Identifying Patterns in Behavioral Public Health Data Using Mixture Modeling with an Informative Number of Repeated Measures

Gary Yu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Public Health
in the Department of Biostatistics
under the Executive Committee
of the Mailman School of Public Health

COLUMBIA UNIVERSITY
2014

ABSTRACT

Identifying Patterns in Behavioral Public Health Data Using Mixture Modeling with an Informative Number of Repeated Measures

Finite mixture modeling is a useful statistical technique for clustering individuals based on patterns of responses. The fundamental idea of the mixture modeling approach is to assume there are latent clusters of individuals in the population which each generate their own distinct distribution of observations (multivariate or univariate) which are then mixed up together in the full population. Hence, the name mixture comes from the fact that what we observe is a mixture of distributions. The goal of this model-based clustering technique is to identify what the mixture of distributions is so that, given a particular response pattern, individuals can be clustered accordingly. Commonly, finite mixture models, as well as the special case of latent class analysis, are used on data that inherently involve repeated measures. The purpose of this dissertation is to extend the finite mixture model to allow for the number of repeated measures to be incorporated and contribute to the clustering of individuals rather than measures. The dimension of the repeated measures or simply the count of responses is assumed to follow a truncated Poisson distribution and this information can be incorporated into what we call a dimension informative finite mixture model (DIMM).

The outline of this dissertation is as follows. Paper 1 is entitled, “Dimension Informative Mixture Modeling (DIMM) for questionnaire data with an informative number of repeated measures.” This paper describes the type of data structures considered and introduces the dimension informative mixture model (DIMM). A simulation study is performed to examine how well the DIMM fits the known specified truth. In the first scenario, we specify a mixture of three univariate normal distributions with different means and similar variances with different and similar counts of repeated measurements. We

found that the DIMM predicts the true underlying class membership better than the traditional finite mixture model using a predicted value metric score. In the second scenario, we specify a mixture of two univariate normal distributions with the same means and variances with different and similar counts of repeated measurements. We found that that the count-informative finite mixture model predicts the truth much better than the non-informative finite mixture model.

Paper 2 is entitled, “Patterns of Physical Activity in the Northern Manhattan Study (NOMAS) Using Multivariate Finite Mixture Modeling (MFMM).” This is a study that applies a multivariate finite mixture modeling approach to examining and elucidating underlying latent clusters of different physical activity profiles based on four dimensions: total frequency of activities, average duration per activity, total energy expenditure and the total count of the number of different activities conducted. We found a five cluster solution to describe the complex patterns of physical activity levels, as measured by fifteen different physical activity items, among a US based elderly cohort. Adding in a class of individuals who were not doing any physical activity, the labels of these six clusters are: no exercise, very inactive, somewhat inactive, slightly under guidelines, meet guidelines and above guidelines. This methodology improves upon previous work which utilized only the total metabolic equivalent (a proxy of energy expenditure) to classify individuals into inactive, active and highly active.

Paper 3 is entitled, “Complex Drug Use Patterns and Associated HIV Transmission Risk Behaviors in an Internet Sample of US Men Who Have Sex With Men.” This is a study that applies the count-informative information into a latent class analysis on nineteen binary drug items of drugs consumed within the past year before a sexual encounter. In addition to the individual drugs used, the mixture model incorporated a count of the total number of drugs used. We found a six class solution: low drug use, some recreational drug use, nitrite inhalants (poppers) with prescription erectile dysfunction (ED) drug use, poppers with prescription/non-prescription ED drug use and high polydrug use. Compared to participants in the low drug use class, participants in the highest drug use class were 5.5 times more likely to report unprotected anal intercourse (UAI) in their last sexual encounter and approximately 4 times more likely to report a new sexually transmitted infection (STI) in the past year. Younger men were also less likely to report UAI than older men but more likely to report an STI.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to thank the following dissertation committee members: Kenneth Ying Kuen Cheung, Ph.D., Sabina Hirshfield, Ph.D., Roger Vaughan, Dr.P.H. (Chair), Melanie M. Wall, Ph.D. (Advisor), and Joshua Willey, M.D.

DIMENSION INFORMATIVE MIXTURE MODELING FOR QUESTIONNAIRE DATA WITH AN INFORMATIVE NUMBER OF REPEATED MEASURES

INTRODUCTION

A common form of questionnaire item used to elicit information on various types of behaviors or preferences includes asking the participant to examine a list of prompts and to 'mark all that apply' while also possibly adding a self-described 'other' category and answering follow up questions for those things that do apply. This type of question leads to triply nested data structure where Y_{idm} is the response of the i^{th} person (i = 1 to n) to the m^{th} follow up question of the d^{th} prompt where d= 1 to the total number of prompts (activities, behaviors, preferences, etc) marked and this total varies (is random) by individual. In this paper we will develop a mixture model for clustering individuals that takes into account their responses to the questions including accounting for the varying number of responses made (i.e. varying number of repeated measures within person). We begin by giving three data examples exhibiting this structure, two collected with questionnaires that will be used throughout this dissertation and one that is hypothetical but represents a paradigm from lab/diagnostic studies that follows the same data structure.

In the Northern Manhattan Study (NOMAS), physical activity was assessed using the questionnaire instrument shown in Table 1 where respondents indicated whether they had participated in each of the 15 different activities and were also allowed to write in other activities and additionally indicated the frequency and duration of all their activities over the last two weeks. Here the nested data structure, Y_{idm} is the response of the i^{th} person (i = 1 to n) to the m^{th} follow up question where m=1,2 indicates follow up questions on frequency and duration for the d^{th} activity where d = 1 to the total number of activities marked. Note the number of activities marked varies randomly by person and each activity also carries its own attribute, i.e. in this example each physical activity has a fixed intensity MET (kcal/kg-hr) score and a label (e.g. walking, jogging, hiking, etc.) associated with it. These attributes of the activity will be treated similarly to follow up questions, so that here m = 1,2,3,4 where the 3rd follow up is the MET score for the activity and the 4th is the type of activity. Table 2 displays example data from the NOMAS. Individuals 1 and 2 conducted one activity, walking, which they did 4-5 times a week, around 40-50 minutes/session, and walking is a moderately active activity [4 MET (kcal/kg-hr)]; individual 3 jogged 3 times/week for 30 minutes/session and played golf [4 MET] once a week for 4 hours/session and jogging is an intense activity [7 MET]; individual 4 walked 6 times/week for 45 minutes/session and hiked once a week for 1.5 hours/session, and hiking is an intense activity [6 MET]; individual 5 did 4 activities including jogging for about 5 times/week for 45 minutes/session, tennis 2 times/week for 2 hours/session and bowling once per week for about 1.5 hours/session and tennis is an intense activity [7 MET] while bowling is a moderately active activity [3 MET].

Another example comes from a study on drug consumption patterns among men who have sex with men (MSM), Table 3. Participants of an internet based sample (Hirshfield et al. 2010) were asked to indicate which drugs they had used prior to or during their last sexual encounter. In the actual study, only the type of drug was asked, but for demonstration here we suppose that also the familiarity/experience with each drug was measured (i.e. first time use, age of onset, partner using drug) and an indicator of whether the partner also used the drug or not. So we see in Table 3 that the 4 individuals differ in their profile of drug use in terms of the 4 attributes listed. Individual 1 is only familiar with alcohol use, while individual 3 is only familiar with ecstasy use. Individual 2 consumes 3 drugs, is experienced with alcohol and marijuana, has a partner that consumes both alcohol and marijuana, and is experimenting with poppers for the first time. Individual 4 consumes 4 drugs, is experienced with alcohol, marijuana, ecstasy, has a partner that uses alcohol, ecstasy and injection heroin and is trying injection heroin for the first time.

Finally, we present an example from a hypothetical breast tumor study (Table 4). In this example of a new treatment regimen for breast cancer, the women under study have varying numbers of tumors. The measurements on each tumor are pre-treatment size, post-treatment size, and tumor stage (I, II, III, IV). So similar to the questionnaire data described above the Y_{idm} represents the measurement for the i^{th} woman on the m^{th} attribute (m = 1,2,3) of the d^{th} tumor, where d = 1 to the number of tumors. Table 4 presents example data. Individuals 1 and 2 have only one tumor of Type I (early stage, slow growth) and are not responsive to treatment. It seems as though individuals 3 and 4 may elucidate certain types of tumors (Type I and Type II, mid-stage, moderate growth) that is immune to treatment where the treatment may exacerbate the size and growth of the 4 tumors. While individuals 5 and 6 with 7 lesions of all types (Type I, II, III and IV, late stage, fast growth) may be more conducive and responsive to treatment with the decrease in the tumor size. The number of tumors can be considered to be useful additional information to help determine where the treatment helps to reduce the average lesion size or not.

In each of these three examples, it may be of interest to aggregate the data across different levels, either aggregating across the attributes, or across the activities/drugs/tumors, or both. For example, in the NOMAS data an aggregation across the duration, frequency, and METS can be can be the total energy expenditure (kcal/wk) which is the product of the total frequency*average duration*fixed intensity for each physical activity. See Table 5. Doing this, the data then becomes only two levels with Y_{id} representing the total energy expenditure for the d^{th} activity done by the i^{th} person. For the drug use example it may only be of interest to model incident (first time) use of certain drugs, hence an aggregation across attributes could be to include the use or not of a particular type of drug that was used for the first time, eliminating the “partner used” and “familiarity” attributes altogether. Thus, Y_{id} would be the d^{th} type of drug used for the first time by person i. For breast tumor size, the two attributes, pre-treatment size and post-treatment size, could be aggregated to create a single attribute representing the treatment effect, which is the difference of the post-treatment minus pre-treatment tumor size. Eliminating the tumor type attribute, Y_{id} then becomes the treatment effect for the d^{th} tumor for the i^{th} woman. Notice that aggregating across attributes, the resulting two level data can be described as a univariate outcome with random, and likely informative, number of repeated measures within person.

Rather than aggregating across attributes, collapsing could also be done across activities/drugs/or tumors, so that summary attributes are created. For example a total frequency, and a total duration and a total METS score as well as a count of total number of different activities done could be made in the NOMAS data. Thus, Y_{im} would be the aggregate of the m^{th} attribute across all the activities. Note an additional attribute is also included representing the total number of activities/drugs/or tumors present. This aggregated data is more easily described as a multivariate outcome for each person i.

So depending on whether aggregation is done across attributes or across activities, the data structure differs. In the present paper we will focus on the first case where two level data are created with a random and informative number of repeated measures within person. Most statistical models of a single attribute implicitly assume that the number of repeated measures for a particular individual is the same or if not the same, that it is uninformative. In each example described above, individuals report a variety of different activities or drugs or are observed to have varying number of tumors and ignoring the actual number of responses or tumors is likely to ignore important information about the individual. Our overall goal will be to cluster individuals who have similar characteristics where the number of responses, i.e. the dimension of responses, is also taken into account. In classical cluster analysis (k-means and hierarchical) and model-based clustering approaches, the number of repeated measures is fixed between subjects forcing the data structure to have a uniform length or dimension.

Given data of the type generated from examples introduced above, with varying number of physical activities, drugs or tumors, these methods for uniform dimensions are likely to be problematic. In the current paper we will propose a Dimension Informative Mixture Model (DIMM) that incorporates the varying number of repeated measures as an informative random component when clustering individuals.

The organization of the article is as follows. Section 2 describes and introduces the dimension informative mixture model (DIMM) where the varying dimension is modeled via a truncated Poisson distribution. For concreteness, the NOMAS example of total energy expenditure per activity example will be used throughout. Section 3 derives the parameter estimation of the DIMM using an Expectation-Maximization algorithm to predict the latent group membership of each individual and also describes how to perform estimation in existing R and Mplus software. Section 4 describes results from a Monte Carlo simulation study comparing the performance of the DIMM to the traditional model that ignores the informative dimension size. Finally, Section 5 demonstrates the model's use for clustering individuals in the Northern Manhattan Study where a multi-ethnic cohort of elderly individuals reported the caloric expenditure (kcal/week) of a variable number of physical activities during leisure time. In summary, a short conclusion section will highlight the main points of this paper and will provide perspective and guidance for future work.

The Dimension Informative Mixture Model (DIMM)

Using the Northern Manhattan Study data example as a paradigm, let $\boldsymbol{y_i} = (\, y_{i\mathbf{1}}, \;\dots\; , y_{i\, d_i})^{\mathbf{T}}$ be the vector of energy expenditure values (kcal/week as a summary measure of frequency*duration*MET intensity) for each of the d_i activities done by person i where i = 1, … , 1971, and max(d_i) = 15 (13 mark-all-that-apply activities plus 2 possible write-ins). Denote all of the responses from all subjects as

$\boldsymbol{Y} = (\boldsymbol{y_1}, \ \dots \ , \boldsymbol{y_N})^{\mathrm{T}}$ which can be seen as a non-rectangular data matrix of dimension 1971 x 15; it is non-rectangular because most individuals do not report 15 activities and thus have structural missing values. Indeed in the NOMAS data, no individual reported more than 7 activities so the observed max(d_i) = 7.

We now propose a dimension informative mixture model (DIMM) for the purpose of clustering individuals based on their activity response profile. A traditional mixture model would assume the dimensions of $\boldsymbol{y_i}$ (i.e. d_i) are non-informative and that any person with d_i < max(d_i) could have had values for the other activities but they are just missing at random. But, this assumption is not appropriate for the current examples where the d_i itself is informative and the "missing" activities are not unobserved, but simply not performed. Thus we construct the DIMM as follows. Assume k is the kth cluster of individuals, k = 1…K, and latent cluster membership status is $\boldsymbol{u_i}$, where $\boldsymbol{u_i} \sim Multinomial(\boldsymbol{\pi})$, such that $\boldsymbol{\pi} = (\pi_1, \dots \ , \pi_K)^{\mathrm{T}}$ and $\sum_{i=1}^{K} \pi_i = 1$. Let d denote each activity from 1…d_i for the ith person, and let $\mathbf{x}_i$ be a set of possible covariates for the ith person including an intercept term representing the overall mean, then the DIMM is:

$$y_{id} | (u_i = k, \boldsymbol{x_i}) = \boldsymbol{x_i^T \beta_k} + \varepsilon_{id}^{k}$$

$$d_i \mid u_i = k \sim truncated\ Poisson(\alpha_k)$$

where ε_{id}^{k} are i.i.d $N(0, \sigma_k^2)$ and the probability mass function of the truncated Poisson distribution is given below:

$$P(d_i \mid u_i = k\ ; \alpha_k) = \frac{e^{-\alpha_k} \alpha_k^{d_i}}{d_i!(1 - e^{-\alpha_k})}.$$

Note that only individuals that report at least one physical activity are included in the data set, thus the truncated Poisson distribution is used to model d_i since values of 0 are not allowed.

The unknown parameters are $\boldsymbol{\theta} = (\beta_1, \dots ,\beta_K, \sigma_1^2, \dots ,\sigma_K^2 , \alpha_1 , \dots ,\alpha_K, \pi_1, \dots ,\pi_K)^{\mathrm{T}}$. The β_k is the fixed effects (or mean) of the k$^{\text{th}}$ cluster and σ_k^2 is the variance in the k$^{\text{th}}$ cluster. The α_k is the mean number of physical activities conducted in the k$^{\text{th}}$ cluster. The $\boldsymbol{\pi_k}$ informs us of the marginal probabilities of being in any one of the clusters. Notice that the parameters noted above can be informative about the clusters. For clustering purposes, the mean number of activities (α_k) can be different and the mean of the underlying distribution of each cluster (β_k) can be different as well. A hypothetical example would be finding two clusters in the data: a cluster with a high mean energy expenditure ($\mathrm{high}\ \beta_k$), yet individuals conduct a low number of activities ($\mathrm{low}\ \alpha_k$) and another cluster with a low mean energy expenditure ($\mathrm{low}\ \beta_k$), yet conduct a high number of activities ($\mathrm{high}\ \alpha_k$).

Estimation

It is common in mixture modeling to utilize the EM algorithm for estimation as it provides a useful way of handling the “missing” underlying clusters. In the following we detail the steps of the EM algorithm for the DIMM where a truncated Poisson distribution is included to account for the informative varying dimension. Treating the $\boldsymbol{u}$ as missing data, we use the EM algorithm to iterate our initial guess and to update our approximation of the MLEs of the parameters. The E-step uses the completely observed data, $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$, and the guess for $\boldsymbol{\theta}^{(t)}$ from the ultimate iterative step. See the complete likelihood function below. By assuming that $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ is linear in u_{ik}, we calculate the expectation of the log likelihood function of $\boldsymbol{\theta}$, which is easier:

The complete data mixture likelihood is as follows:

$$f(\boldsymbol{Y}, \boldsymbol{d}, \boldsymbol{u};\ \boldsymbol{\theta}, \boldsymbol{\pi}) = f(\boldsymbol{Y}, \boldsymbol{d}, \boldsymbol{u};\ \boldsymbol{\theta}) g(\boldsymbol{d} \mid \boldsymbol{u};\ \boldsymbol{\theta}) k(\ \boldsymbol{u}; \boldsymbol{\pi})$$

$$= \left[\prod_{i=1}^{N}\prod_{K=1}^{K}\{f_K(\boldsymbol{y_i}\,;\,\boldsymbol{\theta})\}^{I(u_i=k)}\right]\left[\prod_{i=1}^{N}\prod_{k=1}^{K}\{\frac{e^{-\alpha_k}\alpha_k^{d_i}}{d_i!\,(1-e^{-\alpha_k})}\}^{I(u_i=k)}\right]\left[\prod_{i=1}^{N}\prod_{k=1}^{K}\{\pi_k\}^{I(u_i=k)}\right]$$

$$= \left[\prod_{i=1}^{N}\prod_{k=1}^{K}\{\frac{1}{\sqrt{2\pi\sigma_k^2}}e^{-\frac{\left(y_{ik}-\,x_i^T\,\beta_k\right)^2}{\sigma_k^2}}\}^{I(u_i=k)}\right]\left[\prod_{i=1}^{N}\prod_{k=1}^{K}\{\frac{e^{-\alpha_k}\alpha_k^{d_i}}{d_i!\,(1-e^{-\alpha_k})}\}^{I(u_i=k)}\right]\left[\prod_{i=1}^{N}\prod_{k=1}^{K}\{\pi_k\}^{I(u_i=k)}\right]$$

In terms of the log likelihood:

$l(\boldsymbol{\theta},\boldsymbol{\pi};\boldsymbol{Y},\boldsymbol{u},\boldsymbol{a}) = l_1(\boldsymbol{\pi};\boldsymbol{u}) + l_2(\boldsymbol{\theta};\boldsymbol{Y}) + l_3(\boldsymbol{\theta};\boldsymbol{a})$,

$$l_1 = \sum_{i=1}^{N}\sum_{k=1}^{K} I(u_i = k)\,log(\pi_k)$$

$$l_2 = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{d=1}^{d_i} -\frac{1}{2}\,I(u_i = k)\left\{\log(\sigma_k^2) + \frac{\left(y_{ik} - \,x_i^T\,\beta_k\right)^2}{\sigma_k^2}\right\}$$

$$l_3 = \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{d=1}^{d_i} I(u_i = k)\{-\alpha_k + d_i\log(\alpha_k) - log(d_i!) - log(1-e^{-\alpha_k})\}$$

$$E(u_{ik} = 1 \mid \boldsymbol{y_k}, \boldsymbol{\theta}^{(t)}) = \widehat{u_{ik}} = \frac{f(\boldsymbol{Y},\boldsymbol{d},\boldsymbol{u};\ \boldsymbol{\theta},\boldsymbol{\pi})}{\sum_{k=1}^{K} f(\boldsymbol{Y},\boldsymbol{d},\boldsymbol{u};\ \boldsymbol{\theta},\boldsymbol{\pi})}$$

where notation wise, $u_{ik} = 1$ is equivalent to $I(u_i = k)$

$$\widehat{u_{ik}} = \frac{\pi_k^{(t)}P(\boldsymbol{y_i} \mid\ u_{ik} = 1\,;\ \boldsymbol{\theta}^{(t)})P(d_i\,,\boldsymbol{\theta}^{(t)})}{\sum_{k=1}^{K}\pi_k^{(t)}P(\boldsymbol{y_i} \mid u_{ik} = 1\,;\ \boldsymbol{\theta}^{(t)})P(d_i\,,\boldsymbol{\theta}^{(t)})}$$

Where $P(\boldsymbol{y_i} \mid u_{ik} = 1\,;\ \boldsymbol{\theta}^{(t)}) = \prod_{d=1}^{d_i} f(y_{i\,d} \mid u_{ik} = 1\,; \boldsymbol{\theta}^{(t)}) = \prod_{d=1}^{d_i}\phi\left(\frac{y_{i\,d}-x_i^T\,\beta_k}{\sigma_d^{(t)}}\right)$.

The standard normal distribution density function centered at 0 is denoted by $\phi(.)$. The truncated Poisson distribution is denoted with $P\left(d_i\,, \boldsymbol{\theta}^{(t)}\right)$.

The M-step follows by updating $\boldsymbol{\theta}^{(t)}$, the last guess, with $\boldsymbol{\theta}^{(t+1)}$, which maximizes $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$. The M-step gives the parameter estimates of (β_k, σ_k^2, α_k, π_k) as the weighted average of the sample mean, sample variance, sample number of attributes, sample group membership probability where weight is the predicted probability $\widehat{u_{ik}}$ that each subject belongs to the kth group. Specifically,

$$\widehat{\pi_k} = \sum_{i=1}^{N} \widehat{u_{ik}}$$

$$\widehat{\sigma_k^2} = \frac{\sum_{i=1}^{N} \widehat{u_{ik}} \widehat{\boldsymbol{y}}_{\boldsymbol{i}}^{\boldsymbol{T}} \widehat{\boldsymbol{y}}_{\boldsymbol{i}}}{\sum_{i=1}^{N} \widehat{u_{ik}} d_i}, \widehat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{\beta}_{\boldsymbol{k}}$$

$$\widehat{\beta_k} = \frac{\sum_{i=1}^{N} \widehat{u_{ik}} \boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{y}_{\boldsymbol{i}}}{\sum_{i=1}^{N} \widehat{u_{ik}} \boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{X}_{\boldsymbol{i}}}$$

$$\widehat{\alpha_k} = \sum_{i=1}^{N} \widehat{u_{ik}}\, d_i$$

Once the parameters have been estimated, it is then possible to determine the predicted group membership for the ith individual by finding the maximum group membership probability, we denote this as $\widetilde{u_{ik}}$ = arg max$_k$ $\widehat{u_{ik}}$.

Choosing the optimal number of clusters

The DIMM is fit with a varying number of K clusters and then these models are compared on the Bayesian Information Criteria (BIC). The optimal number of clusters issue is addressed by the best model fit using the minimum BIC value. The Bayesian Information Criteria (BIC) is as follows:

$$BIC_k = 2 \log p(D \mid M_k) \approx 2 \log p(D \mid \widehat{\theta_k}, M_k) - v_k \log(n)$$

For model M_k, a sample set of data D and the sample size n, the υ_k is the number of independent parameters needed to be estimated and $p(D\ |M_k)$ is the probability that the DIMM maximizes the data at the given MLE of the unknown parameters $\widehat{\theta_k}$. Other criteria measures can be used such as AIC (Akaike's Information Criteria) or for simplicity sake only the likelihood value.

Initial Starting Values of the Expectation-Maximization Algorithm

Estimation Using R Software

Due the complex nature of the likelihood function, when utilizing different initial starting values, the computer program may converge to different local maxima for the log likelihood. In order to obtain the global maxima of the likelihood function a range of initial starting values are used to obtain the parameter estimates. The different initial parameters were obtained by the results from the k-means algorithm where we increased the random starting values utilized to 25 in R software. When fitting the actual data from the Northern Manhattan Study, a strategy was adopted consisting of running the EM algorithm in R software for a univariate mixture model to convergence at least 10 times with different k-means initial starting values and then choosing the optimal BIC. For each simulation with the computer program in R software, the EM algorithm can be called up to a maximum of three times: first, using a random starting value for the partioning around mediods method; second, using a random starting value for the k-means method and third, using 25 random starting values for the k-means method. The partitioning around mediods (k-medoids method), which is a variation of k-means clustering, where instead of using an imaginary point in the center of a cluster an actual data point closest to the center is used. Convergence was defined as when the increase in the log-likelihood from each iterative process was within the threshold of 0.01 in R software.

Estimation Using Canned Software - MPlus

Mplus software is capable of fitting multivariate mixture models and performs estimation using the EM algorithm. The DIMM can be programmed in Mplus by treating the varying dimensions as missing data but also adding constraints so that the beta was fixed to be the same across all repeated measures within each individual. The different initial parameters were obtained by the results from the k-means algorithm where we increased the random starting values utilized to 50 in MPlus software. See MPlus code in Appendix. The results for the simulation and application data analysis are reported based on the output in R software and compared to and verified by the MPlus software (5-20 minutes), which required less computational time and quicker speed for model convergence than in R software (12 hours). One limitation of Mplus is that it can only model the varying dimensions as Poisson rather than truncated Poisson. Also, the simulation studies are more cumbersome to conduct in MPLUS due to that fact that the specific maximum number of repeated measures (for a right truncated Poisson distribution) must be specified a priori when running the MPLUS program where as our R program allows for the simulated data structure to be of any length as long as it is less than the maximum number of repeated measures. Although not elaborated herein, it is also possible to perform a multivariate version of the DIMM in Mplus. Appendix B shows example code.

Monte Carlo Experiments

Simulations

The motivation of the simulation is to examine the performance of the DIMM for clustering individuals and estimating cluster characteristics. The true underlying structure is created under 4 scenarios (Table 6): assuming informative betas (different means of the underlying mixture distribution) or uninformative betas (similar means of the underlying mixture distribution); and assuming informative alphas (different average dimension) or uninformative alphas (equal average dimension). DIMM along with a traditional finite mixture model not incorporating a model for the dimension size are fit to all 4 scenarios. We present the simulation section by, first, describing the data setup (how the data was generated); next, specifying the model fitting (how the two models were fit to the simulated data) and finally, the simulation results (a summary of the results for the first 2 scenarios, informative betas, followed by the last 2 scenarios, uninformative betas).

Generation of Data for Different Scenarios

The Monte Carlo experiment consists of creating a sample size of 300 subjects of simulated data from three underlying univariate normal distributions (Scenario 1 & 2) and 300 subjects from two underlying normal distributions (Scenario 3 & 4). In Scenarios 1 and 2, we fix the number of subjects in each of the 3 clusters to be 100 so that π_1 = 0.333, π_2 = 0.333, π_3 = 0.333, and in Scenarios 3 and 4, we fix the number of subjects in cluster 1 to be 120 and in cluster 2 to be 180 so that π_1 = 0.4, π_2 = 0.6.

Given the fixed true cluster membership, the specific parameters β_k, σ_k^2, α_k, are specified in Table 6 and are used to generate the number of repeated measures as well as the observed values. First, the count data (based on α_k) is simulated as a random variable from a Poisson distribution with the true average dimension specified for each cluster. Then, null values and values greater than 10 are excluded and sampled with replacement to simulate the truncated Poisson distribution of varying (informative) or similar (uninformative) average lengths. Once the count data, d_i, is simulated for each person, then d_i repeated measurements are generated for individual i from a univariate normal distribution with mean β_k and variance σ_k^2. The layout of the data matrix is shown in Table 7 and Table 8. The layout shows individuals with randomly generated means (β_k) and repeated measurements (α_k) across 10 different activities given their affiliated k^{th} cluster membership. We see (Table 7) that those individuals coming from true cluster 3 have means around 3 and approximately 7 repeated measures, whereas the individuals from cluster 1 have only 1 measure and have means around 0. Moreover, in Table 8, the mean response value is around 0 for all individuals but those from cluster 1 only have 1 repeated measure while those from cluster 2 have many more, averaging over 7.

In Scenarios 1 and 2, the rationale behind choosing the values of the beta parameter **β = (0, -3, 3)** was that the expression profile values would be standardized to have a mean of 0 and a standard deviation of 1. For example, in the field of physical activity, the measurement of energy expenditure (kcal/week) can be standardized according to the guidelines recommended by the American Heart Association. Subjects around $\boldsymbol{\beta}_1$ = (0) would be around the 50th percentile while subjects around $\boldsymbol{\beta}_2$ = (-3) or $\boldsymbol{\beta}_3$ = (3) would be three standard deviations away from the mean. More importantly, the magnitude of difference in the means of the clusters should make the underlying three clusters relatively easy to identify and to distinguish. In Scenarios 3, the rationale for the choosing values of 1.58 and 7.85 for the α_k was to make the underlying clusters as distinguishable as possible within the range of 1 to 10 possible repeated measures. The rationale for including Scenario 4 where no parameters were informative was to demonstrate the situation where the DIMM breaks down. Essentially, if there is no information in the means or dimension size about clustering, it is of interest to see how the DIMM performs.

Models Fit to Each Scenario

The two models (DIMM and traditional mixture model) will be used to fit the simulated data from each of the four scenarios. For each scenario, 100 experimental datasets are generated. The Expectation-Maximization (EM) algorithm is used to estimate the parameters of 100 simulations for each model. Next, the parameters of interest needed to be summarized across all simulations are as follows: β_k, σ_k^2, α_k, π_k and the predicted value metric score (Tibshirani et al 2005) will be calculated which measures accuracy of clustering.

Due to the label-switching problem where the actual labels of the groups are unknown, the parameter estimates were sorted before summarized by the individual classes. As an unsupervised learning problem, where the labels of each of the three underlying univariate normal distributions are not attached to the betas, the model does not know which group should be presented first when listing the result of the simulation. Thus, the results are not in order whereas when the prediction of the truth was created, there was an order to the simulated data where the first group of individuals was simulated for the first underlying distribution, followed by the next group of individuals for the next underlying distribution.

Simulation Results.

Parameter Estimates

The simulation results show the summary of the estimated mean and the standard error in parentheses in Tables 9 and 10. Scenario 1 summary results show nearly identical results between the DIMM model and the traditional mixture model in terms of all of the parameter estimates being unbiased and near the specified truth. Scenario 2 shows that the regular mixture model does not do as well as DIMM in estimating the betas, estimated to be within 0.039 [=(-3) – (-2.961)] (more than twice as high as 0.016= [0 – (-0.016)] with DIMM) and, in addition, the alphas, estimated to be within 0.873 (=7.006-6.133) (higher than the bias found using DIMM of 0.573=7.006-6.43). The estimated variances are also more biased using the traditional mixture model 0.816 -1 = 0.184 (higher than the maximum bias of 0.012=1-0.988 for DIMM) and the mixing proportions are estimated to be within 0.044=0.333-0.289 (higher than 0.001=0.334-0.333 in DIMM).

Scenario 4 can be considered to be a degenerate case where there are no clusters created that could be differentiated based on the betas and alphas. Therefore, neither method works well and the data is split equally into two groups since there is no information to distinguish the two clusters in Scenario 4. Scenario 3 shows that the regular mixture model does not do as well as DIMM due to the fact that it predicts both clusters equally with a 50%-50% probability (versus 41%-59% for DIMM) and the alpha estimates are of similar values from 4.5 to 4.6 instead of being different. Note that the left truncated MLE alpha parameter value for Class 2 of 8.840 using DIMM for Scenario 3 is close to the truth of 7.854 when a doubly truncated Poisson distribution of a million observations is simulated (by generating a Poisson(10) and then truncating at 1 and at 10 due to the length of the data matrix columns).

Thus, we find the DIMM, which utilizes additional information regarding the distribution of the repeated measures outperforms the traditional mixture model in terms of estimation of the model parameters to reflect the known underlying truth. The differences between these two models will be expounded upon further next in the cluster prediction section.

Predicting the cluster membership

The predictive value using Tibshirani et al 2005 method is used to calculate the how well the predicted group membership compares to the true group membership regardless of the cluster labels. The predicted group membership is defined as the highest (maximum) class membership probability out of all of the groups. If all of the subjects are specified into their correct predicted groups based on the model and compared to their known true assignments, then the maximum value of the predictive value is 0.333 [$0.333^2 + 0.333^2 + 0.333^2$] based on a three cluster partitioning of the data simulated from the truth (Scenario 1 & 2) and 0.520 [$0.4^2 + 0.6^2$] based on two clusters specified in the truth (Scenario 3 & 4).

In Scenario 1 similar values of 0.32 for both models (which reflect accurate clustering as compared to 0.333) show that both models do fine with prediction of clusters when there is no difference in the dimension of repeated measures between the clusters. On the other hand, using a traditional mixture model in Scenario 2 shows a lower predictive value of 0.303 when compared to DIMM, which has a higher predictive value of 0.319, an indicator of more accurate clustering. As expected, Scenario 4 produces similar values of 0.26 (which reflect low ability of clustering as compared to 0.52) for both models, the regular mixture model and DIMM. While in Scenario 3, a lower predictive value of 0.28 for the traditional mixture model as compared to DIMM's predictive value of 0.51 indicate greater misspecification of certain individuals into incorrectly predicted cluster groups based on their known true class assignments. These results are expected when we consider the fact that the simulations were conducted when the beta and alpha parameters are treated as uninformative in Scenario 4 and the alpha parameters are not accounted for in the model prediction for the group membership probabilities in Scenario 3 for the regular mixture model; therefore, limiting accurate prediction of group membership for assigning class membership. This confirms the need to properly take into account the distribution of the repeated measurements in the model to obtain relevant and accurate clustering structures for data sets that are highly variable.

Northern Manhattan Study

The Northern Manhattan Study is a multi-ethnic cohort study of elderly individuals residing in the Northern Manhattan region of New York City. The reported leisure time physical activity provides useful information regarding the varying number and different type of activities consumed where the number and type of activities are random yet informative. A total of 1971 elderly adults answered that they had conducted at least one of the following fifteen physical activities listed and the summary statistics of the total energy expenditure of each activity are provided in Table 11.

According to the guidelines from the World Health Organization regarding leisurely physical activity, a MET (metabolic equivalent in units of [kcal/kg-hr]) score above 0 can be considered to be light activity, above 3 can be considered moderate activity and above 6 can be considered heavy or intense. The MET scale has the unit of 1 kcal/kg-hour. Participants on average expend 300 kcal/week (bowling) – 1400 kcal/week (golf) and a median of 280 kcal/week (gardening) – 850 kcal/week (running) depending on the activity type, the frequency and the duration of conduct.

When the average weekly energy expenditure per activity (kcal/activity/week) of all fifteen physical activity items is summarized per subject, the average weekly energy expenditure is 970 kcal with a standard deviation of 1110 kcal. The distribution is skewed to the right and the median energy expenditure is 650 kcal with an interquartile range of 340 kcal to 1220 kcal. The distribution of the average weekly energy expenditure in kilocalories is shown in Figure 1.

The average weekly total number of physical activity items reported is 1.39 items with a standard deviation of 0.71 items. In the sample 70.2% of individuals conduct just one activity and no one conducts more than 7. The histogram and frequency table are presented in Figure 2. Figure 3 shows that (as expected) there is a positive association between an individual’s total energy expenditure and the reported number of physical activity items. When summarizing information into total energy expenditure and total physical activities conducted, individual patterns of energy expenditures across physical activities items over all fifteen activities can be lost. The goal is to cluster subjects based on their energy expenditure across the different activities they perform while taking into account that they may be performing a different number of activities.

DIMM of the NOMAS Data

The single attribute data structure can be denoted as $Y_{1923 \times 15}$, where Y is the energy expenditure in kcal/week, 1971 is the total individuals that performed at least one physical activity and 15 is the total number of physical activities listed within a two week period. A two week period was chosen to maximize the variability in the count of physical activities reported as opposed to a one week interval. In this data set, every elderly individual has completed at least one physical activity item and the maximum total number of physical activities completed is seven. The goal is to summarize this physical activity profile into groups to elucidate clustering patterns to inform future cardiovascular health outcomes as recommended by the guidelines of the American Heart Association.

Before conducting the model-based cluster analysis using the DIMM, the outcome measurement of energy expenditure (kcal/2 weeks) was transformed on the natural logarithm scale to satisfy the normality assumption. Based on the minimum BIC value (Table 12), our model-based clustering method found two clusters as the optimal solution. We present results for the 1, 2 and 3 cluster model for comparison in Table 13. The one class model results essentially reproduces the sample mean kcal (1.950 = 1950 kcal [=exp(0.116+(1.104/2))], back-transformation of log normal) and average count of activities (1.39 = exp(.330))]. In the two cluster result, each cluster comprises about half of the sample and subjects on average complete 1.43 [=exp(0.355)] and 1.35 [=exp(0.301)] physical activities respectively. We can characterize the first cluster as a low energy expenditure cluster with an average of 1329 [=exp(-0.274+(1.117/2))] kcal/2 weeks while the second cluster as a high energy expenditure cluster with an average of 2524 [=exp(0.580+(0.692/2))] kcal/2 weeks. By taking into account the individual energy expenditure for the fifteen physical activity items, we found two groups that had different energy expression profiles while on average conducting the same number of physical activity items. The sample was equally split into two groups: high energy profile group (54%) and low energy profile group (46%). In fact, it is interesting to note that the group with the low energy profile has a slightly higher number of physical activity items while the group with the high energy profile has a slightly lower number of physical activity items. This may be due to the fact that individuals may conduct fewer physical activity items but those items happen to be more intense on the MET scale than the others who conduct more physical activity items.

The three class result gives similar results to the two class analysis in terms of finding both a low and high energy profile cluster with similar average energy expenditures as in the two class model. The only exception is that 74% of individuals are in the high energy profile cluster while only 25% are in the low energy profile cluster using DIMM. Moreover, individuals exercising at a very high energy expenditure of 2802 kcal/2 weeks [=exp(0.981+(0.099/2))] are filtered out into their own separate group. This high consumption energy group comprises only 0.2% of the total sample, a minutely small group yet having a higher mean energy expenditure.

The NOMAS data were also analyzed using the traditional mixture model, results were similar. For the regular mixture model, the two class solution is exactly identical to the two class solution for DIMM expect for the different proportion of individuals, which shows the added utility of using DIMM to find similar proportions of individuals in the high and low profile clusters. See Table 14 and 15.

According to the American Heart Association, adults are recommended to engage in at least 30 minutes of moderate leisurely physical activity at least 5 times a week. If we assume that subjects are walking, then this can be converted to about 1364 kcal expended per 2 week interval for an elderly individual of 150 lb of body weight. Thus, using the model based clustering method that we propose, we were able to find a two cluster solution in which half of the participants expended on average 1555 kcal per every 2 weeks while the other half on average 2368 kcal per every 2 weeks. This implies that on average half of the participants meet AHA guidelines for recommended exercise while the other half of participants exceed AHA guidelines. This has important public health implications in regards to elucidating subjects that meet guidelines for active physical activity given the complex nature of their consumption patterns in terms of the summary measure of their duration, frequency and MET intensity: their total energy expenditure.

Discussion

The DIMM for a single attribute data structure with an informative number of repeated measures proposed in this paper can be implemented and estimated with the Expectation-Maximization algorithm. The utility of this model can be extended to situations when the informative number of repeated measures are varied and not fixed. The count dimensional informative data structure for DIMM is an extension of the application of current mixture models on a fixed dimensional data structure.

The simulation section shows that taking into account the additional information on the repeated measures can help DIMM accurately predict the given truth based on the group membership when repeated measures are informative and show variability with each of the underlying distributions. Regardless of the informativeness of the beta mean parameters of the underlying distributions, when additional informative information on the count variable is taken into count the DIMM outperforms the regular mixture model in terms of estimating unbiased model parameters and accurately predicting the true group membership. By taking into account the dimensional informative lengths of the repeated measurement of each subject, we can incorporate additional information to help cluster individuals based on their single summary attribute value.

Table 1. NOMAS Survey for Assessment of Physical Activity

60c. Ask the subject if they performed any of the following activities: Record 0 for No and 1 for Yes;
FOR EACH YES, ASK
On the average, how many times in a ***typical*** 14 day period do you do this activity? *FILL IN TIMES*
Estimate how many minutes you actually spend on each occasion? *FILL IN MINUTES*

Activity	Yes/No	Times	Minutes
1. Walking for exercise?	PA1 ____	PA1T _____	PA1M _____
2. Jogging or running?	PA2 ____	PA2T _____	PA2M _____
3. Hiking?	PA3 ____	PA3T _____	PA3M _____
4. Gardening or yard work?	PA4 ____	PA4T _____	PA4M _____
5. Aerobics or aerobic dancing?	PA5 ____	PA5T _____	PA5M _____
6. Other dancing?	PA6 ____	PA6T _____	PA6M _____
7. Calisthenics or general exercise?	PA7 ____	PA7T _____	PA7M _____
8. Golf?	PA8 ____	PA8T _____	PA8M _____
9. Tennis?	PA9 ____	PA9T _____	PA9M _____
10. Bowling?	PA10 ___	PA10T ____	PA10M ____
11. Bicycle riding?	PA11 ___	PA11T ____	PA11M ____
12. Swimming or water exercises?	PA12 ___	PA12T ____	PA12M ____
13. Horseback riding?	PA13 ___	PA13T ____	PA13M ____
14. Handball, racquetball, or squash?	PA14 ___	PA14T ____	PA14M ____
15a. Have you done any other exercises, sports, or physically active hobbies in the past 2 weeks other than the ones listed above?		PA15a __	*IF YES GO TO 15b*
15b. What were they? PA15b ___________________		PA15bT ___	PA15bM ___
PA15c ___________________		PA15cT ___	PA15cM ___
PA15d ___________________		PA15dT ___	PA15dM ___

Table 2. An Example from the Northern Manhattan Study (NOMAS)

	Activity 1				Activity 2				Activity 3					Activity 15	Total Activities Conducted (d)
Individual	Total Frequency (y_{i11})	Average Duration (y_{i12})	METS (y_{i13})	Type (y_{i14})	Total Frequency (y_{i21})	Average Duration (y_{i22})	METS (y_{i23})	Type (y_{i24})	Total Frequency (y_{i31})	Average Duration (y_{i32})	METS (y_{i33})	Type (y_{i34})	…	METS (y_{i153})	
1	5	40	4	Walking	.	.	.		.	.	.		…	.	1
2	4	50	4	Walking	.	.	.		.	.	.		…	.	1
3	3	30	7	Jogging	1	240	4	Golf	.	.	.		…	.	2
4	6	45	4	Walking	1	150	6	Hiking	.	.	.		…	.	2
5	5	45	7	jogging	2	120	7	Tennis	1	150	3	Bowling	…	.	3

Table 3. Drug Use Among Men Who Have Sex with Men (MSM) who use the Internet

Individual	Drug 1				Drug 2				Drug 3				Drug 4				…	Drug 19	Total Drugs (d)
	First Time Use (y_{i11})	Age of Onset (y_{i12})	Partner Use (y_{i13})	Type (y_{i14})	First Time Use (y_{i21})	Age of Onset (y_{i22})	Partner Use (y_{i23})	Type (y_{i24})	First Time Use (y_{i31})	Age of Onset (y_{i32})	Partner Use (y_{i33})	Type (y_{i34})	First Time Use (y_{i41})	Age of Onset (y_{i42})	Partner Use (y_{i43})	Type (y_{i44})		Type (y_{i194})	
1	0	15	0	Alcohol													.		1
2	0	14	1	Alcohol	1	21	0	Poppers	0	16	1	Marijuana					.		3
3	0	18	0	Ecstasy													.		1
4	0	16	1	Alcohol	0	15	0	Marijuana	0	19	1	Ecstasy	1	24	1	Injectable Heroin	.		4

Table 4. An Example from the Treatment of Breast Tumors

Individual	Tumor 1			Lesion 2			Lesion 3			Lesion 4			Lesion 5			Lesion 6			Lesion 7				Lesion 10	Total Lesions (d)
	Pre-Tmt Size (y_{i11})	Post-Tmt Size (y_{i12})	Type (y_{i13})	Pre-Tmt Size (y_{i21})	Post-Tmt Size (y_{i22})	Type (y_{i23})	Pre-Tmt Size (y_{i31})	Post-Tmt Size (y_{i32})	Type (y_{i33})	Pre-Tmt Size (y_{i41})	Post-Tmt Size (y_{i42})	Type (y_{i43})	Pre-Tmt Size (y_{i51})	Post-Tmt Size (y_{i52})	Type (y_{i53})	Pre-Tmt Size (y_{i61})	Post-Tmt Size (y_{i62})	Type (y_{i63})	Pre-Tmt Size (y_{i71})	Post-Tmt Size (y_{i72})	Type (y_{i73})			
1	1	1	I																			.		1
2	2	2	I																			.		1
3	2	5	I	1	4	II	2	5	II	3	6	II										.		4
4	3	6	I	2	7	II	2	4	II	3	6	II										.		4
5	3	0	I	2	0	II	3	1	II	4	1	II	3	0	III	4	1	III	5	2	IV	.		7
6	4	1	I	1	2	II	3	2	II	4	2	II	4	1	III	5	2	III	6	3	IV	.		7

Table 5. Total Energy Expenditure Per Physical Activity Item in NOMAS

Individual	Energy Expenditure 1 (y_{i11})	Type (y_{i12})	Energy Expenditure 2 (y_{i21})	Type (y_{i12})	Energy Expenditure 2 (y_{i31})	Type (y_{i32})	…	Type (y_{i152})	Total Energy Expenditure	Total Activities Conducted (d)
1	1.25	Walking	.	.	.	.	…	.	1.25	1
2	1.00	Walking	.	.	.	.	…	.	1.00	1
3	1.15	Jogging	0.54	Golf	.	.	…	.	1.69	2
4	1.40	Walking	0.32	Hiking	.	.	…	.	1.72	2
5	1.50	Jogging	0.25	Tennis	0.75	Bowling	…	.	2.50	3

Table 6: Simulation Scenarios (1-4) and Summary of Modeling Results: Overview in the Summary Table

		Features of the Discriminating Clusters (k)			**Estimation Performance of Models**	
	k	beta (means)	alpha (dimension)	Description	Traditional Mixture	DIMM
1	3	(0, 3, -3)	(4.075, 4.075, 4.075)	Informative mean, Non-informative dimension	**Good** - unbiased beta, accurate clustering	**Good**- unbiased beta, accurate clustering
2	3	(0, 3, -3)	(1.582, 4.075, 7.006)	Informative mean and Dimension	**Poor -** bias beta, low accurate clustering	**Good** - unbiased beta, accurate clustering
3	2	(0, 0)	(1.582, 7.854)	Non-informative mean, Informative dimension	**Poor** - no ability to cluster	**Good**- good accuracy for cluster prediction
4	2	(0, 0)	(4.075, 4.075)	Nothing distinguishing clusters	**Poor**- no ability to cluster	**Poor** - no ability to cluster

Table 7. Layout of the Data Matrix for the Simulation Study Assuming Informative Betas

Id	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	y_{i6}	y_{i7}	y_{i8}	y_{i9}	y_{i10}	d_i	True Cluster Membership
1	0.01	.	.	.	.	.	.	.	.	.	1	1
2	-0.02	.	.	.	.	.	.	.	.	.	1	1
3	-3.52	-3.03	-3.02	-2.52	.	.	.	.	.	.	4	2
4	-4.02	-3.01	-3.04	-2.01	-2.50	.	.	.	.	.	5	2
5	2.25	2.52	2.75	3.05	3.75	3.51	3.25	3.01	.	.	8	3
6	2.01	2.33	2.66	3.03	3.66	3.33	3.02	.	.	.	7	3

Table 8. Layout of the Data Matrix for the Simulation Assuming Uninformative Betas

Id	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}	y_{i6}	y_{i7}	y_{i8}	y_{i9}	y_{i10}	d_i	True Cluster Membership
1	-0.01	.	.	.	.	.	.	.	.	.	1	1
2	0.03	-0.5	.	.	.	.	.	.	.	.	2	1
3	-0.41	-0.32	-0.21	-0.14	0.02	0.03	0.11	0.23	0.33	0.41	10	2
4	-1.02	-0.75	-0.50	-0.25	0.04	0.01	0.25	0.50	.	.	8	2

Table 9. Results of the Simulation Assuming Informative Betas

	Scenario 1: Non-informative alphas			Scenario 2: Informative alphas		
	TRUTH	Traditional Mixture Model	DIMM	TRUTH	Traditional Mixture Model	DIMM
		Mean (SE)	Mean (SE)		Mean (SE)	Mean (SE)
β_1	0	0.007 (0.052)	0.007 (0.053)	0	0.017 (0.084)	-0.016 (0.093)
β_2	-3	-3.000 (0.046)	-3.000 (0.046)	-3	-2.961 (0.053)	-2.992 (0.053)
β_3	3	3.005 (0.051)	3.010 (0.051)	3	2.980 (0.045)	3.001 (0.044)
π_1	0.333	0.334 (0.005)	0.334 (0.006)	0.333	0.289 (0.024)	0.332 (0.007)
π_2	0.333	0.334 (0.004)	0.334 (0.004)	0.333	0.354 (0.024)	0.334 (0.005)
π_3	0.333	0.332 (0.004)	0.332 (0.004)	0.333	0.358 (0.008)	0.333 (0.004)
σ^2_1	1	0.992 (0.074)	0.992 (0.074)	1	0.816 (0.112)	0.992 (0.132)
σ^2_2	1	0.990 (0.077)	0.990 (0.077)	1	1.001 (0.099)	0.988 (0.069)
σ^2_3	1	1.002 (0.072)	1.002 (0.072)	1	1.024 (0.069)	0.993 (0.066)
α_1	4.075	4.034*(0.181)	4.038 (0.190)	1.582	1.647*(0.047)	1.582 (0.088)
α_2	4.075	4.068*(0.175)	4.066 (0.183)	7.006	6.133*(0.213)	6.433 (0.210)
α_3	4.075	4.052*(0.207)	4.050 (0.211)	4.075	3.881*(0.190)	4.063 (0.183)
Predictive Value	0.333	0.321 (0.004)	0.321 (0.004)	0.333	0.303 (0.008)	0.319 (0.005)
Total Expected n (number of obs on avg/ class)	(400, 400, 400)			(100, 700, 400)		

$\alpha^*_{truncated}$ – this is the posterior predicted (post-hoc) alpha parameter because it is not estimated in Model 1 and it is non-informative because it is not used to calculate the group membership probabilities.

Table 10. Results of the Simulation Assuming Uninformative Betas

	Scenario 3: Informative alphas			Scenario 4: Non-Informative alphas		
	TRUTH	Traditional Mixture Model	DIMM	TRUTH	Traditional Mixture Model	DIMM
		Mean (SE)	Mean (SE)		Mean (SE)	Mean (SE)
β_1	0	-0.007 (0.081)	-0.000 (0.045)	0	-0.002 (0.078)	-0.002 (0.041)
β_2	0	-0.002 (0.094)	0.010 (0.055)	0	-0.006 (0.082)	-0.005 (0.047)
π_1	0.4	0.505 (0.031)	0.408 (0.005)	0.4	0.501 (0.006)	0.501 (0.004)
π_2	0.6	0.495 (0.031)	0.592 (0.005)	0.6	0.499 (0.006)	0.499 (0.004)
σ^2_1	1	0.982 (0.076)	0.989 (0.061)	1	0.992 (0.094)	0.992 (0.045)
σ^2_2	1	0.996 (0.087)	0.997 (0.071)	1	0.984 (0.095)	0.992 (0.046)
α_1	1.582	4.590*(0.493)	1.573 (0.058)	4.075	4.084*(0.163)	4.082 (0.199)
α_2	7.854**	4.620*(0.552)	8.840 (0.112)	4.075	4.064*(0.164)	4.036 (0.241)
Predictive Value	0.520	0.280 (0.036)	0.508 (0.005)	0.520	0.264 (0.009)	0.259 (0.004)
Total Expected n (number of obs on avg/ class)	(120, 1800)			(480, 720)		

$\alpha^*_{\text{truncated}}$ – this is the posterior predicted (post-hoc) alpha parameter because it is not estimated in Model 1 and it is non-informative because it is not used to calculate the group membership probabilities.

** This is the simulated mean of a million values for a doubly-truncated Poisson distribution at 1 and 10.

Table 11. Descriptive summaries of energy expenditure associated with each activity of the NOMAS Sample

Physical Activity Energy Expenditure (10^3 kcal/week)	MET[a] Scale	N[b] = 1971	Mean (SD)	Median (IQR)	Min	Max
1. Walking	4	1636	1.11 (1.19)	0.78 (0.38 – 1.45)	20.30	15.10
2.Jogging or Running	7	51	0.99 (0.84)	0.85 (0.36 – 1.33)	0.10	3.87
3.Hiking	6	11	0.71 (0.57)	0.51 (0.17 – 1.23)	0.06	1.66
4.Gardening or Yard Work	4	32	1.01 (1.66)	0.28 (0.09 – 0.83)	0.04	5.96
5.Aerobics or Aerobic Dancing	5.5	100	0.65 (0.95)	0.41 (0.24 – 0.76)	0.04	8.87
6.Other Dancing	5	67	0.65 (1.02)	0.35 (0.17 – 0.63)	0.02	5.73
7.Calisthenics or General Exercise	5	476	0.54 (0.55)	0.40 (0.22 – 0.65)	0.03	5.94
8.Golf	4	14	1.40 (1.49)	0.70 (0.34 – 1.86)	0.05	5.37
9.Tennis	7	6	0.63 (0.44)	0.66 (0.28 – 0.98)	0.07	1.11
10.Bowling	3	7	0.30 (0.13)	0.32 (0.24 – 0.36)	0.10	0.52
11.Bicycle Riding	5.5	95	0.82 (0.92)	0.57 (0.25 – 0.98)	0.03	5.29
12.Swimming or Water Exercises	6	63	0.79 (0.66)	0.58 (0.34 – 1.18)	0.02	3.25
13.Handball, Racquetball, or Squash	10	2	0.56 (0.38)	0.56 (0.29 – 0.83)	0.29	0.83
14. Other Activity 1[c]	---	171	0.84 (2.15)	0.34 (0.18 – 0.65)	0.00	22.25
15. Other Activity 2[c]	---	11	1.28 (2.61)	0.32 (0.20 – 0.96)	0.04	8.89
Avg kcal		1971	0.97 (1.11)	0.65 (0.34 – 1.22)	0.02	15.10
Avg Total kcal		1971	1.31 (1.52)	0.86 (0.41 – 1.66)	0.02	22.25

[a] Metabolic Equivalent of Task

[b] Total Number of Individuals Reporting the Physical Activity

[c] Includes the following additional activities: arm curls, body sculpting, boxing, carpentry, church activities, cleaning house, climbing stairs, playing dominos, dumbbells, exercise bike, fishing, garbage packing, hand weights, jump rope, lifting weights, mopping at work, Nautilus machine, Nordic track, osteoarthritis rehabilitation, physical therapy, pulling exercise, push-ups, recycling exercise, rehabilitation exercise, repair work, rollerblading, rowing machine, sailing, stationary bicycle, shooting basketballs, sit-ups, skiing, soccer, solitaire, squats, stairmaster, step machine, stretching exercises, tai chi, taking care of mother, therapeutic exercises, treadmill, packing at fish market, yoga.

Figure 1. Distribution of the Average Weekly Energy Expenditure Per Activity (kilocalories/activity)

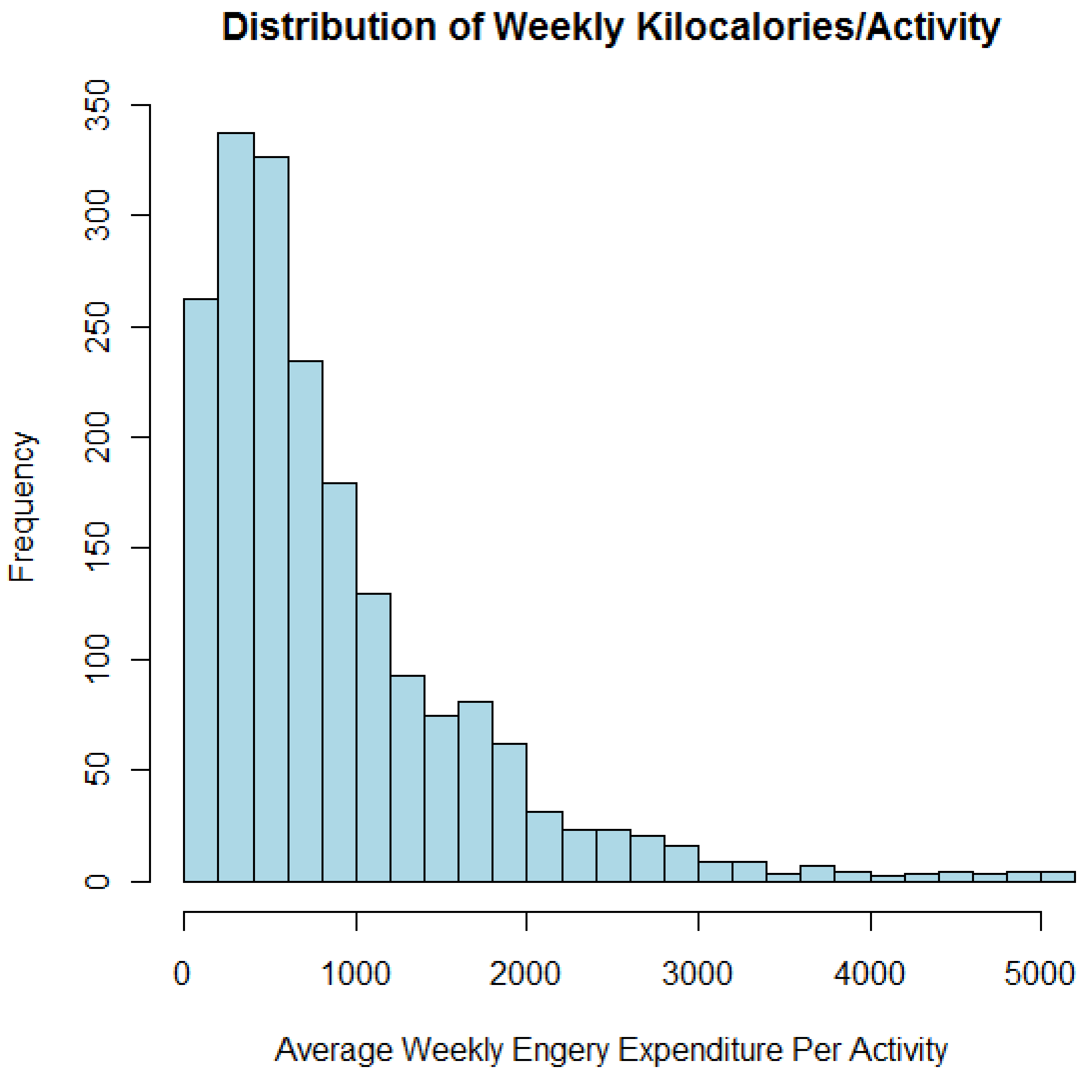
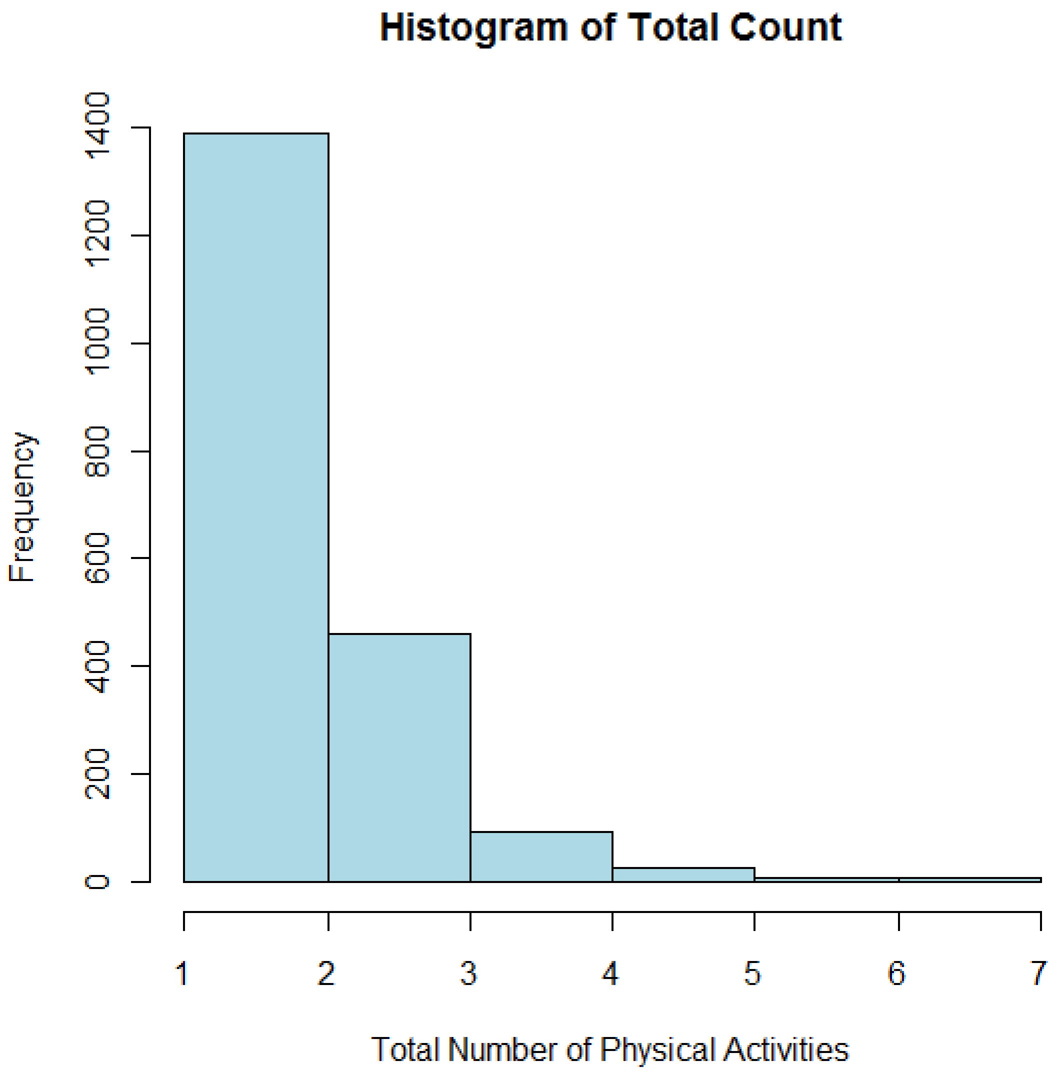
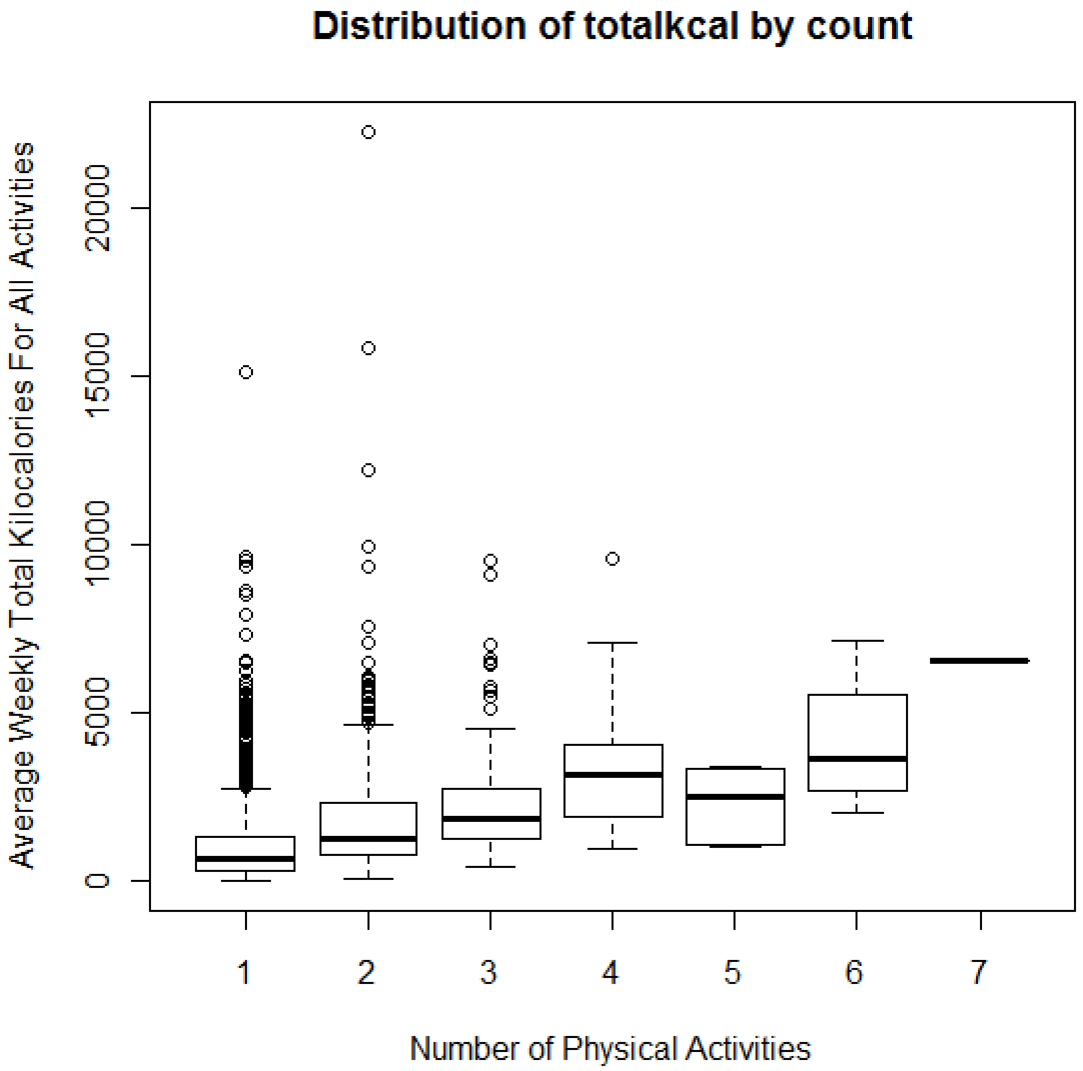


Figure 2. Distribution of the Total Number of Physical Activity Items



Repeated Physical Activity Items	1	2	3	4	5	6	7
Number of Subjects	1383	460	91	26	6	4	1

Figure 3. Total Weekly Energy Expenditure Versus Total Number of Physical Activity Items



Repeated Physical Activity Items	1	2	3	4	5	6	7
Energy Expenditure (10^3 kcal/wk) Mean (SD)	1.01 (1.20)	1.79 (1.88)	2.37 (1.85)	3.36 (1.97)	2.31 (1.07)	4.11 (2.16)	6.57 (---)
Median (IQR)	0.66 (0.31 – 1.29)	1.23 (0.75 – 2.32)	1.82 (1.25 – 2.75)	3.14 (1.91 – 4.03)	2.51 (1.09 – 3.33)	3.64 (2.67 – 5.55)	6.57 (---)

Table 12. Optimal Number of Clusters for the DIMM

K	log-Likelihood	Parameters (p)	BIC = -2*log-L + k*p*ln(1971)
1	-6472.29	3	12967.34
2	**-6445.51**	**7**	**12944.13**
3	-6441.21	11	12965.87

Table 13. Model Parameter Estimates for the 1 Class, 2 Class, 3 Class Solutions using DIMM

	1 Class Solution	2 Class Solution		3 Class Solution		
Ln(kcal) (SE)	0.116 (0.021)	-0.274 (0.150)	0.580 (0.145)	0.406 (0.119)	-0.399 (0.152)	0.981 (0.084)
π	1.000	0.459	0.541	0.745	0.253	0.002
σ^2 (SE)	1.104 (0.031)	1.117 (0.079)	0.692 (0.112)	0.838 (0.073)	1.132 (0.096)	0.099 (0.042)
Ln($\alpha_{truncated}$) (SE)	0.330 (0.011)	0.355 (0.016)	0.301 (0.022)	0.317 (0.018)	0.356 (0.018)	0.265 (0.059)
Total Count of PA	1.391	1.426	1.351	1.373	1.428	1.303
Mean kcal/ every 2 wks from model (10^3)	1.950	1.329	2.524	2.282	1.182	2.802
Mean kcal/ every 2 wks from data (10^3)	1.948	0.691	3.016	2.451	0.462	2.508
Total kcal/ every 2 wks From data (10^3)	2.611	1.157	3.844	3.184	0.864	9.302

Table 14. Optimal Number of Clusters for the Regular Mixture Method

K	log-Likelihood	Parameters (p)	BIC = -2*log-L + k*p*ln(1971)
1	-4007.129	2	8029.431
2	**-3981.065**	**5**	**8000.059**
3	-3976.592	8	8013.871

Table 15. Model Parameter Estimates for the 1 Class,2 Class, 3 Class Solutions for the Regular Mixture Method

	1 Class Solution	2 Class Solution		3 Class Solution		
Ln(kcal) (SE)	0.116 (0.021)	-0.287 (0.152)	0.552 (0.128)	0.398 (0.099)	-0.403 (0.144)	0.978 (0.078)
π	1.000	0.424	0.576	0.757	0.241	0.002
σ^2 (SE)	1.104 (0.031)	1.133 (0.081)	0.707 (0.104)	0.839 (0.068)	1.144 (0.097)	0.096 (0.040)
Ln($\alpha_{truncated}$)* (SE)						
Total Count of PA*						
Mean kcal/ every 2 wks from model (10^3)	1.945	1.322	2.473	2.265	1.184	2.790

*The alpha parameter is not taken into account and is not estimated in the regular mixture model.

Appendix: Latent Class Analysis

Latent class analysis (LCA) is a model based clustering method that utilizes multivariate binary observations. An additional analysis was undertaken to explore whether there were patterns of certain types of activities (i.e. walking, hiking) that were found to cluster together. A 2, 3, and 4 class latent class model was fit to the 15 (yes/no) activities performed. That is each person contributed a 15 x 1 vector indicating which of the 15 activities were performed. Table Appendix1 shows the results. When two clusters are specified then two labels result: a mostly walking class (78.2% of the sample) and a class which is more likely to conduct a variety of activities, conducts all fifteen activities frequently and punctually. When four clusters are specified then four labels result: a walking class, a walking and calisthenics class, an other activities class and a class that conducts all fifteen activities frequently and punctually. With 3 classes there is still the mostly walking class, now 75% of the sample, and there is a class who walks and also does calisthenics making up 19% of the sample, and a cluster of people more likely to write in their own other activity. The LCA shows that the optimal number of clusters should be three based on the minimum BIC (Bayesian Informational Criteria) value. When three clusters are specified then three labels result: a walking class 75% of the sample, a walking and calisthenics class, 19% of the sample and finally, a class more likely to conduct a variety of activities especially writing in their own "other" activities. Subjects in the walking class on average expend 1250 kcal and complete 1 physical activity item, while subjects in the walking and calisthenics class on average expend 1870 kcal and complete 2 physical activity items and subjects in the all inclusive activities group on average expend 1650 kcal and complete 2 physical activity items.

Table Appendix 1. Latent Class Analysis Using Physical Activity as a Binary Outcome (Yes/No)

Exercise Type	Marginal	2 Classes		3 Classes			4 Classes			
1. Walking	82.9	38.4	100.0	43.7	100.0	39.3	100.0	50.2	38.3	65.6
2. Jogging or Running	2.6	6.5	1.1	7.1	1.0	3.8	0.7	13.3	3.1	1.3
3. Hiking	0.6	2.0	0.0	2.3	0.0	0.0	0.0	4.2	0.0	0.0
4. Gardening or Yard Work	1.7	5.3	0.3	6.1	0.3	0.0	0.3	11.4	0.0	0.0
5. Aerobics or Aerobic Dancing	5.1	10.1	3.1	12.1	3.0	0.0	3.0	18.9	0.0	3.7
6. Other Dancing	3.4	7.2	2.0	8.1	1.9	1.7	1.8	15.1	1.4	1.1
7. Calisthenics or General Exercise	24.2	48.8	14.7	60.0	13.4	4.1	0.0	27.5	5.7	100.0
8. Golf	0.7	1.5	0.4	1.4	0.4	1.4	0.4	2.8	0.9	0.1
9. Tennis	0.3	0.8	0.1	0.9	0.1	0.0	0.1	1.5	0.0	0.1
10. Bowling	0.4	0.9	0.1	1.1	0.1	0.0	0.1	1.9	0.0	0.2
11. Bicycle Riding	4.8	12.7	1.8	14.2	1.7	3.7	1.5	22.4	2.4	4.0
12. Swimming or Water Exercises	3.2	7.7	1.4	9.6	1.2	0.0	1.0	14.4	0.0	3.3
13. Handball, Racquetball, or Squash	0.1	0.4	0.0	0.2	0.0	0.9	0.0	0.4	0.8	0.0
14. Other Activity 1	8.6	23.9	2.8	8.6	1.2	100	1.1	11.2	100	3.2
15. Other Activity 2	0.6	2.2	0.0	0.7	0.0	7.6	0.0	1.5	7.0	0.0
Class Probabilities	100.0	21.7	78.3	18.6	74.6	6.7	62.6	8.8	6.5	22.1
Avg kcal (10^3)	0.97	0.73	1.04	0.69	1.06	0.85	1.11	0.84	0.85	0.69
Avg total kcal (10^3)	1.31	1.33	1.30	1.35	1.28	1.45	1.22	1.80	1.43	1.31
Avg total physical activities	1.39	1.25	1.65	1.87	1.25	1.65	1.10	2.18	1.60	1.83
Log Likelihood	-4649.66	-4427.21		-4361.41			-4305.96			
BIC	9413.12	9089.60		9079.38			9089.86			

References

Ainsworth BE, Haskell WL, Leon AS (1993) Compendium of physical activities: classification of energy costs of human physical activities. Medicine And Science In Sports And Exercise, 25:71-80.

Ainsworth BE, Haskell WL, Whitt MC, Irwin ML, Swartz AM, Strath SJ, O'Brien WL, Bassett DR Jr, Schmitz KH, Emplaincourt PO, Jacobs DR Jr, Leon AS. (2000) Compendium of physical activities: an update of activity codes and MET intensities. Medicine and Science in Sports and Exercise, 32:S498-504.

Banfield, JD and Raftery, AE. (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics, 49:803-821.

Beswick AD, Rees K, Dieppe P, Ayis S, Gooberman-Hill R, Horwood J, Ebrahim S. (2008) Complex interventions to improve physical function and maintain independent living in elderly people: a systematic review and meta-analysis. Lancet, 371:725-735.

Borodulin K, Evenson KR, Monda K, Wen F, Herring AH, Dole N. (2010) Physical activity and sleep among pregnant women. Pediatric and Perinatal Epidemiology, 24: 45-52.

Charreire H, Casey R, Salze P, Kesse-Guyot E, Simon C, Chaix B, Banos A, Badariotti D, Touvier M, Weber C, Oppert J-M. (2010) Leisure-time physical activity and sedentary behavior clusters and their associations with overweight in middle-aged French adults. International Journal of Obesity, 34: 1293-1301.

Dempster AP, Laird NM, Rubin DB. (1977) Maximum likelihood for incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B(39): 1-38.

Duda RO, Hart PE, Stork DG. (2001) Pattern Classification. John Wiley and Sons, New York.

Dudoit S and Fridlyand J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology, 3:0036.10036.21.

Fraley C, Raftery AE. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal, 41:578-588.

Fraley C, Raftery AE. (2002) Model-Based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97(458): 611-631.

Fraley, C. and Raftery, A.E. (2006). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report no. 504, Department of Statistics, University of Washington.

Gordon, A. D. (1999). Classification. Boca Raton, Florida: Chapman & Hall Ltd.

Gorely T, Marshall SJ, Biddle SJH, Cameron N. (2007) Patterns of sedentary behaviour and physical activity among adolescents in the United Kingdom: Project STIL. Journal of Behaviour and Medicine, 30: 521-531.

Hu FB. (2002) Dietary pattern analysis: a new direction in nutritional epidemiology. Current Opinion in Lipidology, 12: 3-9.

Jago R, Fox KR, Page AS, Brockman R, Thompson JL. (2010) Physical activity and sedentary behavior typologies of 10-11 year olds. International Journal of Behavioral Nutrition and Physical Activity, 7: 59-69.

Kokkinos P and Myers J. (2010) Exercise and physical activity: clinical outcomes and applications. Circulation, 122:1637-1648.

Kukuljan S, Nowson CA, Sanders K, Daly RM. (2009) Effects of resistance exercise and fortified milk on skeletal muscle mass, muscle size, and functional performance in middle-aged and older men: an 18-mo randomized controlled trial. J Appl Physiol., 107:1864-1873.

Lee IM and Paffenbarger RS, Jr. (1998) Physical activity and stroke incidence: The Harvard Alumni Health Study. Stroke, 29:2049-2054.

Liou, YM. (2007) Patterns of physical activity and obesity indices among white-collar men in Taiwan. Journal of Nursing Research, 15(2): 138-145.

MacQueen J. (1967) Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281-297.

Milligan, G.W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50:159-179.

Moss AJ and Parsons VL. (1985) Current estimates from the National Health Interview Survey. United States, Vital and health statistics 1986:i-iv, 1-182.

Nelson ME, Rejeski WJ, Blair SN, Duncan PW, Judge JO, King AC, Macera CA, and Castaneda-Sceppa C. (2007) Physical activity and public health in older adults: recommendation from the American College of Sports Medicine and the American Heart Association. Circulation, 116:1094-1105.

Qin L-X, Self SG. (2006) The clustering of regression models method with applications in gene expression data. Biometrics, 62:526-533.

Rovniak LS, Saelens BE, Marshall SJ, Sallis JF, Frank LD, Normal GJ, Conway TL, Cain KL, Hovell MF. (2010) Adults' physical activity patterns across life domains: cluster analysis with replication. Health Psychology, 29(5):496-505.

Sacco RL, Gan R, Boden-Albala B, Lin IF, Kargman DE, Hauser WA, Shea S, Paik MC. (1998) Leisure-time physical activity and ischemic stroke risk: the Northern Manhattan Stroke Study. Stroke, 29: 380-387.

Schumacher A, Peersen K, Sommervoll L, Seljeflot I, Arnesen H, Otterstad JE. (2006) Physical performance is associated with markers of vascular inflammation in patients with coronary heart disease. Eur J Cardiovasc Prev Rehabil., 13:356-362.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6:461-464.

Siconolfi SF, Lasater TM, Snow RC, Carleton RA. (1985) Self-reported physical activity compared with maximal oxygen uptake. American Journal of Epidemiology, 122:101-105.

te Velde SJ, De Bourdeaudhuij I, Thorsdottir I, Rasmussen M, Hagstromer M, Klepp K-I, Brug J. (2007) Patterns in sedentary and exercise behaviors and associations with overweight in 9-14 year-old boys and girls – a cross-sectional study. BioMed Central Public Health, 31:7-16.

Tibshirani, R., Walther, G., Botstein, D., and Brown, P. (2005). Cluster validation by prediction strength. Journal of Computational and Graphical Statistics, 14(3):511-428.

Verbeke G, Lesaffre E. (1996) A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association, 91(433): 217- 221.

Villarroel L, Marshall G, Baron AE. (2009) Cluster analysis using multivariate mixed effect models. Statistics in Medicine, 28: 2552-2565.

Vona M, Rossi A, Capodaglio P, Rizzo, S., Servi, P., De Marchi, M., and Cobelli, F. (2004) Impact of physical training and detraining on endothelium-dependent vasodilation in patients with recent acute myocardial infarction. American Heart Journal, 147:1039-1046.

Ward JH Jr. (1963) Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58:236-244.

Wilcox S, Dowda M, Leviton LC, Bartlett-Prescott J, Bazzarre T, Campbell-Voytal K, Carpenter RA, Castro CM, Dowdy D, Dunn AL, Griffin SF, Guerra M, King AC, Ory MG, Rheaume C, Tobnick J, Wegley S.(2008) Active for life: final results from the translation of two physical activity programs. American Journal of Preventive Medicine, 35:340-351.

Willey JZ, Moon YP, Paik MC, Boden-Albala B, Sacco RL, Elkind MS. (2009). Physical activity and risk of ischemic stroke in the Northern Manhattan Study. Neurology, 73:1774-1779.

Zabinski MF, Norman GJ, Sallis JF, Calfas KJ. (2007) Patterns of sedentary behavior among adolescents. Health Psychology, 26(1): 113-120.

PATTERNS OF PHYSICAL ACTIVITY IN THE NORTHERN MANHATTAN STUDY USING MULTI-VARIATE FINITE MIXTURE MODELING

Gary Yu, MPH[1], Melanie M Wall, PhD[1], Joshua Z Willey, MD, MS[2], Ying Kuen Cheung, PhD[1], Ralph L Sacco, MD, MS[3]; Mitchell SV Elkind, MD, MS[2,4]

(1) Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA
(2) Department of Neurology, College of Physicians and Surgeons, Columbia University, New York, NY, USA
(3) Department of Neurology, University of Miami, Miami FL
(4) Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA

Address correspondence and reprint requests to Gary Yu, 722 West 168th St, 6th Floor, New York, NY 10032

Gy2153@columbia.edu

GLOSSARY

Cl = confidence interval; OR = odds ratio; MET = metabolic equivalents

Background

There is limited information on the complex patterns of physical activity (PA) among elderly individuals. We sought to examine and describe the complex patterns of leisure time PA levels among a United States based elderly cohort.

Methods:

The Northern Manhattan Study (NOMAS) is a prospective cohort study of older, urban-dwelling, multiethnic, stroke-free individuals. Baseline measures of leisure-time PA were collected via in-person questionnaires. A multivariate finite mixture modeling approach (MFMM) using the total frequency, average duration per session, total energy expenditure along with the total count of the number of physical activities conducted was used to classify participants into clusters based on reported leisure-time PA. The identified clusters were associated with baseline socio-demographics and vascular disease risk factors, and a comparison was made with a summary measure of total PA.

Results:

NOMAS recruited 3,298 participants with PA questionnaires available – mean age 69, 63% women, 54% Hispanic, 25% black and 21% white. A five cluster solution among those who conducted any PA (n = 1,923) was found: minimal (n = 74, 4%), low (n = 195, 10%), near guidelines (n = 455, 24%), meet guidelines (n = 1015, 53%), highly active (n = 184, 10%). Participants in the clusters that met guidelines and were highly active had meaningful reductions in smoking status, diabetes, obesity, high waist circumference, and hypertension.

Conclusions

The MFMM approach provides a more comprehensive picture of the association between PA and cardiovascular disease risk factors and may allow for better understanding of the impact of exercise on health outcomes.

Introduction

Leisure-time physical activity (PA) is an important component of primary prevention for cardiovascular disease across all age groups [1].The American Heart Association guidelines for primary prevention of cardiovascular disease recommend 150 minutes of moderate intensity or 75 minutes of heavy intensity activity per week [2]. Current recommendations leave several unanswered questions as to how PA should be carried out to achieve optimal health outcomes, such as frequency, duration, and number of different types of activities. Statistical clustering techniques allow for this multidimensional information to be summarized into useful homogeneous subgroups based on PA patterns to then predict the risk of adverse health outcomes. This information could subsequently allow providers to give more specific counseling to patients beyond total weekly activity. Multivariate finite mixture model (MFMM) analysis is a model-based clustering method which is data driven and can aid in producing meaningful patterns from an optimal number of groups in the data [3]. The primary aim of our study was to identify clusters of participants with similar patterns of exercise using four summary variables [total frequency, the average duration, total intensity and the total number of different activities done] derived from a leisure-time PA questionnaire, and describe cross-sectional associations with cardiovascular disease risk factors. We then sought to examine if these clusters would provide meaningful descriptions of leisure-time PA in our cohort compared to a measure of total activity. We hypothesized that these clusters would provide a more detailed description of the association between PA and cardiovascular disease risk factors compared to summary score of total activity.

Methods

Recruitment of the Cohort

The Northern Manhattan Study is a population-based prospective cohort study designed to evaluate the effects of medical, socio-economic, and other risk factors on the incidence of vascular disease in a stroke-free multiethnic community-based cohort. Methods of participant recruitment, evaluation and follow-up have been previously reported [4]. In-person evaluations were performed at Columbia University Medical Center or at home for those who could not come in person (6% were performed at home). The study was approved by the institutional review boards at Columbia University Medical Center and the University of Miami. All participants gave informed consent to participate in the study.

Assessment of Leisure-time Physical Activity

Physical activity was measured by an in-person questionnaire adapted from the National Health Interview Survey (NHIS) of the National Center for Health Statistics [5]. The questionnaire records the duration and frequency of various leisure-time/recreational activities for the 2 weeks before the interview is conducted. The participants were then asked whether they engaged in any PA in the preceding 2 weeks, and those who answered no were coded as physically inactive. For each activity, the participants were asked the duration of activity and the times they engaged in this activity; if the duration of activity was less than 10 minutes, it was coded as “no activity.” The questionnaire has been previously reported as reliable for individuals reporting moderate physical activity and validated in this population, demonstrating a crude concordance rate of 0.69 when proxies of the participants were asked [5]. The same measure also correlated with body mass index, activities of daily living scores, and quality of well-being activity scores [4].

Measures

Fifteen Leisure-Time Physical Activity Items

Participants were included if they had conducted any of the following fifteen physical activities within the past two weeks: walking, jogging or running, hiking, gardening or yard work, aerobics or aerobic dancing, other dancing, calisthenics or general exercise, golf, tennis, bowling, bicycle riding, swimming or water exercises, handball, or any other activity. Two additional activities could be specified if subjects did not find the corresponding activity on the list. We kept the two other activities separate to account for the diversity of the different activities performed. For each activity specifically, each of the following self-reported variables were asked: the participation in the activity, the frequency that each activity was conducted within a two week period and the duration of conduct of the activity at each session. Questionnaires were correlated with compendia of physical activity to allow calculation of metabolic equivalents (MET) [kcal/kg-hour] for the intensity of activity as well as energy expenditure in kilocalories[6]. The MET-score can be calculated by summing the product MET and duration in hours for each activity performed. In previous analyses total physical activity was classified based on quartiles of the MET-score [kcal/kg] as follows: inactive (no reported activity, reported in close to half of the cohort), active (between 1 and 14 MET), or highly active (> 14 MET)[7]. Energy expenditure per week was estimated based on the sum of the MET of each activity times body weight times the number of hours per week it was performed.

Total PA was also summarized using the fifteen reported items with total frequency, average duration, total count and total energy expenditure. In this approach the frequency per two week period measure collapses over all different types of activities; for example walking 4 times in 2 weeks would be the same as gardening, dancing, hiking, and golfing one time each. The total number of different activities performed, however, would be 1 in the first scenario and 4 in the latter. The total frequency variable was the sum of the frequencies per each of the fifteen activities, the average duration variable was the average of the duration per session of each of the fifteen activities [total duration/total frequency], and the total count variable created was the total sum of the number of fifteen physical activity items conducted.

For example, a 150 pound person who only walked 8 times in the 2 week period for 30 minutes each time would contribute an energy expenditure of 1088 kCal/2 weeks [8 sessions/two weeks * 30 minutes/session * 1 hour/60 minutes* 4 kcal/kg-hour * 150 lbs * 1 kg/2.2 lbs]. And a 132 pound person who only aerobically danced twice for 60 minutes each time would contribute an energy expenditure of 661 kCal/2 weeks [2 sessions/two weeks * 60 minutes/session * 1 hour/60 minutes * 5.5 kcal/kg-hour *132 lbs * 1 kg/2.2 lbs].

Statistical Analysis

We sought to identify clusters of individuals with similar patterns of leisure-time PA using the four summary measures (total frequency, average duration, total count, total energy expenditure), and examine associations with baseline cardiovascular disease risk factors. These clusters would be further compared to the MET-score in the associations with baseline demographics and cardiovascular disease risk factors.

Multivariate Finite Mixture Model (MFMM) Analysis

The MFMM analysis [7] was limited to the participants who reported any physical activity (n = 1923) and was conducted using the natural logarithm transformation of the total frequency, average duration, total energy expenditure variables to satisfy assumptions of normality and modeling the total number of different activities as a Poisson distribution. Individuals who reported no physical activity were considered as a separate cluster. The model is assessed and compared based on model convergence and the optimal number of clusters produced. Estimates were obtained using maximum likelihood in Mplus, version 6.11. Choice of number of clusters relied primarily on the Bayesian information criterion (BIC) which balances model fit and parsimony [8]. Analyses were stopped after reaching a maximum of 10 clusters which would limit the qualitative usefulness of attaching descriptive labels to each cluster.

After choosing an optimal number of clusters, qualitative descriptions of the resulting patterns of physical activity clusters were assigned based on in depth examination of increases or decreases in the mean levels of the four summary measures within each cluster as compared to the overall sample average.

Comparisons of the MFMM Clusters and the 3 Intensity Groups

Demographics and risk factor characteristics (age, race, education, marital status, friendship status, hypertension, weight, waist circumference, obesity status, smoking status, alcohol consumption, and cardiac disease) were compared across the sample and within clusters using the clusters to identify similar exercise patterns within each covariate, with no activity as the reference. Chi-squared and ANOVA tests with Tukey correction were also conducted to examine differences in the distribution of these covariates between predicted clusters and the MET score categories for comparison purposes. Given the predicted clusters for each individual, multinomial logistic regression of the categorical demographic and risk factor characteristics on the clusters was performed to examine associations with risk factors across patterns. For comparison purposes multinomial logistic regression was performed across the MET-score categories to demonstrate the added utility of the clusters. Measures of agreement (kappa) were calculated to estimate the level of correspondence between the patterns and the three MET-score categories to ensure validation of the MFMM clusters.

Results

Baseline demographics of the cohort are summarized in table 1. Two-thirds of our cohort are women with a mean age of 69 years (SD = 10 years), a waist circumference of 37 cm (SD = 5 cm) and BMI of 28 (SD = 6). 42% is overweight and 27% is obese. Half of the sample is Hispanic, while the rest self-identified as non-Hispanic white and non-Hispanic black equally. Slightly less than half of the sample has received a high school education and has reported no smoking.

Summary Statistics of the Fifteen Physical Activities

Table 2 summarizes the total activity patterns in the cohort. The majority of participants who reported being active walked a mean of 5 times every week, for a mean of 46 minutes per session and expend a mean of 1110 kcal/week. Walking was the principal activity reported in the cohort (83% of the cohort).

Multivariate Finite Mixture Model Analysis

The MFMM found an optimal five cluster solution based on the minimal value of the BIC. Table 3 reports summary measures of PA based on each cluster. Only 71% of cluster 4 (meet Guidelines) and 97% of cluster 5 (highly active) were within American Heart Association guideline goals with weekly averages of 308 and 533 minutes of moderate exercise respectively. Highly active participants reported 2 sessions per day with lower average minutes per activity 37.8 but higher total energy expenditure of 2865 kcal. Participants in cluster 4 (Meet Guidelines) reported 1 session of activity per day with an average of 44 minutes per activity and 1555 kcal of energy expenditure. Clusters 1 through 3 (minimal, low, near guidelines) comprised 40% of the cohort who reported any physical activity. Participants in clusters 1 - 3 were below recommended guidelines and ranged from 1-6 sessions in a two-week period, 37-52 minutes/session of activity and 140-680 kcal/week of energy expenditure.

There was a statistically significant moderate association between the five clusters [Cluster 0 (inactive), Cluster 1-3 (minimal, low, near guidelines), Cluster 4-5 (Meet

Guidelines and highly active)] and the three MET-score categories (inactive, active and highly active) with a kappa of 0.72.

Association Between the Patterns of Leisure-time Physical Activity and Baseline Demographics

Table 4 outlines the association between clusters 0-5 with baseline demographics and cardiovascular disease risk factors. There was no significant difference in the age distributions of individuals across all clusters. There was a higher proportion of men in cluster 5, (OR = 1.72, 95% CI: 1.26, 2.35) and in cluster 4 (OR = 1.46, 95% CI: 1.23, 1.74) compared to the inactive group. Hispanics were less likely to be grouped into clusters 3 to 5 compared to whites. Individuals in clusters 3 to 5 were also more likely to complete high school.

Association Between the Patterns of Physical Activity, Lifestyle Factors and Cardiovascular Disease Risk Factors

Compared to the inactive group, participants in cluster 2 were less likely to be former smokers versus non-smokers (OR = 0.67, 95% CI: 0.47, 0.94) while those in cluster 3 were less likely to be current smokers than non-smokers (OR = 0.72, 95% CI: 0.52, 0.98). Interestingly, highly active (cluster 5) participants were more likely to be former smokers compared to the inactive (OR = 1.80, 95% CI: 1.29, 2.51). We also found significant differences by BMI status and high waist circumference. Those in cluster 4 and cluster 5 were less likely to be overweight (OR for cluster 4 = 0.78, 95% CI: 0.64, 0.94; OR for cluster 5 = 0.62, 95% CI: 0.44, 0.88) and obese (OR for cluster 4 = 0.61, 95% CI: 0.49, 0.76; OR for cluster 5 = 0.47, 95% CI: 0.31, 0.72) compared to those were inactive (cluster 0), while cluster 3 individuals were more likely to be overweight (OR = 1.42, 95% CI: 1.09, 1.86). Participants in clusters 4 and 5 had lower waist circumferences (OR for cluster 4 = 0.72, 95% CI: 0.61, 0.85; OR for cluster 5 = 0.54, 95% CI: 0.39, 0.75). Participants who met guidelines (cluster 4) and were highly active (cluster 5) had a lower odds of hypertension (OR for cluster 5 = 0.58, 95% CI: 0.42, 0.81) and diabetes (OR for cluster 4 = 0.74, 95% CI: 0.60, 0.90, OR for cluster 5 = 0.59, 95% CI: 0.39, 0.89) as compared to those who were inactive. There were no statistically significant differences in baseline cardiovascular disease risk factors between clusters 4 and 5. Among participants in clusters 1-3 there was no association with hypertension or diabetes when compared to those who reported no activity.

Comparison between cluster analysis categories and total activity summary scores

We compared the association of leisure-time PA categories derived from the MFMM analysis and a raw summary of total weekly activity (MET-score) with baseline cardiovascular disease risk factors. The five cluster MFMM solution differentiated the study participants better than the MET-score in terms of the baseline risk factors. Participants in the active (1-14) and highly active (>14) MET-score categories both had a lower prevalence of baseline cardiovascular disease risk factors when compared to those were inactive and to each other. Participants in clusters 1 to 3 fit into the active group for the MET-score, and yet there was no statistical difference between compared to the inactive, except for African-American ethnicity. Participants in clusters 4 and 5 fit into the highly active group for the MET-score, in addition, the highly active cluster showed a lower prevalence in hypertension which is statistically different than the guideline meeting cluster. The clustering results are more sensitive to a subject's obesity status when comparing active cluster 4 participants to active cluster 3 participants. Active cluster 4 individuals are less likely to be overweight and obese as compared to active cluster 3 individuals.

Discussion

In our MFMM approach we found a dose-response relationship between each cluster and the total frequency and energy expenditure, while the average duration remained constant. Using the MET-score these participants would have been classified as below recommended PA, but by partitioning this category into three clusters we found more subtle associations with Hispanic ethnicity, current versus previous smoking, and educational attainment. In addition a summary score of leisure-time PA, such as the MET-score, indicated that those in cluster 1-3 had a reduction in cardiovascular disease risk factors which was not seen in the MFMM approach. When we compared participants who met guidelines (cluster 4) to those who exceeded guidelines (cluster 5) we found significant reductions in the prevalence of diabetes, hypertension, obesity, and elevated waist circumference. The MFMM approach highlights that it does not appear to be sufficient to perform any type of activity to gain a protective effect, but rather that a certain threshold in minimum leisure-time PA performed is required. Our results support that after meeting recommended targets older individuals continue to have additive effects of more exercise. Previous investigators have found these findings to hold for reduced mortality and extended life expectancy. [15]

In our study we found groups of older individuals (mean age 70.4) who reported being highly active. Though traditionally PA is thought of as having no upper limit, recent literature suggests potential adverse health outcomes with higher levels of PA. One study found that extremely vigorous weight-bearing exercise as compared to its more moderate counterpart resulted in lower bone density with the possibility for osteoporosis in individuals after the age of 50 [12], while others have reported an increased risk of injury with certain activities such as cardiac fibrosis, associated with strenuous excessive exercise [13,14]. Our cluster analysis methods will allow us to describe whether those individuals who reported above recommended levels of activity could have additional harmful effects that offset baseline benefits.

Our study and MFMM approach has some important strengths. Previous results among a prospective cohort reported only beneficial improvement in clinical outcomes (such as reduction in risk of atherosclerosis and hypertension, coronary heart disease, fat deposits in the body and Type II diabetes) among only subjects that engage in regular physical activity in terms of their energy expenditure [10,11]. Our more descriptive approach to classifying PA may allow for detection of more subtle associations with cardiovascular disease outcomes that could translate to more specific recommendations for older individuals who may have difficulties meeting recommended PA guidelines due to other disabilities. The detection of differences in health outcomes among those who would otherwise be labeled as not meeting targets could in turn translate to more realistic exercise recommendations for older patients. The MFMM is data driven which means that it takes into account the high dimensionality of the data (frequency, duration, intensity, count) which is more comprehensive than using a cruder form of categorizing physical activity using only cutoffs of one measure. In addition, the described MFMM method could be generalizable to other datasets as a principled methodological approach when the optimal number of clusters is not specified a priori as needed with more traditional clustering techniques. Our approach could be generalized to other populations so as to account for local variability in life-space, neighborhood characteristics, socio-demographic factors, and could allow for the inclusion of baseline co-morbidities into the information used to define each cluster [17,18]. Our study has some limitations as well. Our analyses with socio-demographic factors and cardiovascular disease risk factors are cross-sectional and as such we cannot conclude on the directionality of association. It may be that for example participants who were free of co-morbidities were able to participate in leisure-time PA more due to the lack of physical impairments. Future analyses will allow us to examine associations with incident cardio-vascular disease risk factors and outcomes to gain the benefit of temporality. Nonetheless the cross-sectional associations may have potential public health benefits in identifying those at highest risk for being below recommended targets. There is incomplete information in our cohort regarding non-leisure time PA, such as occupational and commuting activity, and it may be that participants who are highly active as part of their employment would not perform leisure-time PA. On the other hand several studies have reported an independent protective effect on cardio-vascular disease from leisure-time PA, independent of other forms of PA [19-26]. In our study we did not collect information on time spent on sedentary behavior which may confer additional risk of cardiovascular diseases, though leisure-time PA confers a protective effect regardless of sedentary time [26]. In our study we did collect information on objective levels of PA or fitness, as can be seen from an accelerometer, though in previous studies self-reported PA still shows a protective effect on cardio-vascular diseases [19-26].

The clinical interpretability of the MFMM statistical methodology could be a potential challenge, particularly as it relates to the qualitative labels created for the MFMM clusters. This becomes more apparent when the number of clusters increases and the relevance of the labels diminishes. On the other hand the MFMM approach allows for the description of specific PA patterns within clusters, such as duration, frequency, and number of different activities. Cluster analysis could define differences among those who perform the same total activity depending on how the total is achieved in terms of reduced risk of cardio-vascular disease. For example it is not clear if there is the same benefit from performing 75 minutes of moderate intensity activity in one session per week, as opposed to performing 25 minutes three times per week[16].

Conclusion

The MFMM methodology outlined in our study has potential public health implications and adds to the body of literature on leisure-time PA in older individuals. Despite the commonality of physical inactivity in our cohort the MFMM cluster analysis was able to discern patterns that reflected different levels of PA as compared to American Heart Association recommended targets. The MFMM approach may have potential clinical relevance in allowing to understand the beneficial effects on cardiovascular health of even small amounts of exercise, as well as explore characteristics that are associated with the decision to perform PA but not at recommended targets. In counseling patients and in population based recommendations consideration of the total frequency, average duration, energy expenditure and total number of physical activity items may be more appropriate and clinically useful than summary measures of total PA.

Table 1. Demographic and Vascular Risk Factors of the Sample

	N = 3298	%
Demographic Factors		
Age Mean (SD)	69.25	(10.30)
Gender		
Females	2071	62.95 %
Males	1227	37.05 %
Race-Ethnicity		
Whites	690	21.43 %
Blacks	803	24.91 %
Hispanics	1727	53.65 %
Marital Status		
Married	1042	31.67 %
Single	2254	68.33 %
Education		
High School or More	1511	45.26 %
Under High School	1786	54.74 %
Lifestyle Factors		
Non Smoker	1545	46.83 %
Former Smoker	1191	36.15 %
Current Smoker	560	17.02 %
Moderate Alcohol	1086	32.94 %
Social Support		
Friends	2798	85.01 %
No Friends	500	14.99 %
Cardiovascular Disease Risk Factors		
Diabetes	716	21.94 %
Hypertension	2429	74.00 %
Cardiac Disease	705	21.43 %
Overweight	1366	41.87 %
Obese	919	27.43 %
BMI Mean (SD)	27.85	(5.55)
Waist Circumference Mean (SD)	36.77	(5.02)
High Waist Circumference	1427	42.46 %

Table 2. Descriptive Summary of the 15 Physical Activity Items

	N =	Average Duration (session /minutes)		Total Frequency (times/wk)		Total Energy Expenditure (10^3 kcal/ week)	
Physical Activity	1971	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
Walking	1636	46.24 (37.82)	10 – 420	4.88 (2.49)	0.5 - 21	1.11 (1.19)	0.02 – 15.10
Jogging or Running	51	36.47 (22.03)	10 – 120	3.17 (1.94)	0.5 - 7	0.99 (0.84)	0.10 – 3.87
Hiking	11	164.09 (129.59)	15 – 360	0.68 (0.25)	0.5 - 1	0.71 (0.57)	0.06 – 1.66
Gardening or Yard Work	32	73.75 (89.85)	10 – 480	2.36 (2.32)	0.5 - 7	1.01 (1.66)	0.04 – 5.96
Aerobics or Aerobic Dancing	100	35.76 (27.28)	10 – 180	3.20 (2.21)	0.5 - 7	0.65 (0.95)	0.04 – 8.87
Other Dancing	67	87.16 (91.10)	10 – 400	1.59 (1.80)	0.5 - 7	0.65 (1.02)	0.02 – 5.73
Calisthenics or General Exercise	476	22.64 (16.62)	10 – 150	4.35 (2.58)	0.5 - 14	0.54 (0.55)	0.03 – 5.94
Golf	14	210.00 (137.28)	10 – 480	1.25 (0.85)	0.5 - 3	1.40 (1.49)	0.05 – 5.37
Tennis	6	87.50 (61.62)	15 – 200	0.92 (0.66)	0.5 - 2	0.63 (0.44)	0.07 – 1.11
Bowling	7	74.29 (46.14)	10 – 120	2.07 (2.37)	0.5 - 7	0.30 (0.13)	0.10 – 0.52
Bicycle Riding	95	41.76 (49.86)	10 – 300	3.56 (2.58)	0.5 - 14	0.82 (0.92)	0.03 – 5.29
Swimming or Water Exercise	63	51.75 (35.37)	10 – 180	2.10 (1.31)	0.5 - 6.5	0.79 (0.66)	0.02 – 3.25
Handball/Racquetball/Squash	2	60.00 (0.0)	60 – 60	0.75 (0.35)	0.5 - 1	0.56 (0.38)	0.29 – 0.83
Other Activity 1	171	59.30 (119.04)	3 – 1080	2.64 (1.97)	0.5 - 7	0.85 (2.16)	0.03 – 22.25
Other Activity 2	11	66.09 (64.59)	10 – 180	2.09 (1.83)	0.5 - 14	1.28 (2.61)	0.04 – 8.89
Overall	1971	43.90 (43.60)[1]	5 – 840	5.93 (3.88)[2]	0.5 - 28	1.31 (1.52)[3]	0.02 – 22.25

N is the total number of individuals reporting the physical activity

[1] Average across individuals of the average minutes per session across all the different types of activities done.

[2] Average across individuals of the total sessions of physical activity conducted in a one week period.

[3] Average across people of the total energy expenditure exerted in a one week period.

Table 3. Multivariate Finite Mixture Model (MFMM) Five Cluster solution of the reported physical activity in the Northern Manhattan Study

	Avg	Cluster 0[a]: No Exercise	Cluster 1: Very Inactive	Cluster 2: Somewhat inactive	Cluster 3: Slighly Under-Guideline	Cluster 4: Meet Guilde-line	Cluster 5: Over-exercise
Total Frequency (Sessions/wk)	6.35	0	0.51	1.30	2.91	6.96	14.19
Average Duration (Minutes/session)	42.93	0	52.43	36.62	43.19	44.28	38.57
Weekly Duration (Tot Freq*Avg Dur)	272.42	0	26.83	47.67	125.76	308.39	547.49
Total Energy Expenditure (10^3 kCal/wk)	1.40	0	0.14	0.25	0.68	1.56	3.90
Num of types of physical activity done	1.39	0	1.00	1.04	1.21	1.36	2.43
Proportion	100%	-----	3.75%	9.94%	23.95%	52.51%	9.84%
N	1971	1327	74	196	472	1035	194
MET Categories (n %)							
Inactive (MET = 0)		1322	7	6	11	0	0
		(99.6%)	(9.5%)	(3.1%)	(2.3%)	(0 %)	(0 %)
Active (1 <= MET <= 14)		4	66	188	401	509	11
		(0.3%)	(89.2%)	(95.9%)	(85.0%)	(49.2%)	(5.7%)
Highly Active (MET > 14)		1	1	2	60	526	183
		(0.1%)	(1.4%)	(1.0%)	(12.7%)	(50.8%)	(94.3%)
MET Score (Mean (SD))		0	2.08	3.17	9.29	21.58	37.09
		(0)	(2.70)	(2.60)	(13.44)	(18.54)	(21.73)
Meet Guidelines (>= 150 min/wk moderate exercise)		0%	3%	3%	26%	71%	97%

Dark grey highlights indicate that the amount is above the average

Grey highlights indicate that the amount is around the average

* Average values are obtained from the transformation of the log-Normal distribution parameters

[a] Individuals that do not report any physical activity (Cluster 0) were excluded from the multivariate finite mixture model based cluster analysis

Table 4. Unadjusted Multinomial Logistic Regression of Demographic and Clinical Risk Factors on Each of the 5 MFMM Clusters

OR (95% CI)	Cluster 1 vs Cluster 0	Cluster 2 vs Cluster 0	Cluster 3 vs Cluster 0	Cluster 4 vs Cluster 0	Cluster 5 vs Cluster 0
Baseline Demographics					
Males vs Females	1.56 (0.97, 2.50)	0.99 (0.72, 1.36)	1.17 (0.94, 1.46)	**1.48 (1.25, 1.75)**	**1.73 (1.28, 2.35)**
Blacks vs Whites	2.05 (0.89, 4.73)	1.03 (0.64, 1.65)	0.83 (0.60, 1.15)	0.91 (0.71, 1.16)	**0.56 (0.37, 0.84)**
Hispanics vs Whites	1.47 (0.68, 3.16)	0.85 (0.56, 1.28)	**0.67 (0.51, 0.89)**	**0.43 (0.35, 0.53)**	**0.24 (0.16, 0.35)**
Married	1.34 (0.83, 2.17)	1.12 (0.81, 1.54)	0.95 (0.75, 1.19)	1.03 (0.87, 1.23)	1.01 (0.73, 1.40)
Friends	1.13 (0.60, 2.13)	1.31 (0.86, 2.00)	1.26 (0.94, 1.68)	**1.42 (1.13, 1.78)**	**2.81 (1.60, 4.93)**
High School	0.99 (0.61, 1.60)	0.92 (0.68, 1.26)	**1.43 (1.16, 1.77)**	**1.87 (1.59, 2.21)**	**3.46 (2.51, 4.77)**
Lifestyle Factors					
Former Smoker vs Never	1.05 (0.64, 1.74)	**0.68 (0.48, 0.96)**	0.92 (0.73, 1.15)	1.02 (0.85, 1.22)	**1.75 (1.26, 2.42)**
Current Smoker vs Never	0.64 (0.30, 1.35)	0.76 (0.50, 1.16)	**0.72 (0.53, 0.98)**	1.03 (0.82, 1.29)	0.73 (0.44, 1.21)
Moderate Alcohol vs Light/Never[c]	**1.92 (1.19, 3.10)**	**1.45 (1.05, 1.99)**	**1.35 (1.08, 1.70)**	**1.45 (1.22, 1.73)**	**2.84 (2.09, 3.85)**
Cardiovascular Disease Risk Factors					
Diabetes[a]	0.95 (0.54, 1.65)	0.92 (0.64, 1.32)	1.00 (0.78, 1.28)	**0.74 (0.61, 0.91)**	**0.58 (0.39, 0.88)**
HTN[b]	0.93 (0.54, 1.59)	1.14 (0.80, 1.64)	0.87 (0.69, 1.11)	0.85 (0.70, 1.02)	**0.55 (0.40, 0.75)**
Cardiac Disease	1.04 (0.60, 1.82)	0.87 (0.60, 1.26)	0.93 (0.72, 1.20)	0.92 (0.75, 1.12)	0.91 (0.63, 1.32)
Overweight[d]	1.31 (0.74, 2.31)	1.05 (0.72, 1.54)	**1.37 (1.05, 1.79)**	**0.77 (0.63, 0.93)**	**0.60 (0.42, 0.84)**
Obese[e]	0.90 (0.47, 1.73)	1.19 (0.80, 1.75)	1.14 (0.86, 1.52)	**0.58 (0.47, 0.72)**	**0.42 (0.28, 0.64)**
High Waist Circumference[f]	0.72 (0.45, 1.17)	1.01 (0.75, 1.37)	0.93 (0.75, 1.15)	**0.71 (0.61, 0.84)**	**0.53 (0.38, 0.72)**

Bolded numbers indicate that the value is statistically significant at the 5% alpha level.

[a] Diabetes mellitus - fasting blood glucose of 126 mg/dL, patient self-report of diabetes mellitus, or insulin and/or hypoglycemic agent use.

[b] Hypertension – Systolic blood pressure of 140 mm Hg or diastolic blood pressure of 90 mm Hg based on the average of 2 blood pressure measurements, physician diagnosis of hypertension, or patient self-report of a history of hypertension or antihypertensive use.

[c] Moderate alcohol use – 1 to 2 servings of alcohol per day.

[d] Overweight – BMI between 25 and 30 with normal weight (BMI <= 25) as the reference group.

[e] Obese – BMI greater than or equal to 30 with normal weight (BMI <= 25) as the reference group.

[f] High Waist Circumference – For males, greater than 40 cm; for females, greater than 35 cm.

Appendix Tables

Electronic table 1. Distribution of duration, frequency and energy expenditure by the diversity of the different types of activities

Number of different types of Physical Activity done	N (%)	Average Duration/Session (minutes) Mean (SD)	Range	Total Frequency (times/wk) Mean (SD)	Range	Total Energy Expend (10^3 kcal/ week) Mean (SD	Range
1	1383	45.05 (47.70)	5.00 – 840.00	4.60 (2.61)	0.5 – 21	1.04 (1.34)	0.02 – 22.21
2	460	40.07 (30.69)	9.38 – 333.75	8.44 (3.86)	1.0 – 21	1.79 (1.88)	0.08 – 22.25
3	91	44.44 (36.35)	10.00 – 234.64	10.55 (5.71)	1.5 – 28	2.37 (1.85)	0.39 – 9.54
4	26	53.59 (51.31)	16.25 – 281.25	12.63 (5.61)	2.0 – 26	3.36 (1.97)	0.94 – 9.56
5	6	37.63 (28.08)	17.00 – 86.00	12.17 (7.59)	5.0 – 25	2.31 (1.07)	1.03 – 3.41
6	4	42.14 (5.93)	35.77 – 48.15	15.88 (7.36)	8.5 – 26	4.11 (2.16)	2.03 – 7.11
7	1	41.36 (---)	41.36 – 41.36	22.00 (---)	22 – 22	6.57 (---)	6.57 – 6.57

Additional Table 2. Unadjusted Multinomial Logistic Regression of Demographic and Prevalence of Each Clinical Risk Factors on Each of the 5 MFMM Clusters

Cluster	0	1		2		3		4		5	
	n = 1327	n = 74		n = 196		n = 472		n = 1035		n = 194	
	Prev	Prev	OR (95% CI)	Prev	OR (95% CI)	Prev	OR (95% CI)	Prev	OR (95% CI)	Prev	OR (95% CI)
Baseline Demographics											
Age Mean	69.4	68.2		68.7		68.5		69.4		70.2	
Age SD	10.3	10.4		9.5		10.0		10.5		10.4	
Gender*											
Females	67.1%	56.8%	—	67.3%	—	63.6%	—	58.1%	—	54.1%	—
Males	32.9%	43.2%	1.56 (0.97, 2.50)	32.7%	0.99 (0.72, 1.36)	36.4%	1.17 (0.94, 1.46)	41.9%	**1.48 (1.25, 1.75)**	45.9%	**1.73 (1.28, 2.35)**
Race-Ethnicity*											
White	16.0%	10.8%	—	17.9%	—	20.8%	—	25.8%	—	36.1%	—
Black	20.4%	28.4%	2.05 (0.89, 4.73)	23.5%	1.03 (0.64, 1.65)	22.0%	0.83 (0.60, 1.15)	$30.1\%^{A}$	0.91 (0.71, 1.16)	$25.8\%^{B}$	**0.56 (0.37, 0.84)**
Hispanic	61.3%	60.8%	1.47 (0.68, 3.16)	58.2%	0.85 (0.56, 1.28)	53.6%	**0.67 (0.51, 0.89)**	$42.3\%^{A}$	**0.43 (0.35, 0.53)**	$33.0\%^{B}$	**0.24 (0.16, 0.35)**
Marital Status											
Married	31.2%	37.8%	1.34 (0.83, 2.17)	33.7%	1.12 (0.81, 1.54)	30.1%	0.95 (0.75, 1.19)	32.0%	1.03 (0.87, 1.23)	31.4%	1.01 (0.73, 1.40)
Single	68.7%	62.2%	—	66.3%	—	69.9%	—	68.0%	—	68.6%	—
Education*											
No High School	61.9%	62.2%	—	63.8%	—	53.2%	—	46.5%	—	32.0%	—
High School	38.1%	37.8%	0.99 (0.61, 1.60)	36.2%	0.92 (0.68, 1.26)	46.8%	**1.43 (1.16, 1.77)**	$53.5\%^{A}$	**1.87 (1.59, 2.21)**	$68.0\%^{B}$	**3.46 (2.51, 4.77)**
Lifestyle Factors											
Never* Smoker	46.4%	48.7%	—	55.1%	—	50.4%	—	45.8%	—	38.1%	—
Former Smoker	35.5%	39.2%	1.05 (0.64, 1.74)	28.6%	**0.68 (0.48, 0.96)**	35.4%	0.92 (0.73, 1.15)	$35.7\%^{A}$	1.02 (0.85, 1.22)	$51.0\%^{B}$	**1.75 (1.26, 2.42)**
Current Smoker	18.1%	12.2%	0.64 (0.30, 1.35)	16.3%	0.76 (0.50, 1.16)	14.2%	**0.72 (0.53, 0.98)**	18.5%	1.03 (0.82, 1.29)	10.8%	0.73 (0.44, 1.21)
Light/ Never Alcohol*	72.7%	58.1%	—	64.8%	—	66.3%	—	64.7%	—	48.4%	—
Moderate Alcohol ^{c}	27.3%	41.9%	**1.92 (1.19, 3.10)**	35.2%	**1.45 (1.05, 1.99)**	33.7%	**1.35 (1.08, 1.70)**	$35.3\%^{A}$	**1.45 (1.22, 1.73)**	$51.6\%^{B}$	**2.84 (2.09, 3.85)**

Social Support*										
No Friends	17.9%	16.2%	---	14.3%	---	14.8%	---	13.3%	---	7.2%	---
Friends	82.1%	83.8%	1.13 (0.60, 2.13)	85.7%	1.31 (0.86, 2.00)	85.2%	1.26 (0.94, 1.68)	86.7%[A]	**1.42 (1.13, 1.78)**	92.8%[B]	**2.81 (1.60, 4.93)**
Cardiovascular Disease Risk Factors											
No* Diabetes	75.7%	77.0%	---	77.6%	---	76.1%	---	81.0%	---	84.5%	---
Diabetes[a]	23.8%	23.0%	0.95 (0.54, 1.65)	22.4%	0.92 (0.64, 1.32)	23.9%	1.00 (0.78, 1.28)	19.0%	**0.74 (0.61, 0.91)**	15.5%	**0.58 (0.39, 0.88)**
No HTN*	24.3%	25.7%	---	21.9%	---	26.9%	---	27.5%	---	37.1%	---
HTN[b]	75.7%	74.3%	0.93 (0.54, 1.59)	78.1%	1.14 (0.80, 1.64)	73.1%	0.87 (0.69, 1.11)	72.5%[A]	0.85 (0.70, 1.02)	62.9%[B]	**0.55 (0.40, 0.75)**
No Cardiac Disease	77.8%	77.0%	---	80.1%	---	79.0%	---	79.2%	---	79.4%	---
Cardiac Disease	22.2%	23.0%	1.04 (0.60, 1.82)	19.9%	0.87 (0.60, 1.26)	21.0%	0.93 (0.72, 1.20)	20.8%	0.92 (0.75, 1.12)	20.6%	0.91 (0.63, 1.32)
Normal* Weight	28.1%	25.7%	---	26.0%	---	23.5%	---	36.3%	---	42.8%	---
Over weight[d]	40.8%	48.7%	1.31 (0.74, 2.31)	39.8%	1.05 (0.72, 1.54)	46.8%	**1.37 (1.05, 1.79)**	40.4%	**0.77 (0.63, 0.93)**	37.1%	**0.60 (0.42, 0.84)**
Obese[e]	31.2%	25.7%	0.90 (0.47, 1.73)	34.2%	1.19 (0.80, 1.75)	29.7%	1.14 (0.86, 1.52)	23.3%	**0.58 (0.47, 0.72)**	20.1%	**0.42 (0.28, 0.64)**
Low*	52.8%	60.8%	---	52.6%	---	54.7%	---	61.1%	---	68.0%	---
High Waist Circum ference[f]	47.2%	39.2%	0.72 (0.45, 1.17)	47.4%	1.01 (0.75, 1.37)	45.3%	0.93 (0.75, 1.15)	38.9%	**0.71 (0.61, 0.84)**	32.0%	**0.53 (0.38, 0.72)**

*Global chi-square test significant at the $p < 0.05$ level, df=5

Bolded numbers indicate that the value is statistically significant at the 5% alpha level.

[AB] Post-hoc test that indicates that there is a statistical difference between Cluster 4 (Meet Guidelines) and Cluster 5 (Exceed Guidelines) at the 5% level.

[a] Diabetes mellitus - fasting blood glucose of 126 mg/dL, patient self-report of diabetes mellitus, or insulin and/or hypoglycemic agent use.

[b] Hypertension – Systolic blood pressure of 140 mm Hg or diastolic blood pressure of 90 mm Hg based on the average of 2 blood pressure measurements, physician diagnosis of hypertension, or patient self-report of a history of hypertension or antihypertensive use.

[c] Moderate alcohol use – 1 to 2 servings of alcohol per day.

[d] Overweight – BMI between 25 and 30 with normal weight (BMI <= 25) as the reference group.

[e] Obese – BMI greater than or equal to 30 with normal weight (BMI <= 25) as the reference group.

[f] High Waist Circumference – For males, greater than 40 cm; for females, greater than 35 cm.

Additional Table 3. Unadjusted Multinomial Logistic Regression of Demographic and Prevalence of Each Clinical Risk Factors on Each of the MET Score Categories

MET Score	Inactive Prev	Active Prev	OR (95% CI)	Highly Active Prev	OR (95% CI)
Baseline Demographics					
Age					
Gender*					
Females	66.7%	64.0%	---	54.1%	---
Males	33.3%	36.0%	1.13 (0.96, 1.33)	45.9%	**1.70 (1.42, 2.04)**
Race-Ethnicity*					
Whites	15.8%	21.7%	---	28.6%	---
Blacks	20.0%	27.1%	0.99 (0.77, 1.26)	27.8%	**0.77 (0.59, 1.00)**
Hispanics	62.0%	49.3%	**0.58 (0.47, 0.72)**	40.4%	**0.36 (0.29, 0.45)**
Marital Status					
Married	31.4%	31.7%	1.01 (0.85, 1.20)	31.7%	1.01 (0.84, 1.22)
Single	68.4%	68.3%	---	68.3%	---
Education*					
No High School	62.4%	52.2%	---	42.7%	---
High School	37.5%	47.8%	**1.52 (1.30, 1.78)**	57.3%	**2.23 (1.86, 2.67)**
Lifestyle Factors*					
Never Smoker	46.4%	51.8%	---	40.1%	---
Former Smoker	35.5%	33.5%	0.84 (0.71, 1.00)	41.1%	**1.34 (1.10, 1.63)**
Current Smoker	18.1%	14.7%	**0.73 (0.58, 0.91)**	18.6%	1.19 (0.93, 1.53)
Light/Never Alcohol*	72.5%	67.2%	---	57.4%	---
Moderate Alcohol vs Light/Never[c]	27.5%	32.8%	**1.29 (1.09, 1.53)**	42.6%	**1.95 (1.62, 2.35)**
Social Support*					
No Friends	18.2%	12.6%	---	13.7%	---
Friends	81.8%	87.4%	**1.54 (1.23, 1.92)**	86.3%	**1.40 (1.09, 1.79)**
Cardiovascular Disease Risk Factors					
No Diabetes*	75.5%	77.8%	---	83.1%	---
Diabetes[a]	24.1%	22.2%	0.90 (0.74, 1.08)	16.8%	**0.63 (0.51, 0.80)**
No HTN	24.4%	26.5%	---	29.5%	---
HTN[b]	75.6%	73.5%	0.90 (0.75, 1.08)	70.5%	**0.77 (0.63, 0.94)**
No Cardiac Disease	77.9%	79.1%	---	79.2%	---
Cardiac Disease	22.1%	20.9%	0.93 (0.77, 1.12)	20.8%	0.93 (0.75, 1.15)
Normal Weight*	28.3%	28.2%	---	38.8%	---
Overweight[d]	40.8%	44.1%	1.09 (0.90, 1.31)	38.4%	**0.69 (0.56, 0.84)**
Obese[e]	30.9%	27.7%	0.90 (0.73, 1.11)	22.8%	**0.54 (0.43, 0.68)**
Low*	53.3%	55.5%	---	64.6%	---
High Waist Circumference[f]	46.7%	44.5%	0.92 (0.78, 1.07)	35.4%	**0.63 (0.52, 0.75)**

*Global chi-square test significant at the $p < 0.05$ level, df=5

Bolded numbers indicate that the value is statistically significant at the 5% alpha level.

Additional Table 4. Unadjusted Logistic Regression of Demographic and Prevalence of Each Clinical Risk Factors on the 2 MFMM Guideline Meeting Clusters

	Cluster 4		Cluster 5	
	Prev	OR (95% CI)	Prev	OR (95% CI)
Baseline Demographics				
Age				
Gender				
Females	58.1%	---	54.1%	---
Males	41.9%	---	45.9%	1.17 (0.86, 1.60)
Race-Ethnicity				
Whites	25.8%	---	36.1%	---
Blacks	30.1%	---	25.8%	**0.61 (0.41, 0.91)**
Hispanics	42.3%	---	33.0%	**0.56 (0.38, 0.81)**
Marital Status				
Married	32.0%	---	31.4%	0.98 (0.70, 1.36)
Single	68.0%	---	68.6%	---
Education				
No High School	46.5%	---	32.0%	---
High School	53.5%	---	68.0%	**1.85 (1.33, 2.56)**
Lifestyle Factors				
Never Smoker	45.8%	---	38.1%	---
Former Smoker vs Never	35.7%	---	51.0%	**1.72 (1.23, 2.39)**
Current Smoker vs Never	18.5%	---	10.8%	0.70 (0.42, 1.18)
Light/Never Alcohol	64.7%	---	48.4%	---
Moderate Alcohol vs Light/Never[c]	35.3%	---	51.6%	**1.95 (1.43, 2.66)**
Social Support				
No Friends	13.3%	---	7.2%	---
Friends	86.7%	---	92.8%	**1.98 (1.12, 3.51)**
Cardiovascular Disease Risk Factors				
No Diabetes	81.0%	---	84.5%	---
Diabetes[a]	19.0%	---	15.5%	0.78 (0.51, 1.19)
No HTN	27.5%	---	37.1%	---
HTN[b]	72.5%	---	62.9%	**0.64 (0.47, 0.89)**
No Cardiac Disease	79.2%	---	79.4%	---
Cardiac Disease	20.8%	---	20.6%	0.99 (0.68, 1.45)
Normal Weight	36.3%	---	42.8%	---
Overweight[d]	40.4%	---	37.1%	0.78 (0.55, 1.10)
Obese[e]	23.3%	---	20.1%	0.73 (0.48, 1.11)
Low	61.1%	---	68.0%	---
High Waist Circumference[f]	38.9%	---	32.0%	0.74 (0.53, 1.02)

*Global chi-square test significant at the p < 0.05 level, df=5

Bolded numbers indicate that the value is statistically significant at the 5% alpha level.

Additional Table 5. Unadjusted Multinomial Logistic Regression of Demographic and Prevalence of Each Clinical Risk Factors on the 3 MFMM Clusters That Do Not Meet Guidelines

	Cluster 1	Cluster 2		Cluster 3	
	Prev	Prev	OR (95% CI)	Prev	OR (95% CI)
Baseline Demographics					
Age					
Gender					
Females	56.8%	67.3%	---	63.6%	---
Males	43.2%	32.7%	0.64 (0.37, 1.10)	36.4%	0.75 (0.46, 1.24)
Race-Ethnicity					
Whites	10.8%	17.9%	---	20.8%	---
Blacks	28.4%	23.5%	0.50 (0.20, 1.26)	22.0%	**0.40 (0.17, 0.96)**
Hispanics	60.8%	58.2%	0.58 (0.25, 1.34)	53.6%	0.46 (0.21, 1.01)
Marital Status					
Married	37.8%	33.7%	0.83 (0.48, 1.45)	30.1%	0.71 (0.42, 1.18)
Single	62.2%	66.3%	---	69.9%	---
Education					
No High School	62.2%	63.8%	---	53.2%	---
High School	37.8%	36.2%	0.93 (0.54, 1.62)	46.8%	1.45 (0.87, 2.39)
Lifestyle Factors					
Never Smoker	48.7%	55.1%	---	50.4%	---
Former Smoker vs Never	39.2%	28.6%	0.64 (0.36, 1.16)	35.4%	0.87 (0.51, 1.48)
Current Smoker vs Never	12.2%	16.3%	1.19 (0.52, 2.72)	14.2%	1.13 (0.52, 2.45)
Light/Never Alcohol	58.1%	64.8%	---	66.3%	---
Moderate Alcohol vs Light/Never[c]	41.9%	35.2%	0.75 (0.44, 1.30)	33.7%	0.70 (0.43, 1.16)
Social Support					
No Friends	16.2%	14.3%	---	14.8%	---
Friends	83.8%	85.7%	1.16 (0.56, 2.43)	85.2%	1.11 (0.57, 2.17)
Cardiovascular Disease Risk Factors					
No Diabetes	77.0%	77.6%	---	76.1%	---
Diabetes[a]	23.0%	22.4%	0.97 (0.51, 1.84)	23.9%	1.06 (0.59, 1.89)
No HTN	25.7%	21.9%	---	26.9%	---
HTN[b]	74.3%	78.1%	1.23 (0.66, 2.29)	73.1%	0.94 (0.54, 1.64)
No Cardiac Disease	77.0%	80.1%	---	79.0%	---
Cardiac Disease	23.0%	19.9%	0.83 (0.44, 1.59)	21.0%	0.89 (0.50, 1.60)
Normal Weight	25.7%	26.0%	---	23.5%	---
Overweight[d]	48.7%	39.8%	0.81 (0.42, 1.56)	46.8%	1.05 (0.58, 1.92)
Obese[e]	25.7%	34.2%	1.31 (0.63, 2.73)	29.7%	1.26 (0.64, 2.50)
Low	60.8%	52.6%	---	54.7%	---
High Waist Circumference[f]	39.2%	47.4%	1.40 (0.81, 2.41)	45.3%	1.29 (0.78, 2.12)

*Global chi-square test significant at the $p < 0.05$ level, df=5

Bolded numbers indicate that the value is statistically significant at the 5% alpha level.

Additional Table 6. Unadjusted Logistic Regression of Demographic and Prevalence of Each Clinical Risk Factors on the 2 MET Score Categories (Active and Highly Active) within the MFMM Guideline Meeting Cluster

	Cluster 4 & MET Cat = Active N = 509		Cluster 4 & MET Cat = Highly Active N = 526	
	Prev	OR (95% CI)	Prev	OR (95% CI)
Baseline Demographics				
Age				
Gender				
Females	61.5%	---	54.8%	---
Males	38.5%	---	45.2%	**1.32 (1.03, 1.69)**
Race-Ethnicity				
Whites	24.8%	---	26.8%	---
Blacks	29.9%	---	30.2%	0.93 (0.67, 1.30)
Hispanics	43.8%	---	40.9%	0.86 (0.64, 1.17)
Marital Status				
Married	31.4%	---	32.5%	1.05 (0.81, 1.36)
Single	68.6%	---	67.5%	---
Education				
No High School	48.3%	---	44.7%	---
High School	51.7%	---	55.3%	1.16 (0.91, 1.48)
Lifestyle Factors				
Never Smoker	51.1%	---	40.7%	---
Former Smoker vs Never	33.2%	---	38.0%	**1.44 (1.09, 1.89)**
Current Smoker vs Never	15.7%	---	21.1%	**1.69 (1.20, 2.37)**
Light/Never Alcohol	69.7%	---	59.9%	---
Moderate Alcohol vs Light/Never[c]	30.3%	---	40.1%	**1.54 (1.19, 2.00)**
Social Support				
No Friends	11.0%	---	15.6%	---
Friends	89.0%	---	84.4%	**0.67 (0.47, 3.51)**
Cardiovascular Disease Risk Factors				
No Diabetes	79.2%	---	82.7%	---
Diabetes[a]	20.8%	---	17.1%	0.79 (0.58, 1.07)
No HTN	28.1%	---	27.0%	---
HTN[b]	71.9%	---	73.0%	1.06 (0.80, 1.39)
No Cardiac Disease	79.6%	---	78.9%	---
Cardiac Disease	20.4%	---	21.1%	1.04 (0.77, 1.41)
Normal Weight	33.6%	---	39.0%	---
Overweight[d]	42.2%	---	38.6%	0.79 (0.60, 1.04)
Obese[e]	24.2%	---	22.4%	0.80 (0.58, 1.11)
Low	58.2%	---	63.9%	---
High Waist Circumference[f]	41.9%	---	36.1%	0.79 (0.61, 1.01)

*Global chi-square test significant at the p < 0.05 level, df=5

Bolded numbers indicate that the value is statistically significant at the 5% alpha level.

Additional Table 7. Unadjusted Logistic Regression of Demographic and Prevalence of Each Clinical Risk Factors on the Active MET Score Category within the MFMM Guideline Meeting Cluster and the Slightly Under-Guideline Cluster

	Cluster 3 & MET Cat = Active N = 401		Cluster 4 & MET Cat = Active N = 509	
	Prev	OR (95% CI)	Prev	OR (95% CI)
Baseline Demographics				
Age				
Gender				
Females	65.8%	---	61.5%	---
Males	34.2%	---	38.5%	1.21 (0.92, 1.59)
Race-Ethnicity				
Whites	21.5%	---	24.8%	---
Blacks	24.2%	---	29.9%	1.07 (0.74, 1.55)
Hispanics	51.1%	---	43.8%	0.74 (0.53, 1.04)
Marital Status				
Married	30.9%	---	31.4%	1.02 (0.77, 1.36)
Single	69.1%	---	68.6%	---
Education				
No High School	50.1%	---	48.3%	---
High School	49.9%	---	51.7%	1.07 (0.83, 1.40)
Lifestyle Factors				
Never Smoker	52.1%	---	51.1%	---
Former Smoker vs Never	34.7%	---	33.2%	0.98 (0.73, 1.30)
Current Smoker vs Never	13.2%	---	15.7%	1.21 (0.82, 1.80)
Light/Never Alcohol	65.8%	---	69.7%	---
Moderate Alcohol vs Light/Never[c]	34.2%	---	30.3%	0.84 (0.63, 1.11)
Social Support				
No Friends	14.5%	---	11.0%	---
Friends	85.5%	---	89.0%	1.37 (0.92, 2.03)
Cardiovascular Disease Risk Factors				
No Diabetes	76.1%	---	79.2%	---
Diabetes[a]	23.9%	---	20.8%	0.84 (0.61, 1.14)
No HTN	26.9%	---	28.1%	---
HTN[b]	73.1%	---	71.9%	0.94 (0.70, 1.26)
No Cardiac Disease	78.8%	---	79.6%	---
Cardiac Disease	21.2%	---	20.4%	0.95 (0.69, 1.32)
Normal Weight	23.7%	---	33.6%	---
Overweight[d]	46.9%	---	42.2%	**0.64 (0.46, 0.87)**
Obese[e]	29.4%	---	24.2%	**0.58 (0.41, 0.83)**
Low	53.6%	---	58.2%	---
High Waist Circumference[f]	46.4%	---	41.9%	0.83 (0.64, 1.08)

*Global chi-square test significant at the p < 0.05 level, df=5

Bolded numbers indicate that the value is statistically significant at the 5% alpha level.

REFERENCES

1. Nelson ME, Rejeski WJ, Blair SN, et al. Physical activity and public health in older adults: recommendation from the American College of Sports Medicine and the American Heart Association. *Circulation.* 2007;(116): 1094-1105.
2. Rosamond W, Flegal K, Furie K, et al. Heart disease and stroke statistics–2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation.* 2008;(117):e25– e146.
3. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Stat Med.* 1986; 5:21–7.
4. Sacco RL, Gan R, Boden-Albala B, et al. Leisure-time physical activity and ischemic stroke risk: the Northern Manhattan Stroke Study. *Stroke.* 1998;(29):380-387.
5. Moss AJ, Parsons VL. Current estimates from the National Health Interview Survey: United States, 1985. *Vital Health Stat.* 10 1987:i-iv, 1-182.
6. Ainsworth BE, Haskell WL, Whitt MC, et al. Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc.* 2000;(32):S498-S504.
7. Willey JZ, Moon YP, Paik MC, et al. Lower prevalence of silent brain infarcts in the physically active: the Northern Manhattan Study. *Neurology*. 2011;76(24); 2112-8.
8. Fraley, C., Raftery, A.E. How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis. *The Computer Journal. 1998;*41(8): 578-589.
9. Willey JZ, Moon YP, Paik MC, et al. Physical activity and risk of ischemic stroke in the Northern Manhattan Study. *Neurology.* 2009; (73): 1774-1779.
10. Moore S. Physical activity, fitness, and atherosclerosis. In: Bouchard C,Shephard RJ, Stephens T, eds. Physical Activity, Fitness, and Health: International Proceedings and Consensus Statement. Champaign, Ill: Human Kinetics Publishers; 1994:570 – 578.
11. Fagard RH, Tipton CM. Physical activity, fitness, and hypertension. In: Bouchard C, Shephard RJ, Stephens T, eds. Physical Activity, Fitness, and Health: International Proceedings and Consensus Statement. Champaign, Ill: Human Kinetics Publishers; 1994:633– 655.
12. Michel BA, Bloch DA, Fries JF. Weight-Bearing Exercise, Overexercise and Lumbar Bone Density Over Age 50 Years. *Arch Intern Med.* 1989;(149):2325-2329.
13. Möhlenkamp S, Lehmann N, Breuchmann F, et al. Running: the risk of coronary events. Prevalence and prognostic relevance of coronary atherosclerosis in marathon runners. *European Heart Journal.* 2008;(29):1903-1910.
14. Benito B, Gay-Jordi G, Serrano-Mollar A. Cardiac Arrhythmogenic Remodeling in a Rat Model of Long-Term Intensive Exercise Training. *Circulation*. 2011;(123):13-22.
15. Wen CP, Wai JP, Tsai MK. Minimum amount of physical activity for reduced mortality and extended life expectancy: a prospective cohort study. *Lancet*. 2011;278(9798):1244-53.
16. Hunter GR, Bickel CS, Fisher G. Combined Aerobic/Strength Training and Energy Expenditure in Older Women. *Med Sci Sports Exerc*. 2013.

17. Norton MC, Dew J, Smith H, et al. Lifestyle Behavior Pattern Predicts Incident Dementia and Alzheimer's Disease. The Cache County Study. *J Am Geriatr Soc*. 2012;60(3):405-412.
18. Holtermann A, Marott JL, Gyntelberg F, et al. Does the Benefit on Survival from Leisure Time Physical Activity Depend on Physical Activity at Work? A Prospective Cohort Study. *PLoS One*. 2013;8(1):1-6.
19. Hu G, Eriksson J, Barengo NC, et al. Occupational, commuting, and leisure-time physical activity in relation to total and cardiovascular mortality among Finnish subjects with type 2 diabetes. *Circulation*. 2004;110(6):666-73.
20. Hu G, Jousilahti P, Antikainen R, et al. Occupational, commuting, and leisure-time physical activity in relation to cardiovascular mortality among Finnish subjects with hypertension. *Am J Hypertens*. 2007;20(12):1242-50.
21. Hu G, Tuomilehto J, Borodulin K, et al. The joint associations of occupational, commuting, and leisure-time physical activity, and the Framingham risk score on the 10-year risk of coronary heart disease. *Eur Heart J*. 2007;28(4):492-8.
22. Hu G, Jousilahti P, Borodulin K, et al. Occupational, commuting and leisure-time physical activity in relation to coronary heart disease among middle-aged Finnish men and women. *Atherosclerosis*. 2007;194(2):490-7.
23. Hu G, Sarti C, Jousilahti P, et al. Leisure time, occupational, and commuting physical activity and the risk of stroke. *Stroke*. 2005;36(9):1994-9.
24. Holtermann A, Marott JL, Gyntelberg F, et al. Does the benefit on survival from leisure time physical activity depend on physical activity at work? A prospective cohort study. *PLoS One*. 2013;8(1):e54548.
25. Clays E, De Bacquer D, Janssens H. The association between leisure time physical activity and coronary heart disease among men with different physical work demands: a prospective cohort study. *Eur J Epidemiol*. 2013;28(3):241-7.
26. Sisson SB, Camhi SM, Church TS, et al. Leisure time sedentary behavior, occupational/domestic physical activity, and metabolic syndrome in U.S. men and women. *Metab Syndr Relat Disord*. 2009;7(6):529-36.

COMPLEX DRUG USE PATTERNS AND ASSOCIATED HIV TRANSMISSION RISK BEHAVIORS IN AN INTERNET SAMPLE OF US MEN WHO HAVE SEX WITH MEN

Gary Yu[1], Melanie M Wall[1], Sabina Hirshfield[2], Mary Ann Chiasson[2]

(1) Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W. 168th St., New York City, NY 10032, USA
(2) Public Health Solution, New York City, NY, USA

Corresponding Author: Gary Yu

Email: Gy2153@columbia.edu

ACKNOWLEDGEMENTS

The author was supported as a predoctoral fellow in the Behavioral Sciences Training in Drug Abuse Research program sponsored by Public Health Solutions and National Development and Research Institutes with funding from the National Institute on Drug Abuse (5T32 DA007233). Points of view, opinions, and conclusions in this paper do not necessarily represent the official position of the U.S. Government, Public Health Solutions or National Development and Research Institutes.

ABSTRACT

Little is known about complex patterns of drug use and their association with HIV transmission risk among men who have sex with men (MSM). The aim of this study was to determine whether using a novel statistical method would aid in the detection of individual and polydrug use combinations reported prior to sex, as well as predict HIV transmission risk behaviors, such as unprotected anal intercourse (UAI) in the most recent sexual encounter among MSM. From 2004-2005, an anonymous online survey was conducted among MSM recruited from gay-affiliated websites. Latent class analysis (LCA) clustered participants into drug use groups, incorporating both the specific types and overall count of different drugs used. Analysis was limited to 8,717 U.S. MSM self-reporting drug use prior to sex in a specific encounter within the past year. Men reported average drug use before sex in the past year from a 19-item drug use list. LCA identified six distinct polydrug use classes: 1) low drug use, 2) some recreational drug use, 3) nitrite inhalants (poppers) with prescription erectile dysfunction (ED) drug use, 4) poppers with both prescription and non-prescription ED drug use, 5) all recreational, club drugs and some ED drug use, and 6) high polydrug use. Compared to participants in the low drug use class, participants in the highest drug use class were 5.5 times more likely to report UAI in their last sexual encounter and were approximately 4 times more likely to report new sexually transmitted infections (STIs) in the past year (both $p < 0.01$). Younger MSM were less likely to report UAI than older men but more likely to report an STI (both $p < 0.01$). LCA incorporating overall count of different drugs used detected 6 distinctive polydrug use classes and associated sexual risk among MSM recruited online. Participants in the low drug use class exhibited harm reduction behaviors for UAI and STIs while younger men showed risk reduction behaviors for UAI only.

Keywords: Men who have sex with men; Gay men; Internet; substance use; drug use; sexual health

RESUMEN

No sé conoce mucho de los patrones sobre la asociación con el riesgo transmitido de VIH entre los hombres que tienen sexo con hombres (HSM) y los patrones del uso de drogas. El propósito de esta investigación es para determinar con la ayuda de un método estadístico nuevo para detectar combinaciones de drogas múltiples y individuales que reportaron antes del encuentro sexual último. Y también, predecir comportamientos de riesgo transmitido de VIH, por ejemplo relationes sexuales anal sin protección (SASP). Desde 2004 y 2005, los HSH participaron en una encuesta anónima de páginas de web homosexuales. Análisis de categorías latentes (ACL) incluye el tipo específico y la suma total de drogas diferentes para clasificar consumidores en grupos. Análisis estaba limitado a 8.717 HSH en los EEUU de los que reportaron uso de drogas antes de sexo en un encuentro específico desde el año pasado. Consumidores de drogas reportaron en promedio antes de sexo en el año pasado desde un listo de 19 drogas. ACL identifica seis categorías distintas: 1) uso drogas del nivel bajo, 2) uso drogas recreativas frequentemente, 3) uso poppers y drogas de disfunción erectil (DE) con prescripción,4) uso de poppers y drogas de disfunción erectil (DE) con y sin prescripción, 5) uso drogas recreativas, del club, y Viagra frequentemente, 6) uso drogas del nivel alto. En comparación con los consumidores de drogas del nivel bajo, consumidores de drogas del nivel alto tenían una probabilidad 5,5 veces mayor de reportar SASP en su último encuento sexual ($p < 0.05$) y una probabilidad 4.0 veces mayor de reportar nuevas infecciones transmitidas sexuales (ITS) en el año pasado ($p < 0.05$). Hombres menores disminuyen reportar SASP y aumentan reportar ITS en comparación con hombres mayores. ACL utiliza la suma total y detecta categorías de drogas diversas y riesgo sexual asociado sobre los HSH en el Internét. Consumidores de drogas a nivel bajo muestran comportamientos de riesgo reducido de SASP y ITS mientras hombres menores muestran comportamientos de riesgo reducido en SASP solamente.

Introduction

Relatively little is known about patterns of combined drug use in connection with sexual HIV transmission risk behaviors in men who have sex with men (MSM), as most research has focused primarily on sexual risk behaviors and individual drug use prior to sex [1]. Risky sexual practices, such as unprotected anal intercourse (UAI), can increase the risk of acquiring or transmitting sexually transmitted infections (STI) and HIV [2].

Individual drugs have been found to be highly associated with risky sexual behaviors. Studies of MSM have examined the separate effects of individual drugs associated with sexual HIV transmission risk such as crystal methamphetamine [3-7], cocaine [8, 9], alcohol [10], and other drugs, including marijuana, nitrite inhalants (poppers), Viagra, Ecstasy, GHB, ketamine and downers [11-25]. Crystal methamphetamine use has been consistently associated with an increased risk of UAI and HIV seroconversion [37]. Among African-American MSM, two different studies showed that cocaine was associated with more UAI and HIV seroconversion within sexual networks [8] and higher HIV prevalence among individuals that reported both injection and non-injection drug use [9]. Alcohol use in combination with general non-injection drug use has also been found to be highly associated with UAI among MSM [17].

A variety of drug categories have also been explored for their impact on risky sexual behaviors, including club drugs (e.g., crystal methamphetamine, gamma hydroxybutyrate) [26-28], recreational drugs [29], prescription drugs [30], injection drugs [31], stimulants [32], and erectile dysfunction (ED) drugs [33]. Others have compared multiple drug categories (i.e., club drugs, recreational drugs, enhancement drugs) [34], but have not focused on the independent and additive effects of specific drugs on risk outcomes. These aforementioned drug categories have been associated with UAI as well as non-disclosure of HIV status and lack of knowledge of a sex partner's HIV status [20,24,29,30,32,33].

Although there is a body of literature on the relationship between HIV transmission risk and individual drugs –as well as drug categories – used prior to sex, little information exists on the combination of specific drugs, namely, the individual and additive effects of certain drugs on the likelihood of reporting HIV transmission risk behaviors. Ostrow et al. [35] recently examined the effects of the additive combination of drug categories (poppers, stimulants and ED drugs) on HIV seroconversion and found that men who reported using all three types of drugs together had the greatest risk for HIV seroconversion. However, a limited combination of drug categories was examined and injection drug use was not assessed.

This paper builds upon previous research by identifying patterns of drugs used prior to sex employing a novel modification of latent class analysis that incorporates both the specific types and overall count of different drugs used. The aim of this study was to better understand the underlying patterns and prevalence of a combination of different drugs and the associated probability of engaging in risky sexual behaviors among MSM before their most recent sexual encounter in the past year. We present data from an online sample of adult MSM from the U.S.

Methods

Sample and Study Design

From 2004-2005, MSM were recruited via study banner ads that were posted on eight U.S. and Canadian gay-oriented websites, ranging from sexual networking and chat to news sites. Men who clicked on a study banner ad were automatically directed to the study landing page which briefly described the study and contained the online consent form. Men who clicked consent were then prompted to complete an anonymous survey about sexual, drug- and alcohol-using behaviors in the past year. Participants resided in every U.S. state, Canadian province or territory, and abroad. The survey took 10 to 15 minutes to complete and no incentives were given. This study has been described in detail elsewhere [34,36]. The institutional review board of the principal investigator at Public Health Solutions (a nonprofit organization in New York City) approved all study procedures and granted a waiver of the requirement to obtain documentation of informed consent.

Overall, 19,253 individuals clicked on the survey banner ad and consented to participate; 7,924 respondents (41%) were partial completers as they were missing key outcome variables; 11,329 (59%) completed the survey. Partial completers were significantly more likely than total completers to be under age 30 (age 18-24 odds ratio [OR]: 1.7, 95% CI 1.4-1.9; age 25-29 OR: 1.3, 95% CI 1.1-1.5) [34]. The number of banner ad impressions men were exposed to was not available from the websites, therefore we could not calculate a click-through-rate or response rate. The analytic sample was limited to 8,717 MSM residing in the U.S. who reported having had sex in the last year and were thus prompted to answer questions regarding their drug use before sex within the last year.

Definition of Key Variables

Risky Sexual Behaviors

The main outcome variables were: (1) unprotected insertive and/or receptive anal intercourse during the last sexual encounter within the past year, (2) self-report of a new sexually transmitted infection (STI) diagnosed by a healthcare professional within the past year, which included a checklist: genital herpes, genital warts, anal warts, human papilloma virus (HPV, chlamydia, gonorrhea, syphilis, chancroid, and non-gonococcal urethritis (NGU) (3) knowledge of sex partner’s HIV status at the last sexual encounter within the past year (Did you know this person's HIV status?), and (4) discussion or disclosure of participant’s HIV status with the sexual partner in the most recent sexual encounter within the past year (Did you discuss or disclose your HIV status?).

Substance Use Prior to or During Sex

Respondents were asked if they had used any of the following 19 types of drugs prior to or during any sexual encounter within the past year: ketamine, methamphetamine, injected methamphetamine, ecstasy, gamma hydroxybutyrate (GHB), alcohol, marijuana, poppers, downers, cocaine [smoked, snorted, or swallowed], injected cocaine, heroin [smoked, snorted, or swallowed], injected heroin, prescription and non-prescription erectile dysfunction drugs [Viagra, Levitra, Cialis]. Only subjects that had sex within the past year saw these drug use questions. Our rationale for using past-year drug data before sex, rather than drug data from the last sexual encounter, was due to the robustness of the data, the high response rate and the high correlation with drug data from the most recent encounter. Each drug was coded dichotomously as having been used or not. Within the past year, participants could have cumulatively consumed multiple drugs before or during separate sexual encounters. A total drug count variable was created as a simple sum of all nineteen drug items used (range 0 -19) to reflect the cumulative exposure.

Statistical Analysis

Latent Class Analysis

One of the primary aims of the current work was to identify clusters of individuals reporting similar patterns of drug use prior to sex. The goal being that these distinct and divergent patterns of substance use behaviors may provide meaningful descriptions of individuals and be predictive of risky sexual behaviors perhaps more so than examining the 19 drugs individually. We used latent class analysis (LCA) [37] for this purpose. LCA is a statistical technique that identifies clusters or latent classes by assuming conditional independence between variables (e.g. the 19 dichotomous drug items) given the latent class membership. That is, the latent classes represent the optimal grouping of the data to explain the covariances observed between the variables. The parameters of the LCA model included: 1) the probability (for dichotomous) or mean (for continuous and count) of each variable within each latent class, and 2) the overall proportion of the population in each of the latent classes. The probability that a certain individual belongs to a certain latent class can be computed using Bayes' Rule [38] and the estimated parameters from the model. An individual's predicted membership to a certain latent class is determined by finding the highest class membership probability out of all of the latent classes.

Three different LCA models were fit using maximum likelihood in Mplus, version 6.11 [40] where the dichotomous variables were modeled with a binomial logit link and the count variable was modeled with a log Poisson link. The *first* model was a traditional LCA using only the 19 dichotomous drug items. The *second* was a simplified LCA model where just one observed variable was used which was the total count of different drugs used. This model is also called a univariate finite mixture model [39]. The *third* model was a novel modified LCA using the 19 drug items and additionally including the total drug count as another indicator of the latent classes. This inclusion of the total count as a separate indicator is non-standard for LCA but as described in the results we found it aided in identifying a parsimonious set of classes while also facilitating an ordered dose interpretation. Determination of the optimal number of classes (clusters) relied primarily on the Bayesian information criterion (BIC) which balances model fit and parsimony [38]. Analyses were stopped after reaching a maximum of 10 classes which would limit the qualitative usefulness of the descriptive labels of each class. Qualitative descriptions of the resulting drug profile clusters are based on the prevalence of individual drugs and types of drugs and were labeled as high/low if the prevalence of use within the latent class was above or below the overall sample prevalence by at least 10%.

Comparisons of the LCA Drug use classes and the Risky Sexual Behaviors

Demographic covariates (i.e., age, race, income and self-reported HIV status) were compared across the predicted LCA drug use classes using chi-square tests. Associations between the predicted LCA drug use class for each individual and the four risky sexual behaviors were estimated using logistic regression controlling for demographic covariates.

Results

Polydrug use patterns from the LCA Model

As described above, three different LCA models were fit to the drug use data prior to sex in the past year. The model using the 19 drug items alone did not result in an optimal number of classes found using the BIC comparison statistic. Specifically, the BIC indicated that 9 classes fit better than all smaller number of classes but then for 10 classes, the model would not converge. This model was not considered further. Second, the finite mixture model with only the total drug count as the informative variable resulted in a three cluster solution based on the BIC. The three-cluster solution consisted of 81% of individuals belonging to the low drug use class (mean drugs consumed [range] = 1.6 [0-4]), 16% of individuals belonging to the moderate drug use class (6.4 [5-9]) and 3% of individuals belonging to the high drug use class (11.4 [10-18]). Third, the LCA with 19 drug items and also the total count of different drugs used resulted in an optimal solution of six classes based on the BIC. This hybrid LCA model incorporating specific drugs as well as overall use resulted in six different qualitatively meaningful patterns of drug use (Table 1) for the US based sample.

The overall prevalence of different drugs used and the results of prevalence within the classes identified by the hybrid LCA are shown in Table 1. Overall, men reported an average use of 2.6 drugs, 73% reported alcohol use, 24% reported poppers use and 32% reported marijuana use before their last sexual encounter in the past year. The six latent classes were: low drug use class (**1**) (mean 0.7 drugs); some recreational drug use class (**2**) (mean 2.4 drugs), with higher than average use of marijuana (56.5%), alcohol (96.4%) and poppers (46.9%); poppers with prescription ED drug class (**3**) (mean 3.6 drugs), with higher than average use of poppers (60.2%) and prescription ED drugs (96.6%); poppers with both prescription and non-prescription ED drug class (**4**) (mean 3.9 drugs), with 45.5% using poppers and 86.4%, 46.2%, and 63.6% using non-prescription ED drugs; all recreational drugs, club drugs and some Viagra drug use class (**5**) (mean 5.7 drugs), with higher than average use of all the recreational drugs (i.e. cocaine (52.2%)), club drugs (i.e., methamphetamine (66.1%) and Ecstasy (63.1%)) and ED drugs (31.5%). Latent class **6** was the high polydrug use class (mean 9.7 drugs), with higher than average use of all 19 drug items. The LCA also estimated the proportion of the sample in each class. The low drug use class was the largest (44%); followed by the some recreational drug use class (29%); all recreational, club drugs and some ED drug use class (13%); poppers and prescription drug ED class (8%); high polydrug use class (6%); and poppers with both prescription and non-prescription ED drug class (2%).

Demographic Characteristics associated with the Six Latent Drug Use Classes

Among the 8,717 MSM, median age was 37 (range 18 to 92). Most men were white (71.5%), followed by 12.8% African-American, 9.8% Latino, 1.8% Asian/Pacific Islander, and 3.3% Mixed/Other. Almost half of the sample reported an income greater than $50,000. Among those who answered the HIV testing question, 11.3% reported testing HIV-positive, 67.2% testing HIV-negative, and 21.5% reported an unknown status or were not tested. Over half (53%) of men reported that their last sexual encounter occurred within the last 7 days; 15.5% reported that their last encounter was today (date of the survey interview); 17.2% reported that their last encounter occurred in the past month, with the remainder of the sample reporting their last encounter within the past year.

Each demographic covariate (i.e., age, race, income, and HIV status) was significantly associated with the six LCA drug classes in Table 2 (p < 0.01). Older men tended to be overrepresented in the poppers and prescription ED drug class and the poppers with both prescription and non-prescription ED drug class while predominately white men, men with higher income and HIV+ men tended to be overrepresented in all three of the following classes: the poppers with prescription ED drug class; the all recreational, club drug and some ED drug class and the high polydrug use class.

Associations of Drug Use Classes and Risky Sexual Behaviors
Compared to men in the lowest drug class, men in the higher drug classes in Table 3 and Table 4 were significantly more likely to report UAI and a new STI (both p < 0.01). Most drug use reported was significantly associated with UAI and a new STI in the most recent encounter within the past year.

The relationship between the three latent classes found using just the total count of drugs used and each of the risky sexual behaviors is shown in Table 3. Men in the highest drug count class (using 10-18 drugs) were 4.37 times more likely to report UAI in their most recent encounter than men in the lowest drug use count class (using 0-4 drugs). When considering the total drug count, men consuming the highest number of drugs (i.e., 10-18) were 3.19 times more likely to report a new STI than the lowest drug use class (Table 3). As the polydrug classes increased in terms of the number of drugs used, the odds of engaging in UAI and reporting new STIs increased in direct proportion as well. Younger MSM under age 30 were 0.74 times less likely to report UAI and 4.10 times more likely to report a new STI. African-American MSM were significantly less likely to disclose their HIV status and less likely to know their partner’s HIV status yet more likely to engage in safer sexual practices over all four outcomes (AOR < 1). HIV positive men were 2.01 times more likely to engage in UAI and 3.21 times more likely to report a new STI.

Even stronger association with risky sexual practices of UAI and STI were found across the six drug use classes based on the LCA of the 19 different drugs combined with the count of different drugs used (Table 4). Those in class 6, the high polydrug use category, had odds of engaging in UAI and reporting STI in their most recent encounter (AOR 5.50 and AOR 3.94) compared to the other latent classes (Table 4). Even though class 5 has a higher mean total drug count than class 3, both class 3 and class 5 have similar risks of reporting a UAI. As for HIV status, HIV positive men were more likely to engage in UAI (AOR > 1). African-American men were less likely to display harm reduction behaviors towards knowledge and disclosure of HIV status while less likely to engage in risky practices (AOR < 1). For the two outcomes on HIV disclosure, only Latent Class 5 was associated with a decreased odds of asking about a partner’s HIV status at 0.78. (Table 4).

Discussion

In this Internet sample of U.S. MSM recruited from gay-oriented websites, past-year substance use prior to or during sex and risky sexual behaviors was common. To our knowledge, this is the first U.S. study of MSM that assessed self-reported behaviors of sexual risk-taking with time-related, complex patterns of polydrug use as elucidated through latent class analysis (LCA). We developed a more comprehensive understanding of the complex relationship between polydrug use and sexual risk in this sample of MSM. With the LCA of the 19 drugs, including the total count of different drugs used, we found six distinct patterns of polydrug use: 1) low overall drug use; 2) some recreational drug use; 3) poppers with prescription ED drug use; 4) poppers with both prescription and non-prescription ED drug use; 5) all recreational drug use, club drug and some ED drug use; and 6) high overall polydrug use.

MSM in the low polydrug use class comprised almost half of the sample and also corresponded to the lowest prevalence rates of UAI and new STIs in the past year. MSM in class 2 engaged in recreational drug use, such as marijuana, alcohol and poppers but did not report erectile dysfunction drugs. Respondents in classes 3 and 4 reported poppers with ED drugs, (prescription drugs for class 3 and both prescription and non-prescription drugs in class 4), possibly due to sexual dysfunction side effects attributed to substance use before sex [41, 42]. In the context of the differences between classes 2, 3 and 4, some men may use substances to increase sexual pleasure, some may also experience additional sexual problems because of those same substances and compensate by simultaneous and concurrent use of the ED drugs [43]. Club drugs, such as crystal methamphetamine and ecstasy, can inhibit an erection [44]. Studies of the use of erectile dysfunction medication in conjunction with club drugs to counteract sexual side effects has been associated with HIV and STI transmission risk and riskier sexual behaviors, such as UAI [45-47]. In two online studies of MSM and HIV transmission risk through risky sexual behaviors, risk factors associated with crystal methamphetamine use before sex included young age, having an STI and being HIV-positive [48, 49]. It seems that using both prescription and non-prescription ED drugs (class 4) is associated with an elevated risk for only UAI as compared to only prescription ED drugs (class 3) (OR for class 4 vs class 3=1.47, $p < 0.05$).

Men in classes 5 and 6 reported high polydrug use; these classes are novel as they have not been considered in the literature due to the unique combination of recreational, club, erectile dysfunction and injection drug use. The impact of intravenous drug use, though small in proportion, becomes apparent with its additive effect with recreational, club drug and some ED use (class 5), and with high polydrug use (class 6), which contributed to predicting a subsequent increase in risky sexual behaviors. The sizable proportion of men that fell into classes 5 and 6 (13% and 6%) warrants further exploration, as such high levels of multiple drug use are worrisome in its relationship to HIV transmission risk, with high reporting rates of UAI and new STIs within the past year. The inclusion of intravenous drug use as exemplified by these two classes allowed the assessment of risk taking behaviors that was previously limited in the literature to certain individual drug items and drug groups.

Additionally, demographic trends show differences in reported risky sexual behaviors among young MSM. Younger men were significantly less likely to report UAI than older men but significantly more likely to report an STI. This interesting finding may be a sign of successful harm reduction efforts in terms of the prevention of HIV acquisition through UAI but not newly reported STIs, which may suggest a shift in risky sexual behavior trends in young adults. Future research is needed to examine the relationship of complex drug use patterns and STI transmission among this subgroup of young MSM.

The LCA analyses in this paper clustered individuals by their entire profile of drug use building upon one another in an additive fashion to paint a more complex and diverse picture of the patterns of polydrug use not previously elucidated in prior studies [34, 35]. Also, the six class LCA model found the highest magnitude of association between drug use and risky sexual behaviors as compared to the simpler model using using only total drug count. The LCA provided an overall holistic picture of polydrug use through its six class solution that encompassed combinations of different individual drug items. The LCA also provided a more in-depth look at the variability in polydrug use patterns than simply examining the total count.

Limitations

Limited research exists regarding complex patterns of polydrug use prior to sex in MSM in relation to sexual risk behaviors. This online study sought to measure the prevalence of self-reported risk-taking behaviors for research purposes and the findings were limited to MSM who used the online sites from which participants were recruited. As such, the population studied may be different thus limiting external validity or generalizability of the study findings. Given the study was cross-sectional and used self-report, associations between drug use and sexual risk taking behavior may be hindered by recall bias and or social desirability. Also, we did not ask about the quantity of specific drugs used, and we did not clinically assess substance abuse or dependence. Further, the cumulative combination of reported drug use by participants within the past year of the online survey entry date was time-dependent, meaning that they could have consumed different drugs at different sexual encounters. These limitations should be taken into account for future studies.

Conclusions

A large percentage of U.S. MSM recruited online from gay-oriented sexual networking, chat, or news websites self-reported risky sexual behaviors in connection with drug use in the past year. We did not provide any monetary incentives to complete the survey, yet it is clear that MSM who participated in this online study, as well as in our other online studies [48-53] were willing to report and describe their drug use and sexual risk-taking behaviors [1]. The use of the Internet as a medium for HIV prevention is at an early stage, yet it shows promise as a way to target groups at high risk for substance use disorders and HIV transmission.

The statistical modeling introduced in this paper has implications for future risk-related interventions. The LCA can provide a quick, simple and easy way to identify individuals immediately after completion of the online survey that are at high risk for sexual risk-taking and substance use disorders through their survey profile. Individuals can then be given a risk profile score, as part of a sexual health report card, with referrals to prevention and treatment resources.

Research on the complexity of the patterns of drug use on risky sexual behaviors is limited and more formative work is needed to understand the interplay of a diverse set of drugs among MSM and how they shape and negotiate their subsequent sexual encounters. Increased insight into the diverse combinatorial effects of different classes of substance use can guide researchers and clinicians to more accurately assess, refine and tailor intervention to prevent the transmission of HIV through safer sexual practices and harm reduction techniques in drug use. This content could be provided online or in any offline setting that has access to computers.

Using the LCA enabled us to identify underlying patterns of polydrug use among this sample of MSM recruited online from gay-oriented websites that were not possible using other more commonly used methods of considering drugs separately or grouping similar drugs. The LCA allowed us to elucidate, not only qualitatively meaningful, but also statistically rigorous findings based on a principled methodological approach. The clustering of drug use patterns into six classes with a dose-response gradient indicated distinct subgroups with differing levels of risk-taking behaviors. Future research should investigate these unique patterns in order to develop tailored computer-based assessment and treatment for harm and risk reduction in substance use and sexual risk-taking behaviors in MSM.

ACKNOWLEDGEMENTS

The author was supported as a predoctoral fellow in the Behavioral Sciences Training in Drug Abuse Research program sponsored by Public Health Solutions and National Development and Research Institutes with funding from the National Institute on Drug Abuse (5T32 DA007233). Points of view, opinions, and conclusions in this paper do not necessarily represent the official position of the U.S. Government, Public Health Solutions or National Development and Research Institutes.

Table 1. LCA with Total Drug Counts Model

Latent Class	Average	Low Drug Use (1)	Some Recreational Drug Use (2)	Poppers with Prescription ED Drug Use (3)	Poppers with ED Drug Use (4)	All Recreational, Club Drug and some ED Use (5)	High Polydrug Use (6)
Percentages	%	%	%	%	%	%	%
Recreation Drugs (R)							
Alcohol	72.8%	54.2%	96.4%	68.0%	57.3%	85.5%	77.9%
Poppers	34.2%	4.3%	46.9%	60.2%	45.5%	69.9%	82.9%
Marijuana	31.6%	2.3%	56.5%	27.9%	19.7%	62.4%	69.7%
Cocaine	12.1%	0%	7.8%	4.6%	0%	52.2%	52.8%
Downers	5.7%	0.2%	4.5%	6.3%	0%	17.7%	29.4%
Prescription Drugs (P)							
Viagra	22.0%	3.0%	14.6%	96.6%	30.3%	32.6%	79.3%
Cialisp	8.8%	0.4%	2.6%	49.2%	16.7%	6.4%	53.9%
Levitra	5.9%	0.3%	1.8%	31.6%	13.6%	3.0%	39.7%
Non-prescription Drugs (N)							
Viagra	12.6%	0.7%	12.6%	3.2%	86.4%	31.5%	56.2%
Cialis	4.3%	0.2%	0.7%	1.8%	63.6%	6.9%	37.6%
Levitra	2.5%	0.1%	0%	0.4%	46.2%	2.3%	25.2%
Club Drugs (C)							
Amphetamine	15.3%	0.2%	3.2%	7.9%	15.9%	66.1%	94.6%
Ecstasy	14.4%	0%	5.2%	3.1%	0%	63.1%	85.8%
GHB	10.5%	0.1%	0.6%	3.8%	6.1%	41.2%	85.2%
Ketamine	7.6%	0%	0%	0.7%	0%	28.9%	70.4%
Amphetamine Inj	2.5%	0%	0.2%	0%	2.3%	5.8%	30.7%
Injection Drugs (I)							
Cocaine Inj	0.5%	0%	0.1%	0%	0%	0.9%	7.3%
Heroin	0.4%	0%	0%	0%	0%	0.9%	4.4%
Heroin Inj	0.2%	0%	0.1%	0%	0%	0.2%	3.5%
Avg Different Drug Use	2.6	0.7	2.4	3.6	3.9	5.7	9.7
Proportion in Class	100%	43.5%	29.1%	7.8%	1.5%	12.5%	5.5%
N	8717	3794	2538	681	132	1093	479

Light Grey Shows on Average Prevalence of Drug Use within +/-10%
Dark Grey Shows Greater than Average Prevalence of Drug Use > +10%
No shading indicates lower than average prevalence of Drug Use < -10%

Table 2. Demographic Characteristics of Sample and by the Six Latent Classes

	Overall	Low Drug Use (1)	Some Recreational Drug Use (2)	Poppers with Prescription ED Drug Use (3)	Poppers withED Drug Use (4)	All Recreational, Club Drug and some ED Use (5)	High Polydrug Use (6)
Age**	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
18-29	2252 (26.3%)	1267 (33.9%)	649 (26.0%)	27 (4.0%)	9 (6.9%)	227 (21.1%)	75 (15.8%)
30-39	2759 (32.1%)	1202 (32.2%)	801 (32.1%)	117 (17.4%)	39 (30.0%)	407 (37.8%)	193 (40.6%)
40-49	2532 (29.5%)	901 (24.1%)	762 (30.6%)	309 (46.0%)	53 (40.8%)	350 (32.5%)	157 (33.1%)
50+	1038 (12.1%)	367 (9.8%)	280 (11.2%)	219 (32.6%)	29 (22.3%)	93 (8.6%)	50 (10.5%)
Race/Ethnicity**							
White	6129 (71.5%)	2425 (65.2%)	1796 (71.8%)	584 (86.5%)	100 (76.3%)	837 (78.2%)	387 (82.3%)
African American	1094 (12.8%)	690 (18.5%)	316 (12.6%)	21 (3.1%)	8 (6.1%)	43 (4.0%)	16 (3.4%)
Hispanic	839 (9.8%)	388 (10.4%)	255 (10.2%)	39 (5.8%)	13 (9.9%)	109 (10.2%)	35 (7.4%)
Asian	156 (1.8%)	75 (2.0%)	39 (1.6%)	10 (1.5%)	2 (1.5%)	23 (2.1%)	7 (1.5%)
Mixed/Other	284 (3.3%)	111 (3.0%)	75 (3.0%)	16 (2.4%)	8 (2.8%)	53 (18.7%)	21 (4.5%)
Income**							
< $30 K	1907 (23.9%)	963 (28.2%)	564 (24.1%)	74 (11.7%)	22 (11.7%)	195 (19.2%)	89 (19.9%)
$30-$50 K	2326 (29.2%)	1043 (30.6%)	679 (29.0%)	154 (24.3%)	33 (26.6%)	294 (29.0%)	2326 (27.5%)
>$50 K	3740 (46.9%)	1405 (41.2%)	1098 (46.9%)	407 (64.1%)	69 (55.6%)	525 (51.8%)	236 (52.7%)
HIV Status**							
HIV+	980 (12.3%)	164 (4.6%)	235 (10.0%)	142 (24.0%)	21 (19.4%)	234 (23.9%)	184 (45.2%)

** Statistically significant at the 1% alpha level.

Table 3. Adjusted[a] relationship between classes of total drug count and main outcomes

	UAI		New STIs[b]		Know Partner's HIV Status		Discuss/Disclose Own HIV Status	
	Prevalence	AOR[a] (95% CI)	Prevalence	AOR (95% CI)	Prevalence	AOR (95% CI)	Prevalence	AOR (95% CI)
Count of Number of drugs used								
0-4	19.4%	1.00	9.2%	1.00	64.2%	1.00	63.7%	1.00
5-9	41.2%	**2.44 (2.11, 2.81)***	21.0%	**1.95 (1.62, 2.33)***	62.1%	0.88 (0.76, 1.01)	63.3%	0.97 (0.84, 1.12)
10-18	58.3%	**4.37 (3.20, 5.96)***	33.2%	**3.19 (2.28, 4.46)***	58.9%	0.80 (0.59, 1.09)	63.4%	0.82 (0.60, 1.11)
Age								
18-29	16.9%	**0.73 (0.59, 0.91)***	12.3%	**4.10 (2.88, 5.84)***	58.8%	0.86 (0.72, 1.04)	61.7%	1.20 (0.99, 1.44)
30-39	24.8%	0.94 (0.77, 1.13)	12.3%	**2.89 (2.07, 4.02)***	64.1%	0.97 (0.82, 1.15)	64.7%	**1.22 (1.03, 1.45)**
40-49	28.1%	1.03 (0.85, 1.25)	12.3%	**2.50 (1.79, 3.48)***	66.9%	1.11 (0.93, 1.31)	65.5%	**1.24 (1.04, 1.47)**
50+	27.2%	1.00	5.9%	1.00	65.4%	1.00	60.0%	1.00
Race/ Ethnicity								
White	26.2%	1.00	12.2%	1.00	65.5%	1.00	64.7%	1.00
African American	11.7%	**0.56 (0.45, 0.70)***	7.5%	**0.65 (0.50, 0.86)***	57.3%	**0.75 (0.64, 0.88)***	57.1%	**0.74 (0.63, 0.86)***
Hispanic	22.5%	1.00 (0.82, 1.22)	12.5%	0.99 (0.77, 1.27)	59.9%	**0.84 (0.71, 0.99)**	64.3%	0.98 (0.82, 1.16)
Asian/PI	25.7%	1.03 (0.68, 1.55)	11.1%	0.82 (0.47, 1.43)	58.0%	0.89 (0.63, 1.27)	60.7%	1.05 (0.73, 1.51)
Mixed/ Other	26.3%	0.93 (0.71, 1.22)	12.7%	0.95 (0.68, 1.33)	63.3%	1.08 (0.86, 1.34)	64.8%	0.97 (0.78, 1.21)
Income								
< $30 K	20.9%	0.99 (0.84, 1.17)	12.0%	1.03 (0.83, 1.28)	58.5%	**0.85 (0.74, 0.98)**	60.6%	**0.86 (0.75, 0.99)**
$30-$50 K	23.4%	1.00	10.7%	1.00	63.4%	1.00	63.7%	1.00
>$50 K	26.2%	1.05 (0.91, 1.20)	12.5%	**1.21 (1.01, 1.46)**	66.6%	1.05 (0.93, 1.18)	65.4%	1.03 (0.91, 1.16)
HIV status								
HIV+	44.8%	**2.01 (1.71, 2.35)***	26.9%	**3.21 (2.66, 3.88)***	63.8%	0.97 (0.83, 1.14)	65.1%	1.07 (0.92, 1.26)

[a] Adjusted odds ratios (AOR). Logistic regressions, adjusted by Race, Age, Income, and HIV Status
Bolded numbers indicate that the value is statistically significant at the 5% alpha level.
*Statistically significant at the 1% alpha level.

[b] Defined as a self-report of a new sexually transmitted infection (STI) (n = 948, 10.88%) diagnosed by a healthcare professional within the past year, which included a checklist: genital herpes (n= 97, 1.11%), genital and anal warts (n = 133, 1.53%), human papilloma virus (HPV) (n = 198, 2.27%), chlamydia (n = 262, 3.01%), gonorrhea (n = 340, 3.90%), syphilis (n = 2.26%), chancroid (n = 0.06%), and non-gonococcal urethritis (NGU) (n = 95, 1.09%).

Table 4. Adjusted comparison of main outcomes between individuals in the six latent classes[a]

	UAI		New STIs[d]		Know Partner's HIV Status		Discuss/Disclose Own's HIV Status	
	Prevalence	AOR[b] (95% CI)	Prevalence	AOR (95% CI)	Prevalence	AOR (95% CI)	Prevalence	AOR (95% CI)
Latent Classes[c]								
1	15.4%	1.00	7.1%	1.00	65.0%	1.00	64.0%	1.00
2	20.9%	**1.35 (1.17, 1.56)***	10.9%	**1.48 (1.22, 1.80)***	62.7%	0.91 (0.81, 1.02)	62.3%	0.92 (0.81, 1.03)
3	35.7%	**2.50 (2.02, 3.09)***	12.8%	**1.87 (1.38, 2.53)***	68.2%	1.03 (0.84, 1.27)	66.4%	1.10 (0.90, 1.35)
4	45.0%	**4.10 (2.72, 6.19)***	16.4%	**2.57 (1.48, 4.45)***	61.7%	0.79 (0.52, 1.19)	66.7%	1.00 (0.65, 1.54)
5	35.9%	**2.56 (2.15, 3.04)***	19.9%	**2.46 (1.97, 3.07)***	60.1%	**0.78 (0.66, 0.91)***	62.7%	0.92 (0.78, 1.08)
6	56.2%	**5.50 (4.34, 6.98)***	31.5%	**3.94 (2.99, 5.19)***	63.7%	0.84 (0.67, 1.07)	64.3%	0.93 (0.74, 1.18)
Age								
18-29	16.9%	0.83 (0.67, 1.04)	12.3%	**4.33 (3.03, 6.18)***	58.8%	0.88 (0.73, 1.06)	61.7%	1.22 (1.01, 1.48)
30-39	24.8%	1.03 (0.84, 1.25)	12.3%	**2.94 (2.11, 4.12)***	64.1%	0.99 (0.83, 1.17)	64.7%	1.25 (1.05, 1.49)
40-49	28.1%	1.09 (0.90, 1.32)	12.3%	**2.51 (1.80, 3.51)**	66.9%	1.12 (0.94, 1.33)	65.5%	1.26 (1.06, 1.49)
50+	27.2%	1.00	5.9%	1.00	65.4%	1.00	60.0%	1.00
Race/Ethnicity								
White	26.2%	1.00	12.2%	1.00	65.5%	1.00	64.7%	1.00
African American	11.7%	**0.60 (0.48, 0.75)***	7.5%	**0.70 (0.53, 0.92)***	57.3%	**0.74 (0.64, 0.86)***	57.1%	**0.73 (0.63, 0.86)***
Hispanic	22.5%	1.01 (0.82, 1.23)	12.5%	0.99 (0.77, 1.27)	59.9%	**0.84 (0.71, 0.99)**	64.3%	0.98 (0.82, 1.16)
Asian/PI	25.7%	1.02 (0.67, 1.54)	11.1%	0.81 (0.47, 1.42)	58.0%	0.89 (0.63, 1.27)	60.7%	1.05 (0.73, 1.51)
Mixed/ Other	26.3%	0.93 (0.71, 1.22)	12.7%	0.94 (0.67, 1.32)	63.3%	1.08 (0.86, 1.35)	64.8%	0.97 (0.78, 1.21)
Income								
< $30 K	20.9%	1.01 (0.85, 1.20)	12.0%	1.05 (0.85, 1.30)	58.5%	**0.85 (0.74, 0.98)**	60.6%	0.86 (0.74, 0.99)
$30-$50 K	23.4%	1.00	10.7%	1.00	63.4%	1.00	63.7%	1.00
>$50 K	26.2%	1.03 (0.90, 1.18)	12.5%	**1.21 (1.00, 1.45)**	66.6%	1.05 (0.93, 1.18)	65.4%	1.02 (0.91, 1.15)
HIV status								
HIV+	44.8%	**1.84 (1.57, 2.16)***	26.9%	**2.97 (2.45, 3.59)***	63.8%	0.98 (0.83, 1.14)	65.1%	1.06 (0.90, 1.24)

[a] See Table 2 for the label names of the six latent classes.

[b] Adjusted odds ratios (AOR). Logistic regressions, adjusted by Race, Age, Income, and HIV Status.

Bolded numbers indicate that the value is statistically significant at the 5% alpha level.

[c] Latent class labels: (1) Low Drug use, (2) Some Recreational Drug Use, (3) Poppers with Prescription ED Drug Use, (4) Poppers with ED Drug Use, (5) All Recreational, Club Drug and Some ED Use, (6) High Polydrug Use.

*Statistically significant at the 1% alpha level.

[d] Defined as a self-report of a new sexually transmitted infection (STI) (n = 948, 10.88%) diagnosed by a healthcare professional within the past year, which included a checklist: genital herpes (n= 97, 1.11%), genital and anal warts (n = 133, 1.53%), human papilloma virus (HPV) (n = 198, 2.27%), chlamydia (n = 262, 3.01%), gonorrhea (n = 340, 3.90%), syphilis (n = 2.26%), chancroid (n = 0.06%), and non-gonococcal urethritis (NGU) (n = 95, 1.09%).

References

1. Vosburgh HW, Mansergh G, Sullivan PS, Purcell DW. A review of the literature on event-level substance use and sexual risk behavior among men who have sex with men. AIDS Behav. 2012 Aug;16(6):1394-1410.

2. Gold RS, Skinner MJ, Ross MW. Unprotected anal intercourse in HIV-infected and non-HIV-infected gay men. J Sex Res. 1994;31(1):59-77.

3. Rajasingham R, Mimiaga MJ, White JM, Pinkston MM, Baden RP, Mitty JA. A systematic review of behavioral and treatment outcome studies among HIV-infected men who have sex with men who abuse crystal methamphetamine. AIDS Patient Care STDS. 2012 Jan;26(1):36-52.

4. Forrest DW, Metsch LR, LaLota M, Cardenas G, Beck DW, Jeanty Y. Crystal methamphetamine use and sexual risk behaviors among HIV-Positive and HIV-Negative men who have sex with men in South Florida. J Urban Health. 2010 May;87(3):480-5.

5. Fisher DG, Reynolds GL, Napper LE. Use of crystal methamphetamine, Viagra, and sexual behavior. Curr Opin Infect Dis. 2010 Feb;23(1):53-6.

6. Grov C, Parsons JT, Bimbi DS. In the shadows of a prevention campaign: sexual risk behavior in the absence of crystal methamphetamine. AIDS Educ Prev. 2008 Feb;20(1):42-55.

7. Mimiaga MJ, Fair AD, Mayer KH, et al. Experiences and sexual behaviors of HIV-infected MSM who acquired HIV in the context of crystal methamphetamine use. AIDS Educ Prev. 2008 Feb;20(1):30-41.

8. Tobin KE, German D, Spikes P, Patterson J, Latkin C. A comparison of the social and sexual networks of crack-using and non-crack using African American men who have sex with men. J Urban Health. 2011 Dec;88(6):1052-62.

9. Fuller CM, Absalon J, Ompad DC, et al. A comparison of HIV seropositive and seronegative young adult heroin- and cocaine-using men who have sex with men in New York City, 2000-2003. J Urban Health. 2005 Mar;82(1 Suppl 1):i51-61.

10. Heath J, Lanoye A, Maisto SA. The role of alcohol and substance use in risky sexual behavior among older men who have sex with men: a review and critique of the current literature. AIDS Behav. 2012 Apr;16(3):578-89.

11. Dirks H, Esser S, Borgmann R, et al. Substance use and sexual risk behaviour among HIV-positive men who have sex with men in specialized out-patient clinics. HIV Med. 2012 Oct;13(9):533-40.

12. Outlaw AY, Phillips G 2nd, Hightow-Weidman LB, et al. Age of MSM sexual debut and risk factors: results from a multisite study of racial/ethnic minority YMSM living with HIV. AIDS Patient Care STDS. 2011 Aug;25 Suppl 1:S23-9.

13. Mustanski B, Newcomb ME, Clerkin EM. Relationship characteristics and sexual risk-taking in young men who have sex with men. Health Psychol. 2011 Sep;30(5):597-605.

14. Halkitis PN, Pollock JA, Pappas MK, et al. Substance use in the MSM population of New York City during the era of HIV/AIDS. Subst Use Misuse. 2011;46(2-3):264-73.

15. Mackesy-Amiti ME, Fendrich M, Johnson TP. Symptoms of substance dependence and risky sexual behavior in a probability sample of HIV-negative men who have sex with men in Chicago. Drug Alcohol Depend. 2010 Jul 1;110(1-2):38-43.

16. Hatfield LA, Horvath KJ, Jacoby SM, Simon Rosser BR. Comparison of substance use and risky sexual behavior among a diverse sample of urban, HIV-positive men who have sex with men. J Addict Dis. 2009 Jul;28(3):208-18.

17. Newcomb ME, Clerkin EM, Mustanski B. Sensation seeking moderates the effects of alcohol and drug use prior to sex on sexual risk in young men who have sex with men. AIDS Behav. 2011 Apr;15(3):565-75.

18. Schnarrs PW, Rosenberger JG, Satinsky S, et al. Sexual compulsivity, the Internet, and sexual behaviors among men in a rural area of the United States. AIDS Patient Care STDS. 2010 Sep;24(9):563-9.

19. Semple, S. J., Strathdee, S. a., Zians, J., & Patterson, T. L. (2009). Sexual risk behavior associated with co-administration of methamphetamine and other drugs in a sample of HIV-positive men who have sex with men. Am J Addict. 2009 Jan-Feb;18(1):65-72.

20. Mansergh G, Flores S, Koblin B, Hudson S, McKirnan D, Colfax GN. Alcohol and drug use in the context of anal sex and other factors associated with sexually transmitted infections: results from a multi-city study of high-risk men who have sex with men in the USA. Sex Transm Infect. 2008 Nov;84(6):509-11.

21. Carey JW, Mejia R, Bingham T, et al. Drug use, high-risk sex behaviors, and increased risk for recent HIV infection among men who have sex with men in Chicago and Los Angeles. AIDS Behav. 2009 Dec;13(6):1084-96.

22. Colfax G, Coates TJ, Husnik MJ, et al. Longitudinal patterns of methamphetamine, popper (amyl nitrite), and cocaine use and high-risk sexual behavior among a cohort of San Francisco men who have sex with men. J Urban Health. 2005 Mar;82(1 Suppl 1):i62-70.

23. Beckett M, Burnam A, Collins RL, Kanouse DE, Beckman R. Substance use and high-risk sex among people with HIV : a comparison across exposure groups. AIDS Behav. 2003 Jun;7(2):209-19.

24. Stueve A, O'Donnell L, Duran R, San Doval A, Geier, J. Being high and taking sexual risks: findings from a multisite survey of urban young men who have sex with men. AIDS Educ Prev. 2002 Dec;14(6):482-95.

25. Fendrich M, Mackesy-Amiti ME, Johnson TP, Pollack LM. Sexual risk behavior and drug use in two Chicago samples of men who have sex with men: 1997 vs. 2002. J Urban Health. 2010 May;87(3):452-66.

26. Nettles CD, Benotsch EG, Uban KA. Sexual risk behaviors among men who have sex using erectile dysfunction medications. AIDS Patient Care STDS. 2009 Dec;23(12):1017-23.

27. Drumright LN, Little SJ, Strathdee SA, et al. Unprotected anal intercourse and substance use among men who have sex with men with recent HIV infection. J Acquir Immune Defic Syndr. 2006 Nov 1;43(3):344-50.

28. Prestage G, Grierson J, Bradley J, Hurley M, Hudson J. The role of drugs during group sex among gay men in Australia. Sex Health. 2009 Dec;6(4):310-7.

29. Li Y, Baker JJ, Korostyshevskiy VR, Slack RS, Plankey MW. The association of intimate partner violence, recreational drug use with HIV seroprevalence among MSM. AIDS Behav. 2012 Apr;16(3):491-8.

30. Kelly BC, Parsons JT. Prescription drug misuse and sexual risk taking among HIV-Negative MSM. AIDS Behav. 2013 Mar;17(3):926-30.

31. Ghanem A, Little SJ, Drumright L, Liu L, Morris S, Garfein RS. High-risk behaviors associated with injection drug use among recently HIV-infected men who have sex with men in San Diego, CA. AIDS Behav. 2011 Oct;15(7):1561-9.

32. Gorbach PM, Weiss RE, Jeffries R, et al. Behaviors of recently HIV-infected men who have sex with men in the year postdiagnosis: effects of drug use and partner types. J Acquir Immune Defic Syndr. 2011 Feb 1;56(2):176-82.

33. Benotsch EG, Mikytuck JJ, Ragsdale K., Pinkerton SD. Sexual risk and HIV acquisition among men who have sex with men travelers to Key West, Florida: a mathematical modeling analysis. AIDS Patient Care STDS. 2006 Aug;20(8):549-56.

34. Hirshfield S, Chiasson MA, Wagmiller RL Jr, et al. Sexual dysfunction in an Internet sample of US men who have sex with men. J Sex Med. 2010 Sep;7(9):3104-14.

35. Ostrow DG, Plankey MW, Cox C, et al. Specific sex drug combinations contribute to the majority of recent HIV seroconversions among MSM in the MACS. J Acquir Immune Defic Syndr. 2009 Jul 1;51(3):349-55.

36. Grov C, Hirshfield S, Remien RH, Humberstone M, Chiasson MA. Exploring the venue's role in risky sexual behavior among gay and bisexual men: an event-level analysis from a national online survey in the US. Arch Sex Behav. 2013 Feb;42(2):291-302.

37. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. Stat Med 1986; 5:21–7.

38. McLachlan G, Peel D. Finite Mixture Models. 2nd ed. New York: John Wiley & Sons; 2004.

39. Muthén L, Muthén B. Mplus User's Guide. Los Angeles: Muthén & Muthén; 2004.

40. Fraley C, Raftery AE. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. The Computer Journal. 1998;41(8): 578-589.

41.Lau JT, Kim JH, Tsui HY. Prevalence and sociocultural predictors of sexual dysfunction among Chinese men who have sex with men in Hong Kong. J Sex Med. 2008 Dec;5(12):2766-79.

42. Bhugra D, Wright B. Sexual dysfunction in gay men and lesbians. Psychiatry. 2007;6:125-9.

43. Hurley M, Prestage G. Intensive sex partying: Contextual aspects of sexual dysfunction. J HIV Ther. 2007 Jun;12(2):44-8.

44. Russell-Neary, S. Designer and club drugs: update for the healthcare provider. American Academy of Nurse Practitioners 17th Annual National Conference; June 19-23. Reno, Nevada: published for Medscape; 2002.

45. Prestage G, Jin F, Kippax S, Zablotska I, Imrie J, Grulich A. Use of illicit drugs and erectile dysfunction medications and subsequent HIV infection among gay men in Sydney, Australia. J Sex Med. 2009 Aug;6(8):2311-20.

46. Fisher DG, Reynolds GL, Ware MR, Napper LE. Methamphetamine and Viagra use: Relationship to sexual risk behaviors. Arch Sex Behav. 2011 Apr;40(2):273-9.

47. Fisher DG, Malow R, Rosenberg R, Reynolds GL, Farrell N, Jaffe A. Recreational Viagra use and sexual risk among drug abusing men. Am J Infect Dis. 2006;2(2):107–114.

48. Hirshfield S, Remien RH, Walavalkar I, Chiasson MA. Crystal methamphetamine use predicts incident STD infection among men who have sex with men recruited online: A nested case-control study. J Med Internet Res. 2004 Nov 29;6(4):e41.

49. Hirshfield S, Remien R, Chiasson M. Crystal methamphetamine use among men who have sex with men: Results from two national online studies. J Gay Lesbian Psychother. 2006;10:85–93.

50. Chiasson M, Hirshfield S, Humberstone M, DiFilippi J, Koblin B, Remien R. Increased high risk sexual behavior after September 11 in men who have sex with men: An internet survey. Arch Sex Behav. 2005 Oct;34(5):527-35.

51. Chiasson MA, Hirshfield S, Remien RH, Humberstone M, Wong T, Wolitski RJ. A comparison of on-line and off-line sexual risk in men who have sex with men: An event-based on-line survey. J Acquir Immune Defic Syndr. 2007 Feb 1;44(2):235-43.

52. Hirshfield S, Remien RH, Humberstone M, Walavalkar I, Chiasson MA. Substance use and high-risk sex among men who have sex with men: A national online study in the USA. AIDS Care. 2004 Nov;16(8):1036-47.

53. Hirshfield S, Wolitski RJ, Chiasson MA, Remien RH, Humberstone M, Wong T. Screening for depressive symptoms in an online sample of men who have sex with men. AIDS Care. 2008 Sep;20(8):904-10.

CONCLUSION

The dimension-informative finite mixture model shows potential for adding additional information to contribute to the field of model-based clustering methods. It was found in Paper 1 that if the number of repeated measures is highly informative of cluster membership, then using the DIMM can improve upon the traditional mixture model. Papers 2 and 3 incorporated information about varying dimensions as an additional attribute when performing multivariate finite mixture models. That is, Paper 2 incorporated the number of activities performed as an additional multivariate attribute along with frequency, duration, and energy expenditure. In Paper 3, the number of different drugs used was treated as an additional attribute alongside the vector of dichotomous indicators of the specific drugs used. This additional count or dimension-information proved useful in clustering individuals with complex physical activity and substance use profiles into meaningful clusters which facilitated their association with cardiovascular clinical risk factor and sexual risk behavioral outcomes. Future directions include 1) extending the DIMM from the 2-level model (i.e. repeated measures of a single attribute) examined to the 3-level (multiple attribute) model described in the introduction of Paper 1; 2) exploring the added impact of the recommended guideline meeting classes on clinical cardiovascular outcomes, such as stroke and mortality, in Paper 2; and 3) conducting subgroup analyses, specifically among young adults, to see if the associations between the drug classes and sexual risk outcomes are still present in Paper 3.

APPENDIX

R Program Code

```
# This R program (UVmixture.R) generates a random variable Z from a mixture of univariate normal
# with informative repeated measures (DIMM – dimensional informative mixture models)
# with three underlying distributions
# This simulation assumes equal pi’s (proportion of group n to total N) of all three groups, 100
# individuals are generated per group
# Z1 is from a UVN univariate normal distribution with mu1 = (0)
# Z2 is from a UVN univariate normal distribution with mu2 = (-3)
# Z3 is from a UVN univariate normal distribution with mu3 = (3)
# These three distributions have a compound symmetric variance-covariance structure
# where sigma=1
# The truth is generated from a mixture of 3 univariate normals with informative means and
# informative repeated measures (alpha) [Scenario 1]
# 100 individuals from a normal distribution of mean 0,  sd = 1, alpha = 1
# 100 individuals from a normal distribution of mean -3, sd = 1, alpha = 7
# 100 individuals from a normal distribution of mean 3,  sd = 1, alpha = 4


compute.UVmixture <- function() {

n <- 100

beta_1 <- 0
beta_2 <- -3
beta_3 <- 3

sigma2_1 <- 1
sigma2_2 <- 1
sigma2_3 <- 1

Z <- matrix(rep(0,(3*n*10)),3*n,10)

truegrp <- matrix(rep(0,3*n),3*n,1)

 alpha     <- matrix(rep(0,3*n),3*n,1)

 alpha1    <- rpois(n+1000,1)
 subalpha1 <- subset(alpha1,alpha1!=0 & alpha1<=10)

 alpha2    <- rpois(n+1000,4)
 subalpha2 <- subset(alpha2,alpha2!=0 & alpha2<=10)

 alpha3    <- rpois(n+1000,7)
 subalpha3 <- subset(alpha3,alpha3!=0 & alpha3<=10)

for (i in 1:n) {
                       for (j in 1:subalpha1[i]) {
                               Z[i,j]     <- rnorm(1,beta_1,sqrt(sigma2_1))
                       alpha[i,]   <- subalpha1[i]
                       truegrp[i,] <- 1
                      }
                       for (j in 1:subalpha2[i]) {
                               Z[i+n,j]     <- rnorm(1,beta_2,sqrt(sigma2_2))
                       alpha[i+n,]    <- subalpha2[i]
                               truegrp[i+n,] <- 2
                       }
                       for(j in 1:subalpha3[i]) {
                               Z[i+(2*n),j]     <- rnorm(1,beta_3,sqrt(sigma2_3))
                       alpha[i+(2*n),]    <- subalpha3[i]
                              truegrp[i+(2*n),] <- 3
                       }
               }
return(list(Z=Z,truegrp=truegrp,alpha=alpha))
}

# Creating a list of 100 files. Each file contains 1 simulation experiment for the MPlus program
# to run the simulations.
# data1.txt – Within each text file is the simulated data set
# …
# data100.txt – Within each text file is the simulated data set
#
# The index.txt file should be within the same folder as the list of 100 files (data1.txt to
# data100.txt) and contains just a list of the 100 file names in a column.
# data1.txt – First row of the index.txt file
# …
# data100.txt – Last row of the index.txt file
source("E:/CLMM/UVmixture.R")

n <- 100

for (i in 1:n) {

y <- compute.UVmixture()
data <- matrix(c(y$Z,y$alpha),,11)
write.table(data,paste("E:/CLMM/data",i,".txt",sep=""),sep="\t",row.names=FALSE,col.names=FALSE)

}
# The R program base code below is referenced in Qin and Self (2006) and modified to incorporate
# repeated measures (count) for the DIMM.
# See: http://www.mskcc.org/research/epidemiology-biostatistics/biostatistics/staff/li-xuan-qin
# for the article and detailed documentation on the R base code that was modified.

###########################################################################
#####                                                                 #####
###            R Program : count_informative_UVmixture_model.R          ###
#             Fit a dimensional-informative univariate mixture model      #
#                 (DIMM)                                                  #
#                 and SAMPLE-specific covariates x_i                      #
#           ALLOW FOR CLUSTER-SPECIFIC MEASUREMENT ERROR                  #
###                                                                     ###
#####                                                                 #####
###########################################################################

##### INPUT
###
##     data.y - matrix of observations,
##               data.y[j, i] for sample i and gene j;
###
##     data.x - matrix of covariates,
##               data.x[i, p] for sample i, gene j, and covariate p,
###
##     data.count - vector of count of physical activity items per person (gene j)
###
##     data.z - NULL (place holder)
###
##     n.clst - number of clusters for the beta associated with data.x;
###
##     type.x - type for data.x1, where it takes value
##                     "sample" for sample-specific covariates,
##                     "sample-gene" for sample-gene-specific covariates;
```

```
##
##     n.start - options for starting values (1 - 1 random starting value using pam
##                                                   partitioning around mediods
##                                             2 - 1 random starting value using k-means
##                                             3 - 25 random starting values using k-means
###
##### OUTPUT (a list of)
###
##      theta.hat - regression parameters estimated via EM algorithm;
###
##      data.u - clustering associated with data.x
###

fit.CLM.1u.sigmaK.simple.II   <- function(data.y, data.x, data.count, n.clst, n.start=1){
   ### this will be repeatedly used by M-step
   J                    <- dim(data.y)[1]
   data.x.x           <- compute.x.x.sum.simple(data.x, data.count, J) # t(data.x) %*% data.x
        dim is PxP

   ### try different starting values
   llh                <- -9999999999
   for(s in 1:n.start){
        ### get "start values"
        theta.hat               <-  fit.CLM.1u.simple.start(data.x, data.y, data.count, n.clst,
start=s)$theta.hat
      #llh               <-  fit.CLM.1u.simple.start(data.x, data.y, data.count, n.clst,
start=s)$llh
      #print(llh)
        ### iterate btw E- and M- steps
        est.hat.new    <- fit.CLM.1u.simple.EM(data.x, data.y, data.count, data.x.x, theta.hat)
        if(est.hat.new$theta.hat$llh > llh){
            est.hat    <- est.hat.new
            llh        <- est.hat.new$theta.hat$llh
        }
   }
   return(est.hat)
}

### data.x[i,p]
compute.x.x.sum.simple        <- function(data.x, data.count, J){
#  m <- dim(data.x)[1]
   m <- data.count
   P <- dim(data.x)[2]

   data.x.x.sum                      <- matrix(0, nrow=P, ncol=P) # dim is PxP

  for (j in 1:J) {
#  for(i in 1:m){
   for(i in 1:m[j]){
        data.x.x.sum          <- data.x.x.sum + (t(data.x[i,]) %*% data.x[i,])
   }
  }
   data.x.x.sum               <- data.x.x.sum/J
   return(data.x.x.sum)
}


### find a starting value for "zeta" in CLM
library(cluster)
library(MASS)
fit.CLM.1u.simple.start <- function(data.x, data.y, data.count, K, start){
   # number of genes and covariates
   J                           <- nrow(data.y)
   N                     <- ncol(data.y)
   P                           <- ncol(data.x)
   m                     <- data.count

   u.hat                       <- matrix(0, nrow=J, ncol=K)

   # get beta.hat for each gene-specific model
   #beta.hat           <- data.y %*% data.x %*% t(ginv(t(data.x) %*% data.x))

   beta.hat            <- matrix(0, nrow=J, ncol=P)
   temp                  <- rep(0,J)
   temp.x                      <- data.x[1,]
#  if(m[j]>1){for(i in 2:m[j]){
#       temp.x         <- rbind(temp.x, data.x[i,,])
#  }}
#  temp                        <- ginv(t(temp.x) %*% temp.x) %*% t(temp.x)
   for(j in 1:J) {
       temp.x            <- data.x[1,]
       if (m[j]>1){for (i in 2:m[j]) {
               temp.x      <- rbind(temp.x, data.x[i,])
       }}
       temp              <- ginv(t(temp.x) %*% temp.x) %*% t(temp.x)
        beta.hat[j,]  <- as.vector(t(data.y[j,1:m[j]])) %*% matrix(as.vector(t(temp)),,P)
#      beta.hat[j,]      <- matrix(as.vector(t(temp)),,P) %*% as.vector(t(data.y[j,1:m[j],]))
#      beta.hat[j,]    <- as.vector(t(data.y[j,,])) %*% matrix(as.vector(t(temp)),,P)

   }

   #print(dim(beta.hat))

   # group gene-specific beta.hat's by PAM if "start==1"
   if(start==1) {
        temp           <- pam(beta.hat, K)
        zeta.hat       <- temp$medoids                              # KxP matrix
        temp           <- temp$clustering
   }

   #print(zeta.hat)

   # group gene-specific beta.hat's by K-means if "start==2"
   if(start==2) {
        temp           <- kmeans(beta.hat, K)
        zeta.hat       <- temp$centers
        temp           <- temp$cluster
   }

   # pick group centers randomly if "start>1"
   if(start>2) {
       pi.hat      <- fit.CLM.1u.simple.start(data.x, data.y, data.count, K,
start=2)$theta.hat$pi.hat
      #u.hat       <- fit.CLM.1u.simple.start(data.x, data.y, data.count, K,
start=2)$theta.hat$u.hat
      #pi.hat      <- fit.CLM.1u.sigmaK.simple.II(data.y, data.x, data.count, K,
start=1)$theta.hat$pi.hat
      #u.hat       <- fit.CLM.1u.sigmaK.simple.II(data.y, data.x, data.count, K,
start=1)$theta.hat$u.hat
        #temp   <- sample(J, K, replace=FALSE)
        #zeta.hat      <- as.matrix(beta.hat[temp,])
      temp             <- kmeans(beta.hat, K, nstart=25)
        zeta.hat       <- temp$centers
      #temp        <- temp$cluster
```

```
       temp            <- sample(K, J, replace=TRUE, pi.hat)
   }

   for(k in 1:K) {
       u.hat[temp==k, k]      <- 1
      }
      #print(dim(u.hat))

   # measurement error
   #sigma2.hat                <- rep(10, K)
    b                  <- matrix(rep(beta.hat,K),,K)
    beta.hat.mean       <- apply(b*u.hat, FUN=sum, MARGIN=2)/apply(u.hat, FUN=sum, MARGIN=2)
    sigma2.hat          <-  (apply(((b*u.hat)^2), FUN=sum, MARGIN=2)-apply(u.hat, FUN=sum,
MARGIN=2)*(beta.hat.mean^2))/(apply(u.hat, FUN=sum, MARGIN=2)-matrix(rep(1,K),1,K))



   # frequency of each cluster
   #pi.hat                    <- rep(1/K, K)
    pi.hat              <- apply(u.hat, FUN=sum, MARGIN=2)/J



   # mean number of physical activities for each cluster
   #alpha.hat          <- rep(1,K)

    c                  <- matrix(rep(data.count,K),,K)

    alpha.hat           <- apply(u.hat*c, FUN=sum, MARGIN=2)/apply(u.hat, FUN=sum, MARGIN=2)

   #residuals      <- compute.residuals.simple(data.x, data.y, data.count, zeta.hat, J, N, K)

   #llh            <- compute.llh(residuals, data.count, sigma2.hat, pi.hat, alpha.hat, J, K)

   return(list(theta.hat=list(zeta.hat=zeta.hat, sigma2.hat=sigma2.hat, pi.hat=pi.hat ,
alpha.hat=alpha.hat), u.hat=u.hat))

}


### EM algorithm to fit the CLM
###
## data.x[i,p]
## data.y[j,i]
###
## zeta.hat, sigma2.hat, and pi.hat and alpha.hat are the starting values for the parameters
###

fit.CLM.1u.simple.EM   <- function(data.x, data.y, data.count, data.x.x, theta.hat){
   # number of genes, samples, covariates, and clusters
   J <- nrow(data.y)
   N <- ncol(data.y)
   P <- ncol(data.x)
   K <- length(theta.hat$pi.hat)

   # "log likelihood"
   llh.old      <- -9999999999
   llh      <- -9999999990


   ### iterate btw E- and M- steps
   while(llh-llh.old>0.01){
```

```
   #while(llh-llh.old>0.001){
       # E-step
       u.hat   <- compute.u.hat.simple(data.x, data.y, data.count, theta.hat, J, N, K)

       # M-step
       temp          <- compute.theta.hat.simple(data.x, data.y, data.count, data.x.x, u.hat, J,
N, K, P)

       # update only when llh increases; when some cluster disappears, llh decreases
       llh.old        <- llh
       if(temp$llh > llh){
          theta.hat<- temp
          llh  <- temp$llh
          #print(llh)
       }
   }
    return(list(u.hat=u.hat, theta.hat=theta.hat))

}


### compute "u.hat" - the expected clustering indicator
compute.u.hat.simple   <- function(data.x, data.y, data.count, theta.hat, J, N, K){
   # delist "theta.hat"
   zeta.hat           <- theta.hat$zeta.hat
   sigma2.hat         <- theta.hat$sigma2.hat
   pi.hat                     <- theta.hat$pi.hat
   alpha.hat          <- theta.hat$alpha.hat

   # compute the residuals "gene by gene"
   residuals          <- compute.residuals.simple(data.x, data.y, data.count, zeta.hat, J, N, K)

   # compute the numerator for "u.hat"
   log.u.hat.num              <- matrix(0, nrow=J, ncol=K)
   #temp                 <- matrix(0, nrow=J, ncol=N)
   m                     <- data.count
   for(k in 1:K){
       for (j in 1:J) {
              temp                    <- 0
              for (i in 1:m[j]) {
                     # temp         <- dnorm(residuals[,,k], sd=sqrt(sigma2.hat[k]))
                   temp            <- temp + dnorm(residuals[j,i,k], sd=sqrt(sigma2.hat[k]), log
= TRUE)
              }
                     #log.u.hat.num[j,k]<- log(pi.hat[k]) + temp
                     log.u.hat.num[j,k]<- log(pi.hat[k]) +
log(dpois(data.count[j],alpha.hat[k])/(1-exp(-alpha.hat[k]))) + temp # apply(temp, FUN=prod,
MARGIN=1)
       }
   }

   # compute the denominator for "u.hat"
   u.hat.num            <- exp(t(apply(log.u.hat.num, MARGIN=1, FUN=ceiling.all)))
   u.hat.den           <- apply(u.hat.num, FUN=sum, MARGIN=1)

   #print(u.hat.num/u.hat.den)

   return(u.hat.num/u.hat.den)
}

### substract a constant from a vector to make its max = cutoff
ceiling.all    <- function(aVector, cutoff=600){
    xx         <- max(aVector)
    aVector    <- aVector - xx + cutoff
    return(aVector)
}



### M-step for fitting CLM
###
### INPUT
##
#   data.x.x[j,,] = t(data.x[j,,])%*%(data.x[j,,])
#   data.x.y[j,] = t(data.x[j,,])%*%(data.y[j,])
##
#   u.hat is a J*K matrix of cluster membership probabilities
##
###
compute.theta.hat.simple <- function(data.x, data.y, data.count, data.x.x, u.hat, J, N, K, P){
   # estimtate "zeta" for each cluster
   m                    <- data.count
   zeta.hat           <- matrix(0, nrow=K, ncol=P)
   for(k in 1:K){
       zeta.hat.num   <- 0
       zeta.hat.den      <- 0
       #zeta.hat.den  <- sum(u.hat[,k]) * data.x.x
       for(j in 1:J){
               #for (i in 1:m[j]) {
                 #zeta.hat.num  <-zeta.hat.num + u.hat[j,k]*data.y[j,]
                      zeta.hat.num <- zeta.hat.num +
u.hat[j,k]*(t(data.x[1:m[j],])%*%data.y[j,1:m[j]])
                      zeta.hat.den <- zeta.hat.den +
u.hat[j,k]*(t(data.x[1:m[j],])%*%data.x[1:m[j],])
                    #print(zeta.hat.num)
               #}
       }
       zeta.hat[k,]   <- ginv(zeta.hat.den) %*% zeta.hat.num   # (t(data.x) %*% zeta.hat.num)
   }

   #print(zeta.hat)

   # estimate the measurement error
   sigma2.hat         <- rep(0, K)
   residuals          <- compute.residuals.simple(data.x, data.y, data.count, zeta.hat, J, N, K)
   for(k in 1:K){
       temp                  <- residuals[,,k]^2
       res.sum               <- apply(temp, FUN=sum, MARGIN=1) %*% u.hat[,k]      # numerator
       #res.den               <- N*sum(u.hat[,k])                                   denominator
      res.den          <-   sum(m*u.hat[,k])
       sigma2.hat[k]  <- res.sum/res.den
   }

   # frequency of each cluster
   pi.hat                     <- apply(u.hat, FUN=sum, MARGIN=2)/J

   # mean number of physical activities for each cluster
   c                      <- matrix(rep(data.count,K),,K)
   c_whole                  <- matrix(rep(0,J*K),J,K)
   for (j in 1:J) {
       c_whole[max(u.hat[j,]) == u.hat] <- 1
   }
   alpha.hat            <- apply(c_whole*c, FUN=sum, MARGIN=2)/apply(c_whole, FUN=sum, MARGIN=2)
   # alpha.hat.var      <- (apply(((c_whole*c)^2), FUN=sum, MARGIN=2)/(apply(c_whole, FUN=sum,
MARGIN=2)-matrix(rep(1,K),1,K)))-(alpha.hat^2)

   alpha.hat.var        <- (apply(((c_whole*c)^2), FUN=sum, MARGIN=2)-apply(c_whole, FUN=sum,
MARGIN=2)*(alpha.hat^2))/(apply(c_whole, FUN=sum, MARGIN=2)-matrix(rep(1,K),1,K))



   # compute the "log likelihood" given this MLE
   llh                <- compute.llh(residuals, data.count, sigma2.hat, pi.hat, alpha.hat, J, K)

   return(list(zeta.hat=zeta.hat, pi.hat=pi.hat, sigma2.hat=sigma2.hat, alpha.hat=alpha.hat,
alpha.hat.var=alpha.hat.var, llh=llh))
}


### compute "residuals"
compute.residuals.simple <- function(data.x, data.y, data.count, zeta.hat, J, N, K){
   # compute the fitted values

    m                   <- data.count

   # y.hat                    <- data.x %*% t(zeta.hat)

   # compute the residuals
   residuals          <- array(0, dim=c(J,N,K))
   for(k in 1:K) {
        for (j in 1:J) {
               for (i in 1:m[j]) {
                      y.hat             <- data.x[i,] %*% matrix(as.vector(t(zeta.hat[k,]))) # 1
x 1
                      # residuals[,,k]       <- t(t(data.y) - y.hat[,k])
                      residuals[j,i,k]  <- data.y[j,i] - matrix(as.vector(y.hat))            # 1
x 1
               }
        }
   }
   return(residuals)
}


### compute the "log likelihood" given this MLE
compute.llh    <- function(residuals, data.count, sigma2.hat, pi.hat, alpha.hat, J, K){
   m              <- data.count
   llh         <- 0
   for(j in 1:J){
        temp          <- 0
        for(k in 1:K){
         #for(i in 1:m[j]){
          #temp       <- temp + pi.hat[k]*(dpois(data.count[j],alpha.hat[k])/(1-exp(-
alpha.hat[k])))*prod(dnorm(residuals[j,,k],sd=sqrt(sigma2.hat[k])))
          temp        <- temp + pi.hat[k]*(dpois(data.count[j],alpha.hat[k])/(1-exp(-
alpha.hat[k])))*prod(dnorm(residuals[j,1:m[j],k],sd=sqrt(sigma2.hat[k])))
         #temp        <- temp +
pi.hat[k]*prod(dnorm(residuals[j,1:m[j],k],sd=sqrt(sigma2.hat[k])))
         #print(temp)
         #}
        }
        llh           <- llh + log(temp)
   }
   return(llh)
}
```

```
############                  the end                       ##############


# Running a Monte Carlo Simulation in R for n = 100 experiments
# Truth is informative betas and informative alphas
# Model is a dimensional informative univariate mixture model (DIMM) – Simulation Scenario 1
# The Tibshriani et al. predicted value is calculated at the end of R program

set.seed(1982)

data.x <- matrix(rep(1,1*10),10,1)

source("C:/Users/gary/Desktop/CLMM/count_informative_UVmixture_model")

source("C:/Users/gary/Desktop/CLMM/UVmixture")

group <- matrix(1:3,3,1)

"simulation" <- function(data,K=3,n=100) {

z <- array(rep(0,n*K*K),dim=c(n,K,K)) # beta values
p <- matrix(rep(0,n*K),n,K)           # proportion of group n to total N
#t <- matrix(rep(0,n*K),n,K)
s <- matrix(rep(0,n*K),n,K)           # variance of betas
a <- matrix(rep(0,n*K),n,K)           # alpha values
v <- matrix(rep(0,n*K),n,K)           # variance of the alpha values
g <- matrix(rep(0,300*n),300,n)       # predicted group membership based on the model
m <- matrix(rep(0,n),n,1)             # predicted value

for (i in 1:n) {

y <- compute.UVmixture()
truegrp<- y$truegrp
alpha <- y$alpha

L      <- fit.CLM.1u.sigmaK.simple.II(y$Z,data.x,y$alpha,3,n.start=1)

z[i,,] <- as.array(t(L$theta.hat$zeta.hat))
p[i,] <- as.array(L$theta.hat$pi.hat)
#t[i,] <- as.array(L$theta.hat$D.hat)
s[i,] <- as.array(L$theta.hat$sigma2.hat)
a[i,] <- as.array(L$theta.hat$alpha.hat)
v[i,] <- as.array(L$theta.hat$alpha.hat.var)

U3_whole <- matrix(rep(0,300*3),300,3)

for (j in 1:300) { U3_whole[max(L$u.hat[j,])==L$u.hat] <- 1 }

g[,i] <- (U3_whole %*% group)

SM       <- matrix(rep(0,300*300),300,300)

for (x in 1:300) {

                       for (y in 1:300) {

                              if (truegrp[x] == truegrp[y] & g[x,i] == g[y,i]) {

                              SM[x,y] = 1
```

```
                                    }

                              }
                    }

                    m[i,] <- (sum(SM - diag(1,300,300)))/(300*301)

}

list(z=z,p=p,t=t,s=s,a=a,v=v,g=g,m=m)

}
```

MPlus Program Code

```
! MPlus Program for Paper 1: Non-Dimensional Informative Model for Simulation [Regular Mixture
! Model]
! The truth is generated from a mixture of 3 univariate normals with informative means and
! informative repeated measures (alpha) [Scenario 1]
! 100 individuals from a normal distribution of mean 0,  sd = 1, alpha = 1
! 100 individuals from a normal distribution of mean -3, sd = 1, alpha = 7
! 100 individuals from a normal distribution of mean 3,  sd = 1, alpha = 4
!
! The truth is generated from a mixture of 3 univariate normals with informative means and
! non-informative repeated measures (alpha) [Scenario 2]
! 100 individuals from a normal distribution of mean 0,  sd = 1, alpha = 4
! 100 individuals from a normal distribution of mean -3, sd = 1, alpha = 4
! 100 individuals from a normal distribution of mean 3,  sd = 1, alpha = 4
!
! The truth is generated from a mixture of 2 univariate normal with non-informative means and
! informative repeated measures (alpha) [Scenario 3]
! 120 individuals from a normal distribution of mean 0, sd = 1, alpha = 1
! 180 individuals from a normal distribution of mean 0, sd = 1, alpha = 10
!
! The truth is generated from a mixture of 2 univariate normal with non-informative means and
! non-informative repeated measures (alpha) [Scenario 4]
! 120 individuals from a normal distribution of mean 0, sd = 1, alpha = 4
! 180 individuals from a normal distribution of mean 0, sd = 1, alpha = 4
!
! The index.txt contains 100 simulated datasets (experiments)
! CLM_simulation_data0.inp
DATA: file is E:\CLMM\index.txt;
type is montecarlo;
VARIABLE: NAMES ARE u1-u10 numact ;
USEVARIABLES ARE u1-u10; !numact ;
MISSING ARE ALL (0) ;
CLASSES = c(3) ;
!COUNT = numact ;
ANALYSIS: TYPE = MIXTURE;
ALGORITHM = INTEGRATION;
estimator = MLr ;
STARTS = 50 ;
MODEL:      %OVERALL%
            %c#1%
            [u1-u9](Beta1);
             u1-u9(std);
            %c#2%
            [u1-u9](Beta2);
             u1-u9(std);
            %c#3%
            [u1-u9](Beta3);
             u1-u9(std);

OUTPUT:   TECH1 TECH9 ;
SAVEDATA:
RESULTS
ARE
C:\Documents and Settings\gy2153.RESEARCH-822D.003\Desktop\CLMM\outputIII.txt ;

! MPlus Program for Paper 1: Dimensional Informative Model for Simulation [DIMM]
! CLM_simulation_data.inp
DATA:
file
is
E:\CLMM\index.txt;
```

```
type is montecarlo;
VARIABLE: NAMES ARE u1-u10 numact ;
USEVARIABLES ARE u1-u10 numact ;
MISSING ARE ALL (0) ;
CLASSES = c(3) ;
COUNT = numact ;
ANALYSIS: TYPE = MIXTURE;
ALGORITHM = INTEGRATION;
estimator = MLr ;
STARTS = 50 ;
MODEL:      %OVERALL%
            %c#1%
            [u1-u9](Beta1);
             u1-u9(std);
            %c#2%
            [u1-u9](Beta2);
             u1-u9(std);
            %c#3%
            [u1-u9](Beta3);
             u1-u9(std);

OUTPUT:   TECH1 TECH9 ;
SAVEDATA: RESULTS
ARE
C:\Documents and Settings\gy2153.RESEARCH-822D.003\Desktop\CLMM\output1.txt;


! MPlus Program for Paper 1: LCA on 15 Dichotomous (Binary) Physical Activities
TITLE:      This is an example of a LCA with
            15 physical activity Y/N items
            using automatic random starting values
DATA:       FILE IS E:\physical_activity\LCA_Binary_3_16_13.txt ;
VARIABLE:   NAMES ARE u1-u15 ;
            USEVARIABLES = u1-u15 ;
            CLASSES = c(4) ;
            CATEGORICAL = u1-u15 ;
            !AUXILIARY = u13 ;
ANALYSIS:   TYPE = MIXTURE ;
            STARTS = 100 ;
OUTPUT:     TECH1 TECH8 ;
SAVEDATA:
FILE IS E:\physical_activity\LCA_Only_4_MPlus_u_3_16_13.txt ;
SAVE = CPROBABILITIES ;


! MPlus Program for Paper 1: Dimensional Informative Univariate Finite Mixture Model (DIMM) on   !
the Energy Expenditure of 15 Physical Activities
TITLE:      This is an example of a count-informative Univariate Finite Mixture Model with
            15 physical activity items summarized along the
            Univariate dimension of energy expenditure (Kcal/2 wks)
            and total count using automatic random starting values
            Data follows a log-Normal distribution
DATA:       FILE IS E:\physical_activity\LCA_log_kcal_3_17_13.txt ;
VARIABLE:   NAMES ARE u1-u15 numact ;
            USEVARIABLES = u1-u15 numact ;
            MISSING ARE ALL (0) ;
            CLASSES = c(3) ;
            COUNT = numact ;

ANALYSIS:   TYPE = MIXTURE ;
            STARTS = 50 ;
MODEL:      %OVERALL%
```

```
            %c#1%
            [u1-u15](Beta1);
             u1-u15(std1);
            %c#2%
            [u1-u15](Beta2);
             u1-u15(std2);
            %c#3%
            [u1-u15](Beta3);
             u1-u15(std3);
OUTPUT:    sampstat TECH1 TECH8 ;
SAVEDATA:
!FILE IS E:\physical_activity\LCA_3_MPlus_u_3_17_13.txt ;
!SAVE = CPROBABILITIES ;

! MPlus Program for Paper 1: Regular Univariate Finite Mixture Model on
! the Energy Expenditure of 15 Physical Activities
TITLE:      This is an example of a regular Univariate Finite Mixture Model with
            15 physical activity items summarized along the
            Univariate dimension of energy expenditure (Kcal/2 wks)
            and total count using automatic random starting values
            Data follows a log-Normal distribution
DATA:       FILE IS E:\physical_activity\LCA_log_kcal_3_17_13.txt ;
VARIABLE:   NAMES ARE u1-u15 numact ;
            USEVARIABLES = u1-u15; ¡ numact ;
            MISSING ARE ALL (0) ;
            CLASSES = c(3) ;
            !COUNT = numact ;

ANALYSIS:   TYPE = MIXTURE ;
            STARTS = 50 ;
MODEL:      %OVERALL%
            %c#1%
            [u1-u15](Beta1);
             u1-u15(std1);
            %c#2%
            [u1-u15](Beta2);
             u1-u15(std2);
            %c#3%
            [u1-u15](Beta3);
             u1-u15(std3);
OUTPUT:    sampstat TECH1 TECH8 ;
SAVEDATA:
!FILE IS E:\physical_activity\LCA_3_MPlus_u_3_17_13.txt ;
!SAVE = CPROBABILITIES ;


! MPlus Program for Paper 2: Multivariate Finite Mixture Model (MFMM)
  TITLE:      This is an example of a MFMM with
              15 physical activity items
              Summarized along the dimensions of
              Total frequency/2 wks (t), average duration/session (m),
              Total energy expenditure/2 wks (k)
              using automatic random starting values
              Data follows a log-Normal distribution
  DATA:       FILE IS E:\PatternsofPhysicalActivity
              \TMKI_11_18_12.txt ;
  VARIABLE:   NAMES ARE numact t m k ;
              USEVARIABLES = numact
              t
              m k;
              !MISSING ARE ALL (0) ;
              CLASSES = c(5) ;
              COUNT = numact ;

  ANALYSIS:   TYPE = MIXTURE ;
              STARTS = 50 ;
  MODEL:      %OVERALL%

              t with m k;
              m with k;

OUTPUT:    sampstat TECH1 TECH8 ;
SAVEDATA:  FILE IS E:\MVMixture_5_u_final.txt ;
           SAVE = CPROBABILITIES ;


! MPlus Program for Paper 3: Count-Informative LCA Model
  TITLE:      This is an example of a count-informative LCA model with
              19 drug Y/N items and total count
              using automatic random starting values
  DATA:       FILE IS C:\Users\consultant\Desktop\19drugs_3_30_13.txt ;
  VARIABLE:   NAMES ARE id u1-u19 count ;
              USEVARIABLES = u1-u19
              count ;
              !MISSING ARE ALL (0) ;
              CLASSES = c(6) ;
              CATEGORICAL = u1-u19 ;
              COUNT = count ;

  ANALYSIS:   TYPE = MIXTURE ;
              STARTS = 50 ;
  MODEL:      %OVERALL%

  OUTPUT:    sampstat TECH1 TECH8 ;
  SAVEDATA:  FILE IS C:\Users\consultant\Desktop\19drugs_6_u.txt ;
             SAVE = CPROBABILITIES ;


! MPlus Program for Paper 3: Traditional LCA Model
  TITLE:      This is an example of a traditional LCA model with
              19 drug Y/N items using automatic random starting values
  DATA:       FILE IS C:\Users\consultant\Desktop\19drugs_3_30_13.txt ;
  VARIABLE:   NAMES ARE id u1-u19 count ;
              USEVARIABLES = u1-u19
              !count ;
              !MISSING ARE ALL (0) ;
              CLASSES = c(9) ;
              CATEGORICAL = u1-u19 ;
              !COUNT = count ;

  ANALYSIS:   TYPE = MIXTURE ;
              STARTS = 50 ;
  MODEL:      %OVERALL%

  OUTPUT:    sampstat TECH1 TECH8 ;
  SAVEDATA:  FILE IS C:\Users\consultant\Desktop\19drugs_6_u.txt ;
             SAVE = CPROBABILITIES ;

! MPlus Program for Paper 3: Univariate Finite Mixture Model
  TITLE:      This is an example of a traditional LCA model with
              Total count using automatic random starting values
  DATA:       FILE IS C:\Users\consultant\Desktop\19drugs_3_30_13.txt ;
  VARIABLE:   NAMES ARE id u1-u19 count ;
```

```
            USEVARIABLES = !u1-u19
            count ;
            !MISSING ARE ALL (0) ;
            CLASSES = c(3) ;
            !CATEGORICAL = u1-u19 ;
            COUNT = count ;

ANALYSIS:   TYPE = MIXTURE ;
            STARTS = 50 ;
MODEL:      %OVERALL%

OUTPUT:    sampstat TECH1 TECH8 ;
SAVEDATA:  FILE IS C:\Users\consultant\Desktop\19drugs_6_u.txt ;
           SAVE = CPROBABILITIES ;
```