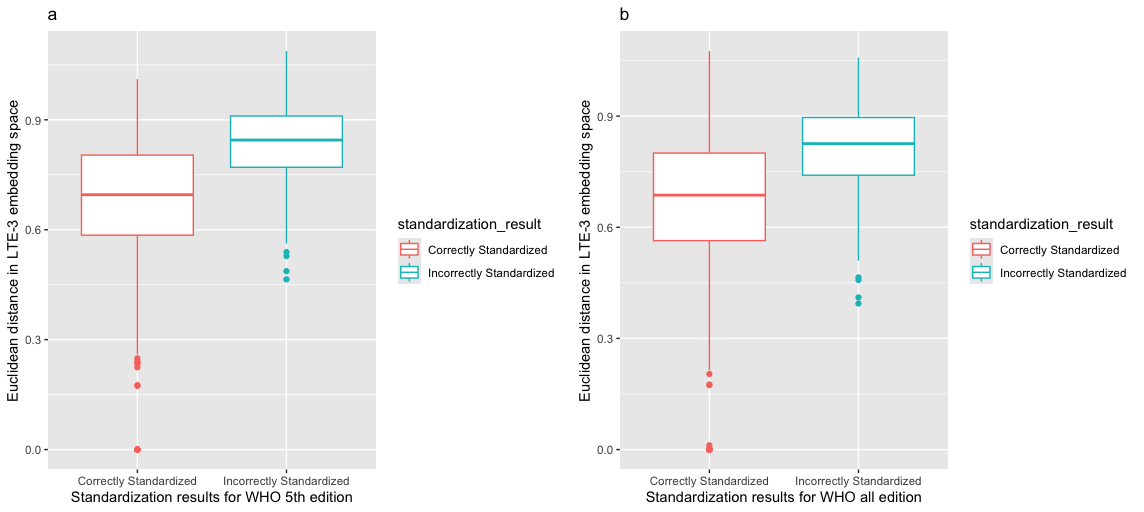
In figure 1, we visualize the standardization performance of the method LTE-3 +Euclidean Dist. Figure 1 shows the distribution of the closest Euclidean distances in the LTE-3 embedding space between the CTR condition names and the standardized tumor names from the WHO system as determined by the method LTE-3 + Euclidean Dist. It should be noted that in figure 1, out of the 1600 randomly sampled CTR terms, only the terms which had a ground truth from the WHO 5th or all edition were considered. Thus, in figure 1a, we have 1,033 CTR terms for which ground truths were identified in the WHO 5th edition and in figure 1b, we have 1,118 CTR terms or which ground truths were identified in the WHO all edition.  In both the 5th and all editions of the WHO System, the LTE-3+Euclidean Dist method correctly standardized a majority of the CTR terms and had a lower median Euclidean distance for the correctly standardized CTR terms than the incorrectly standardized CTR terms. This finding is consistent as it indicates that CTR terms which can be mapped to a standardized term from the WHO system should have a lower Euclidean distance from it in the LTE-3 embedding space. Tables 3 and 4 present the summary statistics for figures 1a and 1b. It is evident from figure 1 and tables 3 and 4, that the closest Euclidean distance in the LTE-3 embedding can distinguish between correct and incorrect standardization, however its performance needs to be improved for better standardization results.  If the median closest Euclidean distance (0.6956 for WHO 5th edition or 0.6866 for WHO all edition) for the correctly standardized term is used as a threshold for determining which terms can be reliably standardized, we will still have a few terms that will be incorrectly standardized as the minimum closest Euclidean distance for the incorrectly standardized terms (0.4646 for WHO 5th edition and 0.3942 for WHO all edition) are lower than the median Euclidean distance for the correctly standardized terms.  While the LTE-3+Euclidean Dist offers a method for standardizing the CTR, the performance can likely be improved with embeddings generated from LLMs specifically trained on tumor nomenclature.



**Figure 1:Distribution of closest Euclidean distances in the LTE-3 embedding space by the method LTE-3+Euclidean Dist**

Illustrates the distribution of the closest Euclidean distances in the LTE-3 embedding space, used by the standardization method LTE-3+Euclidean Dist, to map tumor names from the CTR to the nearest term from the WHO system. This distance is the minimum Euclidean distance between a given tumor name in the CTR and all the standardized terms in the WHO system within the LTE-3 embedding space. For a given tumor name from the CTR, the method LTE-3 + Euclidean Dist uses this distance to identify the nearest standardized term from the WHO system and then maps it to the respective tumor name in the CTR. The box plot in red shows the instances when the LTE-3 + Euclidean Dist method correctly standardized the CTR terms, whereas the blue box plot shows the instances when the method incorrectly standardized the CTR terms. **a.** displays the distribution of the closest Euclidean distances in LTE-3 embedding space when the CTR terms are standardized against the WHO 5th edition, and **b.** displays the distribution of the closest Euclidean distances in LTE-3 embedding space when the CTR terms are standardized against WHO all editions.

**Table 3:** Summary statistics for closest Euclidean distances between the CTR terms and standardized term from the WHO 5th edition as visualized for box plots in figure 1a for the method LTE-3 + Euclidean Dist. 1,033 CTR terms had a ground truth in WHO 5th edition.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| Correctly Standardized | 0 | 0.5852 | 0.6956 | 0.6657 | 0.8037 | 1.0112 |
| Incorrectly Standardized | 0.4646 | 0.7705 | 0.8448 | 0.8322 | 0.9104 | 1.0880 |

**Table 4:** Summary statistics for closest Euclidean distances between the CTR terms and standardized term from the WHO all edition as visualized for box plots in figure 1b for the method LTE-3 + Euclidean Dist. 1,118 CTR terms had a ground truth in WHO all edition

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| Correctly Standardized | 0 | 0.5640 | 0.6866 | 0.6458 | 0.8002 | 1.0759 |
| Incorrectly Standardized | 0.3942 | 0.7405 | 0.8258 | 0.8125 | 0.8962 | 1.0580 |