



Supplementary Figure S2: The following figure provides a detailed view of the standardization workflow employed by CANTOS using ADA002 and LTE-3 Embeddings. CANTOS standardizes by identifying the closest matching WHO and NCIt terms by computing the Euclidean distance in the embedding space. CANTOS also employs a clustering-based standardization workflow using affinity propagation and K-Means clustering. CANTOS performs a cluster size analysis for affinity propagation clusters to determine large clusters. In K-Means, the cluster size is predetermined using average silhouette scores, thus cluster size analysis is not performed. Both the clusters formed using affinity propagation, and K-Means pass through an outlier detection step prior to standardization.