak helps with the analysis (e.g., I can precisely find out metrics for a particular peak). It also helps with visualizing the data (e.g., I can label peaks to understand exactly which peak I am referring to in a home) and in more in-depth analysis leveraging machine learning techniques to run the model over specific peak data. The labelling also helps to troubleshoot the analysis Figure 5 represents sample data for a home with peaks identified (red points) in the MNR and CHD location along with a start (yellow line) and end times (green line).

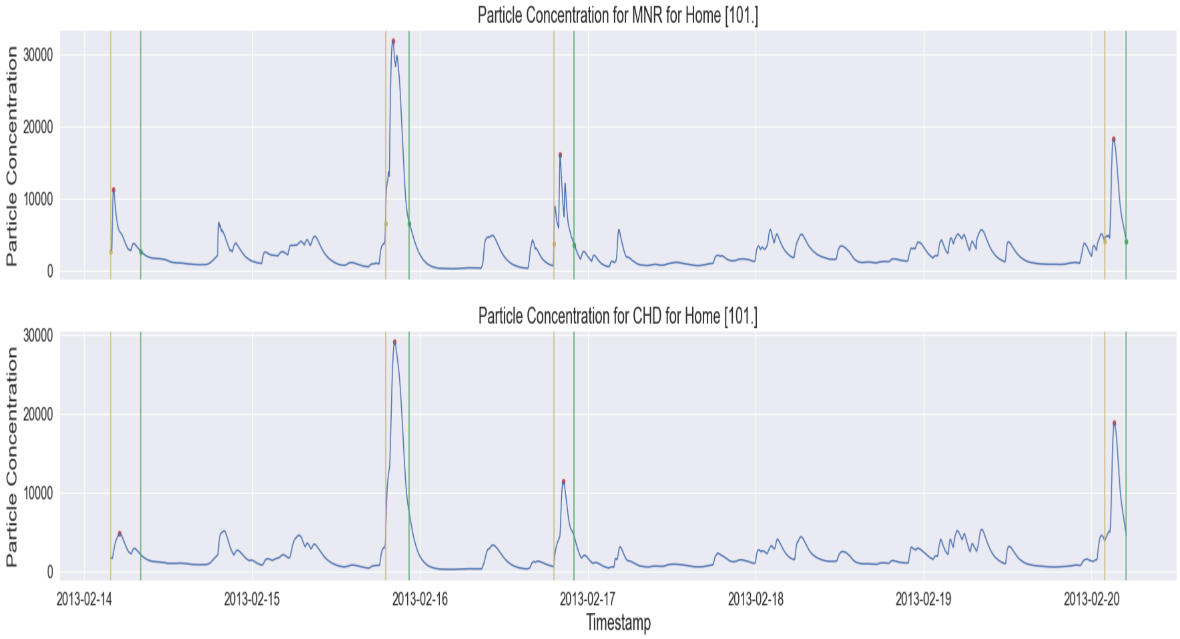


Figure 5: Peak start and end time for MNR and CHD. The red points are peak values in the MNR and CHD along with a start (yellow line) and end times (green line)

2.4 Extraction of CHD location data and defining peak lag

The analysis so far has been focused on the information from the MNR location, but I also need to extract the corresponding data from the CHD location. In our modeling efforts, I focus on defining the time between an MNR peak and the presence of elevated particle levels in the CHD room, which I call the *lag*. I calculate this property by considering the start and end time of a given MNR peak and identifying the maximum CHD air particle concentration value within this time interval. The lag is defined as the time between the MNR peak maximum value and the time of this maximum CHD value. As a representative example, in home 288, peak 2 starts at 07:07:am and ends at 9.36 am with the maximum value at 7.54 am. The corresponding CHD data for the peak 2 has a maximum value is at 8.05 am, so the lag, in this case, is (8.05 am – 7.54 am) i.e., 11 minutes.

2.5 Peak characteristics

Table 2 below provides summary statistics of the peaks extracted from the data. A total of 7,495 MNR peaks identified over all homes, for an average of 261 peaks per home throughout their enrollment. The average peak duration was 310 minutes and median was 188 minutes. The average maximum value of particle concentration in the MNR location was 44,992 and the CHD location was 15,618. The average lag between the timestamp of peak occurrence in the main room and the child room was 38.70 minutes.

Statistic	**Mean**	**Standard Deviation**
# Peaks per home	261.58	106.11
Duration of peaks (Minutes)	310.87	1357.62
Max value of peaks in MNR (Counts)	449,92.62	362,086.60
Max value of peaks in CHD (Counts)	15,618.94	14,429.23
Lag of peak between MNR and CHD location (Minutes)	38.70	96.81

Table 2 : Descriptive statistics of the smoke particle data.

2.6 Computationally efficient algorithm for data extraction

Due to the volume of data (~1 million measures per monitor per home), it is important to identify an efficient way to analyze data. Relevant information from the CHD data was obtained by extracting the air particle data that corresponds to the times for every peak in the MNR. This process takes 4-5 hours to perform for the 298 homes in the study. A Python code was implemented to take advantage of vectorization, a technique of implementing array operations without using for loops. In addition, I used functions defined by various modules which are highly optimized that reduces the running and execution time of code. Vectorized array operations are typically faster than their base Python equivalents, with the biggest impact in any kind of numerical computations since Python is an interpreted language and most of the implementation is slow. The main reason for this slow computation comes down to the dynamic nature of Python and the lack of compiler level optimizations which incur memory overheads. In addition to vectorization, I leverage a Python package swifter that uses the CPU's multiple cores to run the process parallelly. These optimization techniques reduced the runtime from 4-5 hours to approximately 30 minutes, a speedup of a factor of 8-10. All the preprocessed data and derived attributes were then stored in JSON files for future analysis. This approach to store data in key value pairs provide a flexible structure where we can add additional attributes without the need to modify the data structure to store the data.

2.7 Data inclusion and exclusion

I next identified peaks that were outliers in each home's data, which was accomplished by examining the mean and standard deviation of the maximum value for all peaks in a home. Outliers were defined as peaks with the maximum particle concentration greater than three standard deviations away from a home mean for all peaks. I identified these peaks and excluded them from the data analysis. I also found 10 peaks in the MNR location that do not have corresponding air particle data in the CHD location, i.e., there is no data in the CHD location that corresponds to the peak duration of the MNR location. These peaks were eliminated from all subsequent analysis. This process is different than the deletion of outliers mentioned in the 'Preliminary data cleaning' section. The latter functioned over the measurement level, where outliers were identified s based on the mean value of air particle concentration in the MNR location. The former excludes peaks in MNR that do not have corresponding peaks in the CHD location.

3 Quantifying the relationship between MNR and CHD concentrations

3.1 Introduction

The analysis in this chapter is focused on quantifying the overall relationship between PM2.5 in the MNR and CHD and to assess changes in this relationship associated with the introduction of the PFA intervention. More specifically, I would like to understand i.) whether SHS diffuses into the CHD location, ii.) how quickly this occurs, iii.) how much does the intensity of SHS contamination decline when it reaches the CHD location, and iv) what effect does the intervention have on these relationships. To investigate these questions, I implemented the following analytic approaches. The presence of SHS diffusion into the CHD location was investigated by examining Granger causality. The time required for infiltration into the CHD location was answered by the analysis of data and understanding the mean, variance of time difference when peak maximum occurs in the MNR and the corresponding peak maximum in the CHD location. The magnitude of the decrease was investigated by building a linear model that predicts the average concentration in the CHD as predicted by the associated MNR data. The effect of the intervention was assessed by performing a hierarchical linear mixed model. Below, I outline issues for each analysis, describe the methodology used to investigate, and detail results with both data visualizations and results from statistical tests.

3.2 Relationship between the MNR and CHD air particle concentrations

It does not make sense to quantify the diffusion of air particles from the MNR to the CHD locations if there is no relationship between the respective data from these locales. I therefore performed Granger causality tests[102] to determine whether the particle concentration in the main room influences the CHD particle concentration data. The Granger Causality tests the ability of one time series to predict another. In this test, with a time-series $X1$ and $X2$ the auto-regressive models can be written as

$$X_1(t) = \sum_{j=1}^{p} A_{11,j} X_1(t-j) + \sum_{j=1}^{p} A_{12,j} X_2(t-j) + E_1(t)$$

$$X_2(t) = \sum_{j=1}^{p} A_{21,j} X_1(t-j) + \sum_{j=1}^{p} A_{22,j} X_2(t-j) + E_2(t)$$

p is the number of elements that we want to compare in the time series; A is the model's coefficient, and E is the residual. If the variance of E_1 / E_2 is reduced by the inclusion of the $X_2 /$ X_1 terms, then it is said that X_2 / X_1Granger causes X_1 / X_2. The test returns a p-value. If this is less than a significance level of 0.05 null hypothesis is rejected, and it can be concluded that there exists a relationship between the two time series (i.e., the MNR and CHD air particle concentration). When there is no relationship between the data points for the time series in the MNR and the CHD location, it can be concluded that the peaks did not have Granger causal relationship meaning there was no relationship between the data.

I used the *grangercausalitytests* module from the *stats* model package in Python for this analysis. We passed the MNR and CHD data for each peak in each home into the Granger causal model. This process requires the value of maximum lag (i.e., p in the above formula) to be identified, which limits what lags are evaluated as we examine the time series for causality. We set the maximum lag to 1 based on an iterative process where we looked at the failure rate by including different values of this parameter. A value of 1 yielded the fewest failures

I performed the Granger causality test for all sets of peaks in the dataset (7,495 peaks), and 7.6% (576) did not have Granger causal relationship. We excluded these peaks from subsequent analysis since we only want to include peak data where the particle concentration in the MNR location influenced the CHD location concentration.

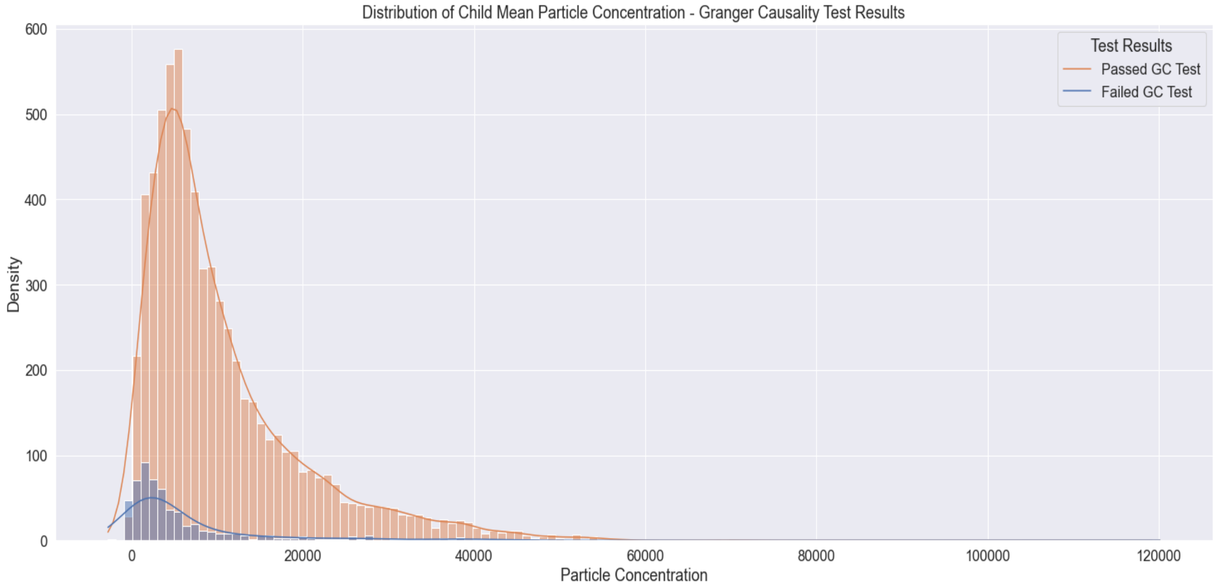


Figure 6: Granger Causality Results, x axis indicates mean air particle concentration in CHD location y axis indicates the density, the blue bars represent peaks for that failed and the brown bars represents peaks for that passed the granger causality test

Figure 6 represents the distribution of the mean value of particle concentration in the CHD location. The data is for each of the peaks that have passed or failed the Granger Causality test. The finding is 93% of the peaks pass the test, with most of the failures being where the peak values are in the lower range of MNR air particle concentration. This result is expected as it indicates less impact to the CHD location for a lower concentration of smoke particles in the MNR. For average particle concentration above 20,000 in the MNR, there was always an impact on the CHD location. The fact that only low MNR air particle concentrations were associated with a non-Granger causality relationship between the MNR, and CHD signals lends face validity to the use of Granger causality to establish a relationship between the two rooms.

3.3 Diffusion between MNR and CHD location

I am interested in understanding the how quickly the particles diffuse from the MNR to CHD location. I measure this by looking at the lag time, i.e., the difference between the timestamp when the peak maximum occurs in the MNR and the corresponding peak in the CHD.

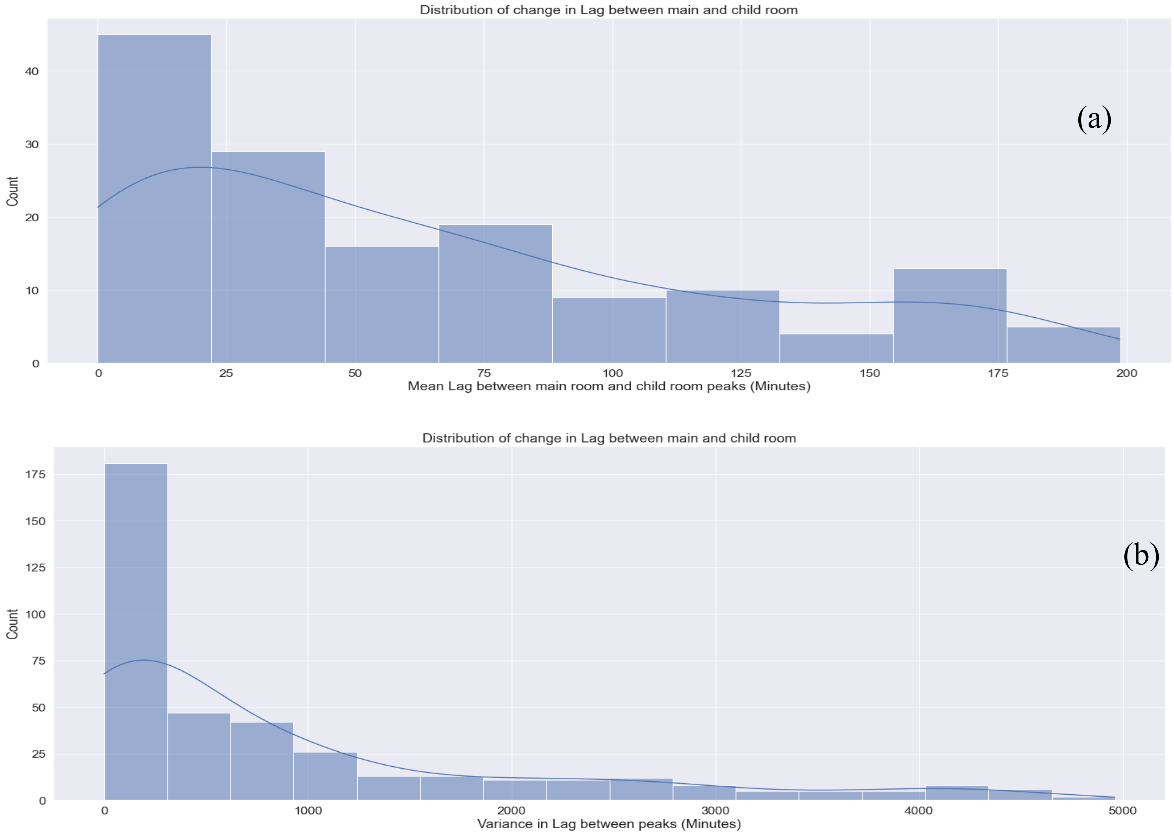


Figure 7: Mean (a) and Variance (b) of lag between MNR and CHD location, x axis indicates the mean and variance in lag between peak of main and child's room y axis indicates the density.

Figure 7 represents the mean and variance of the lag between the main and CHD location for all the peaks across the complete dataset. The mean value of lag across all the homes is 38.70 minutes, meaning that after smoking in MNR, nearly 39 minutes elapses until air particle concentration reaches highest level in CHD.

I also examined select demographic variables that could potentially affect the lag between the MNR and CHD and to understand their impact.The number of rooms in the home could influence the rate at which smoke air particles diffuses from MNR to CHD, the more the rooms the slower the diffusion.The number of adults in a home could provide insight into if highly populated homes have any influence on the smoke air particle difussion. The correlations between the mean lag and the number of adults in the home, the number of rooms in the home, and the presence of a single parent and the mean lag are 0.0122, 0.0144, and 0.00147, respectively. These values indicate that there is low impact of these selected demographic variables to time it takes for the health of the child to be impacted across all the homes in the study.

3.4 Attenuation in maximum values in MNR and CHD location

I was also interested in quantifying the attenuation in air particle concentration that occurs as the smoke diffuses from the MNR to the CHD location. To address this, I fit the following linear regression model:

$$Y = \beta_0 + \beta_1 * X,$$

where beta coefficient (β_1) of the quantifies the rate of the decline. Assume that (β_1) =1. Then the air particle concentration in the CHD location matches the air particle concentration in the

MNR location, up to an additive constant. If (β_1) >1, this means that the air particle concentration in the CHD location is higher than the air particle concentration in the MNR location If (β_1) <1, then air particle concentration in the CHD location is less than the air particle concentration in the main room, which represents attenuation of the peak level. I found that β_0 is the intercept, was 182.91 and a β_1 was 0.684391. This result indicates that on average for a given peak is 68% as high in the CHD room versus MNR, which represents a 32% reduction in mean particle count.

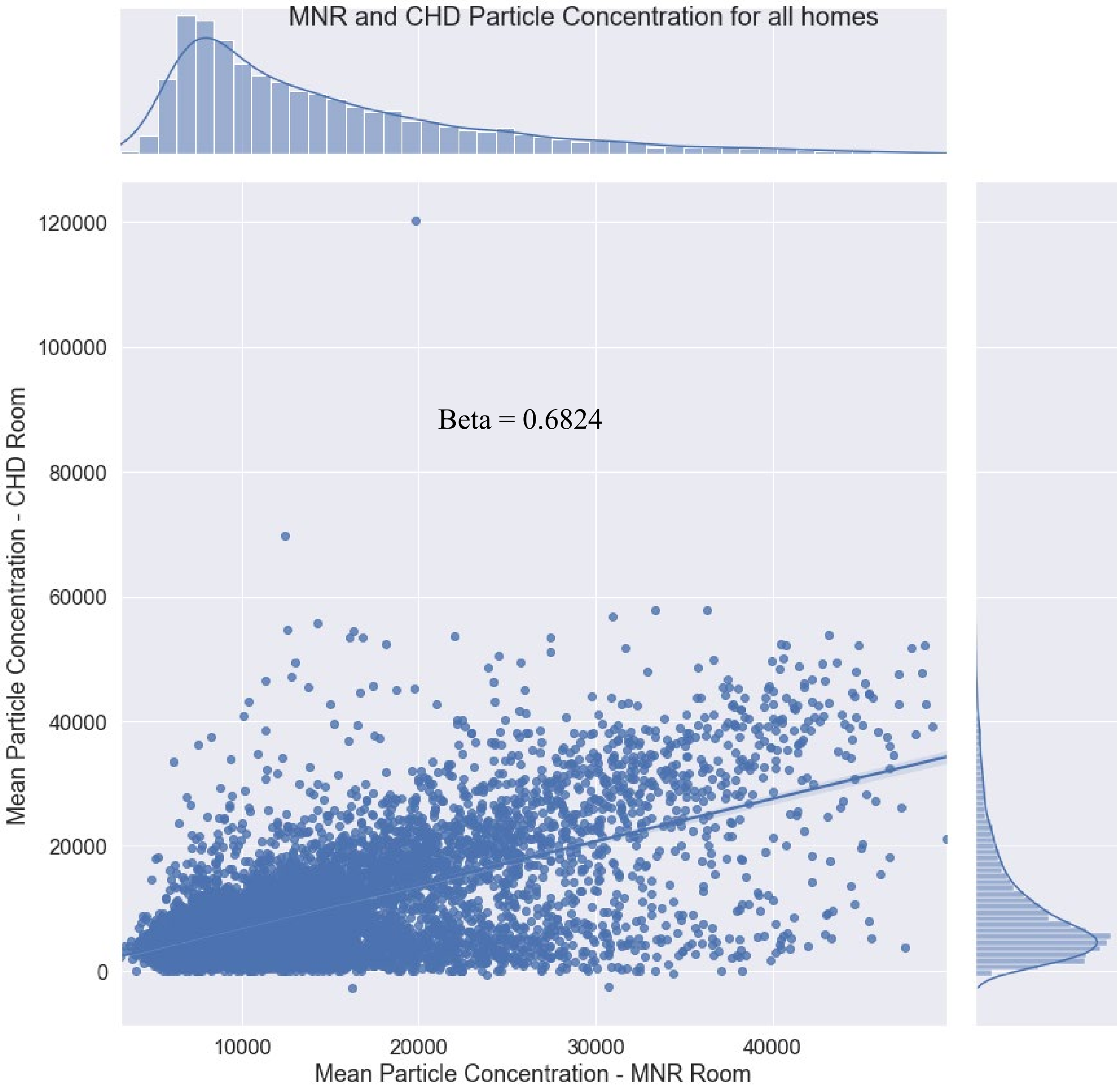


Figure 8: Joint plot to represent relationship between main and CHD location particle concentration. x axis indicates the mean air particle concentration in the MNR location y axis indicates the air particle concentration in the CHD location

Figure 8 displays the results of the regression graphically. represents the particle concentration's relationship and distribution in the MNR and CHD location. The data is for all the peaks in both the control and experimental homes. Each dot in the center plot represents particle concentration's peak value in the MNR and CHD location. The blue line in the center plot represents the linear regression. The linear regression equation is

$$Y = 182.91 + 0.6824 * X$$

3.5 Graphical Assessment of Intervention

Before analyzing the effect of the PFA intervention analytically, I explore visualizations that illustrate features of the data both before and after the intervention. While not as rigorous as statistical analysis, visualization often allows aspects of the data that are missed with statistical analysis to be seen, which in turn allows for a greater understanding of the research findings.

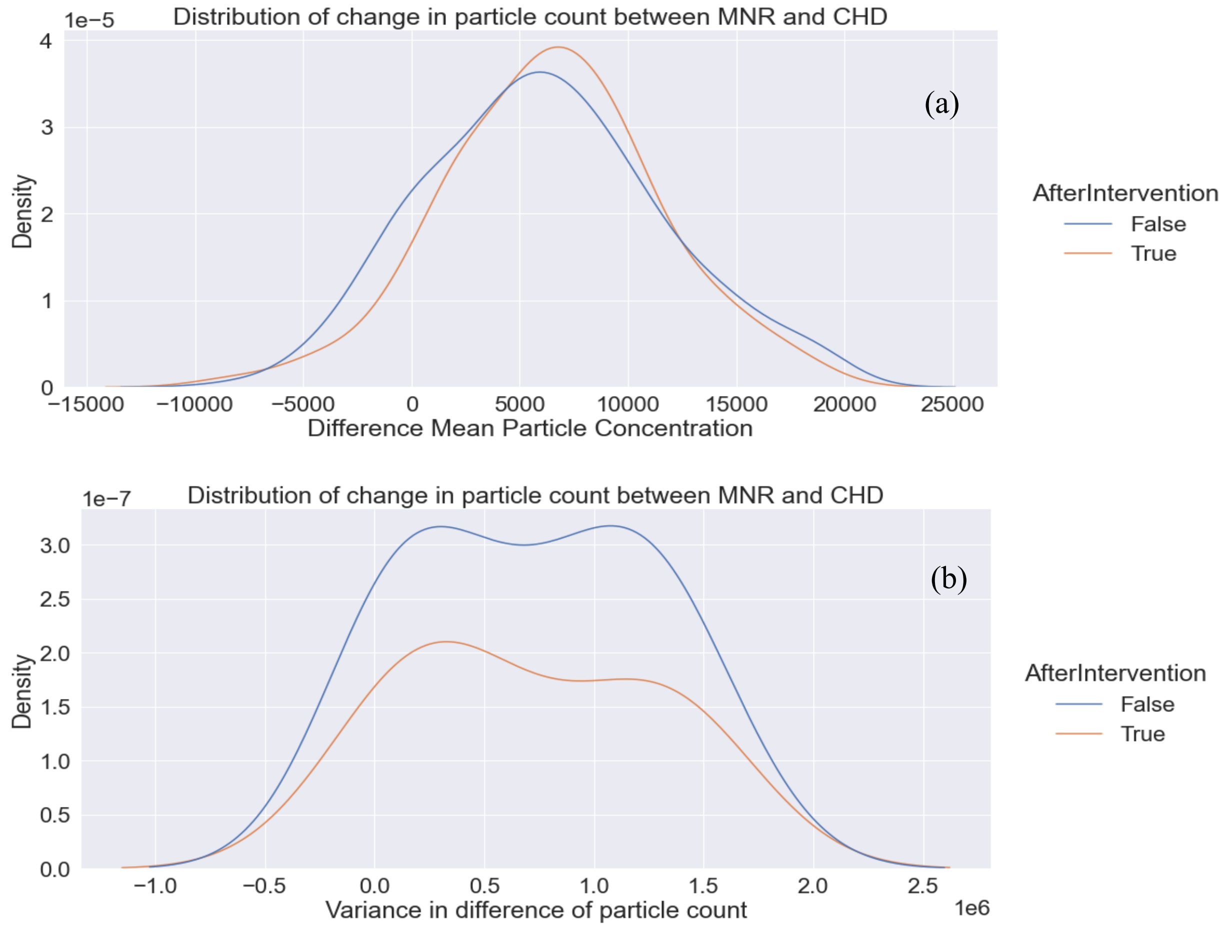


Figure 9: Mean (a) and Variance (b) of particle concentration between MNR and CHD, x axis indicates the mean and variance in air particle concentration y axis indicates the density, the blue line for baseline and the brown line is for treatment group.

In Figure 9, the distribution of the mean particle count difference, calculated for each peak as the mean value of the MNR peak subtracted by the mean value of the CHD peak, is presented for all peaks across all homes. The distribution of the variance of this metric is also provided. The information is presented before (blue line) and after (brown line) intervention. Based on the calculation, positive values represent reduction in air particle concentration in the CHD location and negative values represent increase in air particle concentration in CHD location Concordantly, I would expect an effective intervention to be reflected in a positive value for this metric. The figure shows that distribution has shifted to the right after the intervention, and the variance is trending downwards, with the mean value before intervention being 5532 and after intervention is 5881. This is a small shift, so I ran the Kologromov-Smirnoff test to examine if the differences between the distributions were statistically significant. The calculated statistic was 0.0563 and p value was 8.8368e-05, less than 0.05 so I can conclude than the two distributions are significantly different.

3.6 Quantification of Intervention Effect

To quantify the intervention effect more thoroughly, I now support the above graphical analysis with inferential statistics. For this analysis, I used mixed linear models,[103]which are mainly used when there is non-independence in the data, such as a hierarchical structure or repeated measures s. Linear mixed models are an extension of simple linear models that allow both fixed and random effects. Fixed effect variables are those for which all variable levels of interest are available and compared to each other. For example, in the current trial, study groups (i.e., control vs. experimental) is a fixed effect. Random effect variables, on the other hand, have many possible levels, only a small number of which are represented in the data. Therefore, the level-specific characteristics (e.g., participant means of a repeatedly measured variable) are assumed to be drawn from a normal distribution. In the current trial, participant Id is a random variable, since there are many different individuals who could be included in the study, only a small number of which were recruited. I do not wish to compare one participant versus another, but only model participant specific variables (e.g., mean daily particle counts) being drawn from a common distribution.

As part of the PFA intervention, the homes were randomly assigned to either the control or experimental groups in blocks of two to ensure the same sample size across home conditions. In the experimental home, the study enrollment was easily stratified into "Baseline/Post-Baseline" time frames corresponding to when the intervention was initiated. In control homes, data was passively recorded, and such a delineation did not exist; therefore, the "Baseline/Post-Baseline" designation was assigned to that of its corresponding experimental home. This assignment is a fixed effect and is a vital part of the analysis to understand the intervention event's behavior and impact.

For a two-level mixed linear model, first-level equation is written as

$$Y_{i,j} = \beta_0 + \beta_1 X_{i,1} + \beta_{j,2} X_{i,2} + \varepsilon_{i,j} ,$$

where $Y_{i,j}$ is the i^{th}observation collected from for the j^{th} unit. β_0 is the overall intercept, β_1 is the regression coefficient for the fixed effect associated with the i^{th} measurement of variable X_1, $\beta_{j,2}$ is the regression coefficient for a random effect for the j^{th} unit, and $\varepsilon_{i,j}$ is the residual, such that $\varepsilon_{i,j} \sim N(0, \sigma_0)$. The second level equation is written as

$$\beta_{j,2} = \beta_{0,2} + \gamma_j ,$$

such that $\gamma_j \sim N(0, \sigma_1)$. This example demonstrates that, in a mixed linear model, the effects for each home are assumed to be normally distributed around some grand mean, i.e., $\beta_{0,2}$.

For the existing analysis, the dependent variable is mean CHD location particle concentration (C_{mean}) and the independent variables are mean MNR location particle concentration (P_{mean}), the experimental condition (control/experimental) $H_{c/e}$ and the intervention status (baseline/during intervention) $I_{b/a}$. As stated above, participant ID is a random variable. Within this framework, I implemented bivariate, full regression, and two-way interaction, and three-way interaction between experimental/treatment home, before/after intervention event and particle concentration respectively. The models evaluate the relationship between the mean value of particle concentration in the CHD location as predicted by the mean value of particle concentration in the MNR, control home vs. experimental home, and before vs. after the intervention. The equations for each of these models are as follows (for clarity we eliminated the i and j subscripts in the hierarchical regression equations): -

Model 1. Bivariate

$$C_{mean} = \beta_0 + \beta_1 * P_{mean}$$

$$C_{mean} = \beta_0 + \beta_1 * H_{c/e}$$

$$C_{mean} = \beta_0 + \beta_1 * I_{b/a}$$

Model 2. Full regression

$$C_{mean} = \beta_0 + \beta_1 * P_{mean} + \beta_2 * H_{c/e} + \beta_3 * I_{b/a}$$

Model 3. Two-way interaction

$$C_{mean} = \beta_0 + \beta_1 * P_{mean} + \beta_2 * H_{c/e} + \beta_3 * P_{mean} * I_{b/a}$$

$$C_{mean} = \beta_0 + \beta_1 * P_{mean} + \beta_2 * I_{b/a} + \beta_3 * P_{mean} * H_{c/e}$$

Model 4. Three-way interaction

$$C_{mean} = \beta_0 + \beta_1 * P_{mean} + \beta_2 * H_{c/e} + \beta_3 * I_{b/a} + \beta_4 * P_{mean} * H_{c/e} + \beta_5 * P_{mean} * I_{b/a} + \beta_6 * P_{mean} * H_{c/e} * I_{b/a};$$

where β_0 is the intercept and β_1 , β_2 , β_3 and β_4 are the coefficients of the independent variables. The bivariate models capture each independent variable's individual effects. The full regression model allows capturing the combined effect of the independent variables while controlling for each other. The two-way interaction allows us to examine the combined effect of before/after intervention and experimental/treatment home independent variables. The three-way interaction allows capturing the combined effect of all independent variables.

Lastly, we also controlled for select demographic variables (marijuana usage in the home, adult in the home is single parent, the race and ethnicity of the family, the number of children and adults in the home and the type of home i.e., single family, condo, townhouse) by including them as independent variables during the model evaluation. The hierarchical mixed linear models were run using both random intercept and random slope formulations for the participant ID random effect. The effect of each independent variable constant in the random intercept model and the intercept changes based on the intervention variables' effects. In the random slope model, both the slope and the intercept are varied to capture the intervention's effect.

3.7 Hierarchical Model Results

Table 3 summarizes numerical results of the random intercept hierarchical linear mixed models (bivariate, full regression, two way and three-way interaction) and visualizes results illustrating the relationship between MNR and CHD location monitors for the three-way interaction. In Models 1-3, all regression coefficients were significant, with the exception of a control vs.

experimental home. There was a positive coefficient for MNR particle count, meaning as higher particle concentration in MNR results in higher particle concentration in CHD location. There was a negative coefficient for the effect of switching from the baseline to treatment period, indicating that indicates there is drop in the air particle concentration in the CHD location associated with the onset of the intervention. Both effects signs were in accordance with what we expected to see.

	Model1 (Bivariate)	Model2 (Full Regression)	Model3 (Two Way interaction)	Model4 (Three Way interaction)
	Coefficent(CHD location Particle Count – Mean) (Standard Error)			
MNR Particle Count(Mean)	**0.476 (0.023)**	**0.475(0.018)**	**0.462(0.020)**	**0.420(0.023)**
Experimental Home (ref = Control Home)	-0.012(0.024)	0.010(0.003)	0.006(0.003)	0.022(0.003)
Treatment Phase (ref = BL Phase)	**-0.084(0.022)**	**-0.051(0.024)**	**-0.049(0.020)**	-0.035(0.033)
MNR Particle Count(Mean) * Exp Home			**0.029(0.023)**	**0.112(0.036)**
MNR Particle Count(Mean) * Treatment Phase			**0.009(0.010)**	**0.071(0.031)**
MNR Particle Count(Mean) * Treatment Phase * Exp Home				**-0.146(0.047)**

Table 3 : Hierarchical Linear Mixed Model (Random Intercept) results.

***Values in bold indicate significant values**

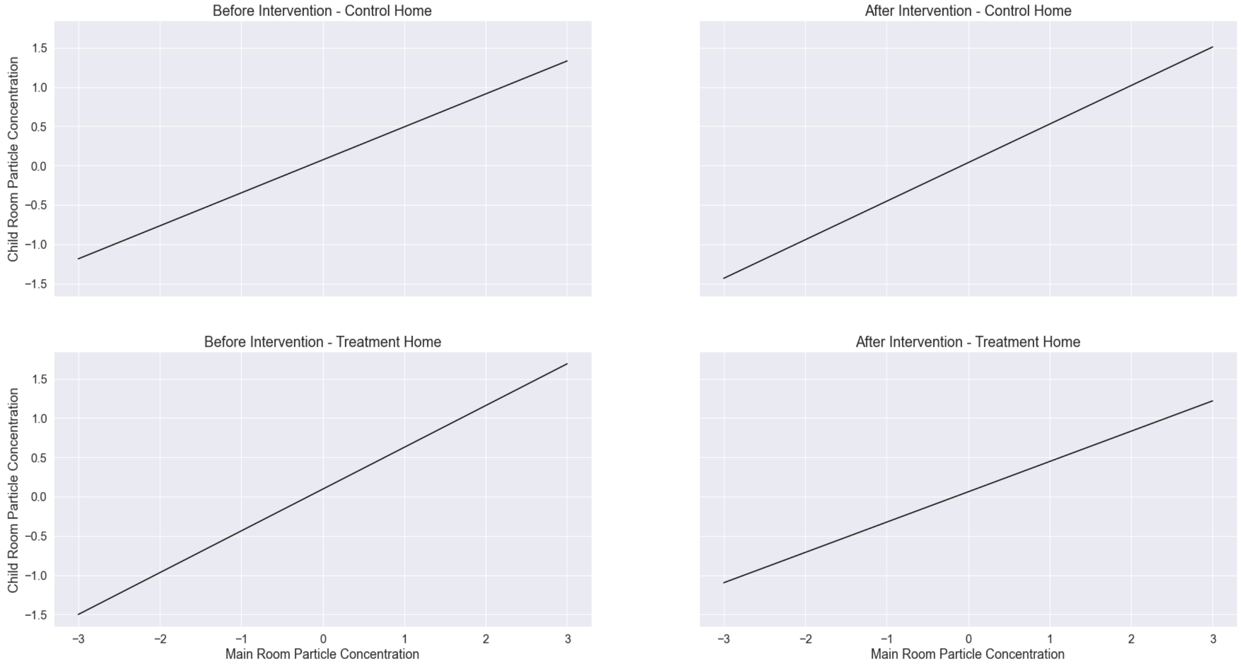


Figure 10 : Hierarchical Linear Mixed Model (Random Intercept) results, x indicates the air particle concentration in the main room and y indicates the air particle concentration in the CHD location. The four quadrants capture the effort of the before and after intervention effect in the control/ treatment home for three way interaction model.

I now provide a detailed description of the Model 4 results, which allow changes in the MNR-CHD location relationship according to the experimental group and treatment phase to be examined. The βvalues quantify the interaction level between MNR and CHD location air particles. The high value (< 0.05) defines a significant relationship. Figure 10 shows the fitted relation between MNR and CHD location air particle concentration before and after intervention in both baseline and treatment homes (based on the $MNR\ Particle\ Count(Mean) * Treatment\ Phase * Exp\ Home$ interaction term and values of categorical variables.). Because I used standardized values, I examined +/-3 SDs from the mean for main room in these figures as reflective on the X-axis. As shown in the top-left panel, the relationship between the

MNR and CHD location monitors for a control home during the baseline phase is captured by the MNR particle count ($B_1 = 0.420$). The overall effect of the transition from the baseline to treatment phase on CHD location mean particle counts for a control home is given by B_3=-0.035, which is not significant. The change in the relationship between MNR and CHD location mean particle count when transitioning from baseline to treatment phase for a control home is given by the non-significant interaction coefficient $B_5 = 0.071$. These effects are shown in the top-right panel of Figure 10. The relationship between the MNR and CHD location monitors for an experimental home during the baseline phase is shown in the bottom-left panel of Figure 10. The overall effect is given by ($B_2 = 0.022$), which is not significant. The overall effect of transitioning from baseline to treatment in an experimental home is defined by the interaction coefficient ($B_4 = 0.112$)., while the effect on the relationship between MNR and CHD location is defined by the interaction coefficient ($B_6 =$ -0.146). This is visualized in the bottom right panel.

Overall, I see that as the air particle concentration in the main room increases, there is a corresponding increase in the CHD location. The slope in the two left panels of Figure 10 is also almost identical ($B_2 = 0.022$ which is not significant), which is expected since during the baseline since there is no difference for control and treatment homes in how the particles traverse between the MNR and CHD locations. The slopes on the top two panels are also similar (B_3=-0.035 which is not significant), which is expected since there was no intervention in the control homes. Any change in smoking behavior could be attributed to the reactive effect[108]. As shown in the bottom two panels, after the intervention has been initiated in the experimental homes, there is a significant change in relationship between MNR and CHD location monitors. This negative change in slope is very important and signifies that there is an impact on the main room's smoking influence on the particle levels in the CHD location because of the intervention.

I can quantify this effect by the three-way interaction term listed in Figure 10. The negative value (B_6 is -0.146 which is significant) of the three-way interaction indicates a slope in the downward direction, I can conclude that there is 14% decrease in the air particle concentration in the CHD location of an experimental home after an intervention event.

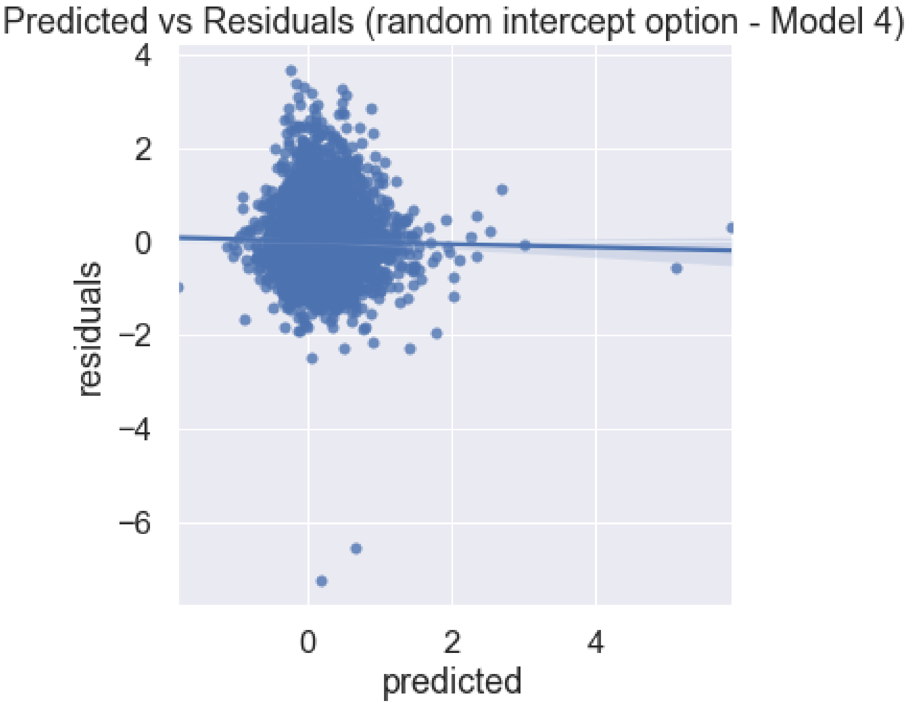


Figure 11 : Hierarchical Linear Mixed Model 4 (Random Intercept) Residual plot, x axis indicates predicted values y axis indicates the residuals.

The residual plot in Figure 11 has most of the predictions centered around 0 and shows few outliers. The dark blue variance line is aligned near-zero, indicating that the residuals are spread evenly across overestimating and underestimating the actual child particle concentration values. This behavior indicates a well-fitted model.

Table 4 illustrates a random slope model result. In Models 1-3, except for the term associated with Control vs. Experimental home, the regression coefficients are significant. Coefficient signs were in accordance with what we expect to see (e.g., positive coefficient for main room particle count and negative coefficient for the effect of switching from the baseline to treatment period). The coefficients for interactive terms are not significant, but they demonstrate expected behavior (e.g., the coefficients are negative from the experimental home and after intervention).

	Model1 (Bivariate)	Model2 (Full Regression)	Model3 (Two Way interaction)	Model4 (Three Way interaction)
	Coefficent(CHD location Particle Count – Mean) (Standard Error)			
MNR Particle Count(Mean)	**0.531 (0.021)**	**0.529(0.046)**	**0.575(0.036)**	**0.600(0.042)**
Experimental Home (ref = Control Home)	-0.016(0.026)	0.015(0.026)	0.027(0.023)	0.026(0.035)
Treatment Phase (ref = BL Phase)	**-0.060(0.027)**	**-0.056(0.031)**	**-0.057(0.019)**	**-0.051(0.031)**
MNR Particle Count(Mean) * Exp Home			**-0.098(0.053)**	-0.084(0.061)
MNR Particle Count(Mean) * Treatment Phase			**-0.056(0.036)**	-0.042(0.052)
MNR Particle Count(Mean) * Treatment Phase * Exp Home				-0.034(0.021)

Table 4 : Hierarchical Linear Mixed Model (Random Slope) results.

***Values in bold indicate significant values**

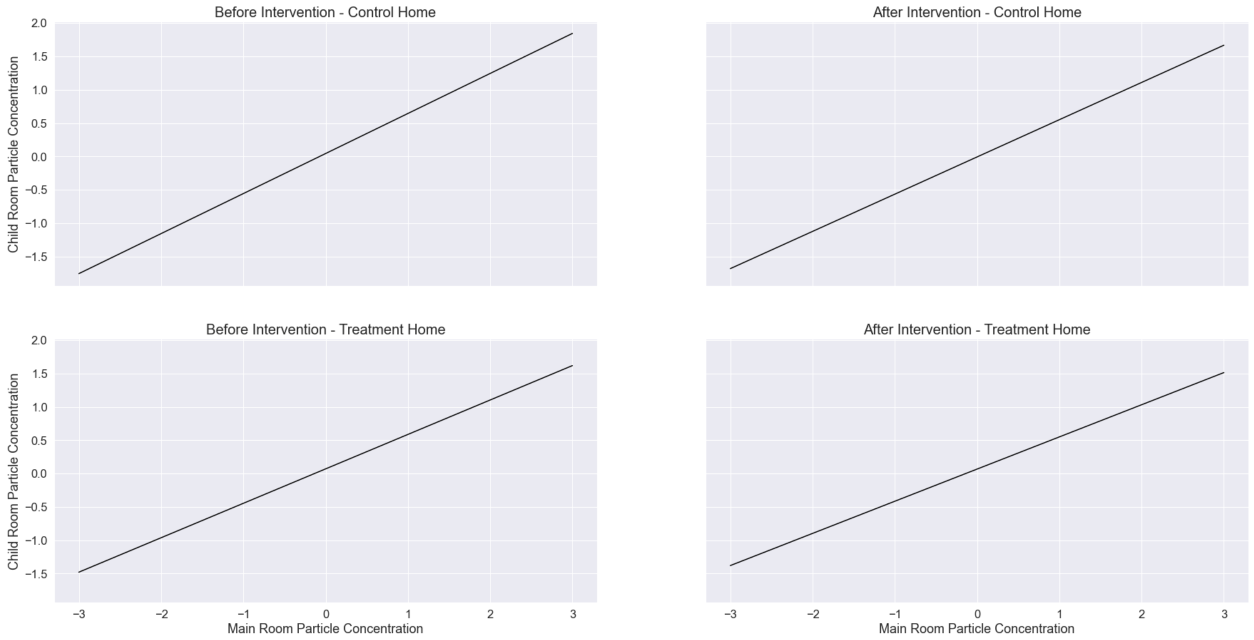


Figure 12 : Hierarchical Linear Mixed Model (Random Slope) Results, x indicates the air particle concentration in the MNR location and y indicates the air particle concentration in the CHD location. The four quadrants capture the effort of the before and after intervention effect in the control/ treatment home. The table below quantifies the effect.

For this model, I also provide a detailed description of the Model 4 results, which allow changes in the MNR-CHD location relationship according to the experimental group and treatment phase to be examined. Similar to Figure 10, the top-left panel of Figure 12 displays the relationship between the main and child room monitors for a control home during the baseline phase. This relationship is captured by coefficient ($B_1 = 0.600$). The overall effect of the transition from the baseline to treatment phase on CHD location mean particle counts for a control home is given by B_3=-0.051. The change in the relationship between MNR and CHD mean particle count when transitioning from baseline to treatment phase for a control home is given by the interaction coefficient $B_5 =$ -0.042, which is not significant. The effects are shown in the top-right panel of Figure 13. The relationship between the MNR and CHD monitors for an experimental home during the baseline phase is shown in the bottom-left panel of Figure 12. The overall effect is given by (B_2 = 0.026), which is not significant and the effect of transitioning from control to experimental home on the MNR-CHD monitor relationship is defined by the interaction coefficient (B_4 = -0.084), which is not significant. The transition from baseline to treatment phase in an experimental home is defined by the interaction coefficient (B_6 = -0.034), which is not significant. This is visualized in the bottom right panel. The slope in all the panels is identical, with a small shift in the lower right panel. While these results are not significant, they are in the expected direction and provide corroborating evidence for the random-intercept results shown above.

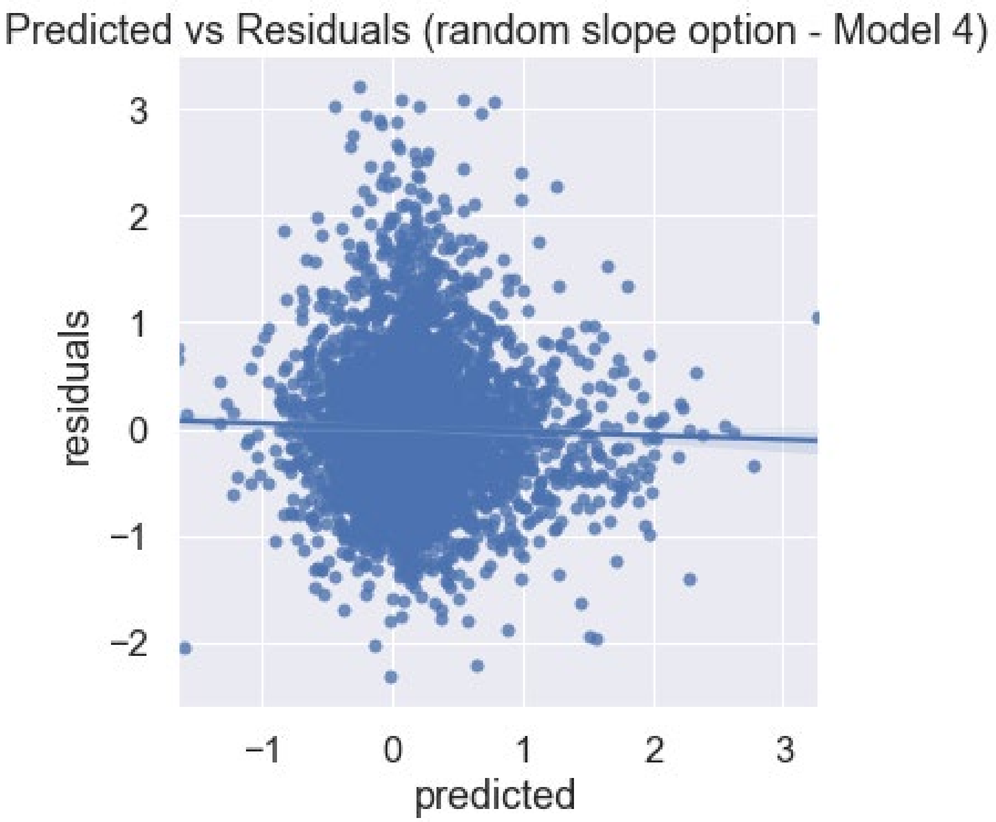


Figure 13 : Hierarchical Linear Mixed Model 4 (Random Slope) Residual plot, x axis indicates predicted values y axis indicates the residuals.

The residual plot in Figure 13 displays a pattern where the model overestimates the CHD location particle concentration when the standardized values are between -1 and 1 and underestimating the values when the standardized value is greater than 1. Overall, the dark blue line is almost flat indicating that we have a well fit model.

Demographic Variable	Random Intercept Model	Random Slope Model
	Coefficient (Standard Error)	
Race Ethnicity (Hispanic)	0.064(0.036)	0.066(0.034)
Race Ethnicity (White)	0.014(0.037)	0.028(0.035)
Race Ethnicity (Other)	0.069(0.037)	**0.076(0.035)**
Single Parent	0.007(0.024)	**0.001(0.023)**
# Of Children in the home	0.009(0.010)	0.004(0.009)
# Of Adults in the home	-0.005(0.011)	-0.002(0.011)

Table 5 : Hierarchical Linear Mixed Model - Coefficient of demographic variables

We also capture the information of demographic variables. Table 5 summarizes the effect of demographic variables for Random Intercept and Random Slope model. These variables are controlled by including them in the hierarchal linear model as independent variables. Single parent and Race Ethnicity (Other) are significant (p value < 0.05).

3.8 Summary

The analysis of quantifying the relationship between the air particle concentration in MNR and CHD location was focused around four goals.

i.) whether SHS diffuses into the CHD location, ii.) how quickly this occurs, iii.) how much does the intensity of SHS contamination decline when it reaches the CHD location, and iv) what effect does the intervention have on these relationships. The analysis (section 3.3) indicates that that SHS does diffuse into the CHD location. The average lag is 38.70 minutes for the existing dataset. For a given peak, there is a 68% reduction in mean particle count in the CHD versus the

MNR location.The hierarchial linear mixed model demonstrates the impact of the intervention event where there is a reduction in the SHS after a intervention event. I am also able to quantify the change as displayed in Figures (10 & 12).

There is evidence from the data analysis the particle concentration is reduced in the CHD location after the intervention event. There are still some questions around if this is due to the air particle monitors' notification or any other factors that cause this behavior change. An example to illustrate this point is the strong relationship between PM2.5 in the MNR and CHD location before the intervention event in some treatment homes. We would expect that there should be a near-identical relationship before intervention in the control home as there is no intervention event, only the monitor is placed in the home. We see from the analysis that that is not the case. This behavior could be attributed to the effect of reactivity[107], where there is a change to the smoking behavior by just placing the monitor in the person's room.

As future research activity several aspects of with work remain to be investigated, including a more detailed account of the impact of the intervention mechanism on smoking behavior. The current study activates the notification (audio or visual) when the particle concentration reaches a specific threshold value (>= 15,000) in the main room and if there is an impact on the CHD location particle concentration if we lower the intervention threshold below 15,000. We know from the Granger causality tests that 8% of the peaks fail the test (i.e., particle concentration in the CHD location does not depend on the particle concentration in the MNR location), and most of these failures are lower particle concentration in the main room. In future work, I will explore the earliest we can predict the threshold will reach 15,000 counts in the CHD location.

The effect of demographic variables should also be further explored in the future. I would like to evaluate whether different behavior exists between homes with marijuana smokers versus the rest. It will also be interesting to evaluate if the number of children at home impacts the smoke particle concentration between the MNR and CHD location. As part of the paper's enhancement, I will look at clustering techniques and understand if these demographic variables impact smoking behavior.

4 Predicting child’s room air particle concentration from main room

4.1 Introduction

The previous analysis demonstrated that the PFA trial successfully increased air particle quality in enrolled homes by deploying an intervention that immediately emitted audio/visual notification upon detecting air particle levels above a 15,000 counts threshold. Earlier analysis (section 3.7) indicated a reduction in smoke particles after the air monitor feedback was activated. While this is undoubtedly a step in the right direction, there are many opportunities to improve upon this intervention system, including that the fact that on an average it will take approximately 39 minutes (section 3.3) before we know that there is a potential impact to the child health in the CHD location. Smokers may be unaware of this risk, especially if they cease smoking in response to monitoring feedback and fail to take steps to protect children's health, either by mitigating the risk associated with the recent smoking event or reducing indoor smoking events. Consequently, I aim to develop a proactive mechanism to quickly identify if/when a smoking event in the main room can potentially impact the child's bedroom environment and possibly their health. The Granger causality tests outlined in Section 3.2 indicates an established relationship between air particle quality in the MNR and the CHD locations. Every home has a varied relationship for diffusion into the child’s bedroom depending on factors such as distance between rooms the square footage of homes. In this chapter, I develop a model that can identify the potential diffusion of SHS into a CHD location as quickly as possible so that appropriate mitigation activities can be initiated to reduce potential risks to the 'child's health.

There are three considerations that I accounted for as I built the model. The first issue revolves around reducing false negatives, which correspond to the model failing to identify an instance of the CHD being impacted by an MNR smoking event. I want to minimize the events where the air particle concentration is above the threshold and model fails to detect this event. It is critical to minimize the occurrence of false negatives. A naïve way to eliminate false negatives would be to score every peak as impacting the child’s environment, but this would reduce the accuracy of the model and likely introduce numerous false positives. This would result in a “boy who cried wolf” effect where caregivers may ignore critical notifications when there are too many false positives. This competition between false negatives and false positives leads to the second consideration in the model building exercise - identifying a set of parameter values that sufficiently balance accuracy and false negatives. The third model building consideration is whether to approach the modeling as a prediction or classification problem. For the former, I can focus on training a model to accurately predict the particle concentration in the CHD location. But this approach may be needlessly complex, caregivers are likely to find a binary impact vs. non-impact outcome based on some particle threshold to be useful [115]. Below, we explore the parameterization that best balances false negatives and false positives and explore both continuous and dichotomous model outcomes.

4.2 Raw vs feature engineering

The data for the study is time series data that can be primarily analyzed in two ways. I can analyze at the raw data or extract features from the raw data and analyze the features. In this approach, the raw data is data as it is collected by the sensors (after being smoothed), but I only consider the peaks that are extracted (i.e., all the data points between start and end of the peak) according to the methodologies outlined in Chapter 2. The input to the models is the raw data set, i.e., all data from an extracted peak will be used to classify whether a CHD location is impacted. In feature analysis I extract characteristics from the peak data, which represent aggregate metrics from the data that are fed into a model. I ran a preliminary analysis using the raw data and features for both continuous and dichotomous outcomes for several feature based approaches + Long Short Term Memory (LSTM) model, which is standard to identify a model that provides the best accuracy and low occurrence of false negatives.

4.3 Feature selection and modeling approach

I explored basic features, like mean, median, standard deviation, minimum, maximum, mode, 25 percentile, 75 percentile of air particle measures over an entire peak, as well as more complex features primary derived from the *tsfeatures* package in Python. Table 6 shows summary of features that were considered.

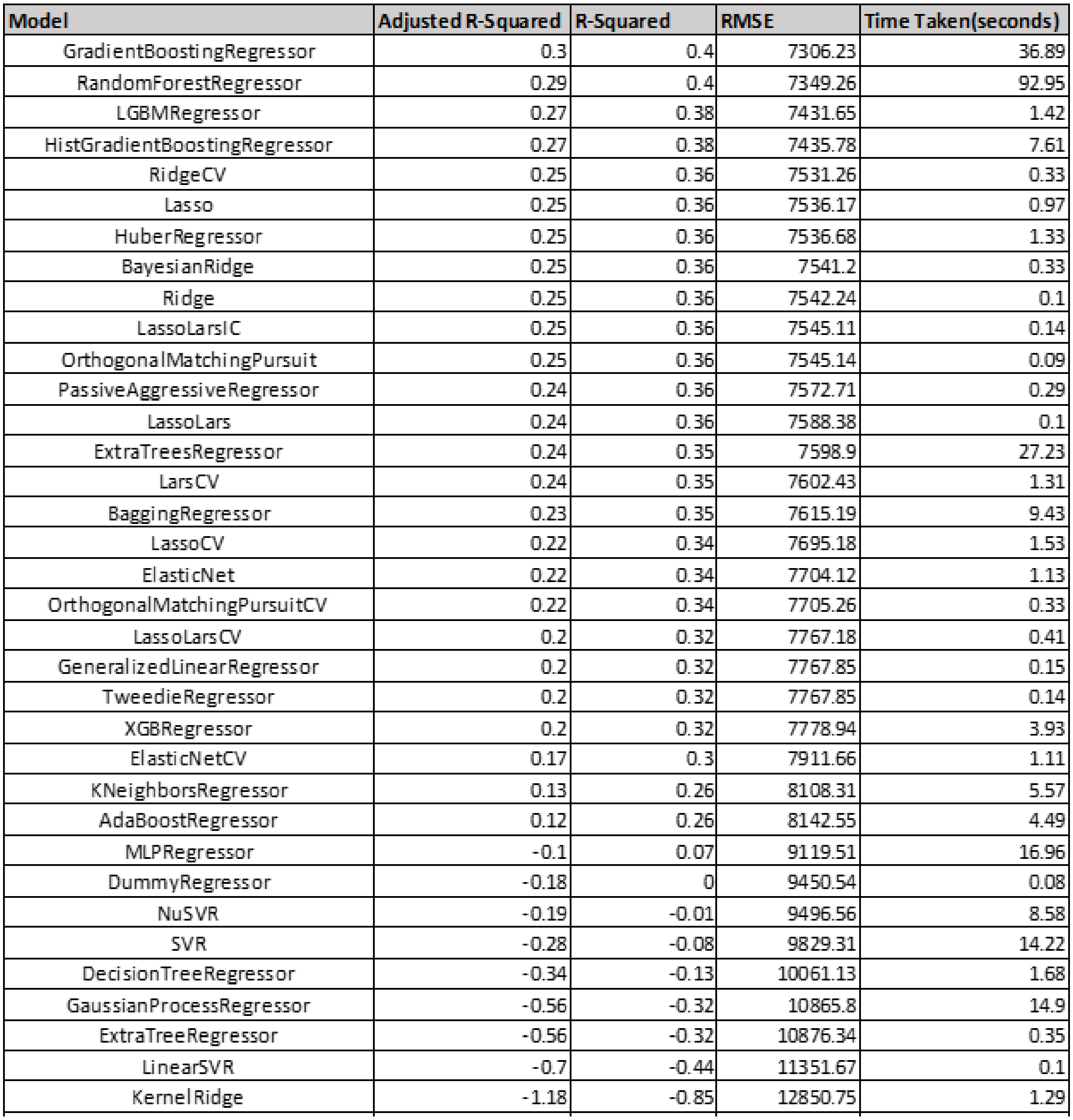
Feature	Description
acf_features	This is a vector that represents the sum of the first ten squared autocorrelation coefficients.
pacf_features	The feature produces a vector of 3 values, which represents the first 5 partial autocorrelation coefficients of the original series, the first differentiated series and the second order differentiated series.
heterogeneity	The variability of time series data is captured in a vector of 4 values.
nonlinearity	The feature measures the linearity of the time series data. The feature has a large value when the data is non linear and trends towards 0 for linear data.
entropy	This feature quantifies the amount of regularity and unpredictability of changes in time series data.
lumpiness and stability	The stability is variance of means and lumpiness is the variance of the variances of time series data that is tiled over non overlapping windows.
max_level_shift, max_var_shift	These features represent the max shift and max variance between times series data that is tiled over non overlapping windows.
crossing_points	The feature captures the numbers of times the time series crosses the median line.
flat_spots	The feature computes the maximum length between each interval where the time series is divided into ten equal intervals.
hurst	The hurst exponent measures if the time series is persistent (value > 0.5) , anti persistent (value < 0.5) and random (value = 0.5).
stl_features	The stl features captures the trend and seasonality of the time series data.
ac_9	The feature captures the auto correlation coefficient at lag 9.
firstmin_ac	The first minimum value in the auto correlation function
firstzero_ac	The first zero crossing of an auto correlation function
binarize_mean	The existing time series is converted into a value of zero(below the mean) and one (above the mean)
outlierinclude_mdrmd	The feature capture the median of the outliers

Table 6 : List of time series features.

As discussed in Section 4.1, I am seeking to build a generic model that will help be proactive to predict the potential impact to health of child based on the information in the MNR location. An ideal model will be highly accurate, minimize the false negatives and provide the prediction quickly with minimal information. I approach the evaluation as both a prediction and classification problem. As a prediction problem I try the model to predict the value of the air particle concentration in the CHD location. As a classification problem I identify a suitable threshold above which the value of air particle concentration will cause potential impact of the child health, so the output from the model is the binary variable who values are Y or N where Y is potential impact to child's health and N where the air particle concentration does not potentially impact the health of child in his/her room. We look at prediction and classification methods to identify the model.

4.4 Preliminary feature engineering analysis – continuous outcomes

In prediction modelling approach, I implemented models that predict the value of the air particle concentration in the CHD based on the air particle concentration in the MNR location. In this section, the approach is to run several standard models and gain a bird's-eye view to the accuracy of the results and the time it takes to run the model. Using the extracted features identified above, I considered all models that are include in Python's *lazy predict* package.[108] I compared models to find the one that that best fits the data set and lends itself to the development of a proactive mechanism to quickly identify if/when a smoking event in the MNR can potentially impact the child's bedroom environment.· It should be noted that each of the models described in this section were run with default parameters and it is likely possible to tweak the parameters to further improve the accuracy of the results.



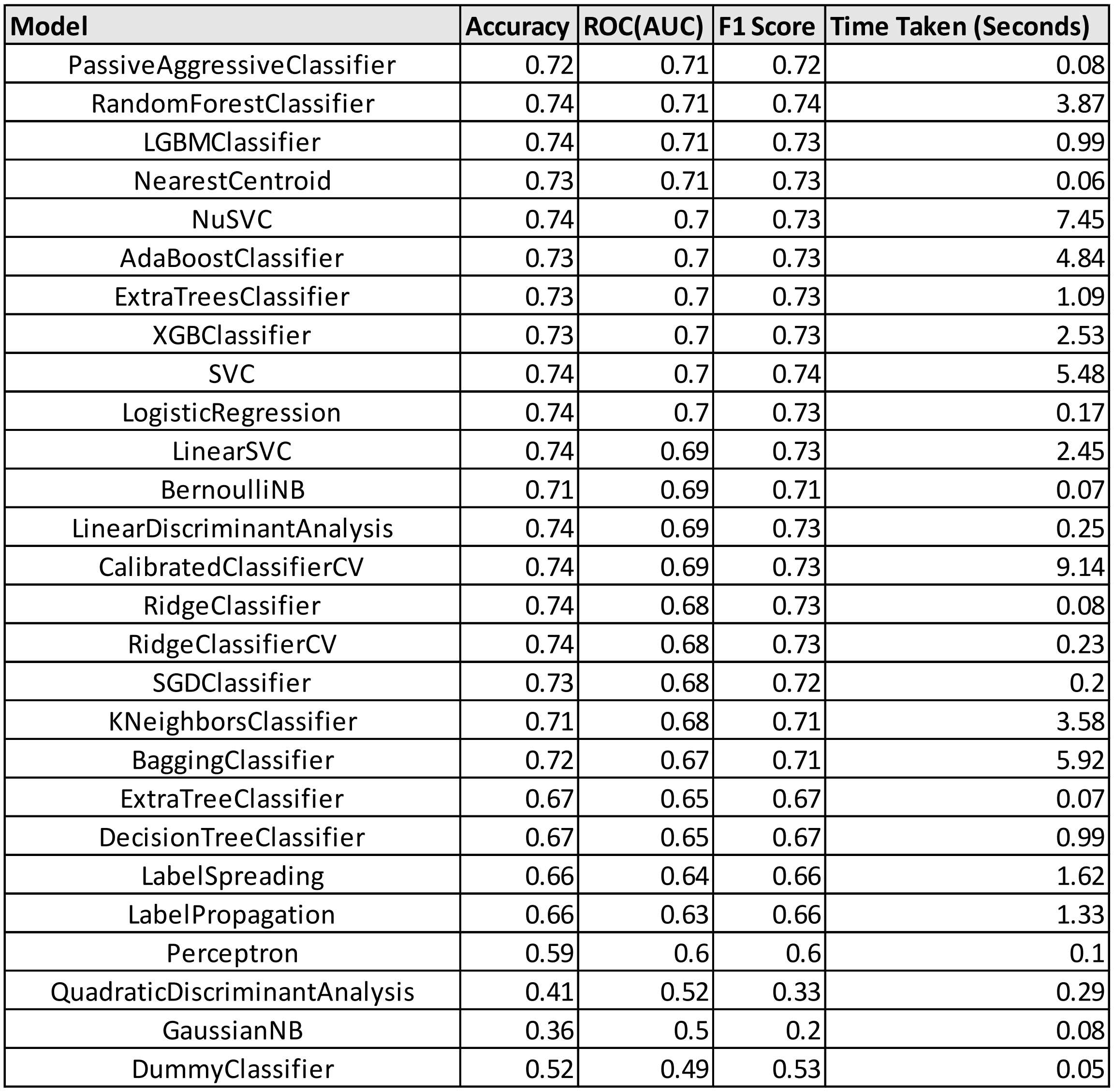
Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken(seconds)
GradientBoostingRegressor	0.3	0.4	7306.23	36.89
RandomForestRegressor	0.29	0.4	7349.26	92.95
LGBMRegressor	0.27	0.38	7431.65	1.42
HistGradientBoostingRegressor	0.27	0.38	7435.78	7.61
RidgeCV	0.25	0.36	7531.26	0.33
Lasso	0.25	0.36	7536.17	0.97
HuberRegressor	0.25	0.36	7536.68	1.33
BayesianRidge	0.25	0.36	7541.2	0.33
Ridge	0.25	0.36	7542.24	0.1
LassoLarsIC	0.25	0.36	7545.11	0.14
OrthogonalMatchingPursuit	0.25	0.36	7545.14	0.09
PassiveAggressiveRegressor	0.24	0.36	7572.71	0.29
LassoLars	0.24	0.36	7588.38	0.1
ExtraTreesRegressor	0.24	0.35	7598.9	27.23
LarsCV	0.24	0.35	7602.43	1.31
BaggingRegressor	0.23	0.35	7615.19	9.43
LassoCV	0.22	0.34	7695.18	1.53
ElasticNet	0.22	0.34	7704.12	1.13
OrthogonalMatchingPursuitCV	0.22	0.34	7705.26	0.33
LassoLarsCV	0.2	0.32	7767.18	0.41
GeneralizedLinearRegressor	0.2	0.32	7767.85	0.15
TweedieRegressor	0.2	0.32	7767.85	0.14
XGBRegressor	0.2	0.32	7778.94	3.93
ElasticNetCV	0.17	0.3	7911.66	1.11
KNeighborsRegressor	0.13	0.26	8108.31	5.57
AdaBoostRegressor	0.12	0.26	8142.55	4.49
MLPRegressor	-0.1	0.07	9119.51	16.96
DummyRegressor	-0.18	0	9450.54	0.08
NuSVR	-0.19	-0.01	9496.56	8.58
SVR	-0.28	-0.08	9829.31	14.22
DecisionTreeRegressor	-0.34	-0.13	10061.13	1.68
GaussianProcessRegressor	-0.56	-0.32	10865.8	14.9
ExtraTreeRegressor	-0.56	-0.32	10876.34	0.35
LinearSVR	-0.7	-0.44	11351.67	0.1
KernelRidge	-1.18	-0.85	12850.75	1.29

Table 7 : Performance statistics of regression models implemented using lazy predict package. R- Squared – statistical measure that determines the proportion of variance in the dependent variable that can be explained by the independent variable. Adjusted R-Squared – modified version of R-squared, which adjusts for predictors that are not significant a regression model. RMSE – Root Mean Square Error. Time taken (seconds) – Time taken to run the model in seconds.

Table 7 illustrates the results of this modeling effort and is sorted by RMSE (Root Mean Square Error), but time taken to run the model is also considered in the evaluation of the model, since I am seeking a model that is accurate and provide the results in the shortest amount of time. It is clear that Gradient Boosting performs the best in terms of fitting a model to the existing data. Depending on the deployment strategy I could also chose the LGBM Regressor (Light Gradient Boosting Machine Regressor) that provides a slight lower accuracy but runs 30x faster than the Gradient Boosting Regressor model. The overall values of r^2 are low indicating that the model is not good at explaining the variability in the existing dataset.

4.5 Preliminary feature engineering analysis – binary outcomes

As opposed to the previous model where I attempted to predict the resulting CHD concentration based on an MNR peak, in the classification modelling approach, I look at a problem from a standpoint of the potential impact to the child's health as a binary yes-no outcome. Based on expertise from thesis involved in the PFA study, I defined a threshold for potential impact to children's health of any air particle level in the CHD location over 15,000 during the entire duration of an MNR peak. As long as the classification model can predict a value accurately if the air particle concentration is above or below this threshold, I have an effective model. I am not interested if the model can predict the accurate value of air particle concentration in the CHD location. This is an effective approach as it gives a flexibility of reduced accuracy in comparison to the performance of the model or minimizing the false negatives. Like the prediction models I run a set of classification models from Python's *lazy predict* package. The results (Table 8) are sorted in the order of model accuracy and the time take to run the model. I also ran the LSTM model (details provided in the next section) that provides an accuracy of 81%.



Model	Accuracy	ROC(AUC)	F1 Score	Time Taken (Seconds)
PassiveAggressiveClassifier	0.72	0.71	0.72	0.08
RandomForestClassifier	0.74	0.71	0.74	3.87
LGBMClassifier	0.74	0.71	0.73	0.99
NearestCentroid	0.73	0.71	0.73	0.06
NuSVC	0.74	0.7	0.73	7.45
AdaBoostClassifier	0.73	0.7	0.73	4.84
ExtraTreesClassifier	0.73	0.7	0.73	1.09
XGBClassifier	0.73	0.7	0.73	2.53
SVC	0.74	0.7	0.74	5.48
LogisticRegression	0.74	0.7	0.73	0.17
LinearSVC	0.74	0.69	0.73	2.45
BernoulliNB	0.71	0.69	0.71	0.07
LinearDiscriminantAnalysis	0.74	0.69	0.73	0.25
CalibratedClassifierCV	0.74	0.69	0.73	9.14
RidgeClassifier	0.74	0.68	0.73	0.08
RidgeClassifierCV	0.74	0.68	0.73	0.23
SGDClassifier	0.73	0.68	0.72	0.2
KNeighborsClassifier	0.71	0.68	0.71	3.58
BaggingClassifier	0.72	0.67	0.71	5.92
ExtraTreeClassifier	0.67	0.65	0.67	0.07
DecisionTreeClassifier	0.67	0.65	0.67	0.99
LabelSpreading	0.66	0.64	0.66	1.62
LabelPropagation	0.66	0.63	0.66	1.33
Perceptron	0.59	0.6	0.6	0.1
QuadraticDiscriminantAnalysis	0.41	0.52	0.33	0.29
GaussianNB	0.36	0.5	0.2	0.08
DummyClassifier	0.52	0.49	0.53	0.05

Table 8 : Performance statistics of classification models implemented using lazy predict package. Accuracy – The accuracy of model. ROC (AUC) – The area under curve represents, how much the model is capable of distinguishing between the classes. F1 Score – Represents the balance between precision and recall of a model. Time taken (seconds) – Time taken to run the model in seconds.

Based on the initial results of the prediction and classification models, I find that the best prediction model Gradient Boosting has an r^2 of 7,306 and take 36 seconds to run for the data set. The best classification model Passive Aggressive Classifier has an accuracy of 0.72. The LSTM model provides and accuracy of 0.81. I am interested in further exploring the LSTM model as it provides better accuracy than the other classification models and try to evaluate additional scenarios and understand the performance of the model.

4.6 LSTM (Long Short Term Memory) model

The prediction and classification learning approaches (Section 4.3) provides encouraging results, so I now turn our attention to the use of machine learning for time series data[109]. Supervised models, where the models require training data, have parameters that are internal to the learning and are estimated from the data during training as the model used tries to learn the mapping between the input features and the labels or targets. The process often involves a tuning process where you start with random parameters values for the model and fine tune as the model trains to fit the outcome in the training data set. I use *Keras Python* package[115] for designing and tuning the LSTM model. These packages have a structured approach to initiating with default values and tuning to fit the model. We just pass the minimum parameters (e.g., epoch i.e., the number of interactions that we want the model to run to identify the optimal parameters and batch size i.e., the number of samples that will be propagated through the model network).

There are data sets (e.g., time series) where the predicted next value in the series is dependent on the prior values. The type of machine learning models that learn from prior data points are called RNN (Recurrent Neural Networks). The RNN are very effective when the model needs to learn from few prior values, but as predictions go further into the future, the RNN model run into issues. During the training of RNN, the information repeatedly loops which results in very large updates to neural network model weights. This is due to the accumulation of error gradients during an update and hence, results in an unstable network. At an extreme, the values of weights can become so large as to overflow and result in NaN values. LSTM models excel at extracting patterns in input feature space where the input data spans over long sequences. The gated architecture has the ability to manipulate its memory state, making them ideal for such problems (Figure 14). LSTMs can almost seamlessly model problems with multiple input variables. This adds a great benefit in time series forecasting, where classical linear methods can be difficult to adapt to multivariate or multiple input forecasting problems.

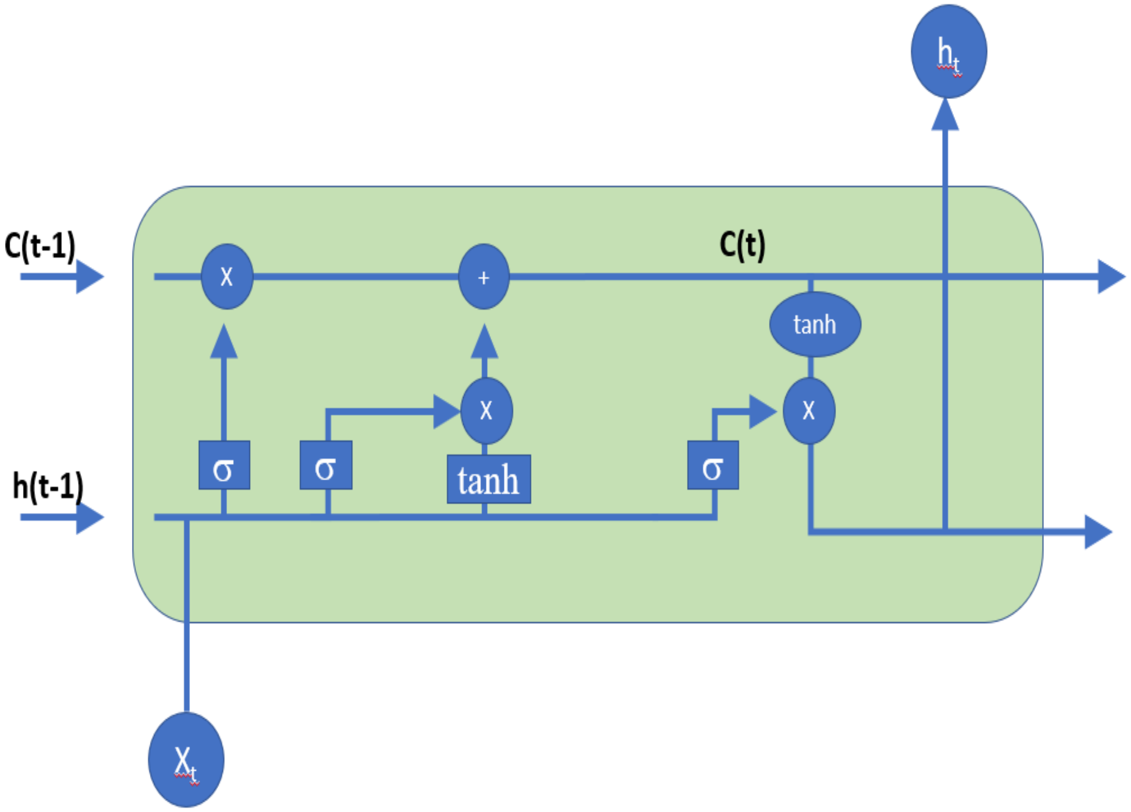


Figure 14 : Schematic of LSTM architecture implemented in this trial. X_t represents the current input, σ indicates the sigmoid layer, X is the scaling of information, tanh is the tanh layer, h(t-1) is the output from last LSTM unit.

4.7 Aggregate model

I first seek to establish a generalized approach that would be able to be deployed in any home and be able to proactively notify if the child's health's would be potentially impacted due the presence of smoke air particles in MNR. An aggregate model is a generic model that combines the different characteristic of homes and predict the potential impact to the health of the child. These models are challenging as there are several variables that could impact the performance of the model such as distance between MNR and CHD, the number of rooms, square footage of the house etc. In the below aggregate model, I pass the input as the air particle concentration data for the main room in all the homes and the output is a classification based on some particle count threshold.

For all homes, I extract the peaks (Section 2.4) and fed the raw data from the main room into the LSTM model. I assumed that an air particle concentration of over 15,000 counts causes impact to child’s health (Section 3.8), but also explored running the model at lower thresholds, from 7,500 to 15,000 counts in increments of 2500. An intervention that is triggered when the air particle concentration is below the recommended threshold would be an overly cautious approach and overprotective of child’s health. I decided to examine the effects of lowering the threshold to be more proactive understanding the potential impact of soke air particle on the health of the child. The only downside of this approach is that the frequency of false alarms may increase at lower threshold, and we may have a false positive impact where the notifications may be ignored.

I am interested in making a determination of the impact to the CHD area as quickly as possible so, for each MNR peak, I considered windows of input data ranging from the first one minute of data to up to 60 minutes, in 5-minute increments. This was done to determine if the accuracy of the model decreased or there is an increase in false negatives as I train the model on less data. For each of these time windows, a separate model was fit, each of which received all data for each peak across all the homes as its input. This data was trained (70% training data, 30% test data) and then used to predict the value of particle count in the child's room. Based on the threshold selected, I convert this to a classification approach where an air particle concentration value is above the threshold. In the scenario where we have existing peak data less than the minimum (e.g., the peak has only 40 minutes of data when we are considering 60 minutes of data to be feed into the model), I pass an air particle concentration of zero for the missing minutes.

The initial LSTM model was designed with 32 units and a dropout rate of 0.2. The accuracy, false negative rate and the distribution of the CHD potential impact were calculated, which are shown in Figure 15. The first plot is the heat map of the accuracy of the model based on the minutes of data fed into the model and the threshold values in the child room. The second plot is the false negative values, which is when the CHD environment is actually impacted, and the model predicts that it is not. It is desirable to have these values to be as close to zero as possible to minimize the situations where the model is not able to detect air particle concentration greater than the set threshold.

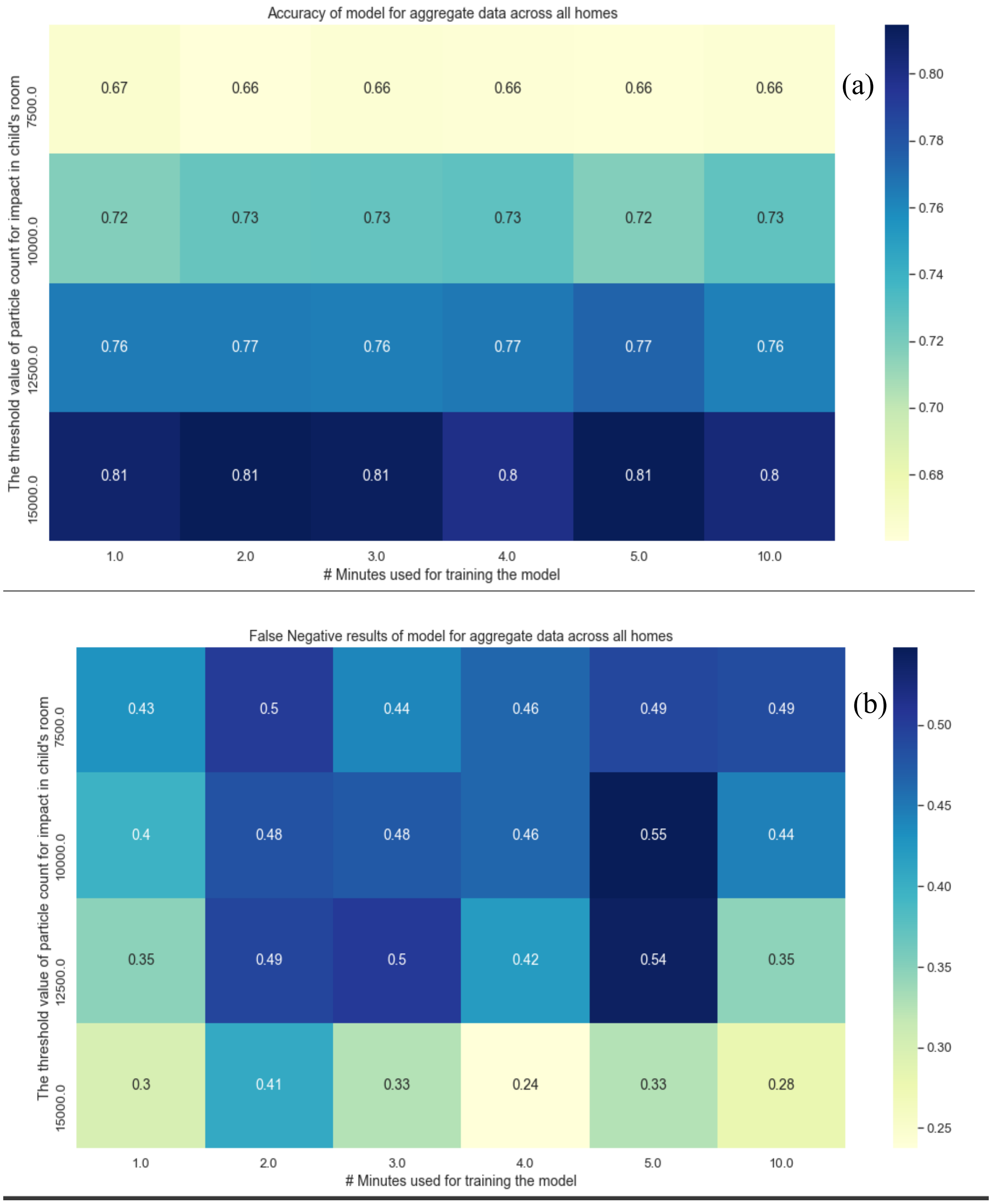


Figure 15 : Summary of model metrics for aggregate data across all the homes. x axis indicates the amount of data as an input to the model and the y axis indicates the thresholds above which the value of air particle concentration indicates impact to the child's health; Color represents the accuracy in the first plot (a) and the percentage of false negatives in the second plot (b). Modeling results are provided for 24 threshold x training minutes parameter combinations represented by the cross product {7,500, 10000, 12500,1500} x {1,2,3,4,5,10}.

Overall, the model performs better at higher thresholds values. This is likely because the distribution of the instances where the child health is potentially impacted (21%) is unbalanced. The model best performs with parameters of 4 minutes of MNR peaks and 15,000 thresholds (80% accuracy and 24% of false negative rate).

The other aspect to note is the accuracy of model does not change much as I feed more data into the model (81% when I have 1 minute of data to 80% when we have 10 minutes of data at 15,000 threshold). I hypothesize that this is due the fact that all data points are within the peak and based on the peak start/end time, the timing and concentration at the beginning of the peak is directly corelated with the peak maximum value, so in practice we are using a transformed value of the peak maximum value to predict the value of particle count in the CHD location.

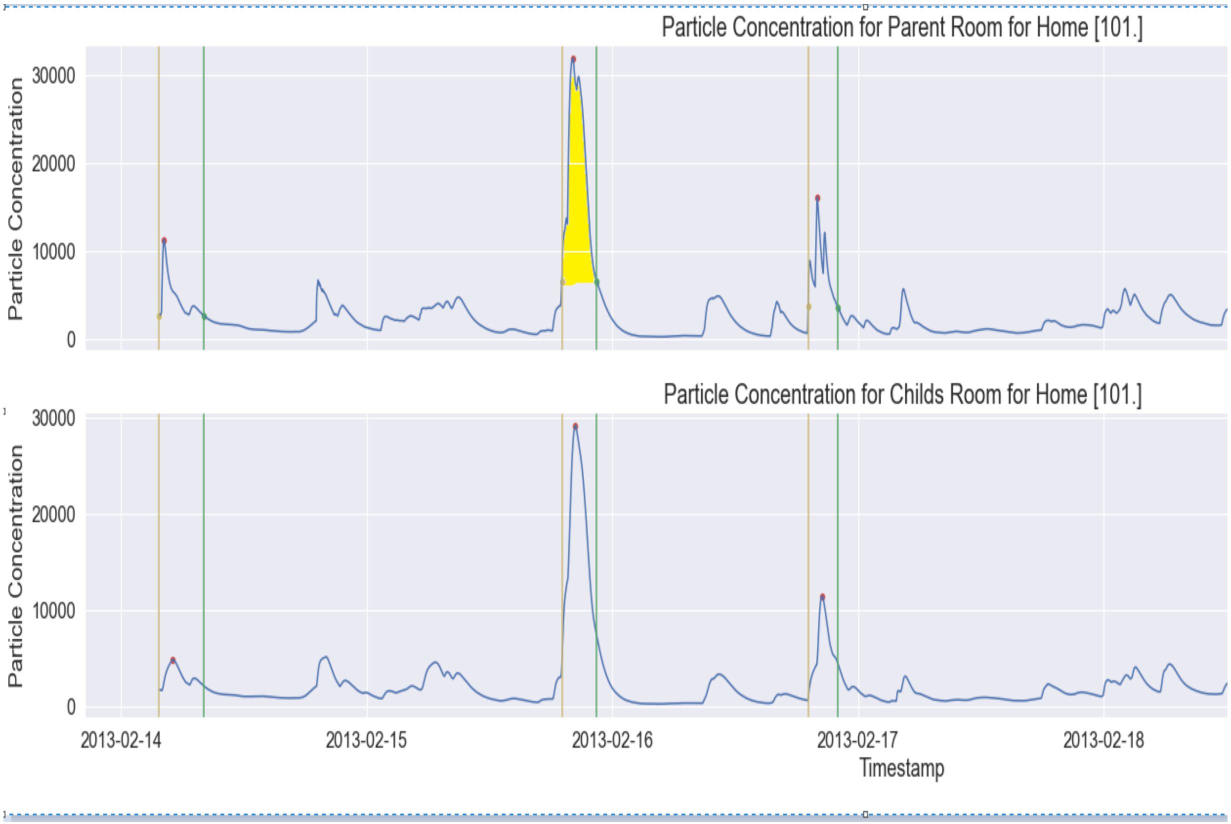


Figure 16 : Particle concentration in the main and child's room, x axis indicates the timestamp of collection of air particle concentration and the y axis indicates the particle concentration, the yellow area in top plot are all the point for particular peak that starts from the golden line and ends with the green line, the bottom plot are the corresponding values of air particle concentration in the child's room.

This point is further illustrated in Figure 16, which shows particle concentrations in the main and child room for a home. The way to identify the start and end of the peak is based on the relative height and prominence (0.8) based on the peak value (Section 2.4). For an area under peak (highlighted in yellow) each point is in the proximity of the maximum value. As I pass more data into the model, I am in essence adding more data that is similar to the maximum value, so I do not see a substantial difference in the accuracy of the model.

4.8 Impact of distance between main and child room

One of the key aspects of the experiment is that air particle concentration needs to travel from the MNR to CHD location. The earlier model was an aggregate for peaks in all the homes. I want the model to be effective across all the homes. It would add validity to the approach to understand if there is an impact to the accuracy of model as distance between MNR and CHD increases. The approach to evaluate the impact of distance between rooms to the effectiveness of the model is by comparing the results of an aggregate model one of the homes where the distance between rooms is less than that of the median of the sample and the other where the distance is greater than the median. I selected median over the mean because there are outliers in the data which I did not want to impact the results. Plus, the median splits the sample into two groups of equal size.

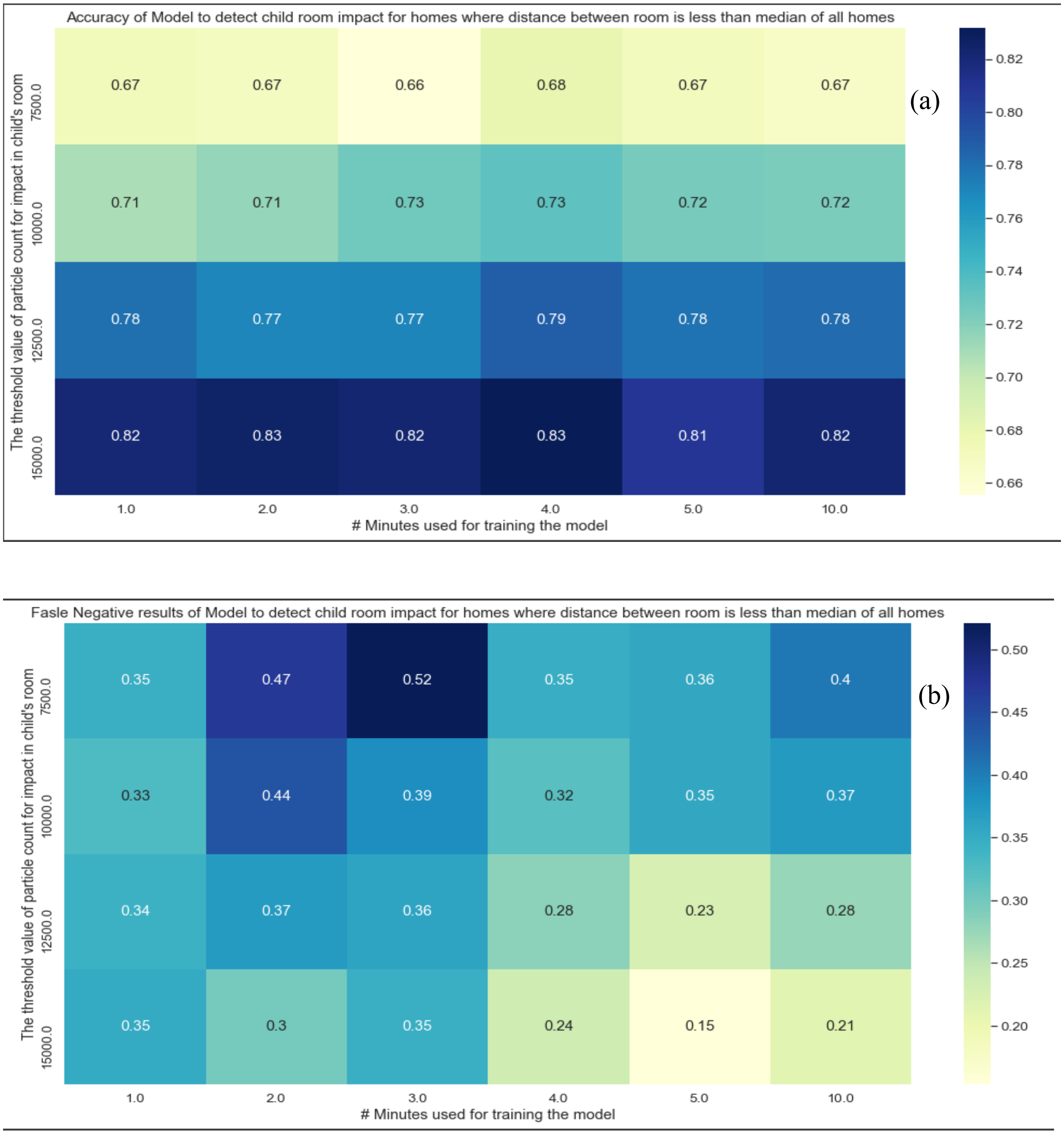


Figure 17 : Summary of model metrics for aggregate data across all the homes where distance between main and child room is less than the median distance for all homes. x axis indicates the amount of data as an input to the model, y axis indicates the thresholds above which the value of air particle concentration indicates impact to the child's health; Color represents the accuracy in the first plot (a) and the percentage of false negatives in the second plot (b). Modeling results are provided for 24 threshold x training minutes parameter combinations represented by the cross product {7,500, 10000, 12500,1500} x {1,2,3,4,5,10}.

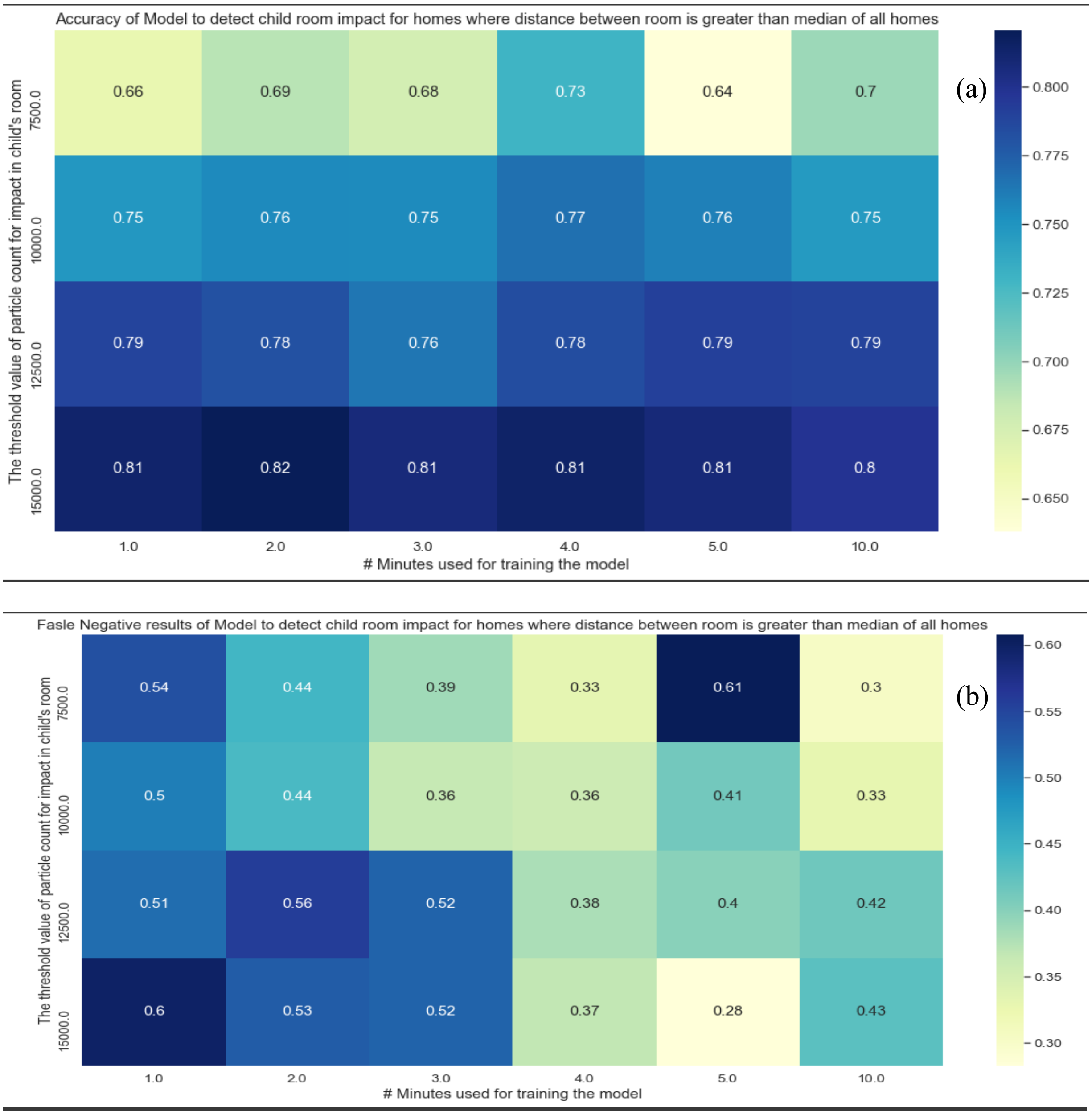


Figure 18 : Summary of model metrics for aggregate data across all the homes where distance between main and child room is greater than the median distance for all homes. x axis indicates the amount of data as an input to the model and the y axis indicates the thresholds above which the value of air particle concentration indicates impact to the child's health; Color represents the accuracy in the first plot (a) and the percentage of false negatives in the second plot (b). Modeling results are provided for 24 threshold x training minutes parameter combinations represented by the cross product {7,500, 10000, 12500,1500} x {1,2,3,4,5,10}.

Figures 17 and 18 represents the accuracy, false negative rate, and the distribution for the models where distance between MNR and CHD is less than the median distance and greater than median distance for all the homes respectively. Interestingly, even when the distance between the rooms is larger, there is not much change in the accuracy (~ 80%) of the model. I would have expected for the model be less accurate as the distance between the room is increased. While the model accuracy does not change, the false negative rates increase (e.g., from 0.13 to 0.38 for 1 minute of training data at 15,000 threshold). This is a consistent pattern across as I increase the amount of data that we use for training the model and is an important finding as it reaffirms the approach of looking at false negatives in addition to accuracy when identifying the right model.

4.9 Single Home Model

I transition from aggregate to single home models, i.e., I build a model for each home rather than the aggregate data for all the homes. While this is more effort, I take this approach as I want to determine if this approach is more effective for each home. Within this approach, it is necessary to consider the number of peaks needed to train the model on for optimal results, which will allow future implementations of this model to gauge how much training data is required to be collected before an accurate early warning system can be activated. In addition, addiction levels are stronger for people who smoke more and/or live with other smokers[116]. Therefore, the dynamics of smoking and children's exposure to SHS may differ according to the frequency of smoking. To investigate this possibility, I examined 25 homes where there were at least 70 peaks. To investigate this characteristic, I trained the model on 10,2,0,20,40 and 50 peaks and examined the accuracy in each case to determine if the number of training peaks impacts the performance of the model. As I train on model with a greater number of peaks, there is less data available for testing and that impacts on how I view the accuracy of the model. For a home with 70 peaks, so I train the model on 10 peaks, I have 60 peaks to test the model against versus if I train on 50 peaks I have only 20 peaks to test the model against. The accuracy numbers (Figure 19 and Figure 20) from a model tested on 60 peaks (accuracy = 72.32%) versus the model tested on 20 peaks (accuracy = 69.59%) will influence on the effective model that I would plan to deploy. I look at more peaks to test the model is as we get further away from the training peaks, I want to evaluate if the accuracy is affected by the predictions made in the future. The total number of peaks in a home limit how far in the future we can test the model. In addition, as the model predicts on a greater number of peaks the time between the training and test peaks is longer and longer and the influence of training data diminishes. A model trained on air particle concentration between 7am to 7.30am will have more influence on prediction at 8.00 am versus 7.00 pm i.e., 12 hours later than the time the training data was captured.

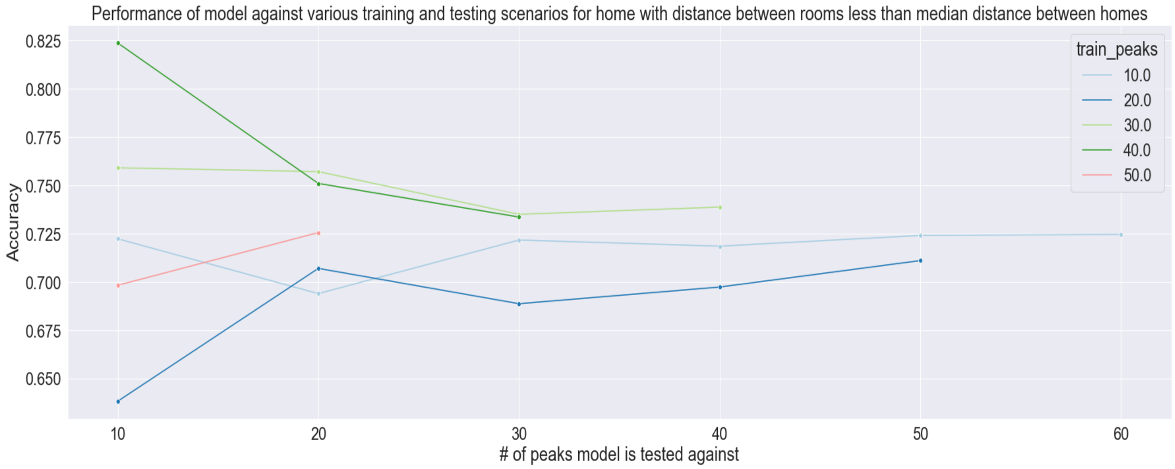


Figure 19 : Accuracy of model where distance between rooms is less than median distance across all homes that have at least 70 peaks, x axis indicates the # of peaks that are we predict the impact in the child room and the y axis indicates the accuracy of the results, each individual lines in the plot indicates the number of peaks that are used to train the model.

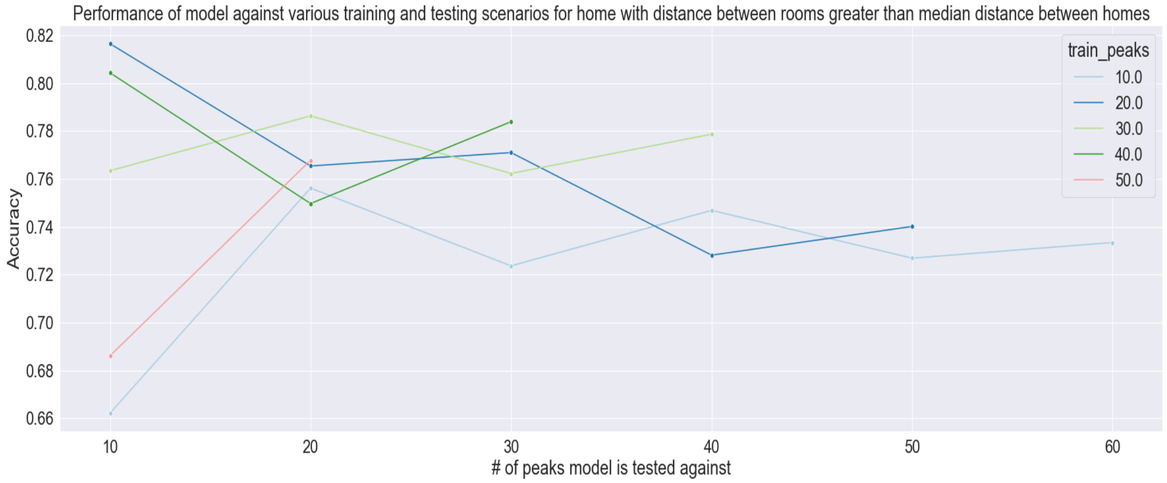


Figure 20 : Accuracy of model where distance between rooms is greater than median distance across all homes that have at least 70 peaks, x axis indicates the # of peaks that are we predict the impact in the child room and the y axis indicates the accuracy of the results, each individual lines in the plot indicates the number of peaks that are used to train the model.

The results for each home are aggregated and presented as one measure per peaks trained on/peaks predicted combination. The model performs as expected with the accuracy improving as I train model with more peaks. The interesting aspect is that even with training on just 10 peaks the model has a high accuracy near 75%. The model with training of 40 peaks improves in accuracy as I test with more peaks. These model behaviors with help us understand how these models can be deployed in various homes. The results indicate that it is possible to predict the impact on the CHD location after seeing only 10 peaks, with relatively small effects on the accuracy.

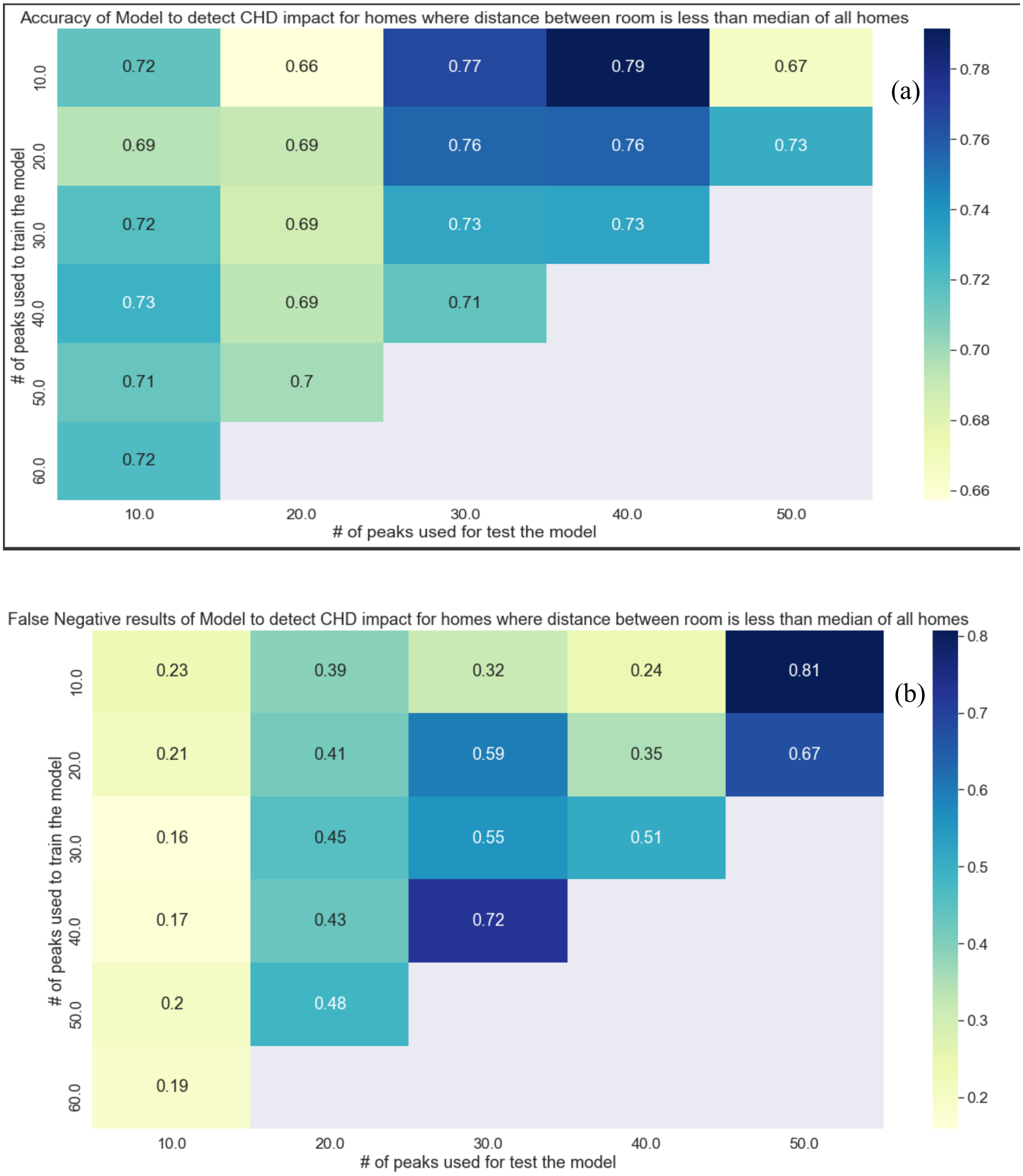


Figure 21 : Model (single home) accuracy (a) and false negative (b) for homes where distance between rooms is less than median distance for all homes.

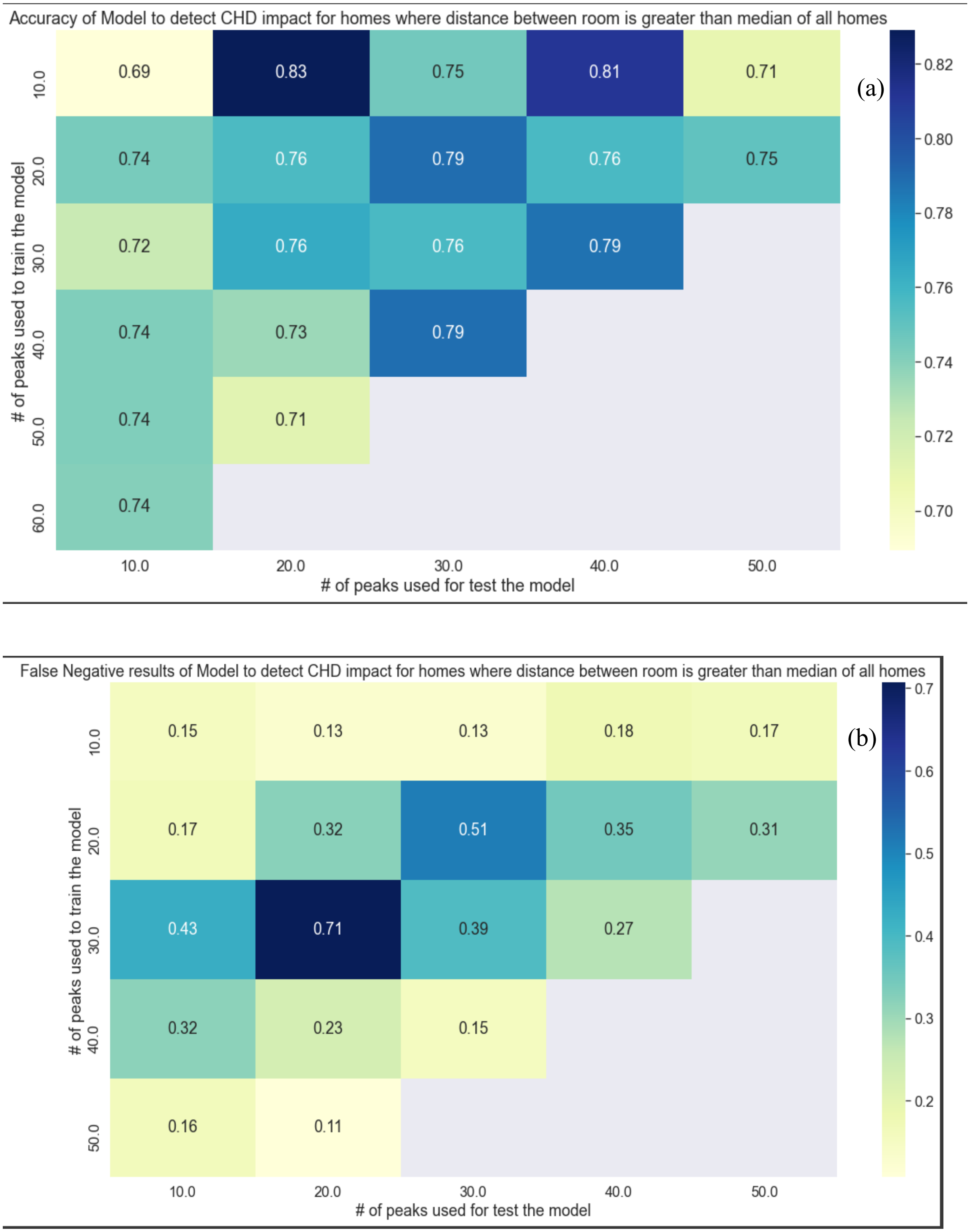


Figure 22 : Model (single home) accuracy (a) and false negative (b) for homes where distance between rooms is greater than median distance for all homes.

Next, the accuracy of the model and false negative rate where the distance between rooms is greater and less than median distance across all homes was investigated. (Figure 21 and Figure 22). The accuracy of the model is overall less for homes where the distance between the MNR and CHD locations is greater than the median distance (66% for model trained on 10 peaks, depending on the number of training peaks) in comparison to homes where the distance is less than the median distance (71%). This is expected as the distance between MNR and CHD location increases there is less smoke particle that reach the CHD location and there could be introduction of noise (other air pollutants) that could impact the accuracy of the model. The other aspect is that is notably visible is the downward trend of the accuracy of the model for model that is trained on 40 peaks. This phenomenon should be investigated before deploying these models in homes where there is a large distance between rooms. Since this model may struggle to predict future smoke events.

4.10 Transformer Model

Over the last decade, there is neural network has gained popularity in the field of Natural language processing (NLP) and computer vision. One of the limitations of the neural nets was the inability to memorize things. This is important in the field of sequence to sequence applications like machine translation, NLP etc. This was overcome by recurrent neural networks (RNN). RNN have limitations like vanishing gradients, exploding gradients, handling long-term dependencies, etc. These limitations have been addressed by LSTM models. Transformers are neural nets that use attention layer are the primary building block. They focus only on the required features instead of focusing on all features.

Transformers with Self-Attention mechanism were introduced in 2017 by a team at Google with Vaswani et al., in a paper entitled Attention is All You Need[116].

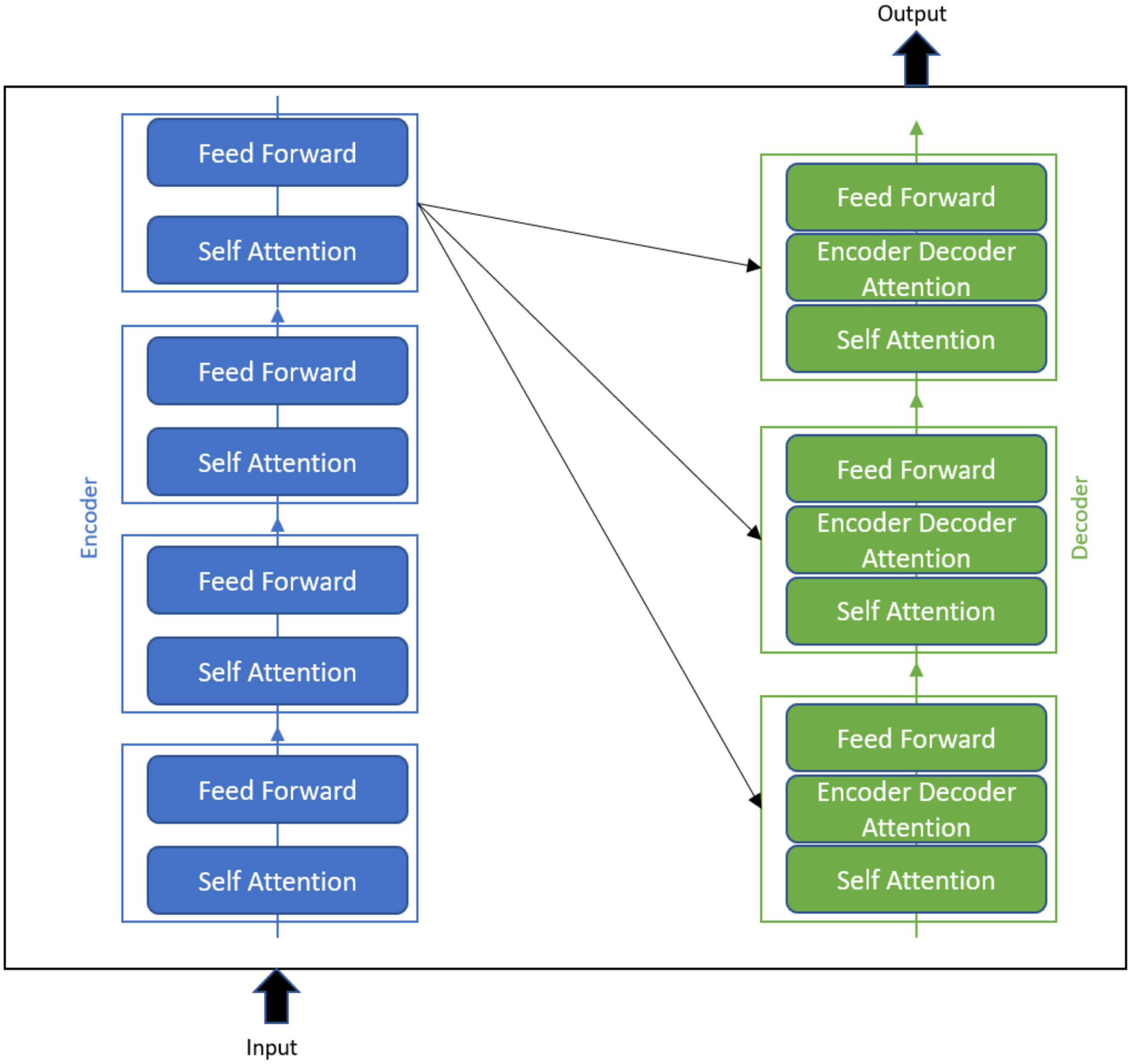


Figure 23 : Structure of a transformer model.

The basis structure of the model is depicted in Figure 23. The architecture uses an encoder, decoder mechanism. Each encoder (blue box) consists of a self-attention and feed forward component. The self-attention component focuses on storing the context in addition to the important features. The feed forward component is similar to the cell of neural network. Each decoder (green box) represents that self-attention, encoder decoder attention and the feed forward component. The self-attention and the feed forward have the same functionality as in the encoder layer. The encode decoder attention will compute the attention between encoder and decoder and tell us how important each encoder vector component is in predicting the next output. One of the major differences between how the traditional neural nets and transformer models are designed, is the way in which information is passed between the layers. In a tradition CNN, the information is passed between each layer, in a transformer model each layer is connect to every other layer creating a global representation of the first layer.

I ran the transformer model on the data set for both the aggregate and single home models. I used the same approach as I used to evaluate the LSTM models. The results (Figure 24 and 25) indicate that the models do not perform well on the existing dataset in comparison to the LSTM approach. The best accuracy is 59% with high false negative rate 0.78 for an aggregate model. In addition, for a single home model we best accuracy of 54% with a false negative rate of 0.27. There could certainly be opportunities to fine tune these models or may perform better on a different data set. For the current study LSTM better performs and would be a recommend approach.

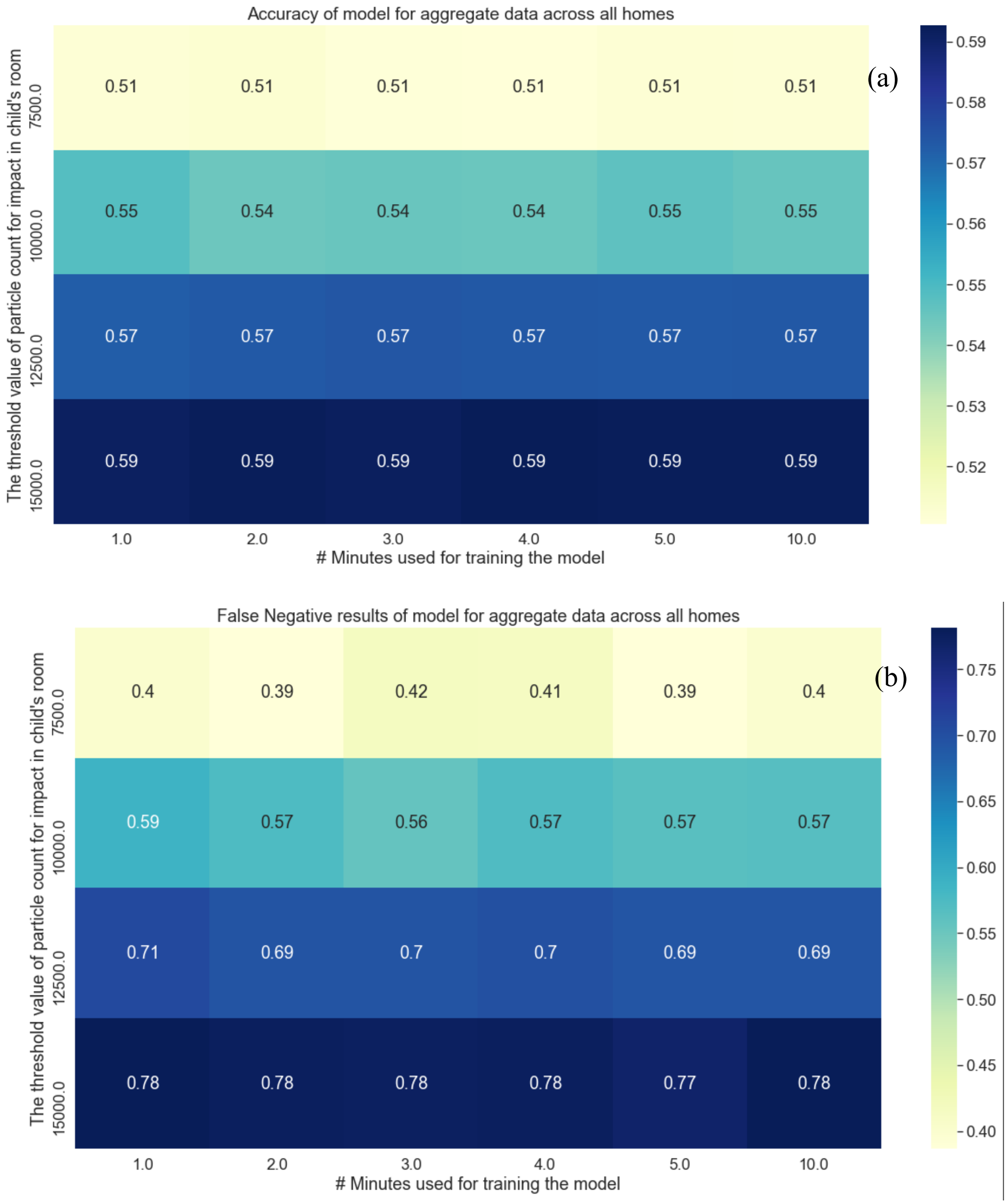


Figure 24 : Summary of model metrics for aggregate data across all the homes. x axis indicates the amount of data as an input to the model and the y axis indicates the thresholds above which the value of air particle concentration indicates impact to the child's health; Color represents the accuracy in the first plot (a) and the percentage of false negatives in the second plot (b).

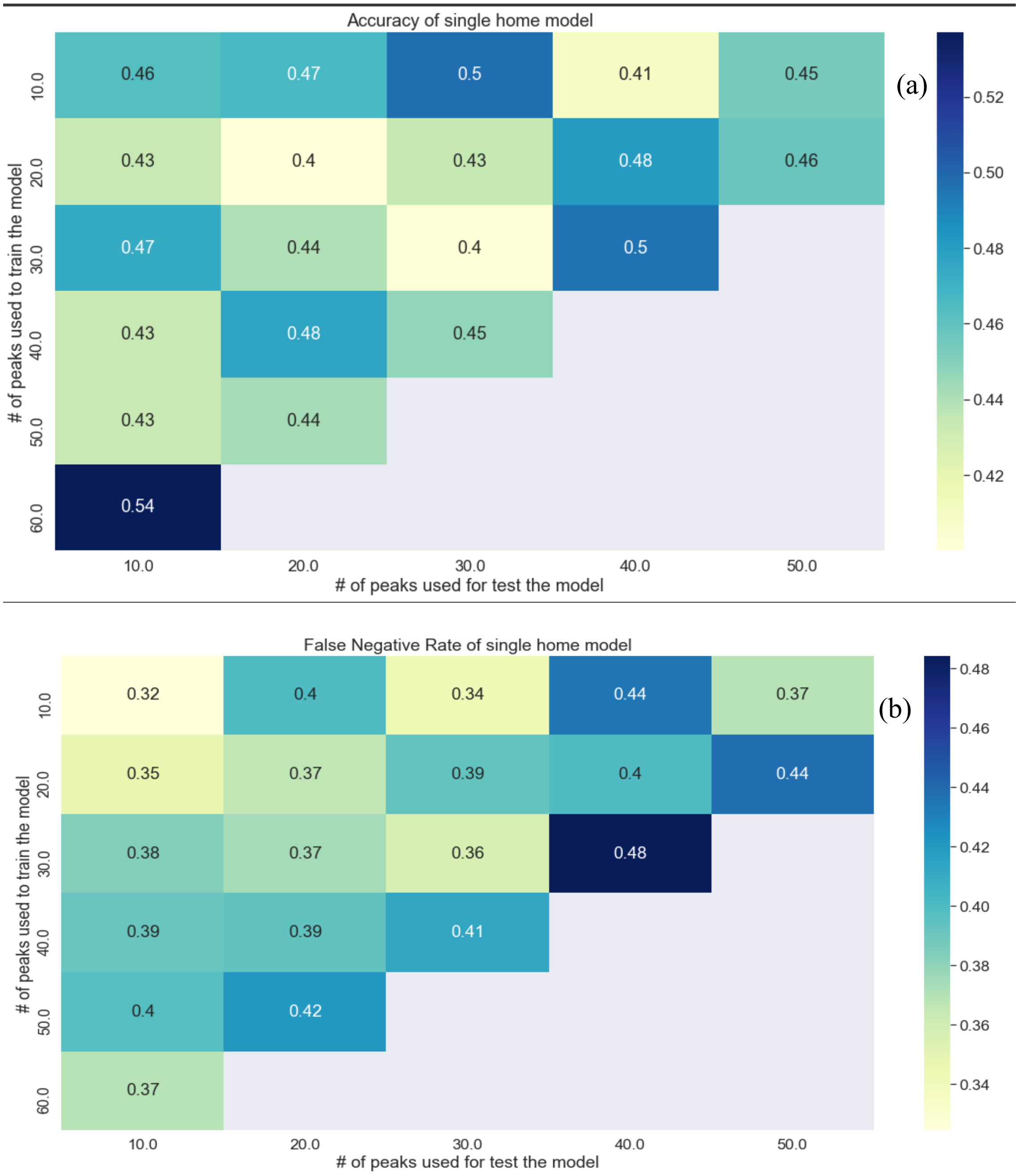


Figure 25 : Summary of single home model metrics for aggregate data across all the homes. x axis indicates the amount of data as an input to the model and the y axis indicates the thresholds above which the value of air particle concentration indicates impact to the child's health; Color represents the accuracy in the first plot (a) and the percentage of false negatives in the second plot (b).

4.11 Summary

In this chapter I evaluated various approaches to identify a model that will address the study goal of quickly and proactively predicting the potential impact of smoke air particle concentrations in the main room on a child's sleeping environment. There are three considerations that we accounted for as we built the model: i.) reducing false negatives, ii.) ensuring the model provides accurate overall results, and iii.) whether to approach the modeling as a prediction or classification problem. I extracted features from the raw data and evaluated various models that provide continuous outcomes. Two models stood out in this analysis. It is clear that Gradient Boosting performs the best in terms of fitting a model to the existing data. It has RMSE 7306 and time taken is 36.89 seconds. Depending on the deployment strategy we could also chose the LGBM Regressor (Light Gradient Boosting Machine Regressor) that provides a lower accuracy but runs 30x faster than the Gradient Boosting Regressor model. I also evaluated models that provide a binary outcome. The best classification model Passive Aggressive Classifier has an accuracy of 0.72. I ran LSTM as an aggregate model on all the homes, which provided best result (80% accuracy and 24% of false negative rate) with parameters of 4 minutes of MNR peak data and a 15,000 threshold value for air particle concentration. This is an important finding as it will allow smokers to know relatively quickly the potential impact smoke air particles to the child health.

I ran the LSTM model on all the home data (aggregate model) and model on each home and aggregated the results. I took this approach as we want to ensure that the model is effective for each home. I found the model to be effective with accuracy above 74% and false negative of less than 0.24. The impact of distance between the MNR and CHD location on the performance of the model was also evaluated to assess if having a CHD location further away from the MNR location can reduce the potential impact of secondhand smoking to the health of the child. Interestingly, even when the distance between the rooms is large, there is not much change in the accuracy (~ 80%) of the aggregate model (Figure 17 and Figure 18). However, in the single home models (Figure 21 and Figure 22) I did see model accuracy is around 72% for homes where distance is less than median distance and 74% when distance between rooms is greater than median distance. While the model accuracy did not change for the aggregate model, the false negative rates increase (e.g., from 0.13 to 0.38 for 1 minute of training data at 15,000 threshold) For single home model the false negative is 0.19 for homes where distance is less than median distance and 0.2 when distance between rooms is greater than median distance.

5 Summary

Smoking combustible tobacco products and exposure to second-hand smoking are known to impact the health of adults and children. Children are particularly at risk due to their biological characteristics. This is a pernicious problem in homes where adults smoke during evening or at night when the child is sleeping, since adults often fail to recognize the potential impact to children's health. Therefore, it would be beneficial to identify an approach that would help to proactively mitigate the impact to children's health. The study focused on understanding and quantifying the relationship between smoking occurring within a home and subsequent impact on the children's bedrooms. This work is foundational and will help us lay the groundwork for future studies that characterize in-home microenvironments so that caregivers have actionable information by which to protect children's health.

The data used for the dissertation was generated by Project Fresh Air (PFA), a multiple baseline/randomized control trials aimed at reducing SHS in the households of smokers from a low-socioeconomic status (SES) population. The data is time series and I had to clean it by removing outliers and preparing the data for analysis. One of the key tasks was to identify peaks within the data i.e., the range of air particle concentration which identifies a smoking event. I smoothed the time series data and identified peaks by implementing an algorithm that uses a combination of threshold value, a horizontal distance between peaks and prominence. I identified peaks in the MNR and using this information as a reference we extracted the corresponding peaks in the CHD location. These are computationally expensive operations, and I leveraged the vectorization techniques and build in python packages (e.g., SciPy) to improve the performance of the analysis. These techniques reduce to the computation time by approximately a factor of 10.

To create an effective solution to reduce the potential impact of SHS on health of the child I need to focus to understand if second-hand smoke diffuses to the child's room, how quickly this happens, is there a loss of intensity and does intervention change these relationships? I performed Granger causality test to determine whether the particle concentration in the MNR influences the CHD particle concentration data. It tests the ability of air particle concentration in the MNR to predict the air particle concentration in the CHD. I found that 92.4% of peaks passed the granger causality test. A linear model with beta coefficient of 0.684 indicates that the maximum particle level from a CHD peak is 68.4% that of an MNR peak. I ran a hierarchical linear model to understand and quantify the impact of intervention event. There was a positive coefficient for main room particle count, meaning as higher particle concentration in MNR results in higher particle concentration in CHD. There was a negative coefficient for the effect of switching from the baseline to treatment period, indicating that there is a drop in the air particle concentration in the CHD associated with the onset of the intervention. The three-way interaction results indicate a negative slope and a 14% decrease in the association between MNR and CHD monitors of an experimental home after an intervention event.

An effective solution would be to quickly identify the air particle concentration in the CHD based on the data in the MNR then an intervention can be proactively triggered, which would mitigate the potential impact to child health An approach that looks at aggregate data for all the homes would provide a flexible model that can be deployed at any home. I examined several continuous and classification model approaches and examined the accuracy of the model, the time it takes to run and false negative ratio in identifying the best model for the study. An aggregate LSTM model provided best result (80% accuracy and 24% of false negative rate) with parameters of 4 minutes of MNR peak data and a 15,000 threshold value for air particle concentration.

I also evaluated if the distance between the MNR and CHD location had any impact on the performance of the model. I compared homes that have distance less than the median distance of the complete data set to the homes that have distance greater than the median distance. Interestingly, even when the distance between the rooms increases, there is not much change in the accuracy of the model. We would have expected for the model be less accurate as the distance between the room is increased. While the model accuracy does not change, the false negative rates increase (e.g., from 0.13 to 0.38 for 1 minute of training data at 15,000 threshold). We also try to understand the model performance by running on individual homes and aggregating the results. To optimally train and test the model I identified homes with at least 40 peaks. The accuracy of the model is overall less (68.92%) for model trained on 10 peaks, in comparison to the rooms that are closer to each other (71.57%). This is expected as the distance between MNR and CHD location increases there is less smoke particle that reach the CHD and there could be introduction of noise (other air pollutants) that could impact the accuracy of the model.

There are limitations to this work that are worth mentioning as we explore to deploy these techniques. The data was collected from low-socioeconomic status (SES) population of 298 homes. The device that was used to capture the air particle concentration data is subject to disruptive behavior by participants (e.g., covering the device with a hat to prevent it for detecting the accurate amount of air particle concentration) The peaks in the CHD are identified by finding the maximum value of air particle concentration in the CHD for the corresponding peaks in the

MNR. This approach limits capturing peaks where the maximum value of air particle concentration in the CHD location is outside the start and end time of peak in the main room. The models we used assume a linear relationship between the input and output variables and are prone to underfitting. In addition, they are sensitive to outliers. Other sources setting off the monitor, smoking not originating in the MNR - could start smoking in CHD location or some other area, other smokers in the home besides the participant may affect results.

The two key outcomes from the study are 1) I was able to quantify the impact of intervention on the flow of air particle concentration between the MNR and CHD location and 2) I was able to develop a modelling approach that can proactively identify the potential impact of SHS to health of the child. The study open doors for several possibilities. For instance, this information can be used by practitioners in counselling session to provide metrics to smoking adults and advice on the potential impact of smoking to the health of the child.

Several opportunities exist for future work. There are devices that can understand the air particle chemical compostion[114] ,which can be leveraged to segregate the smoke air particles from non-smoke air particles (e.g. burning candles, cooking smoke, incense sticks etc.) to further fine tune the models The computational domain in changing rapidly, we have new modern techniques and modelling approaches developed at a rapid scale, we can explore the latest techniques to future improve the accuracy and scalability of the model presented in the paper. The model can also be integrated with existing smart home monitoring systems like carbon monoxide monitoring, Internet of Things to build a capability to real time notification system. Lastly, there is a need to understand and quantify the impact of third hand smoking (e.g., where smoking contamination is present in the environment long after the smoker has left and potentially places household occupants, including children, at risk.)

The study is a small contribution to existing efforts to understand and reduce the potential impact SHS on children's health and broadly discourage the use of harmful smoking products in the society. I hope that practitioners working in this area   can leverage the findings presented in this paper to inform their tobacco control methodologies.

References

1 Gilman, Sander and Xun, Zhou, eds. (2004) Smoke: A Global History of Smoking. London: Reaktion Books.

2 WHO global report on trends in prevalence of tobacco smoking 2000–2025, second edition. Geneva: World Health Organization.

3 Heather L Wipfli, Jonathan M Samet : Second-hand smoke's worldwide disease toll.

4 Tackling second-hand exposure to tobacco smoke and aerosols of electronic cigarettes: the TackSHS project protocol Author: Esteve Fernández,María José López,Silvano Gallus,Sean Semple,Luke Clancy,Panagiotis Behrakis,Ario Ruprecht,Giuseppe Gorini,Ángel López-Nicolás,Cornel Radu-Loghin,Joan B. Soriano,Esteve Fernández,Yolanda Castellano,Marcela Fu,Montse Ballbè et al.

5 Miller MD, Marty M, Arcus A, Brown J, Morry D, Sandy M. Differences Between Children and Adults : Implications for Risk Assessment at California EPA. Int J Toxicol. 2002;21:403-418. doi:10.1080/1091581029009663

6 Russo ET, Hulse TE, Adamkiewicz G, et al. Comparison of indoor air quality in smoke-permitted and smoke-free multiunit housing: findings from the Boston Housing Authority. Nicotine Tob Res. 2015;17(3):316-322. doi:10.1093/ntr/ntu146

7 Arechavala T, Continente X, Pérez-Ríos M, et al. Second-hand smoke exposure in homes with children: assessment of airborne nicotine in the living room and children's bedroom. Tob Control. 2018;27(4):399-406. doi:10.1136/tobaccocontrol-2017-053751

8 Myers V, Shiloh S, Rosen L. Parental perceptions of children's exposure to tobacco smoke: development and validation of a new measure. BMC Public Health. 2018;18(1):1031. doi:10.1186/s12889-018-5928-1

9 California Environmental Protection Agency. Health Effects of Exposure to Secondhand Smoke: Final Report. Sacramento, CA; 1997.

10 U.S. Surgeon General. The Health Consequences of Involuntary Smoking: A Report of the Surgeon General. Rockville, MD; 1986.

11 U.S. Environmental Protection Agency. Respiratory Health Effects of Passive Smoking: Lung Cancer and Other Disorders. Washington, D.C.; 1992.

12 National Research Council (U.S.) Committee on Passive Smoking. Environmental Tobacco Smoke: Measuring Exposures and Assessing Health Effects. Washington (D.C.): National Academies Press; 1986.

13 Jenkins RA, Tomkins B, Guerin MR. The Chemistry of Environmental Tobacco Smoke: Composition and Measurement. Second. Boca Raton, FL: Lewis Publishers; 2000.

14 Louis GB, Damstra T, DíazBarriga F, et al. Principles for Evaluating Health Risks in Children Associated with Exposure to Chemicals. Genevia, Switzerland; 2006. 46. Habre R, Coull B, Moshier E, et al. Sources of indoor air pollution in New York City residences of asthmatic children. J Expo Sci Environ Epidemiol. 2013;24(3):1-10. doi:10.1038/jes.2013.74

15 U.S. Department of Health and Human Services; Centers for Disease Control and Prevention; National Center for Chronic Disease Prevention and Health Promotion; Office on Smoking and Health. The Health Consequences of Smoking-50 Years of Progress: A Report of the Surgeon General.; 2014.

16 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. doi:10.1289/ehp.11848

17 Miller MD, Marty M, Arcus A, Brown J, Morry D, Sandy M. Differences Between Children and Adults : Implications for Risk Assessment at California EPA. Int J Toxicol. 2002;21:403-418. doi:10.1080/1091581029009663

18 Homa DM, Neff LJ, King BA, et al. Vital signs: disparities in nonsmokers' exposure to second-hand smoke--United States, 1999-2012. MMWR Morb Mortal Wkly Rep. 2015;64(4):103-108.

19 Custers K, Van den Bulck J. Television Viewing, Internet Use, and Self-Reported Bedtime and Rise Time in Adults: Implications for Sleep Hygiene Recommendations From an Exploratory Cross-Sectional Study. Behav Sleep Med. 2012;10(2):96-105. doi:10.1080/15402002.2011.596599

20 King BA, Travers MJ, Cummings KM, Mahoney MC, Hyland AJ. Second-hand smoke transfer in multiunit housing. Nicotine Tob Res. 2010;12(11):1133-1141. doi:10.1093/ntr/ntq162

21 Russo ET, Hulse TE, Adamkiewicz G, et al. Comparison of indoor air quality in smoke-permitted and smoke-free multiunit housing: findings from the Boston Housing Authority. Nicotine Tob Res. 2015;17(3):316-322. doi:10.1093/ntr/ntu146

22 Arechavala T, Continente X, Pérez-Ríos M, et al. Second-hand smoke exposure in homes with children: assessment of airborne nicotine in the living room and children's bedroom. Tob Control. 2018;27(4):399-406. doi:10.1136/tobaccocontrol-2017-053751

23 Myers V, Shiloh S, Rosen L. Parental perceptions of children's exposure to tobacco smoke: development and validation of a new measure. BMC Public Health. 2018;18(1):1031. doi:10.1186/s12889-018-5928-1

24 Semple S, Latif N. How long does second-hand smoke remain in household air: Analysis of PM2.5 data from smokers' homes. Nicotine Tob Res. 2014;16(10):1365-1370. doi:10.1093/ntr/ntu089

25 Matt GE, Bernert JT, Hovell MF. Measuring Secondhand Smoke Exposure in Children: An Ecological Measurement Approach. J Pediatr Psychol. 2007;33(2):156-175. doi:10.1093/jpepsy/jsm123

26 Chapman Haynes M, St. Claire AW, Boyle RG, Betzner A. Testing and Refining Measures of Secondhand Smoke Exposure Among Smokers and Nonsmokers. Nicotine Tob Res. 2016;20(2):199-205. doi:10.1093/ntr/ntw315

27 Hughes SC, Bellettiere J, Nguyen B, et al. Randomized Trial to Reduce Air Particle Levels in Homes of Smokers and Children. Am J Prev Med. 2018;54(3):359-367. doi:10.1016/J.AMEPRE.2017.10.017

28 Hovell MF, Bellettiere J, Liles S, et al. Randomised controlled trial of real-time feedback and brief coaching to reduce indoor smoking. Tob Control. February 2019. doi:10.1136/tobaccocontrol-2018-054717

29 Pope CA, Dockery DW. Health Effects of Fine Particulate Air Pollution: Lines that Connect. J Air Waste Manage Assoc. 2006;56(6):709-742. doi:10.1080/10473289.2006.10464485

30 Pope CA. Review: Epidemiological Basis for Particulate Air Pollution Health Standards. Aerosol Sci Technol. 2000;32(1):4-14. doi:10.1080/027868200303885

31 Boyle M. A translational investigation of positive and negative behavioral contrast. ProQuest Diss Theses. 2015.

32 Vahlkvist S, Sinding M, Skamstrup K, Bisgaard H. Daily home measurements of exhaled nitric oxide in asthmatic children during natural birch pollen exposure. J Allergy Clin Immunol. 2006;117(6):1272-1276. doi:10.1016/j.jaci.2006.03.018

33 Allen RW, Mar T, Koenig J, et al. Changes in lung function and airway inflammation among asthmatic children residing in a woodsmoke-impacted urban area. Inhal Toxicol. 2008;20(4):423-433. doi:10.1080/08958370801903826

34 McCreanor J, Cullinan P, Nieuwenhuijsen MJ, et al. Respiratory effects of exposure to diesel traffic in persons with asthma. N Engl J Med. 2007;357(23):2348-2358. doi:10.1056/NEJMoa071535

35 Gong H, Linn WS, Terrell SL, et al. Exposures of elderly volunteers with and without chronic obstructive pulmonary disease (COPD) to concentrated ambient fine particulate pollution. Inhal Toxicol. 2004;16(11-12):731-744. doi:10.1080/08958370490499906

36 Gong H, Linn WS, Terrell SL, et al. Altered heart-rate variability in asthmatic and healthy volunteers exposed to concentrated ambient coarse particles. Inhal Toxicol. 2004;16(6-7):335-343. doi:10.1080/08958370490439470

37 Gong H, Linn WS, Clark KW, et al. Exposures of healthy and asthmatic volunteers to concentrated ambient ultrafine particles in Los Angeles. Inhal Toxicol. 2008;20(6):533-545. doi:10.1080/08958370801911340

38 Penttinen P, Vallius M, Tiittanen P, Ruuskanen J, Pekkanen J. Source-specific fine particles in urban air and respiratory function among adult asthmatics. Inhal Toxicol. 2006;18(3):191-198. doi:10.1080/08958370500434230

39 Li XY, Gilmour PS, Donaldson K, MacNee W. Free radical activity and pro-inflammatory effects of particulate air pollution (PM10) in vivo and in vitro. Thorax. 1996;51(12):1216-1222. doi:10.1136/thx.51.12.1216

40 Bernstein J a, Alexis N, Bacchus H, et al. The health effects of non-industrial indoor air pollution. J Allergy Clin Immunol. 2008;121(3):585-591. doi:10.1016/j.jaci.2007.10.045

41 Ormstad H. Suspended particulate matter in indoor air: adjuvants and allergen carriers. Toxicology. 2000;152(1- 3):53-68. doi:10.1016/s0300-483x(00)00292-4

42 Maciejczyk P, Zhong M, Lippmann M, Chen L-C. Oxidant generation capacity of source-apportioned PM2.5. Inhal Toxicol. 2010;22 Suppl 2(July):29-36. doi:10.3109/08958378.2010.509368

43 Stenfors N, Nordenhall C, Salvi SS, et al. Different airway inflammatory responses in asthmatic and healthy humans exposed to diesel. Eur Respir J. 2004;23(1):82-86. doi:10.1183/09031936.03.00004603

44 Riediker M, Cascio WE, Griggs TR, et al. Particulate matter exposure in cars is associated with cardiovascular effects in healthy young men. Am J Respir Crit Care Med. 2004;169(8):934-940. doi:10.1164/rccm.200310-1463OC

45 Gong H, Linn WS, Sioutas C, et al. Controlled exposures of healthy and asthmatic volunteers to concentrated ambient fine particles in Los Angeles. Inhal Toxicol. 2003;15:305-325. doi:10.1080/08958370390168300

46 Obata H, Dittrick M, Chan H, Chan-Yeung M. Sputum eosinophils and exhaled nitric oxide during late asthmatic reaction in patients with western red cedar asthma. Eur Respir J. 1999;13(3):489-495. doi:10.1183/09031936.99.13348999

47 U.S. Food and Drug Administration. Harmful and Potentially Harmful Constituents in Tobacco Products and Tobacco Smoke: Established List. Silver Spring, MD; 2012.

48 Repace J.L. Exposure to Secondhand Smoke, in Exposure Analysis. (Ott WR, Steinemann AC, Wallace LA, eds.). Boca Raton, FL: Taylor & Francis Group; 2007.

49 Chilmonczyk BA, Knight GJ, Palomaki GE, Pulkkinen AJ, Williams J, Haddow JE. Environmental tobacco smoke exposure during infancy. Am J Public Health. 1990;80(10):1205-1208. doi:10.2105/AJPH.80.10.1205

50 Barbara A. Chilmonczyk, Salmun LM, Megathlin KN, et al. Association between exposure to environmental tobacco smoke and exacerbations of asthma in children. N Engl J Med. 1993;328(23):1665-1669. doi:10.1056/NEJM199306103282303

51 State of California Air Resources Board. Technical support document for the "Proposed identification of second-hand smoke as a toxic air contaminant", Part A-exposure assessment 2006.

52 U.S. Department of Health and Human Services; Centers for Disease Control and Prevention; National Center for Chronic Disease Prevention and Health Promotion; Office on Smoking and Health. The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General.; 2006. doi:10.1088/0953- 8984/27/15/154205

53 California Environmental Protection Agency. Health Effects of Exposure to Secondhand Smoke: Final Report. Sacramento, CA; 1997.

54 U.S. Surgeon General. The Health Consequences of Involuntary Smoking: A Report of the Surgeon General. Rockville, MD; 1986.

55 U.S. Environmental Protection Agency. Respiratory Health Effects of Passive Smoking: Lung Cancer and Other Disorders. Washington, D.C.; 1992.

56 National Research Council (U.S.) Committee on Passive Smoking. Environmental Tobacco Smoke: Measuring Exposures and Assessing Health Effects. Washington (D.C.): National Academies Press; 1986.

57 Jenkins RA, Tomkins B, Guerin MR. The Chemistry of Environmental Tobacco Smoke: Composition and Measurement. Second. Boca Raton, FL: Lewis Publishers; 2000.

58 U.S. Department of Health and Human Services. The Health Consequences of Smoking—50 Years of Progress The Health Consequences of Smoking —50 Years of Progress.; 2014.

59 Louis GB, Damstra T, DíazBarriga F, et al. Principles for Evaluating Health Risks in Children Associated with Exposure to Chemicals. Genevia, Switzerland; 2006. 46. Habre R, Coull B, Moshier E, et al. Sources of indoor air pollution in New York City residences of asthmatic children. J Expo Sci Environ Epidemiol. 2013;24(3):1-10. doi:10.1038/jes.2013.74

60 Wipfli H, Avila-Tang E, Navas-Acien A, et al. Second-hand smoke exposure among women and children: evidence from 31 countries. Am J Public Health. 2008;98(4):672-679. doi:10.2105/AJPH.2007.126631

61 Zhang X, Martinez-Donate AP, Kuo D, Jones NR. "How is smoking handled in your home?": Agreement between parental reports on home smoking bans in the United States, 1995-2007. Nicotine Tob Res. 2012;14(10):1170-1179. doi:10.1093/ntr/nts005

62 Centers for Disease C, Prevention. Vital signs: nonsmokers' exposure to second-hand smoke --- United States, 1999- 2008. MMWR Morb Mortal Wkly Rep. 2010;59(35):1141-1146. doi:mm5935a4 [pii]

63 Mbulo L, Palipudi KM, Andes L, et al. Second-hand smoke exposure at home among one billion children in 21 countries: findings from the Global Adult Tobacco Survey (GATS). Tob Control. 2016;25(e2):e95-e100. doi:10.1136/tobaccocontrol-2015-052693

64 Klepeis NE, Nelson WC, Ott WR, et al. The National Human Activity Pattern Survey (NHAPS): A resource for assessing exposure to environmental pollutants. J Expo Anal Environ Epidemiol. 2001;11(3):231-252. doi:10.1038/sj.jea.7500165

65 Liu X, Liu L, Owen JA, Kaplan DL. Sleep patterns and sleep problems among schoolchildren in the United States and China. Pediatrics. 2005;115(1):241-249. doi:10.1542/peds.2004-0815F

66 Chandra S, Scharf D, Shiffman S. Within-day temporal patterns of smoking, withdrawal symptoms, and craving. Drug Alcohol Depend. 2011;117(2-3):118-125. doi:10.1016/j.drugalcdep.2010.12.027

67 Scharf D, Dunbar M, Shiffman S. Smoking during the night: Prevalence and smoker characteristics. Nicotine Tob Res. 2008;10(1):167-178. doi:10.1080/14622200701767787

68 Wilson I, Semple S, Mills LM, et al. REFRESH--reducing families' exposure to second-hand smoke in the home: a feasibility study. Tob Control. 2013;22(5):e8. doi:10.1136/tobaccocontrol-2011-050212

69 O'Rourke JM, Kalish LA, McDaniel S, Lyons B. The effects of exposure to environmental tobacco smoke on pulmonary function in children undergoing anesthesia for minor surgery. Pediatr Anesth. 2006;16(5):560-567. doi:10.1111/j.1460-9592.2005.01821.x

70 Hovell MF, Meltzer SB, Wahlgren DR, et al. Asthma Management and Environmental Tobacco Smoke Exposure Reduction in Latino Children: A Controlled Trial. Pediatrics. 2002;110(5). doi:10.1542/peds.108.1.18

71 Wahlgren DR, Hovell MF, Meltzer SB, Hofstetter CR, Zakarian JM. Reduction of Environmental Tobacco Smoke Exposure in Asthmatic Children: A 2-Year Follow-up. Chest. 1997;111(1):81-88. doi:10.1378/CHEST.111.1.81

72 Rosen L, Kostjukovsky I. Parental risk perceptions of child exposure to tobacco smoke. BMC Public Health. 2015;15(1):90. doi:10.1186/s12889-015-1434-x

73 Passey ME, Longman JM, Robinson J, Wiggers J, Jones LL. Smoke-free homes: what are the barriers, motivators and enablers? A qualitative systematic review and thematic synthesis. BMJ Open. 2016;6(3):e010260. doi:10.1136/BMJOPEN-2015-010260

74 Lonergan BJ, Meaney S, Perry IJ, et al. Smokers Still Underestimate the Risks Posed by Secondhand Smoke: A Repeated Cross-Sectional Study. Nicotine Tob Res. 2014;16(8):1121-1128. doi:10.1093/ntr/ntu046

75 Rosen LJ, Lev E, Guttman N, et al. Parental Perceptions and Misconceptions of Child Tobacco Smoke Exposure. Nicotine Tob Res. 2018;20(11):1369-1377. doi:10.1093/ntr/ntx169 63. Avila-Tang E, Elf JL, Cummings KM, et al. Assessing second-hand smoke exposure with reported measures. Tob Control. 2013;22(3):156-163. doi:10.1136/TOBACCOCONTROL-2011-050296

76 Lofroth G. Environmental tobacco smoke: multicomponent analysis and room-to-room distribution in homes. Tob Control. 1993;2(3):222. doi:10.1136/tc.2.3.222

77 Klepeis NE, Nazaroff W.W. Modeling residential exposure to second-hand tobacco smoke. Atmos Environ. 2006;40(23):4393-4407. doi:10.1016/J.ATMOSENV.2006.03.018

78 Klepeis NE, Nazaroff W.W. Mitigating residential exposure to second-hand tobacco smoke. Atmos Environ. 2006;40(23):4408-4422. doi:10.1016/J.ATMOSENV.2006.03.017

79 Fabian M, Lee S, Underhill L, et al. Modeling Environmental Tobacco Smoke (ETS) Infiltration in Low-Income Multifamily Housing before and after Building Energy Retrofits. Int J Environ Res Public Health. 2016;13(3):327. doi:10.3390/ijerph13030327

80 Singer BC, Hodgson AT, Guevarra KS, Hawley EL, Nazaroff W.W. Gas-Phase Organics in Environmental Tobacco Smoke. 1. Effects of Smoking Rate, Ventilation, and Furnishing Level on Emission Factors. 2002;36(5):846-853. doi:10.1021/ES011058W

81 Nazaroff W.W., Singer BC. Inhalation of Hazardous Air Pollutants from Environmental Tobacco Smoke in U.S. Residences. J Expo Sci Environ Epidemiol. 2004;14:S71-S77.

82 Kraev TA, Adamkiewicz G, Hammond SK, Spengler JD. Indoor concentrations of nicotine in low-income, multi-unit housing: associations with smoking behaviours and housing characteristics. Tob Control. 2009;18(6):438-444. doi:10.1136/tc.2009.029728

83 Tyc VL, Lensing S, Vukadinovich CM, Hovell MF. Can parents of children with cancer accurately report their child's passive smoking exposure? Nicotine Tob Res. 2009;11(11):1289-1295. doi:10.1093/ntr/ntp129

84 Hovell MF, Zakarian JM, Wahlgren DR, Matt GE, Emmons KM. Reported measures of environmental tobacco smoke exposure: trials and tribulations. Tob Control. 2000;9(iii):22-28. doi:10.1136/TC.9.SUPPL_3.III22

85 Matt GE, Wahlgren DR, Hovell MF, et al. Measuring environmental tobacco smoke exposure in infants and young children through urine cotinine and memory-based parental reports: empirical findings and discussion. Tob Control. 1999;8(3):282-289. doi:10.1136/TC.8.3.282

86 Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol. 1990;43(1):87-91. doi:10.1016/0895- 4356(90)90060-3

87 Borrelli B, McQuaid EL, Wagener TL, Hammond SK. Children With Asthma Versus Healthy Children: Differences in Second-hand Smoke Exposure and Caregiver Perceived Risk. Nicotine Tob Res. 2014;16(5):554-561. doi:10.1093/ntr/ntt180

88 Max W, Sung H-Y, Shi Y, Max W, Sung H-Y, Shi Y. Who Is Exposed to Secondhand Smoke? Self-Reported and Serum Cotinine Measured Exposure in the U.S., 1999-2006. Int J Environ Res Public Health. 2009;6(5):1633-1648. doi:10.3390/ijerph6051633

89 McCarville M, Sohn M-W, Oh E, Weiss K, Gupta R. Environmental tobacco smoke and asthma exacerbations and severity: the difference between measured and reported exposure. Arch Dis Child. 2013;98(7):510-514. doi:10.1136/archdischild-2012-303109

90 Arechavala T, Continente X, Pérez-Ríos M, et al. Validity of self-reported indicators to assess second-hand smoke exposure in the home. Environ Res. 2018;164(January):340-345. doi:10.1016/j.envres.2018.03.014

91 Sexton K, Callahan MA, Bryan EF. Estimating exposure and dose to characterize health risks: the role of human tissue monitoring in exposure assessment. Environ Health Perspect. 1995;103(suppl 3):13-29. doi:10.1289/ehp.95103s313

92 Pérez-Ríos M, Schiaffino A, López M.J., et al. Questionnaire-based second-hand smoke assessment in adults. Eur J Public Health. 2013;23(5):763-767. doi:10.1093/eurpub/cks069

93 Apelberg BJ, Hepp LM, Avila-Tang E, et al. Environmental monitoring of second-hand smoke exposure. Tob Control. 2013;22(3):147-155. doi:10.1136/TOBACCOCONTROL-2011-050301 82. Wallace L. Indoor Particles: A Review. J Air Waste Manage Assoc. 1996;46(2):98-126. doi:10.1080/10473289.1996.10467451

94 Semple S, Turner S, O'Donnell R, et al. Using air-quality feedback to encourage disadvantaged parents to create a smoke-free home: Results from a randomised controlled trial. Environ Int. 2018;120:104-110. doi:10.1016/J.ENVINT.2018.07.039

95 Zakarian J, Hovell M, Sandweiss R, et al. Behavioral counseling for reducing children's ETS exposure: Implementation in community clinics. Nicotine Tob Res. 2004;6(6):1061-1074. doi:10.1080/1462220412331324820

96 Leaderer BP, Hammond SK. Evaluation of vapor-phase nicotine and respirable suspended particle mass as markers for environmental tobacco smoke. Environ Sci Technol. 1991;25(4):770-777. doi:10.1021/es00016a023

97 Van Deusen A, Hyland A, Travers MJ, et al. Second-hand smoke and particulate matter exposure in the home. Nicotine Tob Res. 2009;11(6):635-641. doi:10.1093/ntr/ntp018

98 Buan KDV, Vera K De. Particulate Matter Concentrations At Children And Adult Breathing Heights In Residential Thirdhand Smoke Environments. 2015.

99 Berardi V, Carretero-González R, Klepeis NE, et al. Proper orthogonal decomposition methods for the analysis of real-time data: Exploring peak clustering in a second-hand smoke exposure intervention. J Comput Sci. 2015;11:102- 111. doi:10.1016/J.JOCS.2015.10.006

100 Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. Arch Gen Psychiatry. 2002;59(10):877-883. doi:10.1001/archpsyc.59.10.877

101 Find peaks inside a signal based on peak properties. scipy.signal.find_peaks — SciPy v1.5.2 Reference Guide

102 Granger, C. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica, 37(3), 424-438. doi:10.2307/1912791

103 H. GOLDSTEIN, Multilevel mixed linear model analysis using iterative generalized least squares, Biometrika, Volume 73, Issue 1, April 1986, Pages 43–56, https://doi.org/10.1093/biomet/73.1.43

104 Zhang S, Wang H, Zhou X, et al. A novel peak detection approach with chemical noise removal using short-time FFT for prOTOF MS data. Proteomics. 2009;9(15):3833-3842. doi:10.1002/pmic.200800030

105 Kopczynski D, Rahmann S. An online peak extraction algorithm for ion mobility spectrometry data. Algorithms Mol Biol. 2015;10:17. Published 2015 May 13. doi:10.1186/s13015-015-0045-5

106 Maka, T. Influence of adaptive thresholding on peaks detection in audio data. Multimed Tools Appl 79, 19329–19348 (2020). https://doi.org/10.1007/s11042-020-08780-2

107 Paradis, E. and Sutkin, G. (2017), Beyond a good story: from Hawthorne Effect to reactivity in health professions education research. Med Educ, 51: 31-39. https://doi.org/10.1111/medu.13122

108 Bontempi, Gianluca & Birattari, Mauro. (2003). The lazy Package.

109 Sepp Hochreiter, Jürgen Schmidhuber; Long Short-Term Memory. Neural Comput 1997; 9 (8): 1735–1780. doi: https://doi.org/10.1162/neco.1997.9.8.1735)

110 Giannadaki, D., Lelieveld, J. & Pozzer, A. Implementing the US air quality standard for PM2.5 worldwide can prevent millions of premature deaths per year. Environ Health 15, 88 (2016). https://doi.org/10.1186/s12940-016-0170-8

111 Centers for Disease Control and Prevention. Quitting smoking. cdc.gov. Page updated February 1, 2017. Accessed January 4, 2019

112 Braithwaite, Valerie. "Between stressors and outcomes: Can we simplify caregiving process variables?." The gerontologist 36.1 (1996): 42-53

113 Emery S, Gilpin EA, Ake C, Farkas AJ, Pierce JP. Characterizing and identifying "hard-core" smokers: implications for further reducing smoking prevalence. Am J Public Health. 2000 Mar;90(3):387-94. doi: 10.2105/ajph.90.3.387. PMID: 10705856; PMCID: PMC1446166.

114 Helsen, Lieve. "Sampling technologies and air pollution control devices for gaseous and particulate arsenic: A review." Environmental Pollution 137.2 (2005): 305-315.

115 Chollet, F. & others, 2015. Keras. Available at: https://github.com/fchollet/keras.

116 Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).