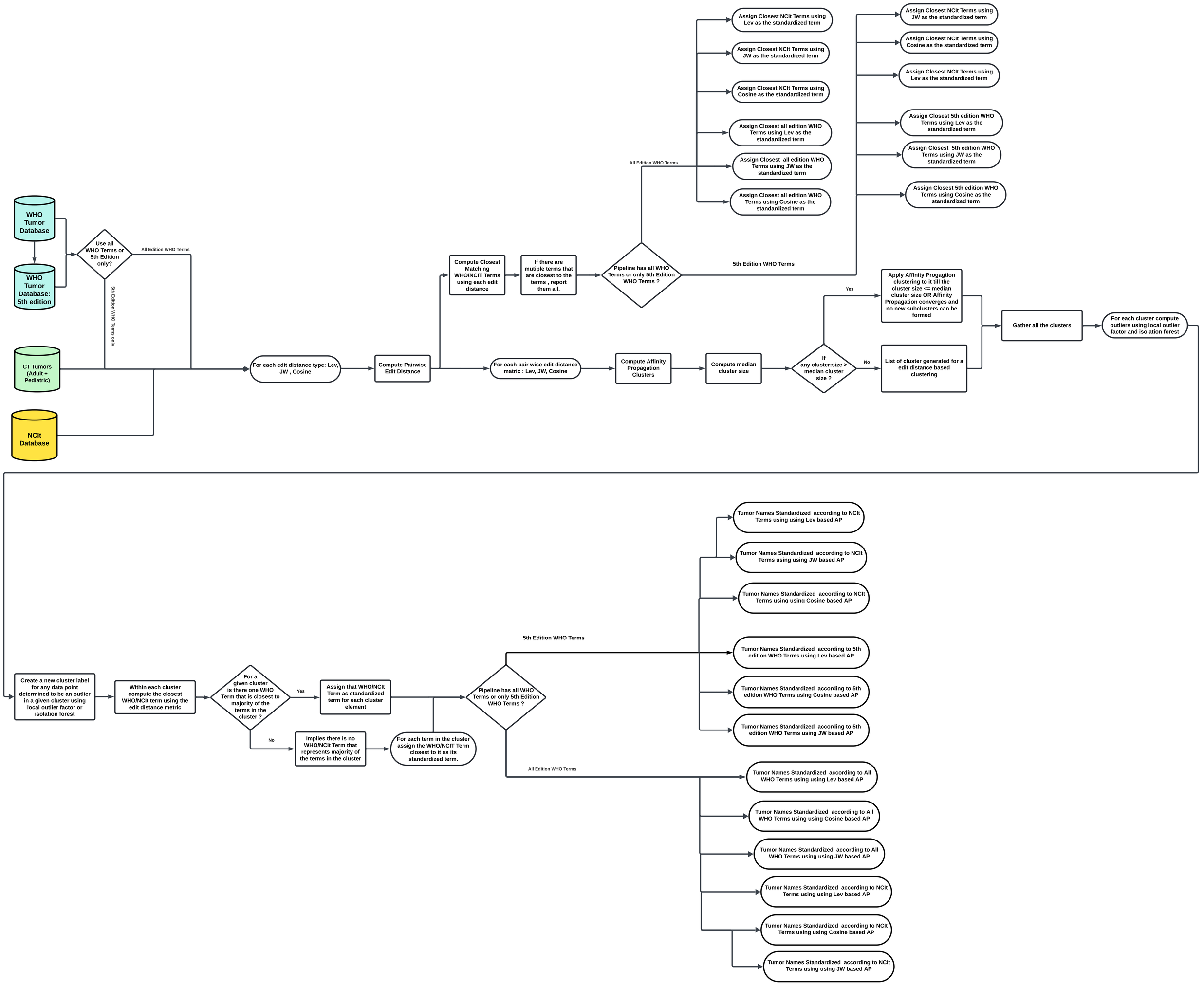

Supplementary Figure S1: The above figure provides a detailed view of the standardization workflow employed by CANTOS using edit distances. Three types of edit distances used were: Levenshtein, Jarro-Winkler, and cosine. In the first approach, CANTOS standardizes by identifying the closest matching WHO and NCIt terms using the edit distances. In another approach, CANTOS performed affinity propagation clustering on all the tumor terms (from CTR, WHO, and NCIt) and then standardized the terms within each cluster. Prior to standardization, CANTOS performs a cluster size analysis and outlier detection on each cluster.