SUPPLEMENTAL MATERIAL

Online Supplemental File

Estimation of the optimal number of clusters (using the NbClust package in R) varied substantially by clustering algorithm (**Supplemental Figure 1**). The optimal number of clusters across all methods was on average 3.5 (median 3, IQR 2-5), with 2 being the most common (32%) followed by 3 (20%) and 4 (19%). For the k-means algorithm, the average optimal number of clusters was 3.3 (median 3, IQR 2-4).

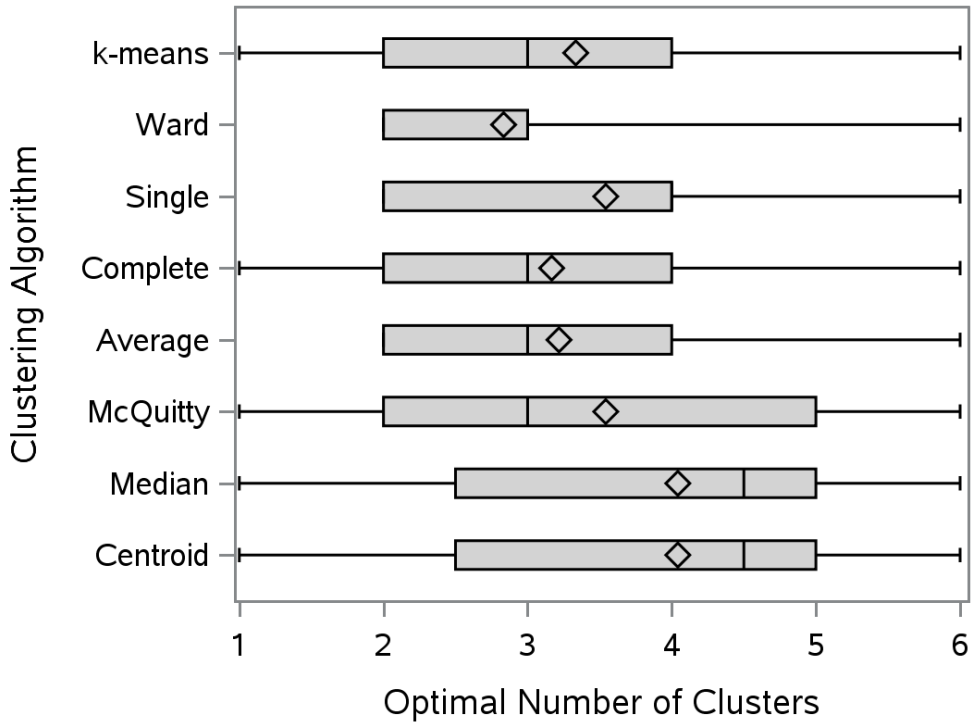
On average, internal validation measures found that connectivity and silhouette width were best for 2 clusters while the Dunn index was best for 3 clusters (**Supplemental Figure 2**), although differences between 2 and 3 clusters were not statistically significant (p=0.07 for connectivity and silhouette, p=0.35 for Dunn). Stability validation, which performs leave-one-out cross-validation by removing one record at a time, found that overlap (APN) and distance between cluster means (ADM) were on average best for 2 clusters while average distance (AD) was best for 6 clusters (**Supplemental Figure 3**). However, differences between 2 and 3 clusters were not statistically significant for APN (p=0.23) or ADM (p=0.22), while AD was better for 3 clusters vs. 2 clusters (p=0.033).

Finally, we compared model fit by testing associations of clusters generated using k-means and hierarchical algorithms (using Ward's method) with our study outcomes (**Supplemental Figure 4**). We selected the hierarchical algorithm for comparison because it showed better measures of internal validation relative to other algorithms (data not shown). Measures of generalized fit were highest for LVH and lowest for mortality, regardless of algorithm (k-means versus hierarchical) and showed modest or no improvement with increasing

numbers of clusters (**Supplemental Figure 4a**). Model discrimination was highest for mortality and lowest for DD, and showed small improvements (c=0.03 or less) as the number of clusters increased, with little difference between algorithms (c=0.01 or less) (**Supplemental Figure 4b**).

Improvements in overall fit (R^2) when adding the 3-cluster k-means variable to the model were modest (1%-3%) and did not reach statistical significance. Model discrimination improved only modestly for DD (from c=0.72 to 0.73, p=0.44), PH (from c=0.73 to 0.74, p=0.27), LVH (from c=0.79 to 0.80) and all-cause mortality (from c=0.81 to 0.82, p=0.28). The Hosmer-Lemeshow goodness of fit test was indicative of good calibration (p>0.2 for all outcomes). Reclassification improvement was minimal for LVH (IDI=0.5%) and modest for other outcomes (PH: 1.1%, DD: 2.4%, mortality: 3.4%).

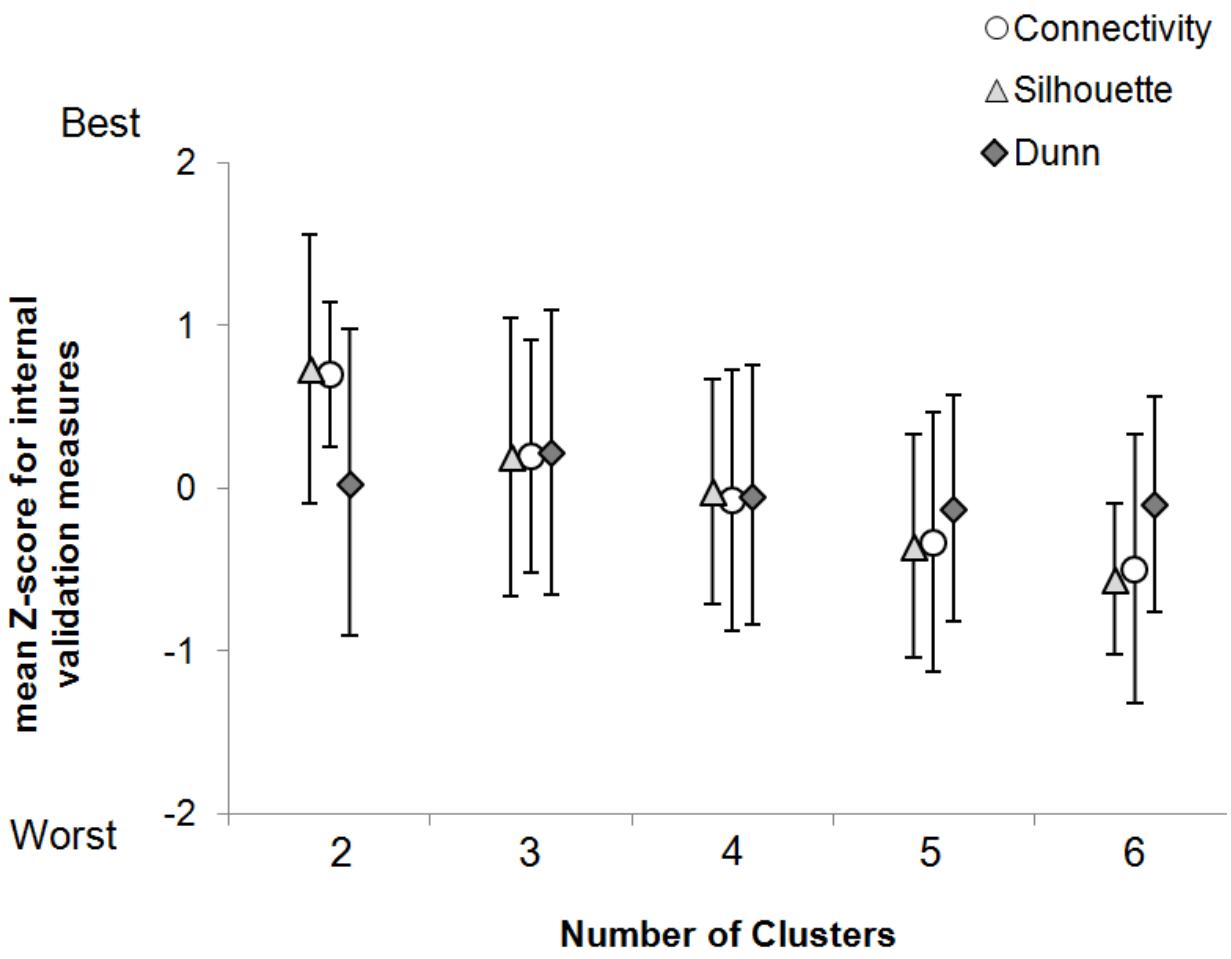
Supplemental Figure 1. Determination of optimal number of clusters



Notes: Comparisons are from NbClust package in R, which determines the optimal number of clusters for each algorithm using 30 indices.

Median is indicated by black center line, mean is indicated by diamond symbol, and the IQR (first and third quartiles) are the edges of the box. Whiskers denote minimum and maximum observations.

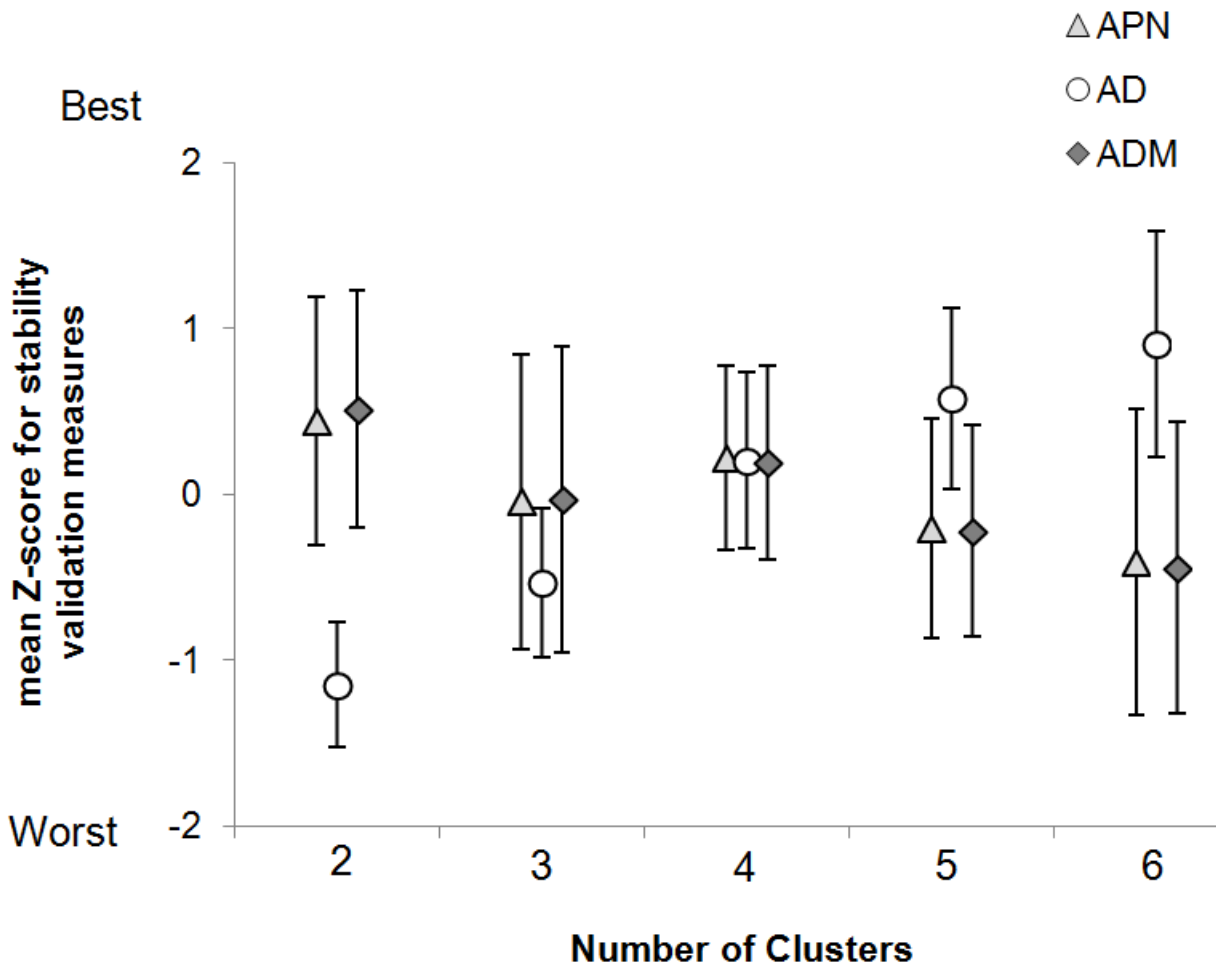
Supplemental Figure 2. Summary of internal validation measures by number of clusters



Notes: internal validation measures are standardized to the same scale (mean=0, SD=1) and reoriented (smallest = worst, largest = best) to faciliate comparison. Figure depicts mean and 95% confidence interval for each measure, averaged over clustering algorithms.

Connectivity measures the extent to which observations are placed in the same clusters as their nearest neighbors. Silhouette width and Dunn Index are measures of cluster compactness and separation.

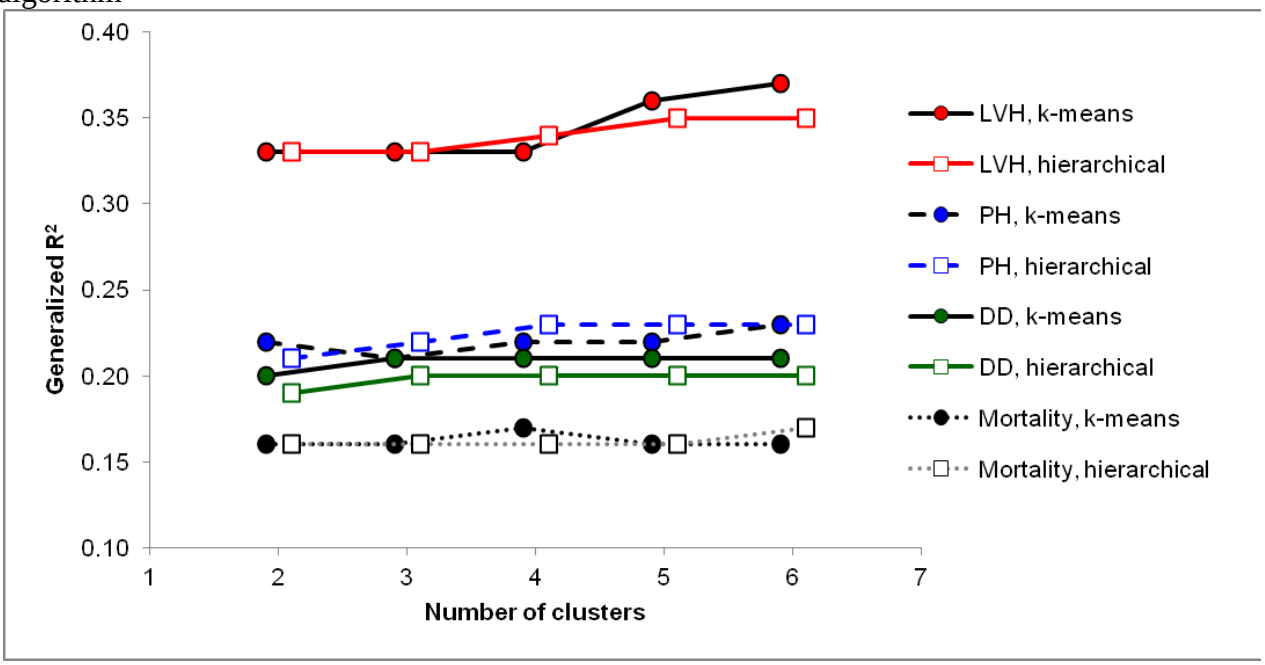
Supplemental Figure 3. Summary of stability validation measures by number of clusters

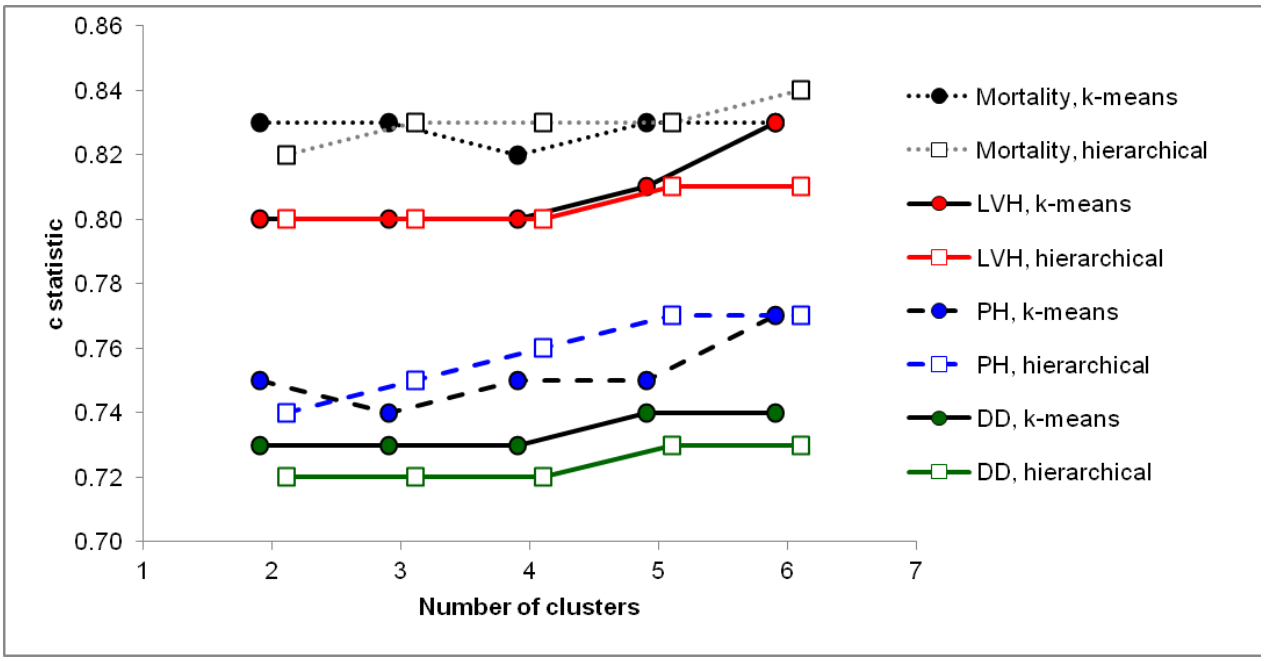


Notes: stability validation measures are standardized to the same scale (mean=0, SD=1) and reoriented (smallest = worst, largest = best) to faciliate comparison. Figure depicts mean and 95% confidence interval for each measure, averaged over clustering algorithms.

Abbreviations: Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance between Means (ADM)

Supplemental Figure 4. Comparison of (A) generalized R^2 and (B) discrimination by clustering algorithm





Abbreviations: DD = diastolic dysfunction, PH = pulmonary hypertension, LVH =left ventricular hypertrophy. Figure shows effect of adding cluster to the base model, which controls for demographics, traditional CVD risk factors, and HIV-related risk factors.Demographics include age, gender, and race/ethnicity. Traditional CVD risk factors include smoking, BMI, DM, HTN, HDL, TG, and LDL. HIV-related risk factors include CD4 count, HIVRNA, HCV, OI, and HAART use.

Supplemental Figure 5. Distributions of biomarkers, stratified by biomarker-derived phenotype