METHODS FOR CLUSTERING TIME SERIES DATA ACQUIRED FROM MOBILE HEALTH APPS

NICOLE TIGNOR[1], PEI WANG[1], NICHOLAS GENES[1,2], LINDA ROGERS[3], STEVEN G. HERSHMAN[4], ERICK R. SCOTT[1], MICOL ZWEIG[1], YU-FENG YVONNE CHAN[1,2], ERIC E. SCHADT[1]

[1]*Department of Genetics and Genomic Sciences*
Icahn School of Medicine at Mount Sinai, New York, NY, USA

[2]*Department of Emergency Medicine*
Icahn School of Medicine at Mount Sinai, New York, NY, USA

[3]*Department of Medicine, Pulmonary, Critical Care and Sleep Medicine*
Icahn School of Medicine at Mount Sinai, New York, NY, USA

[4]*LifeMap Solutions, Inc, New York, NY, USA*

Email: pei.wang@mssm.edu, eric.schadt@mssm.edu

In our recent Asthma Mobile Health Study (AMHS), thousands of asthma patients across the country contributed medical data through the iPhone Asthma Health App on a daily basis for an extended period of time. The collected data included daily self-reported asthma symptoms, symptom triggers, and real time geographic location information. The AMHS is just one of many studies occurring in the context of now many thousands of mobile health apps aimed at improving wellness and better managing chronic disease conditions, leveraging the passive and active collection of data from mobile, handheld smart devices. The ability to identify patient groups or patterns of symptoms that might predict adverse outcomes such as asthma exacerbations or hospitalizations from these types of large, prospectively collected data sets, would be of significant general interest. However, conventional clustering methods cannot be applied to these types of longitudinally collected data, especially survey data actively collected from app users, given heterogeneous patterns of missing values due to: 1) varying survey response rates among different users, 2) varying survey response rates over time of each user, and 3) non-overlapping periods of enrollment among different users. To handle such complicated missing data structure, we proposed a probability imputation model to infer missing data. We also employed a consensus clustering strategy in tandem with the multiple imputation procedure. Through simulation studies under a range of scenarios reflecting real data conditions, we identified favorable performance of the proposed method over other strategies that impute the missing value through low-rank matrix completion. When applying the proposed new method to study asthma triggers and symptoms collected as part of the AMHS, we identified several patient groups with distinct phenotype patterns. Further validation of the methods described in this paper might be used to identify clinically important patterns in large data sets with complicated missing data structure, improving the ability to use such data sets to identify at-risk populations for potential intervention.

1. Introduction

Handheld mobile devices such as the smartphone are increasingly being utilized by app developers to help users better manage their health and chronic disease conditions. These devices and the mobile health apps that run on them have the potential to provide critical, longitudinal components to an individual's health record. In fact companies such as Apple have greatly facilitated this through their HealthKit, ResearchKit, CareKit, and Homekit platforms, which enable acquisition of very high frequency data over long periods of time, thus providing far more detailed phenotypic user profiles than could ever be reasonably generated in a typical clinical or research setting.

Recently, benefiting from advances in mobile health technologies, we successfully conducted the Asthma Mobile Health Study using an iPhone app.[1] Asthma is a common, highly variable and heterogeneous disease, and it has therefore been difficult to characterize patient disease subtypes precisely enough to inform an optimal individualized treatment plan. Less than half of the 25 million people in the United States with asthma have optimal asthma control, significantly contributing to $56 billion in direct and indirect health care costs annually.[2-3] In order to improve outcomes and reduce costs on a population level, it will be important to acquire large data sets to develop individualized models capable of identifying patients at highest risk to better target resources and tailor therapies. Prior efforts at identifying subgroups of asthma patients have been made based on demographics, lung function tests, biopsy results and blood testing, response to therapy,[4] and recently, genetics.[5-6] Our Asthma Health App, however, for the first time, enables one to collect rich time series data on asthma patients' activities on a daily basis. This opens up the possibility to identify at-risk subgroups of patients based on high-resolution time-course symptom data. The ability to identify clinically relevant patterns of disease could potentially allow targeting of resources to at risk patients to improve disease control.

Participants in the Asthma Mobile Health Study (AMHS) were asked to complete daily surveys to record symptoms and presumed triggers for the duration of the study. Taking the *day symptom outcome* as an example, the collected data of one user is a vector of 0's and 1's indicating whether the user experienced any asthma symptom on each day (1 indicating yes and 0 no symptom experienced). Once collected, the day symptom outcome records of all users can be presented as a 0/1 matrix, which can be used to explore whether subgroups of asthma patients with distinct symptom patterns exist. However, one particular challenge with this type of survey results data is that they contain substantial missing values. While most users may respond to daily survey questions or choose to actively input data on their condition when appropriate, for any given subset of days for which data are being collected, the response rates will be highly varied among different users. Further, for formal studies such as AMHS, users enroll in the study on a rolling basis, such that many non-overlapping periods of enrollment among different users must be accounted for. Lastly, even for the same user, survey response rates often varied over time. Users may be more likely to respond on days when they experience disease symptoms, which further complicates analysis of the data.

A crucial step in handling missing data is to characterize the nature of missing-ness. If the probability of missing data does not depend on the missing values, the missing-data mechanism is

referred to as missing-at-random; if so, the mechanism is referred to as not-missing-at-random or non-ignorable. When the proportion of missing values in a data set is large and the missing mechanism is not at random, it is not appropriate to ignore the missing mechanism and perform standard statistical analyses based on the observed values.[7-8] In our AMHS data, since the probability of a user responding to the survey on a particular day depends on the user's asthma symptom on that day, the missing mechanism is non-ignorable. Therefore, in this work, we propose a probability model to characterize the missing mechanism underlying such data and implement a consensus clustering algorithm incorporating multiple imputations. We compare our proposed method with other imputation strategies based on low rank matrix completion procedures.[9] Through extensive simulation studies, we demonstrate the advantage of the probability model based imputation under a range of scenarios reflecting the characteristics of our time series data. While our method is applied to AMHS study and simulated data, the approach can be applied to any time series data in which the missing data mechanism is non-ignorable.

2. Method

Our primary aim was to develop a method that would cluster users in AMHS based on their self-reported day symptom outcome time-series data to identify subgroups of app users with distinct symptom patterns. Given the substantial amount of missing data and that the missing data mechanism is non-ignorable, existing methods were not sufficient for this purpose.

2.1. *A probability based imputation model*

Denote the *day symptom outcome* data matrix as $X_{N\times T} = [\![x_{it}]\!]$, where $i = 1, \cdots, N$, is the index of users and $t = 1, \cdots, T$, is the index of days. Note, since asthma symptoms are often affected by environmental and seasonal changes, we align the profiles of different users according to actual dates instead of arbitrary days in the study. Each x_{it} takes on a value of 1 or 0, depending on whether the i^{th} user reported an asthma symptom on the t^{th} day or not, respectively; x_{it} is set to NA if the i^{th} user did not enroll in the study or did not respond to the daily survey on the t^{th} day.

We further introduce two binary data matrices: $S_{N\times T} = [\![s_{it}]\!]$ to indicate whether users responded to the AMHS survey on each day; and $D_{N\times T} = [\![d_{it}]\!]$ to represent the underlying complete day symptom outcome data. Given these matrices, the observed data $X_{N\times T}$ satisfies: $x_{it} = d_{it}$, if $s_{it} = 1$; and $x_{it} = NA$ if $s_{it} = 0$. If $D_{N\times T}$ was available, existing methods could be employed to cluster users based on this data matrix. However, since we only observe $X_{N\times T}$ and a substantial proportion of $X_{N\times T}$ is NA, we need to impute these missing values first before we can attempt clustering.

The key step for the imputation is to estimate the probability that a given user on a given day had a symptom event that should have been recorded, given the user did not respond to the survey on that day $P(d_{it} = 1|s_{it} = 0)$. In light of the 6-month milestone survey, which is administered to each app user 6 months after the enrollment date in our AMHS, 12% of users indicated that they were more likely to respond to the survey on days when they experienced asthma symptom(s). Given this, we assume that there exists an α_i (≥ 1) for each user, such that $P(s_{it} = 1|d_{it} = 1) = \alpha_i P(s_{it} = 1|d_{it} = 0) = \alpha_i r_{it}^0$, where $r_{it}^0 = P(s_{it} = 1|d_{it} = 0)$. We treat each α_i as a random variable, which takes the value of 1 with probability 0.88, and 2 with probability 0.12, in accordance with feedbacks from AMHS. The choice of 2 is based on the

median level of possible range of α_i ($1< \alpha_i <3$) that ensures realistic scenarios given the observed distribution of user response rates. Sensitivity analysis on choices of α_i is shown in Section 3.

We further denote $\bar{p}_{it} = P(d_{it} = 1)$, $p_{it} = P(d_{it} = 1|s_{it} = 1)$, and $r_{it} = P(s_{it} = 1)$. Thus we have that $r_{it} = P(s_{it} = 1|d_{it} = 1)P(d_{it} = 1) + P(s_{it} = 1|d_{it} = 0)P(d_{it} = 0) = \alpha_i r_{it}^0 \bar{p}_{it} + r_{it}^0(1 - \bar{p}_{it})$; $p_{it} = \frac{P(s_{it}=1|d_{it}=1)P(d_{it}=1)}{P(s_{it}=1|d_{it}=1)P(d_{it}=1)+P(s_{it}=1|d_{it}=0)P(d_{it}=0)} = \frac{\alpha_i \bar{p}_{it}}{\alpha_i \bar{p}_{it}+(1-\bar{p}_{it})}$. it follows that

$$\bar{p}_{it} = \frac{p_{it}}{\alpha_i(1-p_{it})+p_{it}}, \quad \text{and} \quad r_{it}^0 = r_{it}\frac{\alpha_i(1-p_{it})+p_{it}}{\alpha_i}. \tag{1}$$

And we have that

$$P(d_{it} = 1|s_{it} = 0) = \frac{P(s_{it}=0|d_{it}=1)P(d_{it}=1)}{P(d_{it}=0|s_{it}=1)P(s_{it}=1)+P(d_{it}=0|s_{it}=0)P(s_{it}=0)}$$

$$= \frac{(1-\alpha_i r_{it}^0)\bar{p}_{it}}{(1-\alpha_i r_{it}^0)\bar{p}_{it}+(1-r_{it}^0)(1-\bar{p}_{it})} = \frac{(1-\alpha_i r_{it}^0)p_{it}}{(1-\alpha_i r_{it}^0)p_{it}+\alpha_i(1-r_{it}^0)(1-p_{it})}. \tag{2}$$

We then propose to estimate p_{it} and r_{it} based on the observed data in a time window around the t^{th} day such that

$$\hat{p}_{it} = \frac{\sum_{|t'-t|<\delta} I(s_{it'}=1,\ x_{it'}=1)}{\sum_{|t'-t|<\delta} I(s_{it'}=1)}, \text{ and } \hat{r}_{it} = \frac{\sum_{|t'-t|<\delta} I(s_{it'}=1)}{\sum_{|t'-t|<\delta} 1}, \tag{3}$$

where $I(\cdot)$ is the indicator function, and δ defines the size of the time window. If we plug equation (3) into equations (1) and (2), then we can obtain an estimate of $P(d_{it} = 1|s_{it} = 0)$. In the simulation and real data analysis below, we set δ to be 30 days. This choice resulted from a tradeoff between the robustness to estimate empirical response/symptom rates and sensitivity to capture changes within a short time period.

2.2. *Multiple imputation and consensus clustering*

The probability model in section 2.1 provides a convenient framework for integrating the multiple-imputation procedure[8] and the consensus clustering procedure.[10] Specifically, in the b^{th} imputation run, we first simulate a vector of $\{\alpha_i^b\}_i$. Then to impute an unobserved d_{it}, we calculate $\widehat{P^b}(d_{it} = 1|s_{it} = 0)$ based on α_i^b, and randomly sample a value from a Bernoulli distribution with success probability of $\widehat{P^b}(d_{it} = 1|s_{it} = 0)$. We denote the final imputed complete matrix as ${D_{N\times T}}^b = [\![{d_{it}}^b]\!]$.

Naively, we could perform clustering analysis based on ${D_{N\times T}}^b$. However, when we compare the day symptom profiles of two users, it makes more sense to define distance based on their symptom frequencies over a time window instead of based on events on individual days. For example, suppose there are two users: one has symptoms on Monday, Wednesday and Friday in a given week, while the other has symptoms on Tuesday, Thursday and Saturday in the same week. If we considered the 0/1 vectors of daily symptom events of these two users for this week, they would be extremely different. However, if we consider the symptom frequency over the week, these two users actually show a similar pattern. Therefore, we propose to calculate the frequency profile of each user by performing a running average of the symptom profile:

$f_{it}^b = 1/(2h-1)\sum_{|t'-t|<h} s_{it'}^b$. Then, we can derive clusters of users by performing K-means clustering based on the frequency matrix $F_{N\times T}^b = [\![f_{it}^b]\!]$. We can record the clustering result with an adjacency matrix $((A_{ij}{}^b))_{N\times N}$, where $A_{ij}{}^b = 1$ if the i^{th} user and the j^{th} user are assigned to the same cluster; and $A_{ij}{}^b = 0$ otherwise. We repeat the above imputation-cluster process B times. This gives us B adjacency matrices $\{((A_{ij}{}^b))_{N\times N}\}_b$ corresponding to B sets of clustering results. Intuitively, a large value for A_{ij} suggests a high similarity between the i^{th} and j^{th} user. We can define an average adjacency matrix, $\overline{A_{ij}} = 1/B\sum_b A_{ij}{}^b$, over all adjacency matrices, and then perform the final cluster assignment via another round of K-mean clustering based on the $((\overline{A_{ij}}))$ matrix. We refer to the above procedure as the probability based imputation with consensus clustering (PIC) method. For the special case of $h = 1$, clustering is performed on the imputed day symptom matrix $D_{N\times T}{}^b$. We refer to this special case as the PIC.s method.

One variation on the PIC method worth exploring is to first perform Principal Component Analysis (PCA) on the $D_{N\times T}{}^b = [\![d_{it}{}^b]\!]$ matrix, and then select the loading matrix of the leading L principle components to further perform the clustering analysis. We denote this variation of the PIC procedure as PIC.PC.

3. Simulation Studies

In this section, we investigate the performance of the proposed methods through simulation studies under a range of scenarios reflecting real data conditions.

3.1. *Methods to compare*

In addition to the three methods defined above, PIC, PIC.s and PIC.PC, we also consider performing the probability imputation without taking into account the non-random missing pattern (i.e. set $\alpha_i = 1$). We denote this strategy as "PIC($\alpha_i = 1$)". We also include a few low-rank (LR) matrix completion based approaches for comparison. LR matrix completion has been recently demonstrated to be extremely powerful in recovering large scale matrices[9]. Specifically, we employ the R package *softImpute*,[11] which uses convex relaxation techniques to provide regularized low-rank solutions for large-scale matrix completion problems. We considered three strategies to apply the LR matrix completion (referred to as "LR" in below): (1) we directly apply LR on the raw data matrix ($X_{N\times T}$); (2) for each user, we first imputed the missing data based on the probability model of PIC for days within his/her enrollment period, and then apply LR to impute the missing data on days outside the enrollment period; and (3) similar to (2) except that we further derive the frequency matrix following the imputations. Here, enrollment period of one user is defined as the period from the first to the last instance of non-missing observation based on the empirical day symptom data. In all three strategies, after data imputation, consensus clustering is performed in the same way as for PIC. We denote these three strategies as LR, PIC.S.LR, and PIC.LR, respectively.

3.2. *Simulation settings*

To mimic the data from AMHS, in our simulations (see section 4), we set N=334, T=136, and the total number of clusters to be 3. In addition, we assumed 3 roughly equal-sized clusters (n_l=111,

n_2=111, and n_3=112), so the accuracy of clustering result could be more intuitively assessed. We then generated multiple sets of frequency curves representing a variety of hypothetical symptom frequency profiles (i.e. $\{P\ (d_{it} = 1)\}_t$) (see Fig. 1). We assume the samples belonging to the same cluster share the same underlying symptom frequency profile. To generate time-series data for each sample, we simulated symptom events of the t^{th} day by Bernoulli sampling of 0/1 based on the t^{th} point of the corresponding frequency curve. To simulate non-overlapping enrollment periods, we sampled from the empirically observed enrollment period distribution from the AMHS data.

To further generate non-ignorable missing-ness, we used information from the milestone survey results in AMHS. In this survey, users are asked to provide their reasons for not responding to the daily survey during the study period. Based on users who provided milestone survey responses before April 4, 2016, 12% indicated that they tended to skip the daily surveys on days in which they had no symptoms. Thus, in the simulated data we sampled from a Bernoulli distributed random variable I to identify whether a user was among those whose response depended on symptom state, with p (I = 1) = 0.12. For samples assigned to I = 1, we introduced a parameter Δ to modify the rate of missing data depending on symptom state such that $(s_{it} = 1|d_{it} = 1) = P(s_{it} = 1) + \Delta$ and $P(s_{it} = 1|d_{it} = 0) = P(s_{it} = 1) - \Delta$. For other samples assigned to I = 0, we imposed uniformity over time such that $P(s_{it} = 1|d_{it} = 1) = P(s_{it} = 1|d_{it} = 0) = P(s_{it} = 1)$. For each user, $r_{it} = P(s_{it} = 1)$ was set to be a constant r_i, which is either a pre-determined value or is sampled from an empirical distribution of missing rates calculated from the AMHS data. We then used $P(s_{it} = 1|d_{it} = 1)$ and $P(s_{it} = 1|d_{it} = 0)$ to generate missing data within the enrollment period.

We considered various simulation settings to evaluate how the performances of the different methods were affected by various factors including: (1) the shapes of the frequency profiles, (2) the overall missing percentages, (3) the severity level of the non-random missing, (4) alternative scenarios for setting α_i, and (5) we evaluated the power to detect association between a generic simulated covariate and the inferred cluster assignments derived from the application of each method on simulated data. In the following, we varied one factor at a time, where unless specified, the default setting is to use the frequency profile set labeled b in Figure 1, r_i = 0.4 for all users, Δ = 0.3, and α_i is 1 with probability 0.88 or 2 with probability 0.12. For all settings, the window size *h* used to derive the frequency profiles is simply set to be a fix value of 15, as we observed that the performance of all strategies are not sensitive to the different choices of *h* (data not shown).

1. We considered 8 different sets of symptom frequency profiles as illustrated in Figure 1.
2. We considered 4 different ways for setting r_i, where for (1)-(3), r_i = 0.2, 0.4, or 0.6; and for (4) r_i is sampled from an empirical distribution of missing rates calculated from the AMHS data.
3. We varied the value of Δ, where Δ = 0.1, 0.3, or 0.5.
4. We considered 3 alternative scenarios for setting α_i, where for (a1)-(a3): α_i is 1 with probability 0.88 and is 1.5, 2, or 2.5 with probability 0.12.
5. We simulated a binary covariate based on true cluster assignments, where the probability of taking a value of 1 was set to 10% across all clusters (p1), or was set, depending on cluster

assignment, to: (p2) 10%, 15% or 20%, or (p3) 10%, 20% or 30%. For these 3 scenarios, we evaluated the power to detect association between the simulated covariate and the predicted cluster assignments using a p-value cutoff of 0.05 based on Fisher's Exact Test.

3.3. *Simulation results*

For each simulation scenario, we applied each of the strategies in section 3.1 to derive predicted cluster assignments from simulated data sets. True and predicted cluster assignments were compared using the adjusted Rand index.[12] Based on the results from simulation Setting 1, strategies PIC and PICs perform well across a range of symptom profile scenarios (Fig. 1).

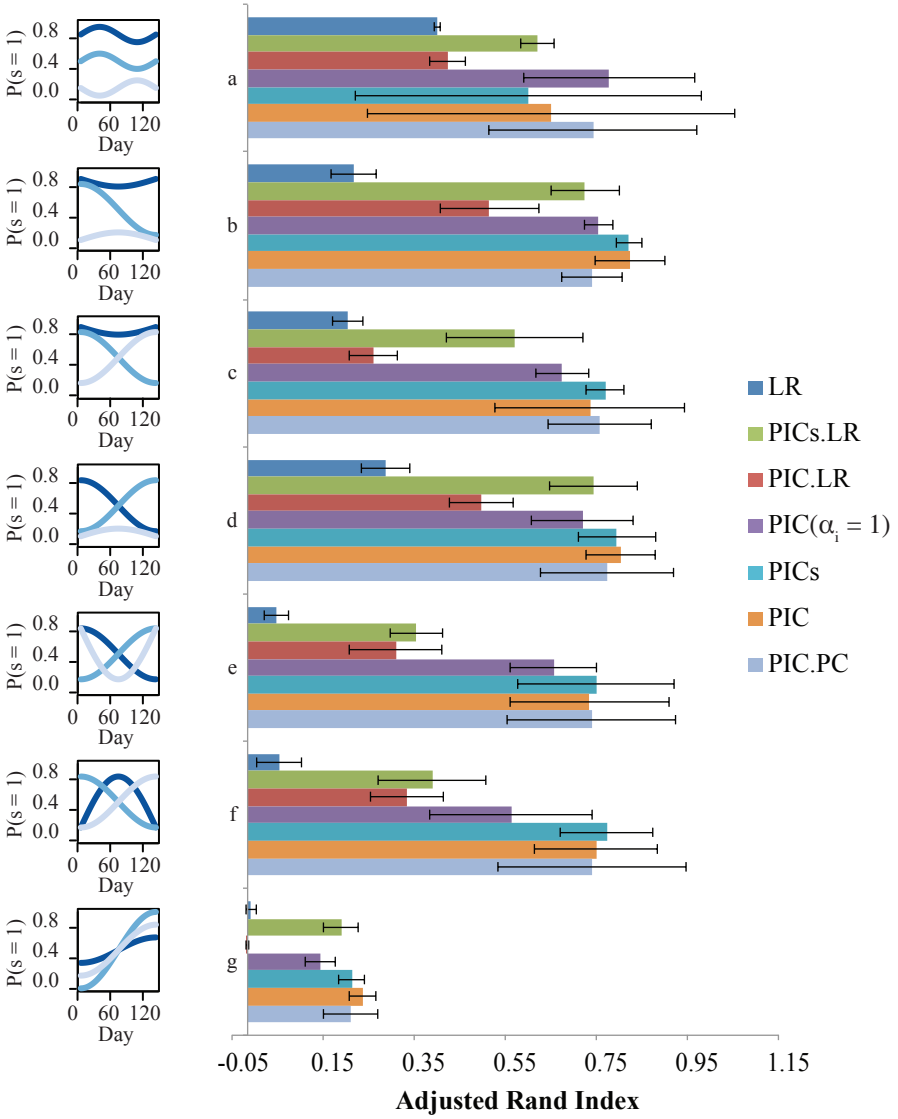

Figure 1. Simulation results for Setting 1, where we consider 8 different sets of symptom frequency profiles while fixing $\Delta = 0.3$, and using $r_i = 0.4$ for all users. Symptom profiles (a-g, left) are defined for sets of 3 clusters, where each cluster is color-coded from highest (dark) to lowest (light blue) overall mean symptom rate. Average adjusted rand indices and their standard deviations across 50 simulations with 100 iterations of imputation each are shown for all strategies.

The strategies involving low rank matrix completion display more variability across symptom profiles, particularly the LR strategy which shows a clear decrease in performance as simulation scenarios become more difficult. The accuracy of all methods tend to decrease with the overall missing rate of the data (Fig. 2A). LR is particularly worse in cases where the overall non-response rate (r_i) or the severity level of non-random missing (Δ) is high (Fig. 2B). We also observe disadvantages of PIC($\alpha_i = 1$) compared to PIC under these same circumstances, due to the lack of treatment of non-ignorable missing (Fig. 2A and Fig. 2B). Most strategies show comparable performance across different α_i scenarios, with the exception of LR, which shows enhanced performance when α_i is set to a2 (Fig. 2C). In the end, Fig. 2D suggests that PIC achieves better power to detect association between covariates and predicted clusters than other clustering strategies when the strength of association is simulated to be more moderate.

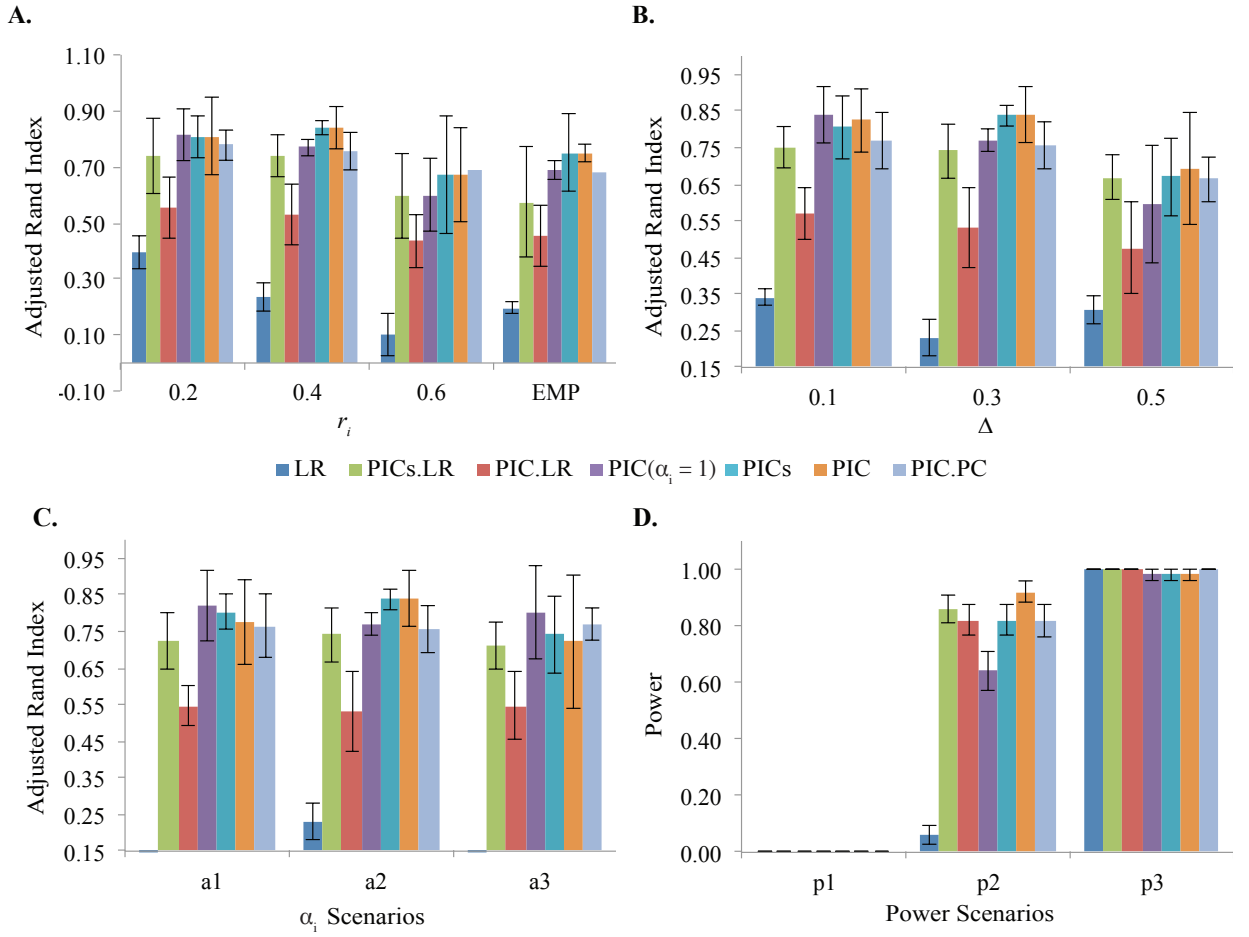

Figure 2. Simulation results based on 50 data simulations with 100 multiple imputations each. A. Results for setting 2, where we consider several values of r_i, including: 0.2, 0.4, and 0.6, where r_i is a constant for all users; and EMP, where r_i is sampled from the empirical distribution of missing rates calculated from AMHS data. B. Results for setting 3, considering several values for Δ. C. Results for setting 4, where we consider different scenarios for assigning values to the random variable α_i, where the maximum fold-difference in $P(s_{it} = 1|d_{it} = 1)/P(s_{it} = 1|d_{it} = 0)$ varies from 1.5 (a1) to 2 (a2) to 2.5 (a3). D. Power analysis based on 3 scenarios of simulated covariate data varying from null (p1) to strongest association (p3) with true cluster assignments.

4. Analysis of the AMHS data using PIC

Clustering analysis was performed for several data types, including daily symptoms and daily self-reports of asthma triggers on air quality, heat, and pollen. Study participants were first clustered into subtypes using daily symptom data collected by the AMHS. To further characterize these subtypes, we tested for associations between predicted cluster assignments and clinical variables (age of diagnosis, GINA control level, smoking status, and weight), demographic variables (gender, income, and ethnicity), as well as self-reported trigger data collected by our app (pollen, heat, and air quality). Tests of association were performed using Fisher's exact tests, where we filtered out categories with fewer than 10 individuals where applicable. Supplemental Table 1 summarizes these results (http://icahndigitalhealth.org/wp-content/uploads/2016/08/Clustering-Supplemental-Data.pdf).

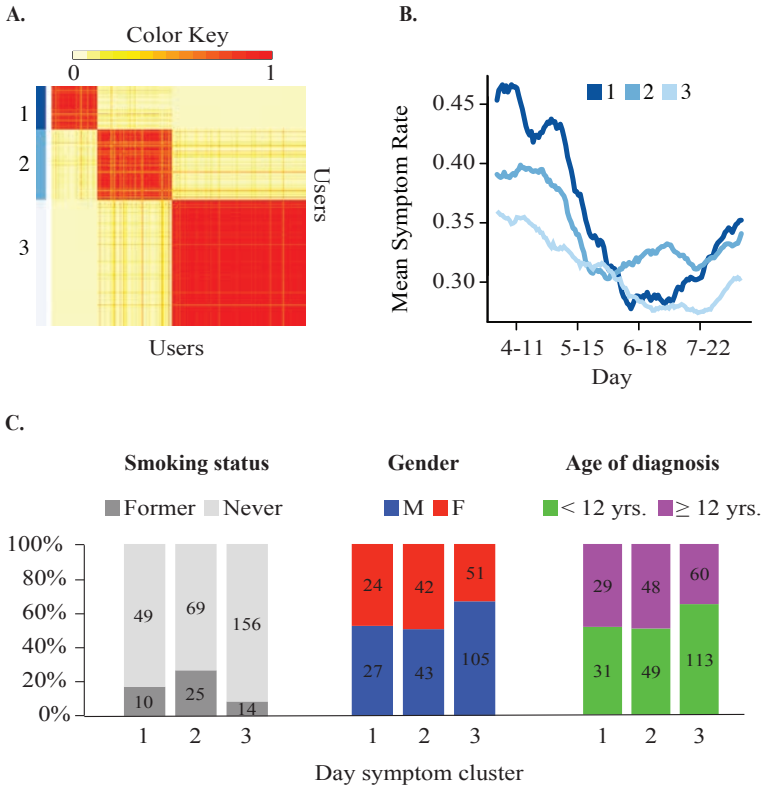

Figure 3. A. Heatmap is based on the adjacency matrix derived from consensus clustering of daily asthma symptoms for 334 users over 136 days using strategy PIC based on 100 iterations of imputation. Three distinct clusters (n_1 = 60, n_2 = 98, and n_3 = 176) are identified by color and enumeration (1-dark, 2-medium, and 3-light blue) where pairs of users most frequently found in the same cluster are found in the red regions along the diagonal. B. Mean curves for the clusters are based on the average of smoothed-imputed data on asthma symptoms. Each curve shows the mean symptom rate for users belonging to each cluster. Clusters are color-coded from dark to light blue by the overall mean symptom rate for each cluster. C. Day symptom clusters are significantly associated with smoking status (p = 0.0005), gender (p = 0.02), and age of diagnosis (p = 0.03) based on Fisher's exact test with simulated p-values based on 2,000 replicates. Barplots show the percentage distribution for each category within each day symptom cluster.

To conduct clustering analysis, we considered daily survey data collected by the app over the 6-month period from March 9, 2015 through August 9, 2015. We restricted our analysis to nonsmokers, defined as either having never smoked or having smoked less than 10 packs per year,

without congestive heart failure or lung diseases other than asthma. We further required that each user have at least 50 survey responses over the entire 6-month period. These filterings led to 334 users in total. To ensure adequate overlap among enrollment periods across these users for comparing among methods in our simulation studies, we restricted our analysis to the 136-day period from April 2, 2015 (early spring) to August 8, 2015 (late summer), which corresponds to 136 days in total.

Based on daily survey data from 334 users over a period of 136 days, the average number of surveys provided per user was 70 (SD = 25), with an average per user enrollment period of 109 days (SD = 25). The average within enrollment missing rate was .4 (SD = 0.2). Clustering on the daily asthma symptom data was performed using the PIC strategy. After running the PIC method separately using different cluster numbers ranging from 2 to 5, we determined that users were well grouped into 3 clusters based on visual comparison of heatmaps derived from the adjacency matrices produced during the consensus clustering step of each run (Fig. 3A). Mean curves based on the average symptom rate for the users belonging to each of these clusters is shown in Figure 3B based on the average of the smoothed imputed data across 100 iterations of imputation, where curves are color-coded from dark to light blue to identify clusters with high, middle, and low symptom rates based on averaging across days.

We first sought to characterize our derived day symptom subtypes by comparing them with clinical and demographic variables. We found a significant association between asthma symptoms and smoking status (Fisher's exact test: p = 5e-4; n = 333), gender (Fisher's exact test: p = 0.02; n = 292), and age of diagnosis (Fisher's exact test: p = 0.03; n = 330). To study the relationship between asthma subtypes and environmental triggers, we used a similar approach to cluster self-reported data on daily asthma triggers collected by the AMHS. In the daily survey, participants were asked to self-report on symptom triggers on a given day. Specifically, users were able to choose from a list of 22 known asthma triggers, including allergens such as pollen, pet dander, and weather conditions. We chose to focus our analysis on air quality, heat, and pollen trigger data based on results from previous validation efforts comparing trigger data with more objective measures (PM2.5, max daily temperature, and pollen counts) using publicly available datasets[1].

Triggers were coded as 0/1 depending on whether a user cited a given trigger on a given day. Although we know that missing data in symptom reports were not random, we have little basis for attributing non-reported symptoms to one trigger over another with greater probability. Therefore, in conducting missing data imputation for trigger data, we used PIC($\alpha_i = 1$). Heatmaps resulting from the application of this method are shown in Supplemental Figure 1A-C (http://icahndigitalhealth.org/wp-content/uploads/2016/08/Clustering-Supplemental-Data.pdf).

Based on these groupings, self-reported asthma triggers were associated with the day symptom cluster groupings. Specifically, with Fisher's exact test, we found highly significant associations between day symptom clusters and clusters derived from self-reported data on pollen (p = 5e-4; N = 333), heat (p = 5e-4; N = 333), and air quality (p = 0.02; N = 333) triggers. As expected, we found a significant association between heat and US climate regions[13] broken down by northern and southern regions (Supplemental Table 2), with users belonging to cluster H1, who reported peak heat trigger complaints in late July, more frequently located in the northern US climate regions (72%) (p = 0.01; N = 288). We found that asthma trigger clusters differentiated by asthma subtype such that users who complain most frequently of pollen and heat are most frequently found in day symptom cluster 1, corresponding to the group with the highest average day

symptom levels (Fig. 4A-B). By contrast, individuals frequently citing air quality as their asthma trigger are more frequently found in cluster 3, corresponding to the lowest overall day symptom rate.

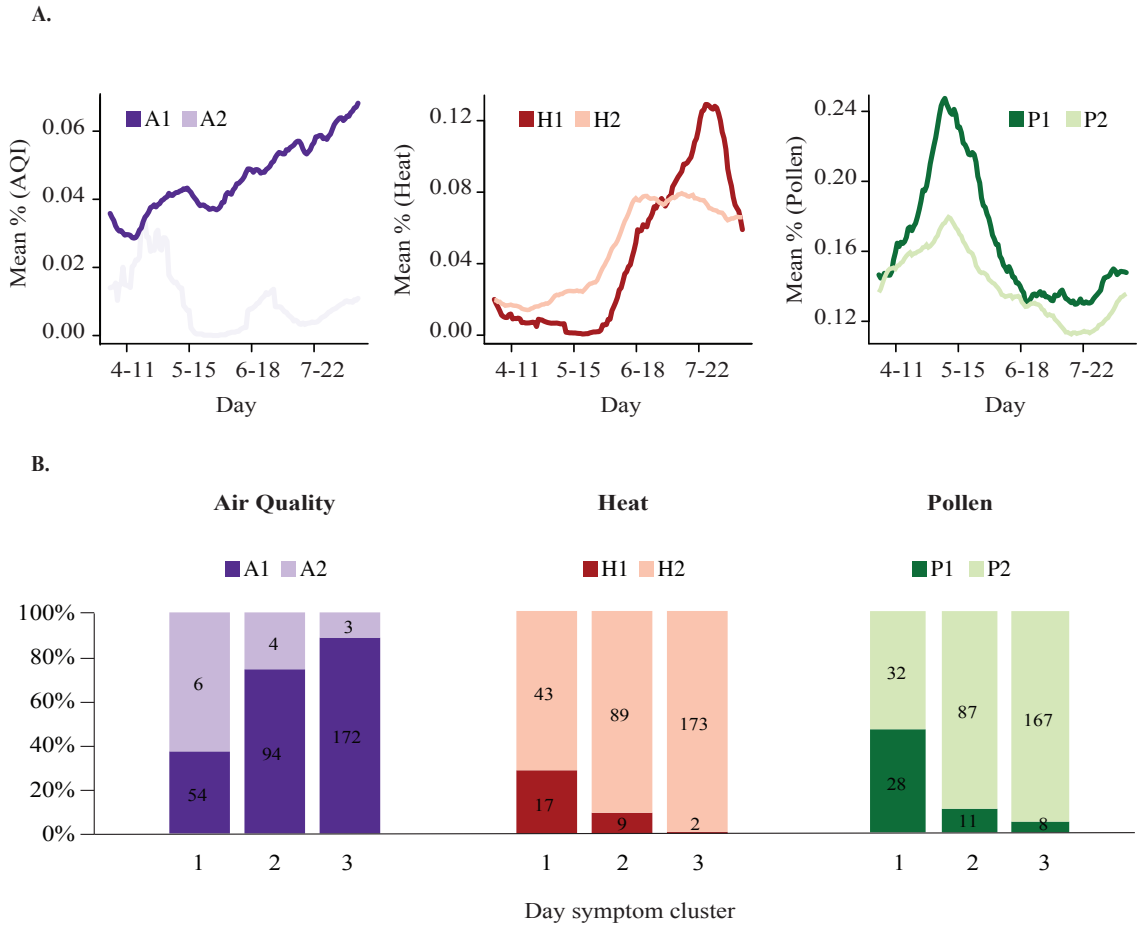

Figure 4. A. Curves depict the mean percentage of users reporting air quality, heat, and pollen for each cluster derived from the application of PIC($\alpha_i = 1$) using 100 multiple imputations. Clusters are color-coded from dark (high) to light (low) according to the overall mean percentage for each cluster averaged across days. B. Day symptom clusters are significantly associated with trigger clusters for air quality (p = 0.02, N = 333), heat (p = 5e-4, N = 333), and pollen (p = 5e-4, N = 333), based on Fisher's exact test with simulated p-values based on 2,000 replicates.

5. Discussion

Here we have considered the problem of clustering time series data collected from mobile health apps in which there is a high proportion of missing data for which the missing data mechanism is at least partially known. For such cases, regular clustering methods cannot be applied directly. To bridge this gap, in this paper, we developed an integrated PIC strategy to both impute the missing data using a probabilistic model and then clustered samples to identify subgroups with distinct patterns. The advantage of our PIC approach over other strategies based on low-rank matrix completion is demonstrated through extensive simulation studies.

When applying PIC on the AMHS data, we identified a unique subgroup of patients who have relatively high symptom rates and are more sensitive to distinct environmental factors with

seasonal changes, such as heat and pollen. Furthermore, we noted relatively lower reported symptom rates associated with air quality, which may be attributed to the multi-factorial, reduced variability, and less well defined nature of this asthma trigger. With further validation, the ability to identify unique disease patterns in data sets with non-random missing data could be extremely useful in the conduct of environmental epidemiologic research as it could be used to track and identify novel environmental risk factors linked to worsening asthma. Moreover, it could enable us to identify at risk populations in large data sets and design targeted interventions to apply to reduce risk and improve outcomes. The ability to monitor asthma symptoms longitudinally by mobile technology, and identify specific subgroups of patients who have destabilization of asthma control based on specific triggers creates the opportunity to intervene early therapeutically. For example, if high heat or high pollen conditions are identified using personalized reports available by mobile technology, personalized alerts regarding presence of triggers would allow patients to seek medical advice and potentially adjust therapy in order to avoid the need for urgent care. R code implementing PIC (probability based imputation and consensus clustering) can be found here: http://icahndigitalhealth.org/wp-content/uploads/2016/10/PIC.R.

6. References

1. Chan, Y.-F.Y., et al., *The Asthma Mobile Health Study, a Large Scale Clinical Study Using ResearchKit.* Nature Biotechnology, submitted., 2016.
2. *Asthma-Data, Statistics, and Surveillance: Center for Disease Control and Prevention* 2015.
3. *GINA guidelines: Global Initiative for Asthma*. 2016.
4. Gauthier, M., A. Ray, and S.E. Wenzel, *Evolving Concepts of Asthma.* American Journal of Respiratory and Critical Care Medicine, 2015. **192**(6): p. 660-668.
5. Kaminsky, D.A., *Systems biology approach for subtyping asthma; where do we stand now?* Current opinion in pulmonary medicine, 2014. **20**(1): p. 17-22.
6. Chung, K.F., *Defining phenotypes in asthma: a step towards personalized medicine.* Drugs, 2014. **74**(7): p. 719-728.
7. Rubin, D., *Inference and missing data.* Biometrika 63 (3), 581-592, 1976.
8. Rubin Donald, B., *Multiple imputation for nonresponse in surveys*. 1987, New York: Wiley.
9. EJ Candès, B.R., *Exact matrix completion via convex optimization.* Foundations of Computational Mathematics 9 (6), 717-772.
10. Filkov, V. and S. Skiena, *Integrating microarray data by consensus clustering.* International Journal on Artificial Intelligence Tools, 2004. **13**(04): p. 863-880.
11. Hastie, T. and R. Mazumder, *softImpute: Matrix Completion via Iterative Soft-Thresholded SVD.* R package version, 2015. **1**.
12. Hubert, L. and P. Arabie, *Comparing partitions.* Journal of classification, 1985. **2**(1): p. 193-218.
13. Karl, T. and W.J. Koss, *Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983*. 1984: National Climatic Data Center.