29 October 2024,

Dear Editors,

We wish to submit our original research article entitled "Automated Tumor Name Standardization in the NIH Clinical Trials Registry Using the CANTOS Pipeline" for consideration by your journal. We declare that we have no competing interests, and all the authors have approved the manuscript for submission.

In our pursuit of studying the landscape of therapeutic agents and drug targets for pediatric tumors, we identified the National Institutes of Health's (NIH) clinical trials registry (CTR, ClinicalTrials.gov) as a critical database for mining such insights. However, the tumor names stored in the CTR are not automatically identifiable and are often unstandardized. Thus, it creates a barrier to integrating this data with another biomedical database and drawing insights. To address this need, in this manuscript, we built a computational pipeline CANTOS (Clinical Trials Automated Nomenclature and Tumor Ontology Standardization) to identify and standardize tumor names from the CTR with respect to the WHO Tumor Classification System (WHO database) and the National Cancer Institute thesaurus (NCIt database). We elected to standardize only tumor names in the CTR as we have access to the WHO database, which is considered the gold standard for tumor nomenclature and is widely used in clinical practice. Furthermore, compared to the WHO database, there is no equivalent reference for standardizing the rest of the conditions and diseases in the CTR. Thus, we limited our focus to only tumors in this study.

To this end, our CANTOS analyzed  801,197 records in the conditions file in CTR and extracted 13,230 unique tumor names, among which it identified 6324 to be pediatric tumors. We manually validated the extraction results and added literature citations for each tumor identified as pediatric. The pipeline then standardizes the 13,230 tumors with respect to both the WHO and NCIt databases 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 tumor names, 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. Furthermore, 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 CTR 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 CTR with respect to both the WHO and NCIt databases in our work.

From our literature survey, we did not identify any prior work focused on standardizing the data contained in the CT registry. Standardizing the tumor names within the CTR will allow the community to integrate this tumor-associated data with 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 the CTR, WHO, and NCIt databases and extensively applies computational and data integration methods, which is why we believe that this work will fall within the scope of this journal and will be of interest to the journal's broad audience.

We sincerely thank the Editors for taking the time to consider our manuscript. We eagerly look forward to receiving your valuable feedback.

Regards,