18 September 2024,

Dear Editors,

We wish to submit our original research article entitled "Standardization of Tumor Names in  NIH-Clinical Trials Registry using Large Language Model Embedding Analysis" for consideration by your journal. We declare that we have no competing interests, and all the authors have approved the manuscript for submission. A preprint of this manuscript is registered under the following DOI: XXXXXXXXXX.

Our work in this manuscript focused on building a computational pipeline to standardize tumor names from the NIH-Clinical Trials Registry (www.clinicaltrails.gov) so that the information from the Clinical Trials Registry (CT Registry) can be integrated with data from other biomedical databases and used for downstream analysis. The CT Registry stores the information regarding the tumors being studied in a particular trial in their conditions data file (conditions.txt). The data in this file suffers from various inconsistencies in the form of extraneous information, typographical errors, missing values, etc. Furthermore, the tumor names are not standardized according to the WHO Tumor Classification System (WHO database) or the National Cancer Institute thesaurus (NCIt database). These discrepancies pose a significant barrier to using the data in the Clinical Trials Registry to study tumors.

To this end, in our work in this paper, we have built a computational pipeline that extracts tumors from the Clinical Trials Registry and standardizes them based on the WHO and NCIt databases. Our pipeline analyzed 801,197 records in the conditions file and extracted 13,230 unique tumors, among which 6324 were identified to be pediatric tumors. The extraction results were manually validated, and literature citations were added for each tumor identified as pediatric. The pipeline then standardizes the 13,230 tumors using several methods based on text-matching (edit-distances) and text-embeddings. The text-embeddings were generated from Open AI's large language models: text-embedding-ada-002 (ADA002) and text-embedding-3-large (LTE-3). We evaluated the standardization accuracy of each method on a subset of 1600 tumors, for which we manually annotated their ground truths from the WHO database as it was not feasible to do manual annotation for all 13,230 tumors. We assessed the accuracies of each of these methods against the WHO database only, as the WHO database is considered the gold standard for tumor nomenclature. We observed that the text-embedding-based methods performed significantly better than the text-matching methods. In particular, we observed that the method that mapped the tumor from the CT registry to the nearest (euclidean distance) WHO term in the embedding space obtained the highest accuracy. We report the standardized terms for each of the 13,230 tumors identified in the CT registry with respect to both the WHO and NCIT databases and report them in the supplementary materials.

In our literature survey, we did not find any prior work focused on standardizing the data contained in the CT registry. Standardizing the tumor terms within the CT registry will allow the community to relate these tumors to other biomedical databases and perform integrative data analysis to obtain a comprehensive understanding of the therapeutic and target landscape for a given tumor. The work discussed in this manuscript utilizes the data from CT Registry and WHO and NCIT databases and implements a pipeline that implements methods that are based on text-matching (edit distances), text-embeddings generated from LLMs, and unsupervised learning methods such as clustering and principal component analysis, which is why believe our work is within the scope of this journal.

Finally, we thank the Editors for taking the time to consider and review our manuscript, and we look forward to receiving your valuable feedback.

Regards,