Published in final edited form as:
J Acquir Immune Defic Syndr. 2012 March 1; 59(3): 325–328. doi:10.1097/QAI.0b013e31823fdbec.

Using cluster heat maps to investigate relationships between body composition and laboratory measurements in HIV-infected and HIV-uninfected children and young adults

Jane C Lindsey[1], **Denise L Jacobson**[1], **Hong Li**[2], **E Andres Houseman**[3], **Grace M Aldrovandi**[4], and **Kathleen Mulligan**[5]

[1]Center for Biostatistics in AIDS Research, Harvard School of Public Health

[2]Dept. of Preventive Medicine, Rush University Medical Center

[3]College of Public Health and Human Sciences, Oregon State University

[4]Saban Research Institute of Children's Hospital Los Angeles, University of Southern California

[5]University of California San Francisco, San Francisco General Hospital

Abstract

Cluster heat maps were used to investigate relationships between body composition, lipid levels and glucose metabolism in HIV-infected and HIV-uninfected children and young adults using data from a cross-sectional study. Three distinct clusters of participants were identified. One group had lower body fat and higher lipid measures and was mostly HIV-infected. The other two groups were a mix of HIV-infected and uninfected participants. Of these, one cluster had more participants with higher body fat and insulin resistance, which are risk factors for future cardiovascular disease, and the other had relatively normal measurements on all outcomes.

Keywords

children and young adults; metabolic abnormalities; heat maps; clusters

Introduction

As HIV-infected infants live longer, it is increasingly evident that HIV or certain antiretroviral therapies (ARVs) can cause distorted body shapes with excess fat in the belly or loss of fat in the extremities, often accompanied by abnormal lipids and glucose metabolism. To better understand the prevalence of these abnormalities during childhood and adolescence, the Pediatric AIDS Clinical Trials Group (PACTG) conducted a cross-sectional study which enrolled HIV-infected and HIV-uninfected participants (P1045)[1]. An array of outcomes was measured, including tests of metabolic function, body composition, lifestyle and diet. To date, analyses have focused on single outcomes [1,2], but an additional goal of the study was to understand the relationship among groups of outcomes and whether these relationships varied by factors such as HIV status, pubertal development, sex and race/ethnicity.

Conflicts of Interest: None declared

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Cluster heat maps [3] are an easy and effective way of visually displaying multivariate (multiple outcome) data, combining the use of color to distinguish the magnitude of measurements and dendrograms (tree-like figures showing a hierarchical grouping) to show clustering [4] of individuals and outcome measurements. Heat maps are often used for DNA microarray data to identify which genes (among hundreds or thousands) are associated with disease. In this paper we use heat maps to investigate clustering of measurements of body shape, lipids and glucose metabolism and to see whether these patterns are related to HIV status and other variables not used to generate the heat map.

Methods

P1045 enrolled 240 HIV-infected participants group-matched by Tanner stage, sex and race/ethnicity to 146 HIV-uninfected participants [1]. The institutional review board at each clinical site approved the study, and appropriate informed consent was obtained before enrollment. Participants ranged in age from 7–24 years (median age 12.4 years, 55% male, 55% African American, 11% white). Dual energy xray absorptiometry (DXA) was performed to estimate total body mass (TBM), total body fat (TBF), extremity (arm + leg) fat (EXF), trunk fat (TRF), leg mass (LEGM) and leg fat (LEGF), with results available on 379 of the 386 participants (236 HIV-infected and 143 HIV-uninfected). Fasting lipids (triglycerides, total and HDL cholesterol), glucose and insulin were measured as well as high-sensitivity C-reactive protein (CRP) which is a measure of inflammation and may be predictive of future cardiovascular disease (n=343). The following outcome measures were calculated for this analysis: Percent body fat (% body fat) = TBF/TBM; percent leg fat (% leg fat) = LEGF/LEGM; trunk-to-limb fat ratio (trunk:limb) = TRF/EXF; total cholesterol to HDL ratio (Chol:HDL) and the homeostasis model of insulin resistance (HOMA-IR [5]). Each of these outcomes depends to a varying extent on age, sex and race/ethnicity, and they differ in magnitude and variability. To put all outcomes on a similar scale, each was standardized into age and sex-adjusted z-scores using the HIV-uninfected as the reference population [1].

Using the standardized outcomes, a heat map was generated using the heatmap.2 function in the gplots library of R [6]. The heat map shows dendrograms (hierarchical tree) for participants on one axis and for the outcomes on the second axis. The vertical length of the branches in the dendrogram represents the degree of separation between individuals or outcomes. Clustering was done using Ward's method, which clusters by minimizing the sum of squared deviations of each point from the mean of its cluster, and which tends to result in spherical clusters [4]. Three colors were used in the body of the heat map representing the magnitude of outcome values with z-scores < -1 in yellow, $-1 \leq z \leq 1$ in blue and $z>1$ in red. The distribution of HIV-infected (green) and uninfected (blue) participants across clusters is shown along the top of the heat map. Choice of the best number of clusters was subjective and based loosely on the vertical length of the branches in the dendrograms. After selecting the number of clusters at the participant level, we calculated the median z-score for each outcome variable within each cluster.

To evaluate associations between clusters of participants with HIV status, sex, Tanner stage, race/ethnicity and CRP (variables not included in generating the heat map), Chi-square tests were used in univariate analyses and multinomial regression for multivariate analyses with cluster as the response variable and including all covariates [7]. To assess robustness of conclusions, 500 datasets were generated where participants were randomly assigned to a cluster. A Chi-square p-value was calculated for each cross tabulation of cluster membership with each external variable. This created a null distribution of p-values against which the observed p value was compared (the permutation p value). Bootstrapping was used to check for consistency of cluster phenotype attributes across samples.

To evaluate associations between high CRP levels and phenotype (cluster), sex, Tanner stage and race/ethnicity, logistic regression models were fit with abnormal CRP as the outcome and with cluster and each of the other three covariates one at a time in HIV-infected and HIV-uninfected participants separately.

P-values less than 5% were used for highlighting statistical significance but these analyses were purely exploratory.

Results

The heat map in Figure 1 illustrates the dendrogram for the seven outcome variables on the horizontal axis and the dendrogram for participants on the vertical axis. There were two distinct clusters of outcome variables at the highest level of separation. The first cluster (A) contains percent leg fat, percent total body fat and HDL cholesterol and the second cluster (B) contains HOMA-IR, trunk:limb fat ratio, Chol:HDL ratio, and triglycerides.

For participants, there were three major clusters named for body phenotype (Table 1). Compared to the other two clusters, cluster 1 (large) had the highest median z-scores for body composition outcomes (%leg fat z=0.64, %body fat z=0.84, trunk:limb z=0.60) and glucose metabolism (HOMA-IR z= 0.58) and second highest medians for lipids (Chol :HDL z=0.71, triglycerides z=0.95 and HDL z= −0.70). This is seen in the heat map as mostly yellow coloration for HDL (for which higher values are better), and blue and red coloration across the other six outcomes. Phenotypically, these were subjects with above average percent fat exhibiting evidence of dyslipidemia and greater insulin resistance. Cluster 2 (thin) had the highest median triglycerides (z=2.27) and Chol:HDL (z=1.38), the lowest median HDL (−0.71), %total fat (−0.89) and %leg fat (−1.22) and moderate to high trunk:limb (0.23). The low percent fat and moderate to high trunk:limb ratio suggest loss of extremity fat along with dyslipidemia, but normal glucose metabolism. Cluster 3 (average) had the highest median HDL (z=0.22) and the lowest median triglycerides (z=−0.31), Chol:HDL (z=−0.41), trunk:limb (z=−0.30) and HOMA-IR (z=−0.45), representing more normal body phenotype, lipids and glucose metabolism.

We then investigated whether membership in one of the three participant clusters was related to HIV status, sex, Tanner stage, race/ethnicity or CRP. Proportions of participants by each characteristic within each cluster are shown in Table 1. In univariate analyses evaluating each predictor separately, only HIV status (p<0.001) and CRP (p<0.001) were related to cluster membership. The high proportion of HIV-infected participants in cluster 2 (thin, 90.9%) as compared to cluster 1 (large, 49.5%) and cluster 3 (average, 55.9%) was evident in the heat map where the sidebar at the top was predominantly green. For CRP, the proportions of participants with high CRP (>2 mg/l) was highest in cluster 1 (large, 32.4%), compared to 13.6% in cluster 2 (thin) and 11.8% in cluster 3 (average). In a multivariate multinomial model that included all five predictors, HIV status and CRP remained significant (p<0.001). Race/ethnicity was also significant (p=0.03) with a greater odds of being Hispanic in cluster 2 relative to the other two clusters. The univariate results were confirmed by the permutation p-values. Consistent with the observed heatmap, the cluster with the lowest median percent leg fat had the highest proportion of HIV-infected subjects in 96% of 500 bootstrapped samples. Similarly, in 92% of the samples, the cluster with the highest median percent leg fat had the highest proportion of subjects with CRP > 2 mg/l.

To investigate whether there were different relationships by HIV status between high levels of CRP (outcome) and cluster, sex, Tanner stage or race/ethnicity (predictors), a logistic regression model was fit for one predictor at a time, separately by HIV status. High CRP was defined as > 2 mg/l. In HIV-uninfected subjects, sex and race/ethnicity were not related to high CRP levels after adjusting for cluster. In a model with Tanner stage and cluster, both cluster (p=0.003) and Tanner stage (p=0.027) were predictive. Cluster 2 (thin) had no subjects with high CRP (0.0%), cluster 3 (average) had 6.9% with high CRP and cluster 1 (large) had 31.1% with high CRP. The prevalence of high CRP increased with Tanner stage (9.9% in Tanner 1–2, 22.2% in Tanner 3–4 and 27.3% in Tanner 5). In HIV-infected subjects, only cluster was predictive. As with the HIV-uninfected subjects, cluster 2 (thin) had the lowest percent of subjects with high CRP (15.8%), followed by cluster 3 (average, 17.9%) and cluster 1 (38.5%). Overall prevalence of high CRP was greater in each cluster for the HIV-infected compared to the HIV-uninfected participants.

Discussion

We have illustrated the use of heat maps to assess relationships among groups of measurements in different domains and to identify participant phenotypes. In this study population, we identified three distinct clusters with characteristic phenotypes. One cluster (thin) consisted mostly of HIV-infected participants with abnormal body shape, characterized by lower extremity fat and dyslipidemia. In contrast, the other two clusters (large and average) had a more equal mix of HIV-infected and uninfected participants, suggesting that host and environmental factors may play more important roles in these phenotypes. The heat maps helped us visualize patterns that support previous observations that some HIV-infected children have a specific phenotype that is probably related to HIV itself or to ARVs. However, they also illustrate that with improved health, HIV-infected children and young adults may have a similar risk as HIV-uninfected for diabetes and cardiovascular outcomes.

Using the phenotypes identified in the heat maps allowed us to explore more complicated relationships with other participant characteristics and see whether patterns were similar by HIV status. For example, we were able to explore the relationship between phenotype and CRP, a predictor of future cardiovascular disease, and found that the large cluster had a higher proportion of subjects with elevated CRP. The HIV-infected group had higher rates of elevated CRP than the HIV-uninfected in all three clusters, suggesting that generalized inflammation is not related only to body phenotype in HIV-infected children and young adults. However, only in the HIV-uninfected subjects did the rates of elevated CRP significantly increase with Tanner stage. These were cross-sectional analyses so we could not determine temporal relationships between body phenotype and CRP.

Heatmap analyses have advantages over standard approaches when investigating differences between groups of subjects. For example, in a dataset with many measurements, variables can be highly correlated (e.g. HDL, LDL, total cholesterol) and models may break down. With heatmaps, there are no such limitations on the number of variables being used, their correlations, or any assumptions of multivariate normality. Less well-known predictive methods such as random forest or support vector machines offer less insight into associations between variables.

Heat maps are based on cluster methodology and care is needed in their use. Data must be on a similar scale, different patterns are observed depending on the distance measure selected, and there is subjectivity in choosing the `best' number of clusters. For example, potentially meaningful distinct color patterns were observed within the three chosen participant sub-clusters. It is also important to assess robustness of assignment to clusters. We illustrated approaches using a permutation test and the use of bootstrapping. Despite these caveats, heat maps and multivariate analyses can be used to further explore the relationships between participant clusters and HIV-specific host and environmental factors, and could include many more outcomes than are illustrated in this example. Heat maps may prove to be particularly useful for exploratory and hypothesis-generating analyses.

Acknowledgments

The authors would like to thank the children and young adults who participated in this study, their families and the entire P1045 protocol team for their contributions and support. Overall support for the International Maternal Pediatric Adolescent AIDS Clinical Trials Group (IMPAACT) was provided by the National Institute of Allergy and Infectious Diseases (NIAID) [U01 AI068632], the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), and the National Institute of Mental Health (NIMH) [AI068632]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work was supported by the Statistical and Data Analysis Center at Harvard School of Public Health, under the National Institute of Allergy and Infectious Diseases cooperative agreement #5 U01 AI41110 with the Pediatric AIDS Clinical Trials Group (PACTG) and #1 U01 AI068616 with the IMPAACT Group. Support of the sites was provided by the National Institute of Allergy and Infectious Diseases (NIAID) and the NICHD International and Domestic Pediatric and Maternal HIV Clinical Trials Network funded by NICHD (contract number N01-DK-9-001/HHSN267200800001C).

Financial Support: This work was supported by the Statistical and Data Analysis Center at Harvard School of Public Health, under the National Institute of Allergy and Infectious Diseases cooperative agreement #5 U01 AI41110 with the Pediatric AIDS Clinical Trials Group (PACTG) and #1 U01 AI068616 with the IMPAACT Group. Support of the sites was provided by the National Institute of Allergy and Infectious Diseases (NIAID) and the NICHD International and Domestic Pediatric and Maternal HIV Clinical Trials Network funded by NICHD (contract number N01-DK-9-001/HHSN267200800001C)

References

1. Aldrovandi GM, Lindsey JC, Jacobson DL, Zadzilka A, Sheeran E, Moye J, Borum P, Meyer WA, Hardin DS, Mulligan K, the Pediatric AIDS Clinical Trials Group P1045 Team. Morphologic and metabolic abnormalities in vertically HIV-infected children and youth. AIDS. 2009; 23:661–672. [PubMed: 19279441]
2. Jacobson DL, Lindsey JC, Gordon C, Moye J, Hardin D, Mulligan K, Aldrovandi GM, the Pediatric AIDS Clinical Trials Group P1045 Team. Total body and spine bone mineral density across Tanner stage in vertically HIV-infected and uninfected children and youth in PACTG 1045. AIDS. 2010; 24(5):687–696. [PubMed: 20168204]
3. Wilkinson L, Friendly M. The history of the cluster heat map. The American Statistician. 2009; 63:179–184.
4. Everitt, B. Cluster Analysis. Heinemann Educational Books; London: 1974.
5. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostatis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia. 1985; 28:412–419. [PubMed: 3899825]
6. R Development Core Team. R Foundation for Statistical Computing. Vienna, Austria: 2009. R: A language and environment for statistical computing. ISBN 3-900051-07-0, URL http://www.R-project.org
7. Agresti, A. Categorical Data Analysis. John Wiley and Sons; New York: 1990.

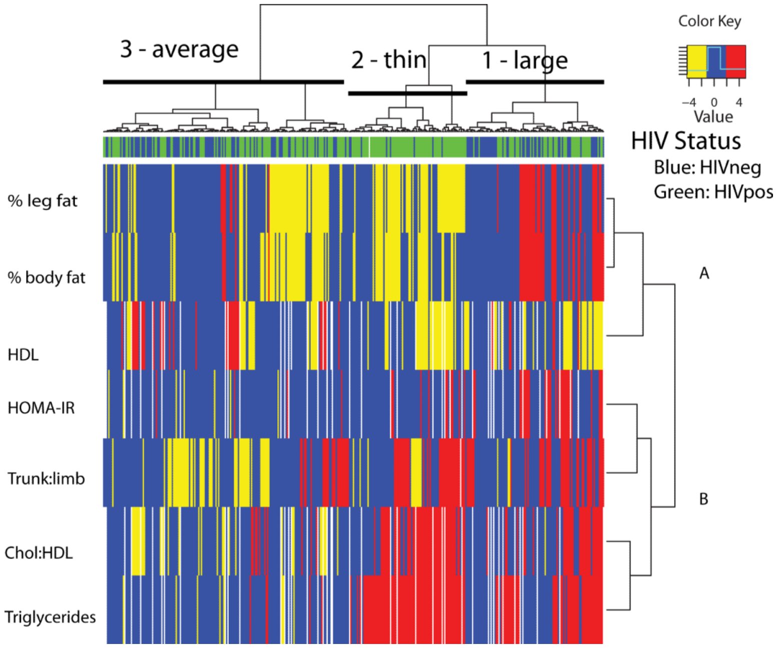
3 - average
2 - thin
1 - large
Color Key
−4
0
4
Value
HIV Status
Blue: HIVneg
Green: HIVpos
% leg fat
% body fat
HDL
HOMA-IR
Trunk:limb
Chol:HDL
Triglycerides
A
B

Figure 1.

Table 1

Cluster Phenotype and Association of External Variables with Cluster Membership

		Cluster 1-large		Cluster 2-thin		Cluster 3-average		p-value	
Cluster phenotype:									
Median z-score for outcomes included in heatmap		N	Median	N	Median	N	Median	ANOVA[1]	
Body	Percent leg fat	105	0.64	88	−1.22	186	−0.66	<0.001	
	Percent body fat	105	0.84	88	−0.89	186	−0.71	<0.001	
	Trunk:limbfat ratio	105	0.60	87	0.23	186	−0.30	<0.001	
Lipids	HDL	93	−0.70	79	−0.71	163	0.22	<0.001	
	Triglycerides	93	0.95	79	2.27	163	−0.31	<0.001	
	Cholesterol:HDL ratio	93	0.71	79	1.38	163	−0.41	<0.001	
Glucose	HOMA-IR	92	0.58	79	−0.09	167	−0.45	<0.001	
Associations of external variables with cluster membership:									
N, Percent with characteristic (%)		N	%	N	%	N	%	Univariate (Permutation)[2]	Multivariate[3]
HIV status	Positive	52	49.5	80	90.9	104	55.9	<0.001 (<0.001)	<0.001
	Negative	53	50.5	8	9.1	82	44.1		
Sex	Male	52	49.5	48	54.5	107	57.5	0.42 (0.44)	0.47
	Female	53	50.5	40	45.5	79	42.5		
Race/ethni city	Black non-Hispanic	57	54.3	41	46.6	109	58.6	0.091 (0.10)	0.033
	Hispanic	36	34.3	37	42.0	48	25.8		
	White/other	12	11.4	10	11.4	29	15.6		
Tanner stage	1–2	56	53.3	50	56.8	93	50.0	0.64 (0.65)	0.27
	3–4	25	23.8	22	25.0	57	30.6		
	5	24	22.9	16	18.2	36	19.4		
CRP	≤2 mg/l	63	60.0	67	76.1	145	78.0	<0.001 (<0.001)	<0.001
	>2 mg/l	34	32.4	12	13.6	22	11.8		
	Missing	8		9		19			

[1] p-value from analysis of variance (ANOVA) for difference in outcome variable across clusters

[2] p-value from univariate Chi-square test of cluster membership by external variable (p-value from permutation test)

[3] p-value from multivariate logistic regression model including all external variables