Title
Statistical Directions for the Analysis of Participatory Mobile Health Asthma Management Data

Permalink
https://escholarship.org/uc/item/95v613tn

Author
Hasenstab, Kyle Andrew

Publication Date
2012

Peer reviewed|Thesis/dissertation
Statistical Directions for the Analysis of Participatory Mobile Health Asthma Management Data

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science in Statistics

by

Kyle Andrew Hasenstab

2012
Mobile technology has become increasingly popular in the past decade through the combination of device portability and advances in network and internet technology. Smart phones, in particular, are at the forefront of these technologies, enabling users to remotely track and further involve themselves in the management of personal health through participatory mobile health applications. Thus far, the majority of participatory applications currently provide users with data visualizations displaying information germane to the user’s medical condition, however, there is still a need for in-depth exploratory and inference-based data analysis using advanced statistical methods to maximize the discernment of potential implications carried in these data streams. This paper provides an overview of the structure of participatory data using an asthma management application as an initial platform and discusses several directions for statistical analysis motivated by three usage cases: individuals using the application, creators of the application, and the scientific community. Methods include functional and semi-parametric data analysis, mixed modeling, and clustering methods to model variables representing asthma wellness as a function of subject-specific, population level, and latent spatiotemporal factors. Societal implications are also discussed.
The thesis of Kyle Andrew Hasenstab is approved.

Ying Nian Wu

Frederic Paik Schoenberg

Mark Hansen, Committee Chair

University of California, Los Angeles
2012
To my best friend and love of my life – Susan
# Table of Contents

1 Introduction .................................................. 1  
  1.1 General Introduction ..................................... 1  
  1.2 Mobile Health ............................................ 2  
  1.3 Applications for Mobile Health Data Streams ............. 3  
    1.3.1 Individual Benefit .................................. 3  
    1.3.2 Benefits for Creators of Wellness Applications ...... 4  
    1.3.3 Contributing to the Scientific Community .......... 5  
  1.4 Thesis Overview ......................................... 6  

2 Previous Research ............................................ 8  
  2.1 Analysis Frameworks ...................................... 8  
    2.1.1 Ecological Momentary Assessment .................... 8  
    2.1.2 Usage .............................................. 11  
    2.1.3 Pollution Studies .................................. 14  
    2.1.4 N-of-1 Trials ..................................... 15  
  2.2 Other Modeling Strategies ............................... 16  
    2.2.1 Longitudinal Modeling ............................... 16  
    2.2.2 Sparsity ........................................... 17  
    2.2.3 Mixture Models and Zero-Inflation .................. 17  
    2.2.4 Other Useful Topics ................................ 18  
  2.3 Remarks .................................................. 18  

3 The Data ..................................................... 19
3.1 AsthmaMD ................................................. 19
3.2 Data and Descriptive Statistics ...................... 20
3.3 Locations of Subjects ................................ 21
3.4 Usage by Major Cities ................................ 24
3.5 Subjects with Heavy Usage ............................ 26
3.6 Age-Gender Analysis .................................. 30
3.7 Comments on Using Great Britain ................... 37

4 Current Research ........................................... 40
4.1 Pollution Exploratory Analysis ....................... 40
4.2 Weather Exploratory Analysis ........................ 41
4.3 Modeling with Pollution and Weather ............... 46
4.4 Clustering and Discretizing Great Britain .......... 49

5 Statistical Directions ...................................... 56
5.1 Usage Cases and Modeling Motivations ............. 56
  5.1.1 Individual Benefit ................................. 56
  5.1.2 Benefits for Creators of Wellness Applications 57
  5.1.3 Contributing to the Scientific Community ....... 58
5.2 Statistical Directions for Achieving Usage Cases .. 58
  5.2.1 Individualized Analysis ........................... 59
  5.2.2 Aggregate Analysis ................................. 61
  5.2.3 Adherence Analysis ................................. 65
5.3 How Do the Usage Cases Relate to the Proposed Models? 66
  5.3.1 Individual Benefit ................................. 66
List of Figures

3.1 Histogram of the log total reports per subject showing zero-inflation with a highly skewed right tail. ........................................ 22
3.2 Global spatial distribution of users showing highest density of subjects in United States and western Europe. ...................... 23
3.3 Spatial distributions of western Europe and United Kingdom. Most subjects are located in Great Britain and Germany. .............. 24
3.4 Spatial distribution of United States showing a higher density of subjects in the Midwest and East Coast. ......................... 25
3.5 Spatial distributions of US time zones. Central and Eastern time zones have larger densities of subjects. ......................... 26
3.6 Regions containing more than 10 subjects. ........................... 27
3.7 Regions having more than 100 reports. ................................. 27
3.8 Reports over time for London for two samples. Plots exhibit seasonal trends and sparsity. ........................................... 28
3.9 Reports over time for various cities. Seasonal trends and sparsity are apparent. Profiles with highly consistent usage across the year, as seen in DC and Dallas, are great candidates for individualized analyses. ............................................... 29
3.10 PFM readings and weekly report counts for the top heavy users in Great Britain. Some correlation is visible between report counts and PFM values for this user, but many subjects do not exhibit this pattern. Report counts may be of better use in aggregate analysis. 30
3.11 PFM readings and weekly report counts for the top 3 heavy users in the United States. .................................................... 31
3.12 Mosaic plot of age and country. ........................................ 32
3.13 Boxplot of log report counts vs age (left) and counts of subjects
in each age bin (right). A cubic-like trend is apparent between log
report counts and age. ...................................................... 33
3.14 CDC 2010 asthma prevalence rates by age group. Teenagers have
high prevalence but low usage. .......................................... 35
3.15 Nonparametric quantile regressions of log report counts vs age ex-
hibiting trends at the upper quantiles. .............................. 36
3.16 Boxplot of log reports by gender. Very little difference in usage
between males and females. ............................................. 37
3.17 Mosaic plot of age vs gender. Proportion of females increases with
increasing age. ............................................................ 38

4.1 Quantile fits of AQI over time (left) and quantile fits of counts over
time (right). Trends are more visible at the upper quantiles. .... 42
4.2 Quantile fits of AQI and report counts over time. .................. 42
4.3 Boxplot of report counts vs daily (left) and weekly (right) median
AQI values. There is evidence that report counts are positively
correlated with air quality. ............................................... 43
4.4 Time series of temperature with Lowess smooth exhibits commonly
seen annually cyclic trend. ............................................. 44
4.5 Scatterplot of report counts vs temperature (left) and boxplot of
report counts vs temperature (right). ................................. 44
4.6 Time series of wind speed with Lowess smooth (left) and scatterplot
of report counts vs wind speed (right) ............................... 45
4.7 Time series of humidity with Lowess smooth (left) and scatterplot
of report counts vs humidity (right) ................................. 46
4.8 Non-parametric fits

4.9 Residuals vs fitted plot (left) and histogram of report counts for the actual data and the fitted values (right) from the GAM model. Both show the lack of fit for a Poisson distributional model assumption and the need to account for zero-inflation and skewness.

4.10 Clustering of Great Britain using weighted dissimilarity matrices of subject locations, pollutants, weather variables, and elevation.

4.11 Barplot of mean counts per week for each subject by cluster. Clustering effectively explains variability in mean counts per week for subjects in different locations.

4.12 Mean counts per subject per week by cluster. Unique trends are visible in the profiles of each cluster.

4.13 FCM clustering for February through June (left) and April through November (right). The plot on the right is typical of many long term report count profiles and the FCM clusters these profiles poorly.

4.14 Cluster means for February through June (left) and April through November (right).

4.15 Alpha projections for February through June (left) and April through November (right). Clustering on the right plot shows an ambiguous separation of clusters and the need for more complex distributional assumptions for the FCM.

5.1 Quantile and variance trends of counts vs temperature. Empirical variance trend matches that of the trends in the upper quantiles.
List of Tables

3.1 Single observation from AsthmaMD data set. 21

3.2 Simple descriptive statistics of AsthmaMD data. Note the large number of missing entries and the skewness of usage. 22

3.3 Subset of the 2010 Child and Adult Asthma Prevalence Tables. 34

4.1 Parametric results of GAM model. Model shows positive correlation between air quality and report counts. Significance of wind parameters support quadratic-like effect. Humidity is significant but has a very small effect. 47

4.2 Non-parametric results of GAM model. Both time and temperature are very significant in the fit. 48
First and foremost, I would like to thank Dr. Mark Hansen, who saw my academic potential during his 201B class and offered me an outstanding project for both a master’s and PhD thesis. Your knowledge and expertise has often guided me through many statistical, administrative, and life challenges. Also, thank you so much for your patience despite all of my silly questions! I would also like to thank my oral examination committee members Dr. Deborah Estrin, Dr. Frederic Paik Schoenberg, Dr. Damla Senturk, and Dr. Ying Nian Wu for making themselves available to me for future assistance and for providing me with additional research resources.

To Dr. Sam Pejham and Salim Madjd of AsthmaMD - Thank you so much for your patience and understanding during those long waits for analysis updates and for providing me with valuable data for such an interesting project. I look forward to working with you in the future for more in-depth research.

To Glenda Jones - I just want to let you know I do not take you for granted! Thank you so much for putting up with all of my administrative inquiries. My program and the entire department would not be able to function without you!

I would also like to thank my childhood friends with whom I am blessed to be around to this day. Although many of you were not interested in going to college, thank you for constantly pushing me to pursue higher education in order to escape the mess that is Carson, CA. Also, thank you to my undergraduate and graduate friends. Without your help and support, passing classes and earning degrees would have taken a lot more effort!
I am extremely grateful to my parents, who sacrificed everything to give me a chance to pursue higher education. Thank you for always being on my case about it and never giving up on me, despite the constant rebellion.

Last but not least - I would like to express by deepest appreciation to Susan. Without your support and love, this would not have been possible...and I mean that in every way possible.

This research has been made possible with partial support by the Center for Embedded Networked Sensing at the University of California, Los Angeles and the National Science Foundation’s Graduate Research Fellowship Program.

Thank you all so much!
CHAPTER 1

Introduction

1.1 General Introduction

Mobile technology (laptops, readers, mp3 players, mobile phones etc.) has become increasingly popular in the past decade through the combination of device portability and advances in network and internet technology, making information and resources easily accessible to many. Smart phones, in particular, are at the forefront of these technologies, enabling users to manage, track, and further engage themselves in lifestyle related activities and interests through the smart phones’ ability to integrate the majority of functions offered by other personalized mobile devices into one cohesive unit. The use of smart phones is widespread and is continuing to grow with 91.4 million smart phones in the United States alone during 2012 [25]. Smart phone penetration in the United States grew 10% in the last three years to 35%, still coming in fourth behind Singapore, Canada, and Hong Kong [9, 25], suggesting its momentum for increasing future usage worldwide. Of the several features offered by smart phones, applications (apps) appear to be a major motivator for smart phone usage with 68% of users using at least one of the over one million apps available [20, 71]. The popular use of mobile applications and the diversity of their functions have provided the scientific community with a plethora of valuable data that may be used to further research in all fields of study.
1.2 Mobile Health

Among the various types of widely used apps available, participatory mobile health (mHealth) or wellness applications have emerged as the innovative tool for personalized health, empowering individuals with the ability to manage and track illnesses such as asthma, diabetes, and post-traumatic stress disorder and improve clinician-patient communication and relations. Users are able to enter data pertaining to their illness, view temporal progress via data visualizations, send data and status to their clinician, and receive feedback from their clinician, all from their smart phone. There are already a large number of individuals using the over 10,000 mHealth applications available with a reported 80% of doctors using smart phones and medical apps and an overall 61% of smart phone users having downloaded at least one mHealth application [12, 47]. Mobile health adoption has doubled between 2010 and 2011 [47], recognizing mHealth as a valuable and pragmatic addition to modern day health care.

Despite the ubiquity of mHealth apps and their widespread use, the abundance of participatory health data produced by wellness applications is not being used to its full potential. The majority of mHealth applications currently provide users with data visualizations displaying information germane to the user’s medical condition, however, there is still a need for in-depth exploratory and inference-based data analysis using advanced statistical methods to maximize the discernment of potential implications carried in these data streams. Although mobile devices and statistical methods have been used to perform health related research [11, 13, 16, 63], the structure of participatory data is entirely different and calls for new mechanisms for analysis. Furthermore, the uncontrolled nature of application memberships and user contributed data offer statistical challenges not normally seen in previous health-related studies. Statistical methods spawned from these data streams will improve the quality of individual feedback by mHealth applica-
tions, create cost-effective alternatives to traditional epidemiological studies, and contribute to the general framework and efficiency of mHealth applications.

1.3 Applications for Mobile Health Data Streams

Before going into the potential applications for mHealth data streams, it is important to make the distinction between participatory mHealth applications and the general field of mobile health. Most studies incorporating mobile devices into their data collection methods are experimentally designed [16, 21, 35, 56, 63, 68], where the structure of data collection is created before the start of the experiment and the participants of the study are instructed on the frequency and types of data that should be entered. Contrary to these mHealth applications, subjects using participatory applications are not required to follow any protocol with regards to data entry. Subjects are able to enter data at anytime and can stop using the application without warning, making analysis of participatory data much more difficult. Despite these challenges, the ease of use and decreased patient responsibility encourages widespread use of the applications, providing statistical power to analyses with decreased financial cost. Future designs of participatory applications may be reformed to allow for data collection methods more similar to that of experimentally designed mobile health studies, but increased effort needed by the individual will produce the need for stronger incentives for subjects to participate, essentially reverting back to experimental design and defeating the purpose of participatory applications.

1.3.1 Individual Benefit

In order to discern the wide array of potential applications stemming from participatory data streams, it will help to divide mHealth data use cases into three groups, the first of which is the individual. The primary purpose of participatory
health applications is to provide helpful information and tools to individuals to assist them in the management of an illness or their general health. As previously mentioned, many of these applications only offer visualizations and simple data summaries on a small number of factors relevant to their ailment. Although these representations are informative, they fail to give subjects and their clinicians a complete overview, ignoring additional variables affecting a subject’s illness. In effect, feedback encompassing a wide variety of related factors, along with their interactions, must be included in an analysis to provide each subject with a complete health profile. This type of feedback necessitates a high resolution analysis of each subject and their surrounding environments and motivates further statistical exploratory analysis and modeling. Approaches may include individualized analyses, especially for subjects with chronic illnesses or populations with inherent subject-level heterogeneity [28], or individual trajectory analysis from aggregate mixed models as commonly seen in many longitudinal studies [36, 48]. These analyses may contribute to treatment selection, dosage optimization, and feedback regarding environmental (pollution, pollen count, etc.) and personal (smoking, dietary intake, etc.) effects, thus, providing individuals and their clinicians with additional treatment options (pending clinician approval) to improve their condition and offer a more extensive understanding of their illness.

1.3.2 Benefits for Creators of Wellness Applications

Another group that will strongly benefit from the analysis of participatory data is the creators of the wellness applications. As previously mentioned, many participatory mHealth applications provide subjects with simple summaries and visualizations which may not paint the entire picture of a subject’s illness profile, thus, ignoring information which may be helpful for illness management. As a result, application frameworks designed around these simple analyses fail to provide high resolution feedback to their users. By analyzing these data streams using ad-
Advanced statistical techniques, contributions can be made to new designs of these interfaces through improvements in the representation and content shown by data visualizations and wellness metrics.

In addition to design implications though advanced statistical techniques, analysis of application compliance is of great interest to application creators. The usage patterns of participating subjects can vary wildly, from one-time use to heavy and consistent patterns to sparse episodic profiles. By gaining an understanding of subject usage through adherence and dropout analysis, individual usage status can be classified into one of many categories, providing creators with information for improving application efficiency for data management and statistical computation. Adherence and dropout analysis may also be incorporated into aggregate or single-subject models to decrease bias, improve accuracy of model estimates, and determine relevance of proposed factors to an illness. As a result, subjects are provided with more informative and accurate feedback. Several studies on subject adherence and dropout have been performed [1, 4, 11, 13, 23, 35, 49] to assess the reliability of incorporating mobile devices into controlled studies, but these are under experimentally controlled conditions, hence, additional analysis of participatory data is needed on these topics to improve the functionality of participatory applications and increase the quality of feedback.

1.3.3 Contributing to the Scientific Community

Although individual feedback is of primary focus, the richness of participatory data allows for inferences to be made on entire populations. Due to the large number of individuals currently using mHealth applications, individual data streams may be pooled together to perform aggregate studies beneficial to the scientific community. Analogues of analyses commonly seen in traditional epidemiological studies may offer cost-effective alternatives or complementary tools for providing insight into the status and dynamics of illnesses throughout most regions of the
world. Mixed and growth mixture modeling, statistical approaches often seen in longitudinal studies [48, 51], may be extended to account for unique characteristics within participatory data to describe the spatiotemporal trends of a certain disease. These extensions can also be used to make demographic comparisons (race, gender, interactions with regions etc.) and inferences on personal and environmental factors believed to be associated with a disease at the population level. Additionally, alternative versions of prevalence and incidence rates may be compared with those found by government agencies to assess plausibility and can be used as an intermediate update between standard surveys or in regions where prevalence and incidence analyses have not yet been performed. These alternative estimates may serve as benchmarks for qualitative comparison or may be used to improve the design of mHealth application data collection methods if significant differences between participatory data statistics and governmental estimates are observed. Observational versions of clinical trials or meta-analyses of N of 1 trials [74] may also be developed using aggregate methods to assess the efficacy of several treatments in relation to subject environment or population demographic. The variety of population level applications fueled by participatory data has provided the scientific community with several additional tools for performing a diverse set of public health studies, enabling researchers to gain a thorough understanding of disease behavior and treatment effectiveness with decreased cost.

1.4 Thesis Overview

The purpose of this paper is to present some preliminary insight into the structure of participatory data streams and propose new statistical directions for analyzing participatory data under the three motivations described above. Data from an asthma management application on the iPhone by Dr. Sam Pejham of AsthmaMD will be used as a participatory mHealth app model. Previous research related to
the three usage cases described in Section 1.3 are discussed in Chapter 2. The overall data set and some preliminary analyses are described in Chapter 3. Chapter 4 gives a detailed explanation of more advanced preliminary results consisting primarily of exploratory analyses used to expose the unique features of the participatory data set and Chapter 5 proposes new directions for participatory research guided by these exploratory analyses. The paper then concludes with some final remarks on the implications of participatory data research on personalized and public health.
CHAPTER 2

Previous Research

The numerous applications motivated by the three communities of interest described in Section 1.3 require an extensive literature review of topics spanning from clinical research and ecological momentary assessment studies to technology usage and statistically intensive analysis. The following review is divided into several subsections, each describing previous research in a particular area and their relevance to achieving the three usage cases.

2.1 Analysis Frameworks

2.1.1 Ecological Momentary Assessment

Ecological momentary assessment (EMA) is an experimental framework designed to allow subjects involved in a research study (many of which appear in psychology and health journals) to enter data on the outcomes of interest in real time (using diaries, phone calls, electronic devices etc.), similar to the real-time nature of data entry in participatory applications. Stone et al. mention three core benefits of using EMA over traditional designs [69]. The first is the avoidance of recall bias by collecting data in real-time. Many studies have shown that subjects may not be able to accurately recall data on past experiences [6, 30] and collecting data in real-time would avoid or at least greatly mitigate this bias. This is supported in papers by Robinson & Clore (2002) [57, 58] and Ross (1989) [60]. The second is the ability to collect data while the subject is in his or her natural environment
and is the ecological aspect of EMA. This is extremely important when attempting to root out noninterventional factors contributing to variability in the outcome within the subjects’ environment. In the context of asthma wellness applications, this can be pollution, weather, and pollen levels. Third, real-time data collection extends opportunities in analysis by allowing researchers to observe dynamic processes over time, a valuable attribute when studying functional outcomes like that of participatory data. Smyth and Stone [67] and Shiffman et al. [66] also describe advantages of using EMA techniques and in the context of behavioral medicine. The reason for exploring EMA methods for the purpose of participatory research is the similarities among their corresponding real-time structures. Methods from EMA may be applied to participatory data, however, there are fundamental differences between the asthma data and data from the majority of EMA studies that require extensions to their widely used modeling techniques.

Many EMA studies appear to use fixed-effects or mixed-effects models and variations thereof to model real-time data as these are the generally accepted methods for modeling trajectories of several subjects over a period of time. Variations of models include hierarchical components, using a generalized response, modeling the covariance structure, and using nonparametric approaches for functional data. Epstein et al. used a logistic repeated-measures model with first-order autoregressive error structure to analyze craving for or use of cocaine or heroin in relation to acute daily life precipitants [16]. The EMA study by Rouse et al. fit eight repeated measures analysis of variance models to model the relationship of eight dependent behavioral variables with day of the week and gender to analyze the physical activity and sedentary behavior patterns of university students in England [61]. Everhart et al. investigated the relationships between asthma-related variables with a quality of life measure to assess the predictability of the Asthma Quality of Life Questionnaire [17]. The EMA study collected data using palmtop computers and peak flow meters. Mixed effects and generalized mixed effects
models with random intercepts were used for continuous and discrete responses, respectively. Their findings suggest that the questionnaire is valuable in assessing patient response to treatment. Other EMA studies used similar model frameworks to model EMA data collected by mobile devices [21, 54, 56, 63]. These techniques may be used as a foundational structure for modeling participatory data for use in the three applications described above.

With the introduction of electronic devices to EMA, there have been several studies questioning the compliance of participating subjects and the validity of their data compared with traditional non-electronic or paper-based methods. Holtz and Whitten tested the feasibility and utilization of tracking asthma symptoms through a mobile phone application [35]. Subjects sent in their daily peak flow meter readings to a web server via short message service (SMS) on their mobile phones. Using simple analyses with a small sample size \((n = 4)\), they concluded the mobile phone messaging system of monitoring asthma was feasible and that mobile phones for asthma management could improve compliance with asthma action plans and reduce adverse asthma events. Anhoj and Moldrup performed a similar analysis, evaluating the feasibility of using SMS for asthma diary collection using mobile phones [1]. Additionally, they investigated patient compliance with the diary using response rates over time. They also concluded that SMS may be a supportive tool for asthma management because of the widespread use of mobile phones in their daily lives. Instead of asthma management web pages where the attrition rate is high, the combination of web pages with mobile phones would increase compliance. Cruz-Correia et al. found that web based asthma management was supported over paper-based methods but compliance was quite variable [13]. Several comments were made on aspects of the web-application that may be improved to increase subject adherence. Another study found similar results using cellular phones to study alcohol consumption [11].

Despite the support for using mobile devices for health management and the
possibility of using statistical methods from EMA studies to produce inferences on participatory data, there are several characteristics of participatory data collection methods which tend to complicate analysis. In contrast to EMA studies where subjects are coached on the frequency and method for data entry, participatory data is voluntary. This causes complications in the context of adherence, where subjects who have dropped out may be confused with subjects whose seasonal exacerbations have passed or with subjects who temporarily dropped out. Moreover, the voluntary nature of participatory data leaves large numbers of subjects with very little entries, introducing skewness, sparsity, and mixture and inflated distributions to the modeling problem. The long term profiles of subjects, lasting years in some cases, produce functional trajectories that must be modeled accordingly. These topics are not often seen in traditional EMA studies and are considered in the sections below.

2.1.2 Usage

A topic of interest when analyzing the usage of an mHealth application is subjects’ adherence to data entry. EMA studies design a structured protocol for data entry which dictates the use of a mobile device by participating subjects. Since EMA researchers know exactly when data is supposed to be entered, whether entry is on a predefined schedule or due to random prompts, they may perform analysis on the adherence of subject’s use of the device. Ryan et al. performed a study researching the adherence of individuals using mobile phone technology for the management of their asthma [62], perhaps the most relevant study to the context presented here. The 91 subjects involved in the study were instructed to enter peak flow meter readings every morning and evening using a hand-held electronic peak flow meter that was connected to their mobile phone. They were interested in analyzing the compliance of subjects to their daily PFM entry instructions. Results suggested that the subject population could be divided into three subgroups.
The first were subjects enduring technical problems, causing large gaps in their data. The second had fewer than 100 readings during the entire study, implying low compliance. The third group was highly compliant and represented 64% of the subject population, suggesting that asthma management via mobile phone based applications is feasible. Mermelstein et al. performed an EMA study to capture adolescent smoking patterns of high school students over the course of 18 months [49]. Four waves of data were collected, each one week long, using mobile devices. The nature of data entry consisted of random prompts and event-based collection. Adherence of subjects to the random prompts was assessed using simple summary statistics and ensured excellent compliance. Event-based entry was assessed using post-study interviews and surveys and comparisons of summary statistics to anecdotal measures. Other studies use similar approaches to perform adherence analysis [4, 23]. In the context of participatory health, adherence is a much more complicated issue. Since a sequence of perfect adherence does not exist, there is no precise benchmark for assessing user compliance. Participatory usage can be interpreted as event based entry, but periodic surveys or interviews may not be feasible. Therefore, the problem of adherence takes on a completely different problem, requiring methods for identifying patterns during periods of non-usage within and between subjects across seasonal cycles to classify when and the degree to which a subject is engaged with the application.

Analyzing usage of a participatory mHealth application is not only important for assessing subject adherence, but may contribute to aggregate modeling as well as to the efficiency of the application itself. Anhoj and Nielsen examine patients’ and health-care providers’ use of an online asthma management tool [2]. Simple descriptive statistics and visualizations were used to assess the rate of usage over a three year period and to construct a demographic breakdown of users. The duration of usage for many subjects was short-lived due to many technical and psychological aspects. Problems included internet access, conflicting user needs,
failure to recognize any immediate diary benefit, and inexperience with technology. Some of these problems are countered by the use of smart phones for asthma management.

Usage studies of modern mobile technologies have also been described in many tech-oriented journals. Falaki et al. examine the use of different features on smart phones (browsing, media, etc.), exposing several methods for improving the efficiency, performance, and power consumption of smart phone communication [18]. Cumulative distribution functions (CDFs) were the primary statistical tool used to find patterns in usage variables of interest. In a second paper by Falaki et al., several aspects of 255 user patterns of smart phones on different platforms were analyzed [19]. Application usage was examined but at the macro level where the use of several applications were of interest versus the use of a single application. Patterns in length of interaction with the smart phone, diurnal trends, and patterns in application popularity were exposed using CDFs and models using an exponential response. Xu et al. concentrated on the diversity of usage within the United States specifically for smart phone applications using CDFs [72]. Other papers use similar methods to explore mobile device use, Wi-Fi network traffic, and platform efficiency [22, 45, 55]. Additional research has also been done on the design of a sampling framework for wellness applications to promote the use of mHealth applications so that the scientific community may better understand a user’s health related habits and observations [34].

Thus far, many of the papers used relatively simple exploratory analyses to examine the usage of smart phones and their applications. Although there are papers on a wide diversity of usage topics, there is an obvious need for a high resolution advanced statistical look into aspects of subject usage and usage patterns of single applications. As a result, this would enhance the quality of the application function and provide improved and customized feedback for the user. Section 2.2 discusses literature that is more statistically intensive and relevant to
modeling asthma participatory health data.

2.1.3 Pollution Studies

When analyzing the condition of asthma within a subject or population, researchers must look at the subjects’ environmental surroundings (pollution, pollen, weather, etc.). Inferences from models incorporating pollution and weather factors are an important aspect of asthma feedback, giving subjects the opportunity to gauge their sensitivity to certain environmental factors. Due to the functional nature of these factors and their collinearity with time and each other, past modeling strategies including environmental factors must be investigated. There have been several studies that attempt to model the relationship between pollution data and morbidity counts. Note that pollution studies are easily extended to pollen counts and weather factors due to their similar functional and annually cyclic behavior. Peng et al. discuss model choice for time series of air pollution and mortality [53]. They perform a Poisson regression with log-link containing parametric terms for variables of interest and a nonparametric term composed of time and weather variables to adjust for confounding between pollution and weather variables. Different spline bases are used along with different methods for selecting degrees of freedom to systematically compare models using mean squared error. They found that optimizing prediction may not produce estimates with small bias and that fully parametric and nonparametric models perform equally well. Dominici et al. use an over-dispersed semi-parametric Poisson model with log-link to model mortality as a function of air pollution [14]. They introduce a closed form estimate of the asymptotically exact covariance matrix, develop a bandwidth selection method to reduce confounding bias, and introduce a sensitivity analysis framework for pollution estimates based on model choice. Other studies used similar methods to make inference on the relationship between mortality, air pollution, and other variables of interest in different contexts [10, 26].
Stieb et al. performed a meta-analysis of several time-series studies on air pollution and mortality around the world using a random effects model [68]. They found that several air pollutants were all positively and significantly associated with all-cause mortality.

2.1.4 N-of-1 Trials

A problem with conducting clinical trials by pooling the entire subject population into one analysis is the heterogeneity among subjects. Inferences on treatment efficacy may not apply to each unit in the population due to inherent differences in the subjects. As a result, recommendations of treatments supported by traditional randomized controlled trials (RCTs) may be ineffective for certain individuals within the population. In this case, N-of-1 trials or single-subject designs (described in many clinical trials journals) may be a more appropriate approach [28]. Guyatt et al. describes N-of-1 trials as RCTs for individual patients where each patient undergoes a pair of treatments (one with actual treatment and one with placebo). They claim N-of-1 trials to be useful for chronic, stable conditions, making this method relatively applicable to asthma trials. Further details such as when to perform the trial and ethical concerns are also covered in the paper. Thus far, several different statistical analyses have been proposed to make inferences on data produced by N-of-1 trials. Campbell describes and compares four methods for obtaining effect estimates in single-subject designs and notes the limitations of each [7]. These methods include mean baseline reduction, percentage non-overlapping data, percentage of zero data, and a regression based statistic. Given small amounts of data, he prefers the non-regression based methods over the regression-based method as they produce approximately the same results with more efficient calculation. Kinugasa et al. describes multiple single-subject research designs and data analysis techniques for athletic conditioning [40]. Data analysis methods include visual inspection, time-series analysis, C-statistics, ran-
domination tests, and rank tests. Other articles also perform time-series analysis to model N-of-1 data [24, 59]. In addition to analyzing subjects separately, meta-analyses may be performed to find aggregate effects of treatments. Zucker et al. describe previous attempts at combining N-of-1 trials including pooled treatment effects and linear mixed models, and propose a Bayesian hierarchical model for meta-analysis [74]. Although N-of-1 studies have a variety of methods for making inferences on individual factor effects, they are experimentally designed and are structurally different from participatory data. Despite the controlled design, aspects of N-of-1 trials may be extended to assist with inferences on individual subjects with heavier usage. These subjects offer a large amount of information on their condition and feedback on their asthma triggers and medications based on their data alone may be plausible.

2.2 Other Modeling Strategies

2.2.1 Longitudinal Modeling

Since aggregate modeling of participatory data involves data from several individuals over a period of time, longitudinal modeling is the ideal approach for modeling participatory data. Furthermore, the functional nature over large intervals of time of the covariates involved in the asthma data support the use of functional longitudinal methods. Guo discusses the use of functional mixed effects models using smoothing splines, its relation to specific cases, and its applications [27]. Li and Wong examine the selection of covariance patterns for longitudinal data in semi-parametric models [44], an important aspect of longitudinal modeling due to the autocorrelation within subjects. Variations such as the varying-coefficient models discussed by Hastie and Tibshirani allow the estimated coefficients to be time-varying [31]. Other longitudinal methods have been developed which allow for the analysis of sparse data under functional or varying coefficient frameworks.
2.2.2 Sparsity

Sparsity is a prominent characteristic among participatory data. Half of subjects in the asthma data set only have 4 entries throughout 2010 but collectively contribute a large amount of information to aggregate analysis and should be included in modeling. A study by Yao et al. proposed a nonparametric method to perform functional principal components analysis for sparse longitudinal data under the assumption that subjects follow a subject-specific trajectory with some measurement error [73]. James and Sugar develop a model-based procedure for clustering functional data using random effects [36]. They decompose the standard random effects model into components that enable a low dimensional representation of the clusters under normality assumptions.

2.2.3 Mixture Models and Zero-Inflation

Due to the complicated structure of subject usage, simple distributional assumptions are insufficient and mixture models may be more appropriate. Many are familiar with Diane Lambert’s Zero Inflated Poisson model where count data is considered to originate from a mixture distribution of 0 point mass and a Poisson distribution with respective probabilities \( p \) and \( 1 - p \) [43]. Hall and Min and Agresti both extend this model by adding random effects. Hall also extends the framework to binomial data with random effects [29, 50]. Vermunt and Magidson examine a hierarchical extension of the finite mixture model for analysis of nested data structures [70]. This may be useful for the asthma data where there are groups of individuals each consisting of a different type of mixture distribution. Methods involving over-dispersion and Poisson regression are discussed in Berk and MacDonald [5]. A separate topic utilizing mixture modeling methods
is group-based trajectory modeling (Nagin [51]), where growth or finite mixtures may be used to make inferences on group trajectories.

2.2.4 Other Useful Topics

As will be seen in later sections, usage counts of the asthma application are extremely right skewed, making it difficult to accurately model. Distributional assumptions may be made via mixture models as an attempt to capture this skewness, but alternative methods may be better in terms of accuracy and simplicity. Quantile regression [41] is a great way of capturing trends at the outer tails of a time series, especially when dealing with rare events. Additionally, trends in the upper quantiles of highly skewed distributions may be easier modeled by their variances. Hedeker et al. use mixed-effects location-scale models to model variability in EMA data [32, 33].

2.3 Remarks

The past research on the topics previously mentioned may be combined to model the complicated structure of participatory data. The real-time characteristics of participatory data are similar to the structure of data from EMA studies, but the sparsity, complex distributional assumptions, and functional attributes of covariates of interest provide additional modeling challenges. Aggregate modeling accounting for these challenges can contribute to larger scope studies in the fields of epidemiology. N-of-1 studies supply a valuable starting point for providing individual specific feedback to subjects who are frequently engaged with mHealth applications, but departures of participatory data from the structured nature of N-of-1 trials require extensions to these methodologies. Moreover, subject specific inferences made from aggregate models may be compared with that of the single subject analysis to identify the optimal method for feedback.
CHAPTER 3

The Data

3.1 AsthmaMD

Before describing the actual data set used in the analysis, it will help to gain an understanding of the structure and capabilities of the mHealth application providing the data streams. The participatory wellness app of interest is an asthma management mHealth application called AsthmaMD which allows a user to input information pertaining to their asthma symptoms, view visualizations of their condition over time, and send vital information to their clinician for feedback and support completely from their smart phone. User data is stored in the form of a diary to be sent to a clinician. The diary allows users to view their entire temporal profile through charts and graphs, enabling them to recognize patterns between asthma triggers, symptoms, and exacerbations. Among the types of data being entered are peak flow meter readings (a measure of lung capacity), factors causing exacerbations (exercise, pollen, perfume etc.), symptoms (cough, shortness of breath etc.), age, height, and gender. Users also have the option of entering information on their medications, dosage, and frequency of use, however, these data are not in the raw data set but are anticipated to be included in future data sets. In addition to the subject entered data, each subject is assigned an anonymous subject ID and observations for each subject are geo-coded and time-stamped.

In the following analyses, the goal is to familiarize the reader with the structure of participatory data and declare any arguments for using the subsettted data
3.2 Data and Descriptive Statistics

A single observation from the AsthmaMD data frame can be seen in Table 3.1. It is immediately apparent that not all subjects have complete observations. Several subjects are missing symptom and trigger entries and a large number of subjects are missing their latitude-longitude locations. Subject height was either entered in centimeters or inches and was converted into inches using a clear split in the height distributions. Table 3.2 gives a series of summary statistics after the data frame was subsetted to only include reports observed during 2010. The reason for subsetting the data in this way is due to an upgrade in the AsthmaMD application toward the end of 2010. As a result, our data only include subjects who did not consent to the upgrade. There are a large number of reports (30,811) coming from 1,691 subjects with slightly more females than males and the majority of the subjects being adults. The standard deviation of total counts per subject (49.27) is enormous with the median number of reports being 4 and the maximum number of reports being 676, suggesting a highly right skewed distribution. The is due to almost half of subjects engaging with the application once or twice and halting use thereafter. Figure 3.1 illustrates this skewness in a histogram using log number of overall reports per subject. There is an inflation of zeros from the large num-
ber of one-report subjects followed by the smooth right skewed count distribution.

<table>
<thead>
<tr>
<th>Obs#</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>User ID</td>
<td>46024</td>
</tr>
<tr>
<td>Date of Birth</td>
<td>1980-06-11-12-00-00+0000</td>
</tr>
<tr>
<td>Gender</td>
<td>MALE</td>
</tr>
<tr>
<td>Height (cm/in)</td>
<td>173</td>
</tr>
<tr>
<td>Join Date</td>
<td>2010-01-10-13-40-07+0000</td>
</tr>
<tr>
<td>Country</td>
<td>DE</td>
</tr>
<tr>
<td>ZIP Code</td>
<td>22453</td>
</tr>
<tr>
<td>Record Date</td>
<td>2010-01-09-04-10-08+0000</td>
</tr>
<tr>
<td>PFM</td>
<td>420</td>
</tr>
<tr>
<td>Max PFM</td>
<td>637</td>
</tr>
<tr>
<td>Triggers</td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>Worse</td>
</tr>
<tr>
<td>Symptom</td>
<td>Can do usual activities</td>
</tr>
<tr>
<td>Lat Coordinate</td>
<td>53.61</td>
</tr>
<tr>
<td>Long Coordinate</td>
<td>9.98</td>
</tr>
</tbody>
</table>

Table 3.1: Single observation from AsthmaMD data set.

### 3.3 Locations of Subjects

Figure 3.2 shows the worldwide distribution of AsthmaMD users. Since subjects have multiple report locations, subject locations were calculated by taking the average of their entire longitude and latitude entries. Any subjects without latitude-longitude coordinates were omitted. We can see that the majority of subjects are located in western Europe and the United States with Great Britain.
Table 3.2: Simple descriptive statistics of AsthmaMD data. Note the large number of missing entries and the skewness of usage.

<table>
<thead>
<tr>
<th>Description</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of reports</td>
<td>30,811</td>
</tr>
<tr>
<td>Total number of subject IDs</td>
<td>1,691</td>
</tr>
<tr>
<td>Number of females</td>
<td>971</td>
</tr>
<tr>
<td>Number of subjects under 18</td>
<td>334</td>
</tr>
<tr>
<td>Median number of reports per subject</td>
<td>4</td>
</tr>
<tr>
<td>Max number of reports for subject</td>
<td>676</td>
</tr>
<tr>
<td>Standard deviation of total reports per subject</td>
<td>49.27</td>
</tr>
<tr>
<td>Number of subjects missing lat-long coordinates</td>
<td>452</td>
</tr>
<tr>
<td>Number of missing trigger entries</td>
<td>25,769</td>
</tr>
<tr>
<td>Number of missing symptom entries</td>
<td>19,099</td>
</tr>
</tbody>
</table>

Figure 3.1: Histogram of the log total reports per subject showing zero-inflation with a highly skewed right tail.
appearing to have the highest density of subjects per unit area.

Figure 3.2: Global spatial distribution of users showing highest density of subjects in United States and western Europe.

A closer look at the spatial distribution in western Europe (Figure 3.3) verifies this claim with most subjects located in Germany and Great Britain. Figure 3.3 also displays subject locations for the United Kingdom, along with all cities containing populations exceeding 450,000. Intuitively, population is commensurate to number of subjects using AsthmaMD. Figure 3.4 shows the spatial distribution of subjects for the United States. The cities displayed contain populations greater than 1.5 million and appear to be cluster centers for many AsthmaMD users. The maps in Figure 3.5 show the spatial distribution of subjects for each of the four mainland US time zones (states containing 2 time zones may be entirely included in one plot). The number of subjects in the pacific and mountain time zone regions are sparse compared to that of the central and eastern time zones. The Asthma and Allergy Foundation of America (AAFA) performed a study that ranks the worst cities according to their pollen scores, allergy medications taken, and number of allergy specialists per patient in its 2011 asthma report [3]. Cities located in the central and eastern United States are among the highest ranked, potentially explaining the increase in the number of subjects in these areas. This
conjecture is very important because it shows that this participatory data and the very use of the application tend to align with conclusions and reports made by organizations and controlled studies.

Figure 3.3: Spatial distributions of western Europe and United Kingdom. Most subjects are located in Great Britain and Germany.

Although AsthmaMD has users in over 40 countries, more than 20 have only one user. Figure 3.6 is a bar plot showing the number of subjects for regions containing at least 10 subjects and Figure 3.7 is a bar plot showing the total counts for regions that have at least 100 total entries. The United States and Great Britain have the largest number of subjects, containing over 80% of the user population. The bar plot of total reports per region shows similar results. Great Britain has the largest number of counts per unit area of land and appears to have more consistent application usage, therefore, data from Great Britain during 2010 will be used in further analysis and modeling in Chapters 4 and 5.

3.4 Usage by Major Cities

In addition to the spatial distribution of subjects, the temporal usage trend of subjects is also of interest. Exposing and accounting for as much latent spatiotemporal
Figure 3.4: Spatial distribution of United States showing a higher density of subjects in the Midwest and East Coast.

Variation is important in modeling to reduce error and model complexity. Figures 3.8-3.9 are plots of the application usage of subjects in seven major cities around the world over 2010. Samples of 50 subjects were taken from a population within an arbitrarily specified radius of each major city. Each horizontal set of points is a subject with each point representing a report by that subject. Subjects are ordered so that subjects higher on the y-axis have more reports. We see that the majority of the subjects have less than five reports, but the subjects displaying heavier usage have clusters of recurring usage, possibly conveying the separation between subjects having chronic or episodic asthma. London appears to have both short and long bursts of seasonal activity spread throughout different parts of the year. Dusseldorf has a subject using the application year round and has generally strong usage during late spring, possibly due to pollen or pollution levels. Other cities have similar temporal distributions with seasonal usage occurring at different times of the year.
3.5 Subjects with Heavy Usage

A peak flow meter reading is an important piece of information immediately notifying a subject on the condition of his or her asthma and breathing, therefore, observing peak flow meter values, along with subjects’ weekly report counts, can offer some additional information on the legitimacy of using report counts as an indicator for aggregate asthma health. Unfortunately, we do not have data for when subjects last took their medications, which would bias the value of the peak flow meter reading. Subjects in the middle of an exacerbation will tend to take their medication before engaging with the application. As a result, report counts may be a better alternative to peak flow meter readings for aggregate modeling as
Figure 3.6: Regions containing more than 10 subjects.

Figure 3.7: Regions having more than 100 reports.
Figure 3.8: Reports over time for London for two samples. Plots exhibit seasonal trends and sparsity.

It only records the instance of a record, however, subjects with heavier usage have extensive peak flow meter profiles and these may be used for individual feedback.

Figures 3.10-3.11 are time series of peak flow meter readings and weekly report counts for subjects with large total report counts. The red line in the first plot for each figure is the Lowess smooth of peak flow meter readings versus time. Both curves in the second plot of each figure are also Lowess smooths. Report counts were calculated by summing the number of counts each week during 2010. Rug plots are included to show the density of points at a point in time. There are several different temporal profiles with many appearing to be seasonal. Environmental factors such as pollution, pollen, and weather may help to explain this temporal variability. The rug plots indicate that some subjects tend to use the application when their peak flow meter readings are low and usage subsides as their condition improves. Although there appears to be a small inverse correlation between peak flow meter readings and report counts, there are several cases of positive or no correlation at different intervals of time, suggesting that report counts can be partly generated by a separate process. This is under the assumption that a peak flow meter reading is an indicator of a subject’s condition at a point in time and that the reading is accurate (may be biased as previously
Figure 3.9: Reports over time for various cities. Seasonal trends and sparsity are apparent. Profiles with highly consistent usage across the year, as seen in DC and Dallas, are great candidates for individualized analyses.
described). Other wellness indicators may exist which might explain the variation in the amount of usage. Despite weak correlation between peak flow meter readings and report counts, a mere “blip” or increase in usage should be a sufficient indicator for increases in asthma exacerbations in a specific area.

Figure 3.10: PFM readings and weekly report counts for the top heavy users in Great Britain. Some correlation is visible between report counts and PFM values for this user, but many subjects do not exhibit this pattern. Report counts may be of better use in aggregate analysis.

3.6 Age-Gender Analysis

Age and gender are important benchmark variables in most analyses and are most certainly relevant in this analysis, especially in the context of peak flow meter readings, where age and gender are necessary functional inputs to determine healthy peak flow meter values. Additionally, looking at age and gender in the context of application usage can help mHealth application creators better understand their consumer base, allowing them to cater their product to the appropriate demographic for higher quality feedback and enhanced app features.

Figure 3.12 is a mosaic plot of the distribution of the three countries with the highest total reports against ten disjoint age groups. Cuts are at ages 0, 5, 10, 15, 18, 25, 35, 45, 55, 65, and 100 and were determined using the 2010 Adult and Child
Figure 3.11: PFM readings and weekly report counts for the top 3 heavy users in the United States.
Data Prevalence tables by the Center for Disease Control and Prevention (CDC) [8]. The vertical length of each box is the proportion of subjects (among the three countries) in the corresponding age group given a country. The horizontal length is the proportion of subjects in each country. The United States has the most users with many of its subjects in the 25-35 and 35-45 age groups. The number of subjects shrinks for the older age groups. Great Britain has a total number of subjects comparable to the United States with most subjects between the ages of 18 and 45. The youngest age group is much smaller than that of the US and Germany. Germany has the smallest number of subjects of the three countries with most subjects in the 25-45 age range.

![Subject Distribution of Top Three Countries by Age](image)

Figure 3.12: Mosaic plot of age and country.

A boxplot of log number of reports for the same age partition can be seen on the left in Figure 3.13. The right skewness of the distribution for most age bins is visible with their upper quantiles much larger than their respective medians.
Wider boxes indicate a higher number of subjects in that age bin relative to other age groups. Counts are higher for ages 5-10 and drop down for ages 10-18 and there is a gradual increase in the number of counts continuing on to the oldest age bins. Ages 35-55 appear to have the highest median log number of reports. The last two age bins have a combined total of less than 100 subjects where there are zero subjects between 80 and 90 and one subject who is 91. The number of subjects for each age bin is seen on the right in Figure 3.13 where ages 25-45 are shown to have the highest number of reports and the older two age groups are quite sparse. This may be due to the lack of technology familiarity of the older generation.

Figure 3.13: Boxplot of log report counts vs age (left) and counts of subjects in each age bin (right). A cubic-like trend is apparent between log report counts and age.

A subset of the 2010 child and adult asthma prevalence tables is shown in Table 3.3. Each observation is a summary of prevalence rates, standard errors, and confidence intervals for the entire United States where the prevalence rate is the rate of the number of individuals in a given population with asthma over the total number of individuals with or at risk for asthma. Most of the age groups
have approximately equivalent prevalence rates except for a low rate for ages 0-4 (5.90%) and a high rate for ages 18-24 (10.30%). Standard errors decrease as age increases most likely due to the increasing sample sizes for increasing age. Figure 3.14 shows a bar plot of these prevalence rates. The low prevalence for ages 0-4 agrees with the low number reports to the mHealth application, however, the high prevalence rate for ages 18-24 is contradicted by their low number of app reports. This may be explained by the latent effect of the desire to manage one’s own health. Group 0-4 may have a low report count due to its low prevalence in this age bracket. Although asthma is much more prevalent for ages 18-24, young adults may have less of a desire to or have not been properly informed of how to manage their asthma. Given that this theory is correct, implications to public health may be to implement campaigns to improve young adult asthma health management. AsthmaMD may focus more on the middle age bracket as these individuals appear to be their primary consumer.

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Sample Size</th>
<th>Prevalence Rate (%)</th>
<th>SE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>15,111</td>
<td>5.90</td>
<td>0.45</td>
<td>(5.1 - 6.9)</td>
</tr>
<tr>
<td>5-9</td>
<td>16,661</td>
<td>9.80</td>
<td>0.44</td>
<td>(9.0 - 10.7)</td>
</tr>
<tr>
<td>10-14</td>
<td>19,275</td>
<td>9.40</td>
<td>0.44</td>
<td>(8.6 - 10.3)</td>
</tr>
<tr>
<td>15-17</td>
<td>14,839</td>
<td>9.00</td>
<td>0.44</td>
<td>(8.1 - 9.9)</td>
</tr>
<tr>
<td>18-24</td>
<td>12,106</td>
<td>10.30</td>
<td>0.44</td>
<td>(9.4 - 11.2)</td>
</tr>
<tr>
<td>25-34</td>
<td>34,141</td>
<td>8.60</td>
<td>0.25</td>
<td>(8.1 - 9.1)</td>
</tr>
<tr>
<td>35-44</td>
<td>57,449</td>
<td>8.10</td>
<td>0.20</td>
<td>(7.7 - 8.5)</td>
</tr>
<tr>
<td>45-54</td>
<td>84,547</td>
<td>8.50</td>
<td>0.16</td>
<td>(8.2 - 8.8)</td>
</tr>
<tr>
<td>55-64</td>
<td>100,611</td>
<td>9.30</td>
<td>0.16</td>
<td>(9.0 - 9.6)</td>
</tr>
<tr>
<td>65+</td>
<td>149,024</td>
<td>8.10</td>
<td>0.12</td>
<td>(7.8 - 8.3)</td>
</tr>
</tbody>
</table>

Table 3.3: Subset of the 2010 Child and Adult Asthma Prevalence Tables.

In order to formally assess statistical significance, a modified Tukey-Kramer
Figure 3.14: CDC 2010 asthma prevalence rates by age group. Teenagers have high prevalence but low usage.

Pairwise multiple comparison test for unequal sample sizes and variances was run to determine any significant differences between counts in each pair of age groups [15]. Group 0-5 was found to be significantly different from group 35-45 at the 5% level, and groups 10-15, 15-18, and 18-25 were significantly different from ages 35-45 and 45-55. Adolescent and young adult subjects tend to use the application less than middle-age subjects, probably due to adolescents being under parental care or their lack of incentive to manage personal health as previously mentioned. Middle-age subjects may have higher usage due to aging or sedentarism.

Nonparametric quantile regressions are run to further explore the relationship between application usage and age at different quantiles [42] and are shown in Figure 3.15. The bottom, middle, and top lines represent the regression of log report counts on age using 5%, 50%, and 95% quantile smooths, respectively. Subjects over 70 years of age are omitted to avoid bias in estimation. The previously discussed trend of low adolescent counts and high middle-age counts are
supported, especially for the higher quantiles, where the pseudo-cubic trend is more pronounced. Lower quantiles are not as informative due to the extreme right skewness of the data. This fit suggests that quantile regressions are of interest when attempting to expose patterns visible at upper quantiles.

Figure 3.15: Nonparametric quantile regressions of log report counts vs age exhibiting trends at the upper quantiles.

Figure 3.16 shows a boxplot of log report counts by gender. There appears to be no difference in the number of reports between males and females. The female boxplot is almost entirely encompassed by the male boxplot which has higher variability. A two-sample t-test for unequal variances suggests that log reports for males and females are significantly different at the 10% level (pval=0.078). This weak statistical significance is due to the strong skewness of each distribution exhibiting large influence on their respective means. Needless to say, gender is still an important factor when performing peak flow meter based analyses.

Despite little or no difference in usage between different genders, usage among males and females can differ for separate age groups. Figure 3.17 is a mosaic plot
displaying subject proportions of males and females among different age groups. The majority of subjects are within ages 25-45 while other age groups seem to contain about the same number of subjects. The distribution of males and females is nearly the same in all groups except for groups under 25 years of age. Subjects using the application who are in the youngest age groups have a much higher proportion of males. This proportion decreases below the female proportion until ages 25-35, where male and female proportions approximately equalize. An ANOVA regressing log report counts on age, gender, and age*gender produced statistically significant results at the 5% level for all factors. This age-gender interaction will be important in explaining variability during modeling.

3.7 Comments on Using Great Britain

The purpose of Chapter 2 was only to provide the reader with a preliminary look at the structure of participatory data and general usage patterns of the mHealth
Figure 3.17: Mosaic plot of age vs gender. Proportion of females increases with increasing age.
application. The data that will be used for the remainder of the paper only includes reports from Great Britain during 2010. Great Britain has the highest density of subjects and usage is relatively consistent across the majority of the region, providing more power for modeling. Additionally, Great Britain has widely available pollution and weather data, variables highly relevant to modeling asthma exacerbations of subjects. Preliminary analysis of these environmental factors are discussed in Chapter 4.
CHAPTER 4

Current Research

4.1 Pollution Exploratory Analysis

Pollution is a known causal factor of asthma exacerbations. The Natural Resources Defense Council states that the primary pollutants affecting individuals with asthma are ground level ozone, sulfur dioxide (SO\textsubscript{2}), particulate matter (PM2.5 or PM10), and nitrogen oxides (NO\textsubscript{x}) \cite{52}. Hourly data on these five pollutants in Great Britain were obtained from the Department for Environment, Food, and Rural Affairs (DEFRA). The data contain monitor locations, monitor names, dates and times of entry, and levels for each pollutant. The levels of the air pollutants are converted into air quality indexes (AQI) as they are a more effective way of representing the association between pollution levels and pollution related health risks. Defined air quality indexes for each pollutant are located on the DEFRA website and are partitioned in the following way: 1-3 is low health risk, 4-6 is moderate, 7-9 is high, 10 is very high. Individuals with asthma are at risk for exacerbations with AQI values as low as moderate (4-6). Since distributions of pollution and counts are extremely right skewed, upper quantiles of their distributions are more appropriate for exploring their trends and correlations.

The left plot of Figure 4.1 shows a quantile regression fit for the 98\% and median quantiles of overall daily AQI over time. Overall daily AQI was calculated by taking the median of all five pollutants for each day of 2010 for each of the 23 pollution collecting monitoring sites collecting data on all five pollutants. Other
variations were attempted (different quantiles, using all pollutants separately etc.) but trends using these variations were not found. The plot shows that there is a gradual decrease in the upper quantiles of AQI over time with a sharp increase toward the end of the year. The right plot of Figure 4.1 shows the 90% and median quantile fits of weekly report counts over time. Weekly report counts were found by counting the number of reports within a specified radius of one of the 23 pollution collecting monitoring sites. There is a gradual decrease in the 90th quantile of the number of reports and a flat trend toward the end of the year, but this is highly variable. The flat trend is due to the large number of zero counts, however, higher quantiles suggest an increasing trend toward the end of the year. Figure 4.2 shows the 98% quantile fit for AQI and 90% quantile fit for reports on the same plot. Overall trend directions for AQI and reports align with the exception of the terminal values where AQI drastically increases and counts level off. Figure 4.3 shows boxplots displaying relationships between report counts and AQI for daily and weekly partitions of AQI. AQI is divided into three or four parts, each displaying a level of pollution severity. Note that the magnitude within the partitions does not exactly reflect the levels defined by DEFRA and these AQI levels are a result of median calculations. Both daily and weekly AQI partitions show a positive correlation between AQI and report counts, suggesting subjects tend to use the application more during periods of higher pollution.

4.2 Weather Exploratory Analysis

Temperature, wind, and humidity are climatic factors that may affect asthma symptoms. Weather data on these climate measures was obtained from Weather Underground’s database of hourly weather reports from airport locations in Great Britain. The following analyses attempt to understand time trends of weather variables and how they relate to subject reports. Report counts were found by
Figure 4.1: Quantile fits of AQI over time (left) and quantile fits of counts over time (right). Trends are more visible at the upper quantiles.

Figure 4.2: Quantile fits of AQI and report counts over time.
Figure 4.3: Boxplot of report counts vs daily (left) and weekly (right) median AQI values. There is evidence that report counts are positively correlated with air quality. Counting the number reports within a specified radius of each weather data collecting location for every climate factor during each day of 2010.

Figure 4.4 is a time series of temperature over 2010 with a Lowess smooth indicated in red and displays the standard seasonal trend. The left plot of Figure 4.5 is a scatterplot of the number of daily report counts against daily median temperatures along with the median and 95% quantile smooths. The vertical bars contain the middle 95% of temperature values. Daily report counts corresponding to temperature monitoring sites were calculated in the same manner as the report counts for pollution. The median of the smooth is strongly influenced by the inflation of zero counts. Colder temperatures (30-40 °F) tend to have larger numbers of counts and higher variability. There is a downward trend in report counts toward average temperatures and an upswing toward the higher temperatures. Extremely low median temperatures have very small counts, likely due to the smaller number of engaged subjects or smaller populations in areas with such low temperatures. Figure 4.5 shows another representation of counts versus temperature showing the higher upper quantiles for low and high temperatures.

Figure 4.6 shows the time series of wind and report counts against daily median
Figure 4.4: Time series of temperature with Lowess smooth exhibits commonly seen annually cyclic trend.

Figure 4.5: Scatterplot of report counts vs temperature (left) and boxplot of report counts vs temperature (right).
wind values with the median and 95% quantile smooths, respectively. The Lowess smooth in the times series suggests average wind speed remains fairly constant over time but with high variability within and across time. In Figure 4.6 on the right, the median trend, like for temperature, is strongly influenced by the right skewness of the count distribution. Increasing wind speeds suggest decreasing counts for upper quantiles. Higher wind speeds may blow pollution away from highly populated cities whereas lower wind speeds can trigger allergy symptoms and may explain the upward trend in the upper quantiles for lower wind speeds. This may be investigated through the interaction of wind and discretized parts of Great Britain. Finally, Figure 4.7 is the time series for humidity and scatterplot of report counts vs humidity, respectively. Average humidity is lower during the late spring and early summer months. The red lines in the scatterplot show the 99.9%, 95%, and median quantile smooths. The lower quantiles have a downward trend, but variability increases for increasing humidity. The extreme upper quantiles display this increase in variability with a positive correlation between counts and humidity. The 99.9% quantile aligns with claims that humidity is a cause of asthma exacerbations. Quantile trends that are in agreement with claims made through scientific studies may be used in modeling aggregate trends of asthma.

Figure 4.6: Time series of wind speed with Lowess smooth (left) and scatterplot of report counts vs wind speed (right)
4.3 Modeling with Pollution and Weather

Using the exploratory analysis on pollution and weather, a model is run to quantify the relationship between aggregate localized report counts and pollution and weather factors based on the paper by Peng et al. [53]. Equation 1 shows the form of the model assuming the response has a Poisson distribution.

\[
Y_t \sim \text{Poisson}(\mu_t) \quad (1)
\]

\[
\log(\mu_t) = \sum_{i=1}^{4} \beta_i x_{it} + f(t) + q(\text{temp}_t)
\]

The \(x_{it}\)’s are air quality index, wind, wind\(^2\), and humidity. \(f(t)\) and \(q(\text{temp}_t)\) are non-parametric functions of time. Time and temperature are included as smoothing parameters in order to adjust for seasonal and long term trends. This accounts for the confounding effect between pollution and time and temperature [14]. The model is fit using penalized regression via the `gam` function in the `mgcv` package in R. Degrees of freedom for the cubic spline terms (time and temperature) were set to 50, much larger than needed (but smoothed using penalized methods) to ensure the removal of seasonal and long-term variation in pollution [53].

Tables 4.1-4.2 display the results from the model fit for the parametric and
non-parametric terms, respectively. All predictors are very significant at the 5% level. $e^{\beta_{AQI}} = 1.36$ suggests the number of counts for a given day are 1.36 times larger than days with one unit smaller AQI. Note that this is a one unit increase in the AQI averaged across a week. This result supports the positive correlation seen in the boxplots. The significance of the wind parameters implies that there is a linear relationship between log counts and a quadratic wind trend such that counts tend to decrease for higher wind speeds. $e^{\beta_{Humidity}} = 0.99$ suggests the number of counts on a given day are 0.99 times smaller than on a day with one unit smaller humidity, however, although significant, this effect is minuscule. Figure 4.8 shows plots of the smoothing terms with the empirical counts over time. They closely resemble the empirical time series with a large amount of variability conveyed by temperature.

|                  | Estimate | Std. Error | z value | Pr(>|z|) |
|------------------|----------|------------|---------|----------|
| (Intercept)      | 0.82     | 0.073      | 11.22   | < 2e-16  |
| AQI              | 0.31     | 0.012      | 26.85   | < 2e-16  |
| Wind1            | -6.24    | 0.79       | -7.92   | 2.32e-15 |
| Wind2            | -3.35    | 0.82       | -4.081  | 4.49e-05 |
| Humidity         | -0.0082  | 0.00086    | -9.49   | < 2e-16  |

Table 4.1: Parametric results of GAM model. Model shows positive correlation between air quality and report counts. Significance of wind parameters support quadratic-like effect. Humidity is significant but has a very small effect.

In order to understand the goodness of fit by the additive model, it helps to view a scatter plot of the residuals versus fitted values and the distributions of the fitted and actual report counts. The residuals in the left plot of Figure 4.9 exhibit a highly variable fanned out trend for smaller fitted values and decreases in magnitude and variability as the fitted values increase. This is due to the incorrect Poisson distributional assumption made on the report count data (the necessity of
Table 4.2: Non-parametric results of GAM model. Both time and temperature are very significant in the fit.

<table>
<thead>
<tr>
<th></th>
<th>edf</th>
<th>Ref.df</th>
<th>Chi.sq</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>s(Time)</td>
<td>48.94</td>
<td>49.00</td>
<td>5074.1</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>s(Temp)</td>
<td>46.16</td>
<td>48.33</td>
<td>412.7</td>
<td>&lt;2e-16</td>
</tr>
</tbody>
</table>

Figure 4.8: Non-parametric fits

more complicated distributions like mixtures have already been suggested). The large number of weeks with zero counts forces the fitted distribution to be pulled toward the lower range of counts, but the larger value outliers of report counts pull the distribution to the right. This causes the fitted distribution to negotiate between these two influences. The result is a fitted distribution which incorrectly predicts the lower values of report counts. Furthermore, although residuals decrease, the model consistently under-predicts larger report counts. This is due to the smaller skewed right tail of the Poisson which considers these values extremely rare. Non-parametric methods or methods incorporating distributions with extremely skewed tails which can account for these extreme values must be considered. These results are also shown in the histogram of Figure 4.9 where one can see the lack of fit for the zero-inflated portion of the actual data distribution and the model’s failure to make accurate predictions on the right tail. This lack of
fit conveys the need for more complicated methods accounting for highly skewed and zero-inflated distributions and are discussed in Chapter 5.

Figure 4.9: Residuals vs fitted plot (left) and histogram of report counts for the actual data and the fitted values (right) from the GAM model. Both show the lack of fit for a Poisson distributional model assumption and the need to account for zero-inflation and skewness.

4.4 Clustering and Discretizing Great Britain

Due to the heterogeneous nature of subjects’ asthma symptoms and application usage, discretizing Great Britain into disjoint parts may simplify model assumptions and contribute to more accurate model inferences. Below are two methods for dividing subjects into discrete parts, one using partitioning around medoids (pam) [37] to cluster subject locations with pollution and weather factors and the other using a functional clustering model by James and Sugar [36] to cluster individual weekly report count trajectories. Discretizing subjects can offer additional random effects for hierarchical modeling and may provide an extension to N of 1 analyses by clustering subjects into homogeneous groups.

Figure 4.10 shows the result of running the pam function in the cluster package for 10 clusters. The dissimilarity matrix used in the clustering algorithm is a weighted combination of dissimilarity matrices for subject location, pollutants,
weather variables, and elevation. Subject locations were found by calculating the mean latitudes and longitudes for each subject. Average weekly pollution or weather values of the three closest monitors for each subject were used to calculate dissimilarity matrices. Pollutants included ozone, NO$_2$, PM2.5, PM10, and SO$_2$ and weather variables included temperature, wind, and humidity. The number of clusters was chosen using visual inspection of clustering results. The weights used in the weighted dissimilarity matrix are 0.5 for location, 0.4/5 for each pollutant, 0.1/3 for each weather variable, and 0 for elevation. Some dissimilarity matrices were multiplied or divided by a factor of 10 to maintain a consistent order of magnitude across all factors, making the weighting process more tractable. Elevation was given a weight of zero due to the small amount of topographical variation across Great Britain causing poor location-wise clustering. Given the weights, the clustering produced desirable results. Subjects are divided first by location and then by micro-climates, implicitly incorporating some of Great Britain’s topography. Figure 4.11 shows a bar plot of the mean of the average report counts per week per subject for each cluster. Notice that there are no neighboring clusters with similar counts, implying a subjective conclusion that the number of clusters is sufficient. More clusters were attempted with undesirable results. The profiles of mean reports per week by cluster are shown in Figure 4.12. The first plot is saturated with each profile but gives a sense of the variability and common trend of each cluster. The heterogeneity defined by the separate spikes in usage is of interest. The second plot separates the profiles over the y-axis with clusters one through ten represented by the lower through upper profiles, respectively. There are clear differences in the profiles, each having separate spikes and and flat areas, supporting the effectiveness of using weighted dissimilarity matrices to separate heterogeneous populations in the asthma data.

Using pam produced successful results, but the clustering used subjective approaches to identify the weights and numbers of clusters. Another piece of im-
Figure 4.10: Clustering of Great Britain using weighted dissimilarity matrices of subject locations, pollutants, weather variables, and elevation.
Figure 4.11: Barplot of mean counts per week for each subject by cluster. Clustering effectively explains variability in mean counts per week for subjects in different locations.
Figure 4.12: Mean counts per subject per week by cluster. Unique trends are visible in the profiles of each cluster.

Important information missing from the clustering is the incorporation of subject counts. A model-based approach to clustering subject report count trajectories would be useful in accounting for distributional assumptions made on the asthma data. Model-based clustering is performed using the functional clustering model (FCM) where normality is assumed [36]. Although the counts are obviously non-Gaussian, an initial look into the structure of the algorithm and its results might shed some light on future attempts at modeling. Figure 4.13 shows the profiles of subjects for the months of February through June and April through November after clustering by FCM using two clusters, respectively. The algorithm for the first interval clusters by those who have higher or lower usage and appears to be successful when trajectories are relatively smooth and consistent over time. In contrast to this profile structure, the profiles in Figure 4.13 on the right are more characteristic of participatory data for longer intervals of time, where sporadic spikes followed by stretches of zero use are common. The algorithm attempted to cluster by high and low counts, but the temporal pattern toward the end of the time period influenced the results. Clusters larger than two appeared to be ineffective at improving results. The cluster means shown in both plots in Figure 4.14 are similar, except for the terminal zero mean values of the April-November
interval. An important aspect of the FCM was the ability to visualize the clusters using a projection into a lower-dimensional space. FCM appears to have separated subjects with higher maximum counts as seen in the one-dimensional projection of clusters in Figure 4.15 on the left, but the poor clustering for the April-November interval is reflected by the plot on the right. The profiles of subject-specific trajectories produced by FCM improved with increasing degrees of freedom of the spline basis, however, fitted curves fell below zero and were inaccurate at tails of the curve. Clustering may be improved by introducing new distributional assumptions to the FCM clustering framework composed of zero-inflated mixtures. These and other extensions are discussed in Chapter 5.

Figure 4.13: FCM clustering for February through June (left) and April through November (right). The plot on the right is typical of many long term report count profiles and the FCM clusters these profiles poorly.
Figure 4.14: Cluster means for February through June (left) and April through November (right).

Figure 4.15: Alpha projections for February through June (left) and April through November (right). Clustering on the right plot shows an ambiguous separation of clusters and the need for more complex distributional assumptions for the FCM.
CHAPTER 5

Statistical Directions

Future Research is divided into four major sections. The first lists several possible questions from each of the three usage cases and are used to motivate modeling. Section 5.2 describes general modeling techniques that may be used to answer these questions and provides statistically intensive arguments for their use based on data characteristics. Modeling is given its own section since many of the proposed models may be applied across all usage cases. These approaches can be divided into three subsections: individualized, aggregate, and adherence analyses. Statistical challenges using these methods are also considered. Section 5.3 discusses how the usage cases further relate to modeling approaches described in Section 5.2. Finally, the last section lists some limitations to the analyses and proposed methods described in this paper.

5.1 Usage Cases and Modeling Motivations

5.1.1 Individual Benefit

In order to propose relevant statistical methods for producing feedback for individuals, we must first ask ourselves how a subject would want to benefit from using the AsthmaMD application and the type of content they would like to be provided. These questions may be answered by performing various forms of individualized and aggregate analysis described in Section 5.2. Questions include:

1. What times of the year do my exacerbations most often occur?
2. What environmental factors tend to worsen my asthma condition?

3. Which personal activities help or hurt my asthma?

4. What times of year am I affected by factors the most?

5. Is my current medication better than my previous medication?

6. Am I following the most efficient treatment regimen in terms of dosage?

7. Has my asthma been improving since I started using AsthmaMD?

8. How does my current condition compare with others over the course of a year?

5.1.2 Benefits for Creators of Wellness Applications

Application creators seek to improve their design framework by finding the best possible approaches for supplying subjects with information to help facilitate their asthma management. With this in mind, they ask:

1. Which factors are most relevant to specific individuals?

2. What type of feedback will be most helpful to subjects given an illness?

3. Are there overall increases in asthma wellness for subjects who are more engaged with the application?

4. How do we define whether a subject is being compliant or engaged with the application?

5. What types of visualizations and statistics are more user friendly?
5.1.3 Contributing to the Scientific Community

Aside from subject-specific feedback and application framework design, participatory data contain information on population level trends which may contribute to several aspects of public health. Common questions are:

1. What environmental factors affect subjects with asthma at the population level?
2. At what times of the year are they affected?
3. Which regions are most adversely affected by these factors?
4. How do illness conditions in specific regions compare to the same time in the previous year?
5. How do “pseudo-” prevalence and incidence rates produced by these data compare to those calculated using traditional methods? (I say “pseudo” since these statistics traditional incorporate information on subjects at risk of an illness as well. This is information not included in the data, therefore, relative statistics may be used.)
6. Can treatment efficacy be assessed using these data?
7. Are personal habits, treatments, or symptoms common to a specific area?

5.2 Statistical Directions for Achieving Usage Cases

I propose that the modeling techniques needed to answer the majority of the questions asked above can be divided into three categories. Individualized analysis is concerned with modeling single-subject asthma profiles. Aggregate analysis pools the data using aggregate modeling methods to find population level patterns.
Adherence analysis is concerned with investigating user compliance and may play a role in the other analyses previously described.

5.2.1 Individualized Analysis

The goal of individualized analysis is to provide feedback to subjects on effective treatments, usage of medications, and other factors affecting their illness through the application of statistical methods to their profile history. N of 1 studies have discussed several simple approaches for determining differences in a response (such as PFM readings) at disjoint intervals of an individual’s time series, some using only visual confirmation to determine the significance of a factor of interest. These methods may be applied to a specific set of subjects in the asthma data, mainly those with somewhat heavy and consistent usage over time, but a large number of observed and latent factors must be accounted for in order to ensure the validity of an inference for a particular factor. Hence, time series models are more appropriate for making inferences on asthma related factors since they allow the investigator to control for additional effects such as temperature and pollution. Since PFM, pollution, and weather appear to have functional relationships with time, semi-parametric time series methods including time and/or temperature as a non-parametric component and several other factors as parametric components are being considered. Non-parametric components are included to adjust for confounding effects between time and weather variables and other time-sensitive covariates such as pollution [53]. Additional parametric components would include pollen counts, demographic variables, personal risk factors, and medications. Parametric estimates from these models can provide information on the sensitivity of a subject’s asthma condition to these factors. Peak flow meter readings will serve as the response under a Gaussian assumption. Other model components may include interactions between pollution and time or temperature in order to account for varying seasonal effects on PFM readings.
Time-covariate interactions may reveal trends in the effects of factors on asthma conditions at different portions of the year.

Although the semi-parametric time series model described above may be effective in accounting for seasonal variability and seasonal interactions with factors of interest, the number of required parameters to be estimated is quite high, especially when there are several interactions involved with non-parametric components. A more efficient way of performing this type of analysis is to use varying coefficient models, where time-dependent model parameters can explain effects of factors varying across time or temperature [31]. These are legitimate assumptions for the asthma data since pollution and medications may have varying effects on PFM during the year. In addition to these assumptions, PFM values at specific times may have a dependence on values of factors at other times during the year. The dependence can range from the deep past in the case of personal factors to the recent past in the case of pollutants. For example, when a factor causes an asthma exacerbation, PFM readings decrease and remain at a lower level until the lungs heal or the damaging factor subsides. The lingering effect of these factors suggest that histories of these factor processes may help predict PFM values. The historical functional linear model and recent history functional linear model are most appropriate in these cases [38, 46]. Methods incorporating information during relevant seasons in previous years must also be investigated as seasonal history may help to explain current PFM status. Senturk and Muller (2010) combine aspects of the recent history functional linear model and the varying coefficients model to create a functional varying coefficients model for longitudinal data [64]. This method may be simplified to account for analysis of individual time series.

Thus far, PFM readings have been used as an indicator for the wellness of subjects suffering from asthma. Additional indicators such as the variability of PFM readings represent instability in an asthma condition and may also be considered as a meaningful response. Variances of PFM over time calculated using a time
A binning procedure can be used as a new response, but this procedure treats these variances as fixed values and does not account for uncertainty in the variance estimates. Moreover, a bin size must be determined for each specific user profile as rates of entry and densities of entry vary from subject to subject. Methods extended from location scale models by Hedeker et al. may be used to model the variance in terms of relevant covariates [32, 33].

All of these models may be compared by prediction accuracy, model simplicity, and computational cost in order to find the most appropriate approach to be incorporated into participatory mHealth applications’ data analysis frameworks. Inference from these models should include effects of environmental or personal factors on PFM readings and how they vary over time. Medication efficacy and dosage suggestions at different times of the year may also be inferred. Contrasts producing numerical summaries and functional plots of parameter estimates along with their confidence bands may be used as feedback. Unfortunately, these analyses can only be performed for subjects contributing larger amounts of data consistently across longer intervals of time. As previously mentioned, 50% of subjects entered less than 5 observations during 2010. Although this does not account for the start dates of application memberships, it is observed that sparsity is a dominating attribute of participatory data. Simulations and clustering may be used to find a usage threshold to identify subject profiles with sufficient data to be subjected to these methods. Subjects with insufficient data may receive feedback based on simple summary statistics and asthma conditions of nearby populations through aggregate modeling described in the next section.

5.2.2 Aggregate Analysis

The majority of subjects in the asthma data have ultra-sparse profiles defined by small numbers of observations over separate intervals of time. As a result, these subjects do not supply enough information on their asthma condition to be
modeled individually. A solution would be to use asthma status of nearby populations to estimate a mean trajectory of the sparse subject and the surrounding population to produce feedback for the sparse individual. As described in Section 5.1, the functional nature of the response and factors of interest, as well as the possible varying effects of covariates on asthma wellness over time, require the use of varying coefficients and functional linear modeling methods or a combination of the two. Senturk and Nguyen developed a variant of the varying coefficients model which accounts for sparsity by borrowing information from all subjects in the data [65]. Kim et al. extend recent history functional linear models to account for sparsity in longitudinal data [39]. The functional varying coefficients model [64] must also be explored as recent and seasonal history of predictor processes and factor effects varying with time may have significant effects on PFM. Challenges with these sparsity methods include the different times at which subjects start using the application. This implies that each subject is defined on a different support, complicating the covariance analysis. Analyses can be performed assuming subjects have identical supports and deeming periods before joining the application as sparse or missing, but data on personal levels will be missing completely. Effects on the validity of model inferences under these assumptions may be assessed through simulations. Feedback from these models to the entire population of subjects, including those exhibiting sparsity, can include summaries and trajectories of the asthma wellness of nearby populations, allowing subjects to assess their asthma relative to the status of the surrounding subjects. In addition to user feedback, contributions can be made to the scientific community through real-time updates on the status of asthma in different regions. Trajectories of PFM readings and functional coefficients can shed light on the dynamics of asthma over time among population level factors. Analogues of prevalence and incidence may be created using contrasts of asthma conditions between locations and across time to complement summary statistics commonly used in epidemiological studies.
Analyzing subject report counts over time is another effective way to assess asthma status as a function of population level factors since aggregate level counts are robust to periods of non-usage by subjects. The response of interest for these types of models are the average number of counts per person within a pre-specified radius around several focal points spanning Great Britain. Values of each population level factor must be associated with each focal point. Since factors such as pollutants and weather are collected at several different monitoring sites, there is difficulty in imputing levels for all factors at a single focal point. Smoothing methods may be ineffective due to micro-climates. In this case, wunderground.com may be helpful in providing data on these factors at a high resolution. As seen in Chapter 4, the relationship between population level factors and report counts are extremely skewed, exhibiting visible correlation only at the upper quantiles. I propose two methods for modeling these relationships. The first is under a quantile regression framework. I suggest the use of a hierarchical mixed-effects quantile regression model where random effects are defined for focal points and disjoint regions containing focal points. Alternatively, disjoint regions may be included in the model as fixed effects. The disjoint regions must be decided through clustering algorithms discussed later in this section. The model will be run for several quantiles to explore the varying relationships between report counts and population-level factors. Inferences from these models will provide insight into the effects of environmental variables on the asthma subpopulations of Great Britain over time, assuming that the counts somewhat accurately represent the asthma status of a population. The second method is to use generalized mixed effects location and scale models to model the variances of relevant factors. Figure 5.1 compares quantile trend lines in the scatterplot of report counts vs temperature seen in Figure 4.6 to a trend line defined by empirical report count variability across temperature. The trends are very similar suggesting variance may be a legitimate alternative to quantiles in the case of highly skewed data. The effec-
tiveness of these two approaches may be compared through simulation. Qualitative inferences from these models should be used for epidemiological research as quantities of counts may be meaningless. Creators of applications may find quantitative inferences useful for data management and application efficiency.

Figure 5.1: Quantile and variance trends of counts vs temperature. Empirical variance trend matches that of the trends in the upper quantiles.

In order to maximize the statistical power of the asthma data before modeling, latent spatial and temporal factors explaining variability in PFM and counts must be rooted out using mixture modeling and clustering methods. Clustering subjects based on PFM readings and report counts over time will provide insight into the types of usage profiles that exist in the asthma data. Clusters may include subjects suffering from chronic asthma or episodic asthma and subjects with mild asthma whose compliance is poor. This information is useful when deciding the proper methods for analyzing a specific profile and is especially useful in cases of automated feedback where classifying thousands of profiles by hand is impossible. Thresholds may be decided subjectively, but may be inconsistent due to the ever changing types of usage by subjects. Identification of these clusters can suggest
additional ways for analyzing user adherence and will allow application creators to focus on subjects who would benefit most at a specific time. In Section 4.4, the functional clustering model was used to cluster subject profiles using their report counts per week. Since the FCM assumes normality, the cluster fit was poor for profiles exhibiting a variety of zero-inflated trends over time. I propose investigating different mixture model approaches involving zero-inflated and skewed count distributions to improve clustering of application usage profiles over time. The normality assumption of the FCM may be more appropriate when clustering subject profiles using PFM readings. In addition to temporal clustering, spatial clustering of subjects based on location and several population level factors is important in identifying homogeneous groups of individuals who share similar environmental conditions. The partitioning around medoids clustering using a weighted sum of dissimilarity matrices of relevant factors in Section 4.4 worked relatively well in separating groups with different mean count profiles. These groups may be entered as fixed effects or random effects in a hierarchical model or as indicators within the semi-parametric methods described above. Inferences from these models can serve as localized aggregate feedback, informing subjects of their asthma status relative to the immediate surrounding population. Although pam was successful at separating subjects into groups that are heterogeneous by count profiles, micro-climates across Great Britain suggest clustering should be performed at an even higher resolution. This requires additional research into the spatial breakdown of relevant environmental factors and automated methods for identifying spatial clusters at a much smaller scale.

5.2.3 Adherence Analysis

Aside from the feedback produced by the analyses proposed above, the effectiveness of the application must be assessed in order to validate the use of the application for asthma management. This is a question of great interest by the
creators of AsthmaMD since a relationship between improved asthma conditions and use of the application suggest that participatory mHealth apps may be used as an effective preventive health care tool. Use of the application must first be analyzed under an adherence analysis framework. Adherence of a subject to the application is defined on two dimensions, volume of reports and consistency of entries over time. Consistency does not have to span the entire year as many subjects suffer from episodic asthma. Since data entry reminders are not sent to subjects during the 2010 year, subject profiles should be separated using the clustering methods proposed in Section 5.2. These disjoint groups may be incorporated into the models proposed above and temporal contrasts within groups or across groups may be used to observe increases in PFM or decreases in PFM variability while controlling for predictor levels. Temporal contrasts through individualized analyses can inform a subject of improvements in their condition or offer advice if their asthma fails to improve. Although clustering may be somewhat effective, missing data mechanisms such as missing at random and missing not at random also define the missingness of the data set, suggesting the possibility of false inferences from models not taking this into account. This motivates further exploration into missingness modeling methods. Recently, AsthmaMD has incorporated a report notification system based on the usage of a subject. This additional information in future data sets is predicted to drastically simplify adherence analysis as the absence of reports following notifications are strong indicators of poor compliance.

5.3 How Do the Usage Cases Relate to the Proposed Models?

5.3.1 Individual Benefit

Questions 1-7 of 5.1.1 are primarily concerned with the individual, therefore, the individualized analyses described in Section 5.2.1 are highly appropriate, espe-
cially for subjects with sufficiently dense data. Subjects are interested in knowing which factors, environmental or personal, cause asthma exacerbations. These factors may be included as covariates in each model, however, each approach offers a different type of feedback to the individual. Simpler methods such as semi-parametric time-series analysis may provide single estimates for relevant covariates, suggesting a sensitivity to a particular factor. Interactions with time can provide additional information on the effect of covariates during different portions of the year. These results may be displayed graphically, but these representations may not accurately represent the functional nature of many of the covariates of interest. In this case, varying coefficients models and other functional models are more appropriate as functional representations of parameter estimates more accurately depict the effect of a factor on an asthma condition over time. Contrasts may be investigated to find differences in asthma wellness during different times of the year while controlling for covariate values and to determine the effectiveness of using the application. Additionally, contrasts may also be used to assess differences in response (PFM or PFM variability) under separate treatment regimens, essentially allowing the subject to perform his or her own clinical trial. Note that the historical functional linear models and functional varying coefficients models include a single predictor process. Since a large number of covariates are of interest, the varying coefficients model may be more appropriate. If response processes are also functions of past predictor values, lag values may be included to gauge the extent to which history affects current wellness conditions.

In addition to the feedback from individualized analyses, subjects would like to know how their asthma condition compares to the surrounding population at a given point in time. Aggregate modeling described in Section 5.2.2 may be used to provide summaries on population PFM readings for a given demographic. Additionally, spatial representations of sensitivity to environmental factors such as pollution can supply users with a breakdown of nearby “hot spots” for asthma
exacerbating factors. Since feedback using individualized methods for subjects with ultra-sparse data is not feasible, updates on surrounding aggregate asthma conditions may be used.

5.3.2 Benefits for Creators of Wellness Applications

Creators of wellness applications can benefit from all of the analyses described in Section 5.2. Individualized and aggregate analyses can inform creators on which factors are most relevant to specific individuals and regions during portions of the year. Moreover, these analyses will inform them on the effectiveness of the application at helping to improve asthma management and wellness, possibly using contrasts over time. Using adherence analysis, creators can separate subjects with dense profiles from subjects who are not engaged with the application, allowing for more efficient data management and improved user feedback. Additional methods for adherence analysis may be explored once data including entry notification is obtained. Aside from the content of feedback, how feedback is represented is of utmost importance. Since data visualization can vary for different modeling techniques, modeling strategies, the information they offer, and methods for transforming their results into user friendly representations must also be considered. By combining all aspects of these analyses, mHealth app creators will be provided with a foundational framework structure for improving and creating new applications for personalized preventive health care.

5.3.3 Contributing to the Scientific Community

Inferences from many of the aggregate analyses described can be used to complement epidemiological studies. Spatial and temporal clusters incorporated into aggregate models (as simple as indicators) can offer information on seasonal and spatial trends of an illness in relation to environmental and climatological stres-
sors. Meta-analyses of medication regimens using random effects or functional methods can provide insight into treatment efficacy at the population level. Using aggregate modeling techniques, real-time analogues to the traditional prevalence and incidence rate statistics can be used to flag areas of concern through contrasts and ratio-based statistics across regions of interest. Collectively, these results can provide the scientific community with a profile of illnesses in several regions in real-time.

5.4 Limitations

It should be noted before moving on to the next chapter that there are several limitations to the analyses performed in Chapter 4 and the methods proposed in Chapter 5. Participatory data contain a large amount of noise and there is a large amount of additional data (pollen counts, individual habits, etc.) needed to provide accurate inferences from the signal in the data streams. Moreover, even when provided with additional data, statistical models and their inferences may only be applied to the iPhone AsthmaMD subject population. Consult with public health researchers within this field of study must be sought before attempting to generalize any inferences from these methods to subjects suffering from asthma within the general population. In addition to these limitations, I remind the reader that the analyses performed here are in their beginning stages and the purpose of their results are only to provide preliminary insight into the structure of participatory data and potential methods for modeling them. That being said, after extensive research and testing of the proposed modeling methods mentioned throughout this paper, it is my hope that conclusions from these techniques may be used as supplemental tools for individualized health management and epidemiological studies.
CHAPTER 6

Closing Remarks

The potential implications of AsthmaMD data to individuals, the designers of participatory mHealth applications, and the scientific community are plentiful. Feedback on the asthma conditions of individuals and surrounding populations based on the analyses proposed enable subjects to make informed decisions on how to manage their asthma health. Determining factors relevant to an asthma analysis for a specific group or region will help app creators improve the quality of feedback with high resolution statistical summaries. App creators will also benefit from conclusions produced by adherence analysis, helping to determine application efficacy for illness management to increase mHealth app popularity for preventive health care. Inferences made by aggregate models can provide real-time updates on the status of asthma across regions in relation to population level factors, providing the scientific community with additional tools for epidemiological analysis and clinical trials. The use of participatory mHealth apps, evidenced by the application of statistical methods proposed here, may deem mHealth a cost-effective approach to preventive health care, motivating individuals to take action in managing personal health.
References


