# Automated Tumor Name Standardization in the NIH Clinical Trials Registry Using the CANTOS Pipeline

Aditya Lahiri[1], Sangeeta Shukla[1], Ben Stear[1], Taha Mohseni Ahooyi[1], Katherine Beigel[1], Elizabeth Margolskee[2], Deanne Taylor[1,3]

**Affiliations:**

1. The Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia PA
2. Department of Pathology & Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia PA
3. Department of Pediatrics, University of Pennsylvania Perelman Medical School, Philadelphia PA

## Big Picture

The National Institutes of Health's (NIH) clinical trials registry (CTR, ClinicalTrials.gov) is a public database that provides a wide range of information regarding a trial, such as its outcomes, interventions used, conditions studied, experimental design, etc. This information is distributed in various text files in the database, each containing the National Clinical Trial Identifier number (NCT ID) to identify and aggregate data for a given clinical trial. In pursuit of studying the landscape of therapeutic agents and drug targets associated with pediatric tumors, the CTR serves as a valuable resource for mining such insights. However, the tumor names contained in the conditions file of the CTR are not automatically identifiable from the rest of the conditions and are often unstandardized; this poses a barrier to integrating the tumor-associated information from CTR with other biomedical databases to generate insights. Thus, there is a need to develop methods to identify and standardize tumor names from CTR.


## Summary

This study aimed to identify and standardize tumor names from the NIH's clinical trials registry (CTR, ClinicalTrials.gov) using the World Health Organization (WHO) Tumor Classification System and the National Cancer Institute Thesaurus (NCIt). We developed CANTOS (Clinical Trials Automated Nomenclature and Tumor Ontology Standardization), a computational pipeline that identified tumor names from the CTR and standardized them using twelve methods based on text matching and embedding. We evaluated the accuracies of these methods only against the WHO database, as it is considered the gold standard for tumor nomenclature. In general, text-embedding-based methods outperformed text-matching-based methods with standardization accuracies ranging from 60.5%-65.6% and 63.5%-68.5% for the 5th edition and all editions (3rd-5th) of the WHO database. The LTE-3+Euclidean Dist method, which mapped tumor names to the nearest WHO term, achieved the highest standardization accuracy, with 65.6% for the 5th edition and 68.5% for all editions of the WHO database.



## Keywords

NIH Clinical Trials Registry, WHO Tumor Classification, Text Embedding, Clinical Text Standardization


## INTRODUCTION

Cancer is a significant global health issue and the second leading cause of death in the United States [1]. In 2024, over 2 million new cancer cases and 611,720 deaths are expected in the U.S. alone [2]. While pediatric cancer survival rates have improved to 80% over the last five decades [3,4], which can primarily be attributed to successes in the treatment of common childhood hematological malignancies such as acute lymphoblastic leukemia [5], unfortunately this success is not uniformly shared [6], and certain cancers, particularly those of the brain and nervous system [2], remain difficult to treat. Pediatric cancers are rarer than adult cancers and face unique challenges in clinical trials, including limited therapeutic agents, difficulty recruiting diverse populations, and tumor heterogeneity [7,8]. Therefore, to understand the therapeutic landscape of adult and pediatric tumors, it is critical to extract and analyze data from various biomedical databases, particularly from the NIH's Clinical Trials Registry (ClinicalTrials.gov, CTR) which contains data from over 482,529 research studies across all 50 states in the US and 223 countries [9]. While the CTR has established protocols and guidelines for data the submission process,

the data with respect to tumor names contained in the "conditions" data file (conditions.txt) of the CTR database contains inconsistencies in the form of extraneous information, typographical errors, missing values, non-standard nomenclature, etc, which create barriers for data integration and downstream analysis of this data. Although the CTR mandates standardized terminology like Medical Subject Headings (MeSH) [10], these terms often fail to capture the specific details of many tumor types outlined in the conditions data file. In supplementary document SD1, we provide examples to compare the differences between the terms in the condition file and the MeSH terms. Thus there is a need to standardize the tumor names in the CTR with respect to standardized nomenclature from the WHO Classification of Tumors system (referred to as the "WHO database" in the rest of the paper, https://tumourclassification.iarc.who.int/welcome/) or National Cancer Institute Thesaurus (referred to as the "NCIt database" in the rest of the paper https://ncithesaurus.nci.nih.gov/ncitbrowser/). The WHO and NCIt standardized nomenclature are available in supplementary tables ST1-ST3.

To this end, we developed a computational pipeline CANTOS (Clinical Trials Automated Nomenclature and Tumor Ontology Standardization) that standardizes tumor names in the CTR's "conditions" file using methods based on text-matching (based on edit distances) and text-embedding derived from OpenAI's Large Language Models. This pipeline maps tumor names to their standardized counterparts in the WHO and NCIt databases. Standardizing these tumor names in the CTR enables integration of this data with other biomedical databases, such as Open Targets (OT, https://www.opentargets.org/) and Illuminating the Druggable Genome (IDG, https://druggablegenome.net/) , facilitating a more comprehensive understanding of tumor biology, drug-targets, and therapeutic agents.

## RESULTS

In this paper, we developed the CANTOS pipeline to extract and standardize tumor names from the CTR according to WHO and NCIt databases. CANTOS extracted 13,230 tumors from the CTR, with 6,324 identified as pediatric, and then implemented 12 methods based on text-matching (edit-distances) and text-embedding to standardize the tumor names from the CTR. To evaluate these methods, we needed ground truth (i.e., standardized terms) for all the 13,230 tumor names. Since these ground truths were unavailable and it was not feasible to manually annotate all the CTR tumor names, we randomly sampled 1,600 tumor names from the 13,230 for ground truth annotation to assess our methods' performance.

For these 1,600 tumor names, we annotated ground truths using only the 5th edition (latest) and all editions (3rd, 4th, and 5th) of the WHO database (and not NCIt), as the WHO is the gold standard for tumor nomenclature. During annotation, we found some CTR tumor names that had multiple ground truths or none from the WHO database. Tumors without ground truth were excluded from accuracy evaluations. When standardizing using the WHO database 5th edition, we identified 567 CTR tumor names that did not have a ground truth, while 482 CTR tumor names did not have a ground truth when WHO database all editions were used. This difference arises because the WHO database all editions contain more standardized tumor terms than the 5th edition. Consequently, we evaluated accuracy for 1,033 terms when using the WHO database 5th edition and 1,118 terms for the WHO database all editions. Accuracy for each method was calculated by counting instances where a method identified at least one ground truth associated with a CTR tumor name and dividing by the total number of terms (1,033 or 1,118). We report accuracy for all editions in Table 1 and the 5th edition in Table 2.

Tables 1 and 2 indicate that text-embedding-based methods consistently outperform text-matching methods, irrespective of the edition of the WHO database used for standardization. Generally, LTE-3 embeddings performed better than ADA002, except for the LTE-3+K-means method, which slightly underperformed compared to ADA002+AP and ADA002+Euclidean Dist when using WHO database all editions (Table 1). Notably, LTE-3+K-means ranked higher than its ADA002 variant method, ADA002+K-means, for both editions of the WHO database. We attribute the superior performance of LTE-3 methods to the higher dimensionality of LTE-3, which likely captures the complexity of the input data better.

Among the edit distances, Levenshtein distance outperformed both Jarro-Winkler and cosine distances or any of their implementations using AP clustering. Jarro-Winkler is effective for minor text discrepancies and common prefixes, while cosine distance relies on word frequency ("bag of words") and ignores word order. In contrast, Levenshtein distance counts the minimum edit operations needed to transform one text into another while maintaining word order and without considering prefix similarity, which likely contributed to its superior performance. Ground truth annotations for the 1,600 CTR tumor names and their standardization results are available in supplementary tables ST4 (WHO database all editions) and ST5 (WHO 5th Edition). Standardized terms for each tumor in the CTR are reported in supplementary tables ST6-ST9.

**Discussion:**

The CTR provides information that can be integrated with external databases to gain insights into therapeutic and drug-target landscapes for tumors. However, raw tumor names from the CTR cannot be directly linked to other databases due to inconsistencies such as extraneous information, typographical errors, missing values, and non-standard nomenclature. Since tumor names are not automatically identifiable in the CTR, they also need to be extracted. To address this, we developed CANTOS, a computational pipeline to extract and standardize tumor names from the CTR according to the WHO (5th edition and all editions) and NCIt databases.

CANTOS classified tumor names from the CTR as adult or pediatric, and we manually validated each classification. For pediatric tumors, we assigned citations that confirmed the classification (Supplementary table ST10). CANTOS identified 13,230 unique tumors, of which 6,324 were pediatric tumors. It standardized these names according to the WHO and NCIt databases using 12 methods based on text-matching (edit distances) or text embedding. We evaluated these methods only against the WHO database, which is considered the gold standard for clinical tumor nomenclature. To assess accuracy, we randomly selected 1,600 tumor names from the 13,230 standardized by CANTOS and manually annotated their ground truths using the WHO 5th edition and all editions. We found more tumors could be annotated using WHO database all Editions, as the 5th edition has fewer terms. Overall, text-embedding methods proved more accurate than text-matching methods because text embeddings capture the semantic and contextual meaning between texts and map similar texts accordingly, while text-matching techniques focus only on syntactical differences between texts.

We generated text embeddings using OpenAI's LLM: text-embedding-ada-002 (ADA002) and text-embedding-3-large (LTE-3). LTE-3 methods generally outperformed ADA002, except for LTE-3+K-means, which slightly underperformed compared to ADA002+Euclidean Dist and ADA002+AP when standardized against all editions of the WHO database. Regardless of the WHO database edition, LTE-3+Euclidean Dist achieved the highest accuracy, followed by LTE-3+AP. The LTE-3+Euclidean Dist method standardizes clinical trial tumors by identifying the closest WHO term based on Euclidean distance in the LTE-3 embedding space. This method is simpler and faster than LTE-3+AP, as it does not require additional steps like clustering, cluster size analysis, or outlier detection.

**Limitations of the study**

As more CTR samples are annotated with ground truth, the performance accuracies of each method will approach their true values. While we expect accuracy changes, our annotation of 1,600 CTR tumor names suggests that text-embedding methods will likely outperform text-matching methods. Expert annotation of CTR tumor names is crucial for accurately evaluating these methods and is a limitation of our study. Additionally, as the CTR updates, new tumor names must be identified and standardized, necessitating rerunning CANTOS and manually validating the findings, which is time-consuming. Beyond expert annotation, there are computational costs associated with running the CANTOS pipeline, generating embeddings, and storing data, which can become expensive over time. Another limitation is that the embeddings generated by OpenAI may require switching to other LLMs if the models are updated or discontinued. Furthermore, OpenAI's models are not specifically trained on a medical corpus, so an LLM trained on such data would likely perform better and differentiate tumor names more precisely[11].

The CANTOS pipeline uses embeddings from LLMs that may inadvertently produce outputs reflecting biases inherent in their training data. It is important to recognize that LLM outputs are not always neutral or free from error, and results should be interpreted with caution. Importantly, the CANTOS pipeline is designed solely for research analysis purposes and should not be applied in clinical or diagnostic settings. This tool is not intended for use in contexts where decisions may directly affect human health, treatment, or care. Any conclusions drawn from the use of CANTOS in research should be further reviewed and validated by clinical professionals and subjected to rigorous peer-reviewed testing before any potential medical or therapeutic applications.

The CANTOS pipeline facilitates the extraction and standardization of tumor names from the CTR. While guidelines exist for submitting data to the CTR to ensure basic integrity, there are no enforced protocols for standardizing tumor names in the conditions file. Although studies have emphasized the need for standardization in clinical trials regarding study design [12,13] and evidence reproducibility [14], none have focused on standardizing tumors or condition names in the CTR. Standardizing tumor names will enhance their searchability in other biomedical databases, enabling researchers to quickly gather information on associated therapeutic agents, drug-targets, and clinical outcomes for specific tumors.

## Experimental procedures

This section outlines the design and methods of the CANTOS computational pipeline, which performs two main tasks: (i) identifying tumor names from the CTR and (ii) standardizing CTR tumor names. CANTOS takes condition names from CTR and standardized tumor names from the WHO and NCIt databases as inputs. After identifying tumor names, CANTOS iterates through the standardization step twice to account for different editions of the WHO database. The WHO database is considered the gold standard for tumor nomenclature and has multiple versions due to updates over the years. We used the 3rd, 4th, and 5th editions of the WHO database, which are all publicly available online. In the first iteration, CANTOS used only the latest 5th edition, while the second iteration incorporated all editions. This approach aimed to standardize tumors from clinical trials based on both the latest WHO 5th edition (referred to in this text as the WHO database 5th edition) and the combined version of all editions (referred to in this text as the WHO database all editions). Throughout these iterations, tumor names from the CTR were also standardized against the NCIt database.

### Identification of Tumor Names from CTR

The CTR provides information on various aspects of clinical trials, such as outcomes, interventions, conditions, experimental design, and study sponsors. The CTR data is distributed across multiple text files, and each of these files contains the National Clinical Trial Identification Number (NCTID). The NCTID serves as a unique identifier for each trial, which allows for referencing and aggregating related information distributed across the files. In this study, we focused on the conditions ("conditions.txt" or conditions file) and interventions ("intervention.txt" or interventions file), which detail the conditions and drugs used in each trial, respectively. The conditions file contains 801,197 records and is annotated with the fields: "*id*", "*nct_id*", "*name*" , and "*downcase_name*" . The "*id*" represents the record number within the conditions file, while the "*nct_id*" is the foreign key. The "*name*" and "*downcase_name*" fields include the names of the studied conditions, with "*downcase_name*" in lowercase format. We identified 105,483 unique conditions based on string uniqueness in the conditions file.

The conditions file does not categorize the condition names into specific subtypes or classes, consequently the tumor names contained in the file are not easily identifiable and need to be extracted. Furthermore, we are focused on extracting tumor names that are associated with a therapeutic agent (i.e., has a drug-target) registered in the CTR, so that it enables us to obtain a clear view of the therapeutic and drug-target landscape for a given tumor in future studies. To address this need , we designed the CANTOS pipeline to extract tumor names from the rest of the conditions, and then annotate each unique tumor as pediatric or adult tumors (Figure 1). The tumor name identification step implemented by CANTOS is segregated into three phases.

In Phase 1, CANTOS extracted condition names (and consequently tumor names) linked with CTR registered therapeutic agents that potentially had an associated drug-target.  It achieves this by subsetting the condition names in the conditions file using the types of drugs from the interventions file. The interventions file contains 786,898 records and includes the fields: "id", "nct_id", "intervention_type", "name", and "description". The "id" represents the record number, while the "nct_id" serves as the foreign key. The "name" and "description" fields detail the interventions, and "intervention_type" classifies them into eleven categories listed in supplementary document SD2. CANTOS extracts condition names (consequently tumors names) linked with therapeutic agents

potentially associated with drug-targets by joining the intervention file with the conditions file via "nct_id" and then filtering for the intervention types: "Drug", "Biological", "Combination Product", and "Genetic". Filtering therapeutic agents by these intervention types increases the likelihood that each agent has a corresponding drug target. Consequently, any tumor name extracted this way can potentially highlight the therapeutic and drug-target landscape associated with that tumor. After filtering, CANTOS extracted 50,410 unique condition names, which served as input for tumor identification in Phase 2 of the pipeline. The tumor extraction process is illustrated as Phase 1 in **Figure 1.**

In Phase 2, CANTOS identified tumor names from the Phase 1 extracted condition names using two independent protocols (Figure 1, “Phase 2”) .The first protocol consisted of checking if each condition name contained a tumor key word listed in supplementary document SD3. If the condition name contained a tumor keyword, that condition was annotated as a potential tumor. The second protocol of CANTOS matched each condition name in the CTR to tumor names in the WHO database 5th edition using a fuzzy string matching algorithm. If a condition name from the clinical trials exactly matched a term in the WHO database 5th edition (supplementary table ST3),, it was annotated as a tumor. If there were no exact matches, we performed a fuzzy match by calculating the generalized Levenshtein edit distance between the CTR condition name and each term in the WHO database 5th edition. This was implemented using the agrepl function in R [15], with a maximum distance set to 0.2. If a WHO database 5th edition term fell within this distance, the clinical trial condition was flagged as a potential tumor. Once flagged, we manually validated these findings to confirm that the conditions marked as potential tumors were indeed tumors. A total of 13,230 unique (by string uniqueness) tumor names were extracted from the CTR at end of Phase 2.

During Phase 3, CANTOS focused on determining which extracted tumor names were pediatric (see Figure 1, Phase 3) by applying a fuzzy string matching algorithm similar to that used in Phase 2. However, this time, the algorithm compared the tumor names against only the pediatric tumor names (supplementary table ST11), listed in the WHO database 5th edition. After CANTOS flagged the tumors as pediatric or adult, we manually validated the results. For tumors confirmed as pediatric, we added citations from peer-reviewed literature, government websites, or articles from research institutions that specify the tumor as pediatric. If the tumor was found in the WHO database, we reported it as “Listed in WHO Ped Tumor” instead of providing a literature reference. From the 50,410 conditions, 13,230 were identified as tumors, with 6,324 classified as pediatric tumors. These annotations are recorded in the supplementary file ST10.

A cursory analysis of the identified tumor names revealed various discrepancies, including typographical errors, extraneous information, missing values, drug names instead of condition names, and multiple tumor names. Furthermore, many tumor names in the disease file did not adhere to standardized nomenclature from the WHO or the NCIt databases. These unstandardized names hinder integration with other biomedical databases like IDG or OT. Additionally, they obstructed the manual annotation of tumors as pediatric or adult, leading to 144 tumors being labeled as “DA” (Do not Annotate) in the pediatric tumor designation field (“PedCanTumor”) in supplementary table ST10. Supplementary document SD4 highlights common discrepancies in tumor names. To address these issues, standardization is needed to link tumor names to external databases and gain insights into

associated drug targets and therapeutic agents. The standardization methods used by CANTOS are detailed in the following section.

## Standardization of CTR Tumor Names

We implemented various methods in CANTOS to standardize the tumor names in the CTR. These methods are based either on text-matching (edit-distances) or text-embedding. It also combines unsupervised clustering with these approaches for improved standardization. In total, CANTOS implemented 12 standardization methods, and their performance accuracies are discussed in the results section. The following subsections will detail the text-matching methods first, followed by the text-embedding techniques.

### *Text Matching Technique: Closest match using edit distance*

The aim of standardizing tumor names from the CTR to corresponding WHO or NCIt terms is to ensure each unstandardized term aligns accurately with its standardized equivalent while preserving the original meaning of the CTR tumor name. Edit distances measure the similarity between text strings by calculating the minimum number of edit operations (deletions, substitutions, insertions, etc.), q-grams, or heuristics needed to transform one string into another. The larger the edit distance between two strings, the further apart the strings are, thus, two strings with minimal edit distance could potentially convey the same meaning. CANTOS employs normalized Levenshtein distance, Jaro-Winkler distance, and cosine distance to compute edit distances. An example of using edit distances for string comparison is provided in supplementary document SD5, followed by brief descriptions of each method.

*Normalized Levenshtein distance*: Levenshtein distance measures the minimum number of single-character edits (insertions, deletions, substitutions) needed to transform one string into another. Levenshtein distance between two strings can be normalized by dividing it by the length of the longer string thus the distance would range from [0,1] and allow comparison ofmultiple strings on the same scale. We defined the normalized Levenshtein distance between two strings S1 and S2 as follows:

$$Distance_{Normalized\ Levenshtein} = \frac{Distance_{Levenshtein}}{max(|S1|,|S2|)}$$

In the above equation , |S1| and |S2| represent the respective lengths of strings S1 and S2 between which we are computing the normalized Levenshtein distance. We calculate Levenshtein distance using the stringdist library in the R programming language [16,17] . Following the calculation of the Levenshtein distance, we compute the normalizing factor (i.e. divide the Levenshtein by the longest string size) for distance between each pair of strings and normalize the Levenshtein distance.

*Jarro-Winkler distance*: The Jarro-Winkler distance is a normalized edit distance between two strings. It is a variant of the Jarro similarity measure which is defined as follows between two strings S1 and S2 respectively:

$$Sim_{jarro} = 0\ ,\ if\ m = 0$$

$$= \frac{1}{3}\left(\frac{m}{|S1|} + \frac{m}{|S2|} + \frac{m-t}{m}\right),\ otherwise$$

Where S1 and S2 are lengths of the strings S1 and S2 respectively , m is the number of matching characters and t is the number of transpositions. It should be noted that when estimating 'm', each character in S1 compared to the characters in S2, and match is counted only when the characters are the same and if the characters are within a certain distance of each other typically defined as half the length of the longer string, rounded down, minus one, i.e. $\frac{max(|S1|, |S2|)}{2} - 1$.

The Jarro-Winkler similarity measure builds on top of the Jarro similarity measure and introduces two more parameters for rewards and favorable scales the Jarro similarity score if the two strings share similar prefixes. The Jarro-Winkler similarity is defined as follows:

$$Sim_{jarro-winkler} = Sim_{jarro} + lp\,(1 - Sim_{jarro})$$

Where *l* is defined as the length of the common prefix at the start of the string (maximum of 4 characters) , whereas *p* is a scaling factor that rewards the score for having common prefixes. Typically *p* is set to 0.1 and should not exceed 0.25 (or ¼ as the maximum length of prefix being considered is 4). With the above definition of Jarro-Winkler similarity, the Jarro-Winkler distance is defined as follows:

$$Distance_{jarro-winkler} = 1 - Sim_{jarro-winkler}$$

We calculate the Jarro-Winkler distance using the stringdist package in the R-programming language.

*Cosine Distance*: To define cosine distance , we first need to establish the concept of cosine similarity. For two non-zero vector vectors, cosine similarity is defined as the dot product of the two vectors divided by the product of their lengths. Cosine similarity ranges from [-1,1], with -1 representing total opposition, 0 representing complete dissimilarity, and 1 representing full similarity between the vectors. Cosine similarity between two vectors A and B is defined as follows:

$$Sim_{cosine} = cos(\theta) = \frac{A \,.\, B}{|A|\,|B|}$$

However, to use cosine similarity in the context of strings, the vectors A and B represent the frequencies of unique words in strings S1 and S2. Since frequencies cannot be negative, the cosine similarity ranges between [0,1]. Thus, there is no need to normalize this metric, and cosine distance is defined simply as

$$Distance_{cosine} = 1 - Sim_{cosine}$$

We calculate the cosine distance using the stringdist package in R. Based on the three edit distances, CANTOS computed the pairwise distances between each tumor name identified in the CTR and the standardized tumor terms with respect to the WHO (5th edition and all editions) and NCIT database. For each CTR tumor name, CANTOS selects the nearest standardized terms under each edit distance. If more than one term qualified as the closest term, CANTOS reported them all by separating individual terms with a semicolon.

**Text Matching Technique: Edit Distance combined with Affinity Propagation Clustering**

CANTOS implements another set of standardization methods based on edit distances and clustering techniques. These methods consist of a clustering step followed by a mapping step. CANTOS applied affinity propagation (AP) clustering to form the clusters, where the divergence matrix is computed by calculating the pairwise edit distance between the tumor names in CTR, WHO, and NCIt databases. CANTOS computes three divergence matrices using normalized Levenshtein distance, Jarro-Winkler distance, and cosine distance. Using each of these divergence matrices, CANTOS runs the AP clustering and forms three sets of clusters. We selected the AP algorithm for clustering as it automatically determines the number of clusters instead of requiring the number of clusters (unknown to us) to be a user-defined hyperparameter. AP is also not dependent on the initialization conditions and is deterministic [18]. AP works by recursively passing real-valued messages between each data point until they converge, and based on these converged values, the algorithm establishes the clusters and assigns each cluster an "exemplar data point" which serves as an ideal representative of that cluster [19]. Furthermore, AP clustering methods have shown success in clustering textual data [20,21]. Once the clusters were computed using AP, CANTOS performed a cluster size analysis to check if any clusters were large and may contain members that should not belong together. This was done by determining the median cluster size, and clusters larger than the median cluster size were identified and designated as large clusters. On each of these large clusters, CANTOS performs nested AP clustering until their sizes drop below the previously determined median cluster size or if the AP clustering algorithm converges and no more new clusters can be performed.

CANTOS then detects outliers in each cluster using isolation forest and local outlier factors (LOF) algorithms. Outliers identified by either method are removed from their clusters and reassigned as single-member clusters. CANTOS implements isolation forest using the R isolation.forest package [22], where we set the hyperparameter number of trees (ntrees) to 100 as recommended by Lie et al.2008 [23] in their original introduction of the isolation forest algorithm and the dims argument to 3, as suggested for numeric datasets in the package documentation[22]. The standardized outlier scores are calculated for each data point within a cluster, a score close to 1 indicates the data point is a likely outlier, while a score close to 0 indicates the data point is likely a member of the cluster. The CANTOS pipeline uses an outlier score of 0.5 as the threshold and any data point with an outlier score greater than 0.5 is deemed as an outlier. Similarly, CANTOS uses the lof function within the dbscan package [24] in R to compute the LOF values to determine outliers in each cluster. The LOF value for a data point *p* is defined as the local reachability density *p* and the local reachability density of “minPts”-nearest neighbors of *p* [25–27]. The hyperparameter “minPts” specifies the minimum number of nearest data points around *p* that need to be considered for calculating LOF value. An LOF value close to 1 indicates that the data point *p* is in a region with a relatively uniform density, whereas a LOF > 1 indicates the data point *p* has a lower density than its neighbors and is likely an outlier [27]. To compute the LOF values for each cluster member in each cluster, the CANTOS pipeline requires the users to define the minPts parameter. We set minPts to be integers ranging from 2 (clusters need to have more than one element to have an outlier) to $Size(cluster) - 1$. Iterating through each value of minPts, CANTOS then computes LOF values for each cluster element and then computes their median LOF value. If the median LOF value is above 1 for a data point *p* in a cluster, then it is designated as an outlier.

Upon completing the outlier analysis, CANTOS implements the mapping step, in which each cluster member is mapped to a standardized term. To achieve this, CANTOS iterates through each cluster and identifies the closest standardized term from the WHO and NCIt databases based on the edit distance implemented in the pipeline. If a majority of the tumor names in a cluster are close to a specific standardized term, then all the tumor names are mapped to that standardized term. In case there is a tie, where two or more standardized terms are equally represented in a cluster, then each tumor name within that cluster is assigned to its closest (based on the edit distance used so far) matching standardized tumor name. The text-based matching pipeline is described in detail in supplementary figure SF1 and the summarized version is shown in figure 2. The following section will discuss the pipeline that standardizes the clinical trial tumors based on text embeddings.

***Text Embedding Analysis: Closest match in Embedding Space***

The methods in the previous section employed edit distances to compare texts. These methods primarily focus on syntactical differences to quantify the differences between texts. In this section, the standardization methods deployed by CANTOS are based on text embeddings (or word embeddings), which can also be used for comparing texts. Text embeddings are low dimensional numeric vector representations of unstructured text data. Unlike edit distances, text embeddings focus on capturing the semantic and contextual meaning of the input text they encode; consequently, in the embedding vector space, texts with similar meanings should have embeddings close to each other and texts which differ in meaning should be further apart [28–31]. Text-embeddings have been used in various applications such as developing search engines [32,33], text clustering [34] and classification [35], recommender systems [36], and anomaly detection [37]. Text-embeddings can be generated by natural language processing models such as Word2Vec, GloVE, FastText or through large language models (LLM) such as BERT, GPT, ELMO [31]. In this paper, we generated text embeddings using two OpenAI models: text-embedding-ada-002 (ADA002) and text-embedding-3-large (LTE-3) [38]. LTE-3, the newer model, produces 3072-dimensional embeddings, while ADA002 generates 1536-dimensional ones. We used both models to create embeddings for tumor names in the CTR, WHO, and NCIt databases. Utilizing these embeddings, CANTOS standardizes the tumor names in CTR by calculating Euclidean distances in the embedding space (LTE-3 or ADA002) between each tumor name in the CTR and standardized terms in the WHO and NCIT databases and mapping each tumor name to its nearest standardized term. CANTOS applied this method with both types of embedding to standardize tumor names according to the WHO 5th edition, all WHO editions, and the NCIT database.

## Text Embedding Analysis: Embeddings and Clustering

Similar to CANTOS's approach of using edit distances to compute divergence matrices for AP clustering, in this approach, CANTOS calculated the pairwise Euclidean distances in the embedding space (LTE-3 or ADA002) among tumor names in the CTR, WHO and NCIt databases. However, this method is computationally expensive due to the high dimensionality of the embedding space: ADA002 (1536) and LTE-3 (3072). To address this issue, CANTOS applied principal component analysis (PCA) to each set of embeddings (ADA002 and LTE-3) and based on the iteration of pipeline that depended on the edition of WHO (5th edition or all editions) database used in the pipeline, generated four sets of PCA-transformed embeddings. The dimensions of the four PCA-transformed embedding spaces are

summarized in Table 3. For each case, CANTOS retained the minimum number of principal components necessary to explain 80% of the variance in the data.

With the four sets of PCA-transformed embeddings, CANTOS computed the divergence matrices for AP clustering using pairwise Euclidean distance. After clustering, CANTOS performs cluster size analysis by computing the z-scores for each cluster based on the number of cluster members. Using the default z-score of 2.5 as a threshold, CANTOS determined the maximum cluster size and designated any cluster with a z-score greater than 2.5 as a large cluster. For each large cluster, CANTOS performed nested AP clustering until their sizes were below or equal to the maximum number of cluster members or the AP clustering algorithm converged. The use of z-score for determining large clusters differs from the edit distance-based AP clustering method in the previous section, where the median cluster size was used as the threshold for determining the large clusters in CANTOS. The median cluster size for embedding-based AP clustering was lower than the maximum cluster size established by z-score-based clustering. This consequently led to more clusters with relatively homogeneous cluster members being labeled as large clusters and flagged for nested clustering. The maximum cluster size determined using the z-score was larger, and fewer such clusters were flagged as large, hence it was selected as the threshold for deciding large clusters in CANTOS. It should be noted that CTR has a diverse range of tumor names; thus, clustering these tumor names will produce clusters of nonuniform cluster size. Hence, estimating a maximum cluster size is not trivial when considering the variation in the entire dataset. With more curation and classification of the tumors within the CTR, one can get a more accurate estimate of a reasonable cluster size for each tumor type.

After clustering, CANTOS performed outlier detection as in the previous section using isolation forest and LOF analysis. The hyperparameters for isolation forest and LOF analysis were kept the same as they were for edit-distance-based AP clustering. Following the outlier detection step, CANTOS iterates through each cluster and determines the standardized term from the WHO (5th edition and all editions) and the NCIt databases that are closest to each cluster member by computing the Euclidean distance in the embedding space (not the PCA-transformed space). If a majority of the tumor names in a cluster are close to a specific standardized term, then all the tumor names are mapped to that standardized term. In case there is a tie, where two or more standardized terms are equally represented in a cluster, then each tumor name within that cluster is assigned to its closest matching standardized term. Table 4 compares the number of AP clusters formed when we use text-embeddings and text-matching (edit distances).

In addition to AP clustering, CANTOS implemented K-means clustering on PCA transformed embeddings to standardize the tumor names in CTR. Unlike AP, K-means requires users to define the number of clusters, "K," as a hyperparameter[39]. Without prior information on the types of tumors present in CTR, it is difficult to estimate "K" qualitatively and thus we used a computational method to decide on an optimal value for “K”. To determine “K”, we implemented a commonly used cluster performance metric known as the silhouette coefficient [40] within CANTOS. For a given value of “K”, when K-means clustering is completed, the silhouette coefficient is computed for each data point. The silhouette

coefficient has a range of [-1,1]. A silhouette coefficient of 1 signifies that the data point is well-matched to other elements in its own cluster and poorly matched to members of neighboring clusters [41]. A silhouette coefficient of 0 indicates that the data point is at the decision boundary of neighboring clusters and a score of -1 indicates that the data point is poorly matched with other cluster members and likely assigned to an incorrect cluster [41]. Intuitively, a higher silhouette coefficient for a data point represents high cohesion of that data point with rest of the cluster members and high separation from members of neighboring clusters. CANTOS iterates through several values of "K" and performs K-means clustering and calculates the silhouette coefficient for each data point. For each cluster, CANTOS calculates the average silhouette coefficient which indicates the performance for that cluster. CANTOS then computed the average of each of the averaged silhouette coefficients per cluster, this produces a metric that can be used to evaluate the overall clustering performance of the K-means algorithm for a given value of "K". We refer to this metric as the mean silhouette score which ranges from [-1,1] and a higher positive value indicates better clustering performance. Using this method, CANTOS evaluated the mean silhouette score for various values of "K" under both ADA002 and LTE-3 embeddings for each iteration (WHO database 5th edition and WHO database all editions) . In Figure 3, we plot the mean silhouette score for each of these cases.

After K-means clustering is completed, CANTOS follows the same steps for outlier detection (isolation forest and LOF analysis) and standardization (compute euclidean distances with standardized term and identify the closest match) as it did for embedding based AP clustering, while maintaining the exact hyperparameter configurations. A detailed view of the text-embedding based standardization pipeline is displayed in supplementary figure SF2, and a brief summary is presented in figure 2.

## Conclusion:

The CTR records various aspects of clinical trials, including conditions studied, and shares this information publicly in text file format. However, it lacks a mechanism for identifying tumor names, which are often unstandardized, making it challenging to integrate them with other biomedical databases for downstream analysis. In this paper, we present CANTOS, a computational pipeline that extracts tumor names from the condition file in CTR and classifies them as adult or pediatric tumors. It then standardizes the extracted tumor names with respect to the WHO and NCIt databases using 12 standardization methods based on text-matching and text-embedding. Our findings show that embedding-based methods outperform text-matching methods, with the LTE-3+Euclidean Distance method achieving the highest accuracy. We standardized all 13,230 tumors in the CTR against the WHO and NCIt databases and reported the results.

## Resource availability

**Data and Code Availability**. Users can download the data used in this paper from the CTR website (https://clinicaltrials.gov/) [9], Clinical Trials API or from the Aggregate Analysis of ClinicalTrials.gov-Clinical Trials Transformative Initiative (AACT-CTTI) website (https://aact.ctti-clinicaltrials.org/download). We downloaded a copy of the database from the

ACCT-CTTI website on August 22, 2023 which is available as a zip file titled “20230822_export.zip” under the section titled “Monthly Archive of Static Copies”. The AACT-CTTI website is updated daily with content from ClinicalTrials.gov and provides a static database at the beginning of each month. This database includes information on all registered studies in the CTR, with details about clinical trials, such as experimental design, conditions, and interventions, available in separate pipe-delimited text files within the zip file. The process for downloading the dataset used in this study is outlined in supplementary document SD6. The code files for building and running CANTOS, along with the instructions for downloading the embeddings generated using Open AI are publicly available at our GitHub repository: https://github.com/TaylorResearchLab/CANTOS.

**Lead contact.** Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Deanne Taylor (taylordm@chop.edu).

**Materials availability.** The WHO and NCIt standardized tumor names generated by CANTOS are available in supplementary tables ST6-ST9. The code for implementation of the CANTOS pipeline is available in our public GitHub repository: https://github.com/TaylorResearchLab/CANTOS.

**Acknowledgements**
We thank Dr. Susan Furth and CHOP’s Department of Biomedical and Health Informatics for their ongoing support, and Dr. Sarah Tasian for her guidance on cancer nomenclature review.

**Author Contributions**
AL and DT conceived and designed the study. AL wrote the code. AL,SS, BS, TM performed the analyses. AL wrote the manuscript. DT, EM, and KB made critical revisions to the manuscript. All authors interpreted the content, made revisions to the manuscript, and had final approval of the completed version. DT oversaw the project. DT acquired funding for the project.

**Declaration of Interests**
The authors declare no competing interests.

**Declaration of generative AI and AI-assisted technologies in the writing process**
During the preparation of this work the authors used ChatGPT 4.0 in order to improve clarity and conciseness of the written text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

**References**

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., and Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *74*, 229–263.

2. Siegel, R.L., Giaquinto, A.N., and Jemal, A. (2024). Cancer statistics, 2024. CA Cancer J. Clin. *74*, 12–49.

3. Matt, G.Y., Sioson, E., Shelton, K., Wang, J., Lu, C., Zaldivar Peraza, A., Gangwani, K., Paul,

R., Reilly, C., Acić, A., et al. (2024). St. Jude Survivorship Portal: Sharing and Analyzing Large Clinical and Genomic Datasets from Pediatric Cancer Survivors. Cancer Discov. *14*, 1403–1417.

4. Aristizabal, P., Winestone, L.E., Umaretiya, P., and Bona, K. (2021). Disparities in Pediatric Oncology: The 21st Century Opportunity to Improve Outcomes for Children and Adolescents With Cancer. Am Soc Clin Oncol Educ Book *41*, e315–e326.

5. Hunger Stephen P., and Mullighan Charles G. Acute Lymphoblastic Leukemia in Children. N. Engl. J. Med. *373*, 1541–1552.

6. Laetsch, T.W., DuBois, S.G., Bender, J.G., Macy, M.E., and Moreno, L. (2021). Opportunities and Challenges in Drug Development for Pediatric Cancers. Cancer Discov. *11*, 545–559.

7. Renfro, L.A., Ji, L., Piao, J., Onar-Thomas, A., Kairalla, J.A., and Alonzo, T.A. (2019). Trial Design Challenges and Approaches for Precision Oncology in Rare Tumors: Experiences of the Children's Oncology Group. JCO Precis Oncol *3*. https://doi.org/10.1200/PO.19.00060.

8. Rivers, Z., Hyde, B., Ronski, K., Stearns, D., Toll, S., Ritt, K., Cooney, M., Nimeiri, H., Federman, N., and Kaneva, K. (2023). Exploring Barriers to Pediatric Cancer Clinical Trials: The Role of a Networked, Just-in-Time Study Program. Clin. Ther. *45*, 1148–1150.

9. National Institutes of Health Clinical Trials Registry ClinicalTrials.gov. https://clinicaltrials.gov/.

10. Zarin, D.A., Tse, T., Williams, R.J., Califf, R.M., and Ide, N.C. (2011). The ClinicalTrials.gov Results Database — Update and Key Issues. N. Engl. J. Med. *364*, 852–860.

11. Stanford CRFM https://crfm.stanford.edu/2022/12/15/biomedlm.html.

12. Canonica, G.W., Baena-Cagnani, C.E., Bousquet, J., Bousquet, P.J., Lockey, R.F., Malling, H.-J., Passalacqua, G., Potter, P., and Valovirta, E. (2007). Recommendations for standardization of clinical trials with Allergen Specific Immunotherapy for respiratory allergy. A statement of a World Allergy Organization (WAO) taskforce. Allergy *62*, 317–324.

13. Katz, M.H.G., Marsh, R., Herman, J.M., Shi, Q., Collison, E., Venook, A.P., Kindler, H.L., Alberts, S.R., Philip, P., Lowy, A.M., et al. (2013). Borderline resectable pancreatic cancer: need for standardization and methods for optimal clinical trial design. Ann. Surg. Oncol. *20*, 2787–2795.

14. Dickersin, K., and Mayo-Wilson, E. (2018). Standards for design and measurement would make clinical research reproducible and usable. Proc. Natl. Acad. Sci. U. S. A. *115*, 2590–2594.

15. Snášel, V., Keprt, A., Abraham, A., and Hassanien, A.E. (2009). Approximate String Matching by Fuzzy Automata. In Man-Machine Interactions (Springer Berlin Heidelberg), pp. 281–290.

16. van der Loo, M. (2013). Stringdist: Approximate string matching, fuzzy text search, and string distance functions. (The R Foundation). https://doi.org/10.32614/cran.package.stringdist https://doi.org/10.32614/cran.package.stringdist.

17. R Core Team (2024) *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing,* Vienna, Austria. https://www.R-project.org/.

18. Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. Science *315*, 972–976.

19. Kitahara, Y.F.G.I. Fast Algorithm for Affinity Propagation. In Twenty-Second International Joint Conference on Artificial Intelligence, T. Walsh, ed. (IJCAI/AAAI), pp. 2238–2243.

20. Shi, X.H., Guan, R.C., Wang, L.P., Pei, Z.L., and Liang, Y.C. (2009). An incremental affinity propagation algorithm and its applications for text clustering. In 2009 International Joint Conference on Neural Networks (IEEE). https://doi.org/10.1109/ijcnn.2009.5178973.

21. Shailendra Kumar Shrivastava, J.L.Rana, R.C.Jain (2013). Text document clustering based on phrase similarity using affinity propagation. International Journal of Computer Applications *61*. https://doi.org/10.5120/10032-5077.

22. Cortes, D. (2019). isotree: Isolation-Based Outlier Detection. (The R Foundation). https://doi.org/10.32614/cran.package.isotree https://doi.org/10.32614/cran.package.isotree.

23. Liu, F.T., Ting, K.M., and Zhou, Z.-H. (2008). Isolation Forest. In 2008 Eighth IEEE International Conference on Data Mining (IEEE), pp. 413–422.

24. Hahsler, M., and Piekenbrock, M. (2015). Dbscan: Density-based spatial clustering of applications with noise (DBSCAN) and related algorithms. (The R Foundation). https://doi.org/10.32614/cran.package.dbscan https://doi.org/10.32614/cran.package.dbscan.

25. Alghushairy, O., Alsini, R., Soule, T., and Ma, X. (2020). A review of Local Outlier Factor algorithms for outlier detection in big data streams. Big Data Cogn. Comput. *5*, 1.

26. Ding, H., Ding, K., Zhang, J., Wang, Y., Gao, L., Li, Y., Chen, F., Shao, Z., and Lai, W. (2018). Local outlier factor-based fault detection and evaluation of photovoltaic system. Solar Energy *164*, 139–148.

27. Xu, H., Zhang, L., Li, P., and Zhu, F. (2022). Outlier detection algorithm based on k-nearest neighbors-local outlier factor. J. Algorithm. Comput. Technol. *16*. https://doi.org/10.1177/17483026221078111.

28. Morris, J., Kuleshov, V., Shmatikov, V., and Rush, A. (2023). Text embeddings reveal (almost) as much as text. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics). https://doi.org/10.18653/v1/2023.emnlp-main.765.

29. Mikolov, T. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

30. Incitti, F., Urli, F., and Snidaro, L. (2023). Beyond word embeddings: A survey. Inf. Fusion *89*, 418–436.

31. Khattak, F.K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., and Rudzicz, F. (2019). A survey of word embeddings for clinical text. J. Biomed. Inform. *100S*, 100057.

32. Gökçe, O., Prada, J., Nikolov, N.I., Gu, N., and Hahnloser, R.H.R. (2020). Embedding-based scientific literature discovery in a text editor application. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (Association for Computational Linguistics). https://doi.org/10.18653/v1/2020.acl-demos.36.

33. Mai, G., Janowicz, K., and Yan, B. (2018). Combining text embedding and knowledge graph embedding techniques for academic search engines. In Joint proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering over Linked Data challenge (QALD-9) co-located with 17th International Semantic Web Conference (ISWC 2018) CEUR Workshop Proceedings., Key-Sun Choi, Luis Espinosa Anke, Thierry Declerck, Dagmar Gromann, Jin-Dong Kim, Axel-Cyrille Ngonga Ngomo, Muhammad Saleem, Ricardo Usbeck, ed. (CEUR), pp. 77–88.

34. Mehta, V., Bawa, S., and Singh, J. (2021). WEClustering: word embeddings based text clustering technique for large datasets. Complex Intell Systems *7*, 3211–3224.

35. Stein, R.A., Jaques, P.A., and Valiati, J.F. (2019). An analysis of hierarchical text classification using word embeddings. Inf. Sci. *471*, 216–232.

36. Musto, C., Semeraro, G., de Gemmis, M., and Lops, P. (2016). Learning word embeddings from Wikipedia for content-based recommender systems. In Lecture Notes in Computer Science Lecture notes in computer science. (Springer International Publishing), pp. 729–734.

37. Pande, A., and Ahuja, V. (2017). WEAC: Word embeddings for anomaly classification from event logs. In 2017 IEEE International Conference on Big Data (Big Data) (IEEE), pp. 1095–1100.

38. New embedding models and API updates https://openai.com/index/new-embedding-models-and-api-updates/.

39. Wu, J. (2012). Cluster Analysis and K-means Clustering: An Introduction. In Advances in K-means Clustering: A Data Mining Thinking, J. Wu, ed. (Springer Berlin Heidelberg), pp. 1–16.

40. Shahapure, K.R., and Nicholas, C. (2020). Cluster Quality Analysis Using Silhouette Score. In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) (IEEE), pp. 747–748.

41. Shutaywi, M., and Kachouie, N.N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. Entropy *23*. https://doi.org/10.3390/e23060759.

## TABLES

**Table 1: CANTOS Accuracies for Standardization Methods when WHO database all editions are used.**

| Ranking | Basis | Methods | Accuracy WHO All Editions |
|---|---|---|---|
| 1 | Embedding | LTE-3 + Euclidean Dist | 0.6851521 |
| 2 | Embedding | LTE-3 + AP | 0.6708408 |
| 3 | Embedding | ADA002 + Euclidean Dist | 0.6618962 |
| 4 | Embedding | ADA002 + AP | 0.6466905 |
| 5 | Embedding | LTE-3 + K-means | 0.6449016 |
| 6 | Embedding | ADA002 + K-means | 0.6359571 |
| 7 | Text Match | Levenshtein | 0.3246869 |
| 8 | Text Match | Levenshtein + AP | 0.2924866 |
| 9 | Text Match | Jarro Winkler | 0.2549195 |
| 10 | Text Match | Jarro Winkler + AP | 0.244186 |
| 11 | Text Match | Cosine | 0.2388193 |
| 12 | Text Match | Cosine + AP | 0.2271914 |

**Table 2: CANTOS Accuracies for Standardization Methods when the WHO database 5th edition was used.**

| Ranking | Basis | Methods | Accuracy WHO 5th Edition |
|---|---|---|---|
| 1 | Embedding | LTE-3 + Euclidean Dist | 0.6563408 |
| 2 | Embedding | LTE-3 + AP | 0.6456922 |
| 3 | Embedding | LTE-3 + K-means | 0.6360116 |
| 4 | Embedding | ADA002 + Euclidean Dist | 0.6292352 |

| | | | |
|---|---|---|---|
| 5 | Embedding | ADA002 + AP | 0.6263311 |
| 6 | Embedding | ADA002 + K-means | 0.6050339 |
| 7 | Text Match | Levenshtein | 0.3059051 |
| 8 | Text Match | Levenshtein + AP | 0.286544 |
| 9 | Text Match | Jarro Winkler | 0.2342691 |
| 10 | Text Match | Jarro Winkler + AP | 0.232333 |
| 11 | Text Match | Cosine + AP | 0.2197483 |
| 12 | Text Match | Cosine | 0.2178122 |

**Table 3: PCA dimensions used for each embedding method based on WHO database editions**

| **Tumor Terms** | **Dimensions for PCA+ADA002** | **Dimensions for PCA+LTE-3** |
|---|---|---|
| CT + NCIT + WHO 5th Edition | 136 | 178 |
| CT + NCIT + WHO All Edition | 141 | 185 |

**Table 4: Number of clusters from AP clustering under embedding and text-matching based methods**.

| **Basis** | **Affinity Propagation Clustering Divergence Metric** | **Number of Clusters for CT Terms, NCIt Terms, WHO database all editions** | **Number of Clusters for CT Terms, NCIt Terms, WHO database 5th edition** |
|---|---|---|---|
| Text Match | Cosine | 1040 | 967 |
| Text Match | Levenshtein | 2020 | 1808 |
| Text Match | Jarro Winkler | 1965 | 1785 |
| Embedding | ADA002 + Euclidean Dist | 3790 | 3456 |
| Embedding | LTE-3 + Euclidean Dist | 3894 | 3427 |

**FIGURES**

**Figure 1: Tumor Annotation Process in the CANTOS Pipeline: The annotation workflow within the CANTOS pipeline, specifically for categorizing tumor names extracted from the NIH Clinical Trials Registry. The process involves filtering extracted tumor names, matching them to the WHO database, and categorizing them as adult or pediatric tumors. The conditions flagged as tumors were manually reviewed and citations were added for the pediatric tumor types.**

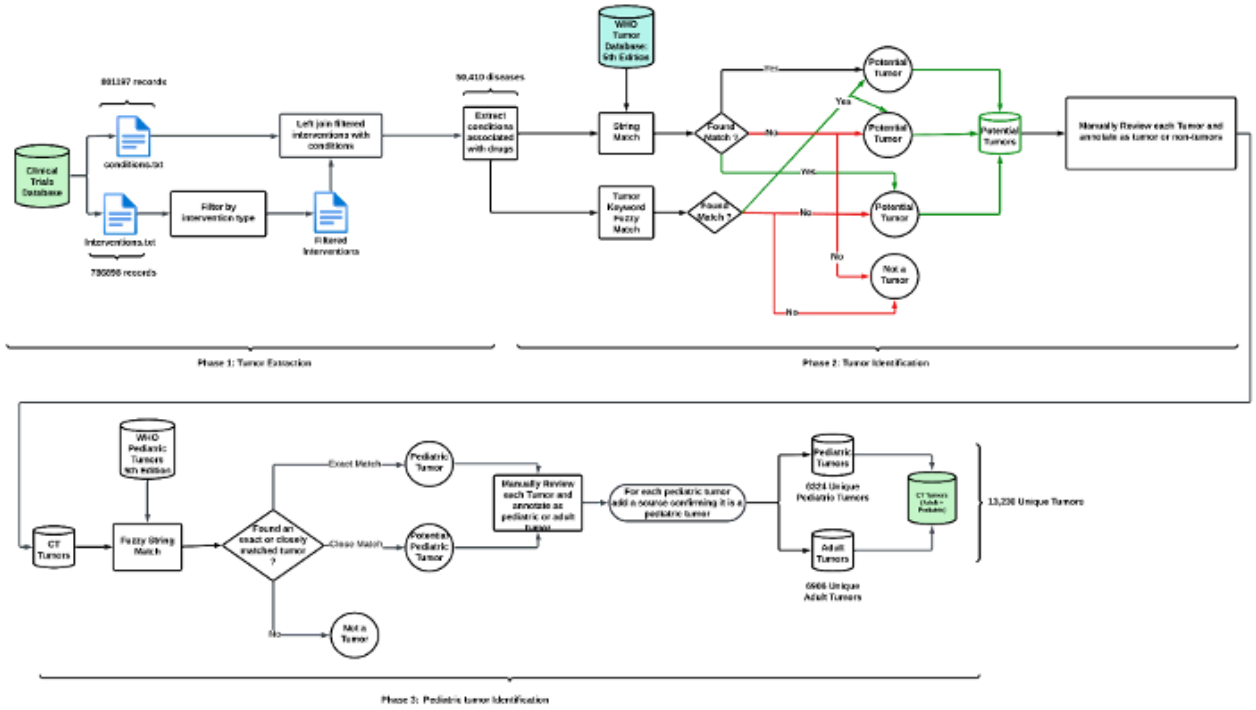


**Figure 2: Text Matching and Embedding Pipeline for Tumor Standardization in CANTOS. Designed to standardize tumor names extracted from the NIH Clinical Trials Registry. The pipeline uses both edit-distance-based text matching and embedding-based methods to standardize the registered tumor names with entries in the WHO and NCIt databases.**

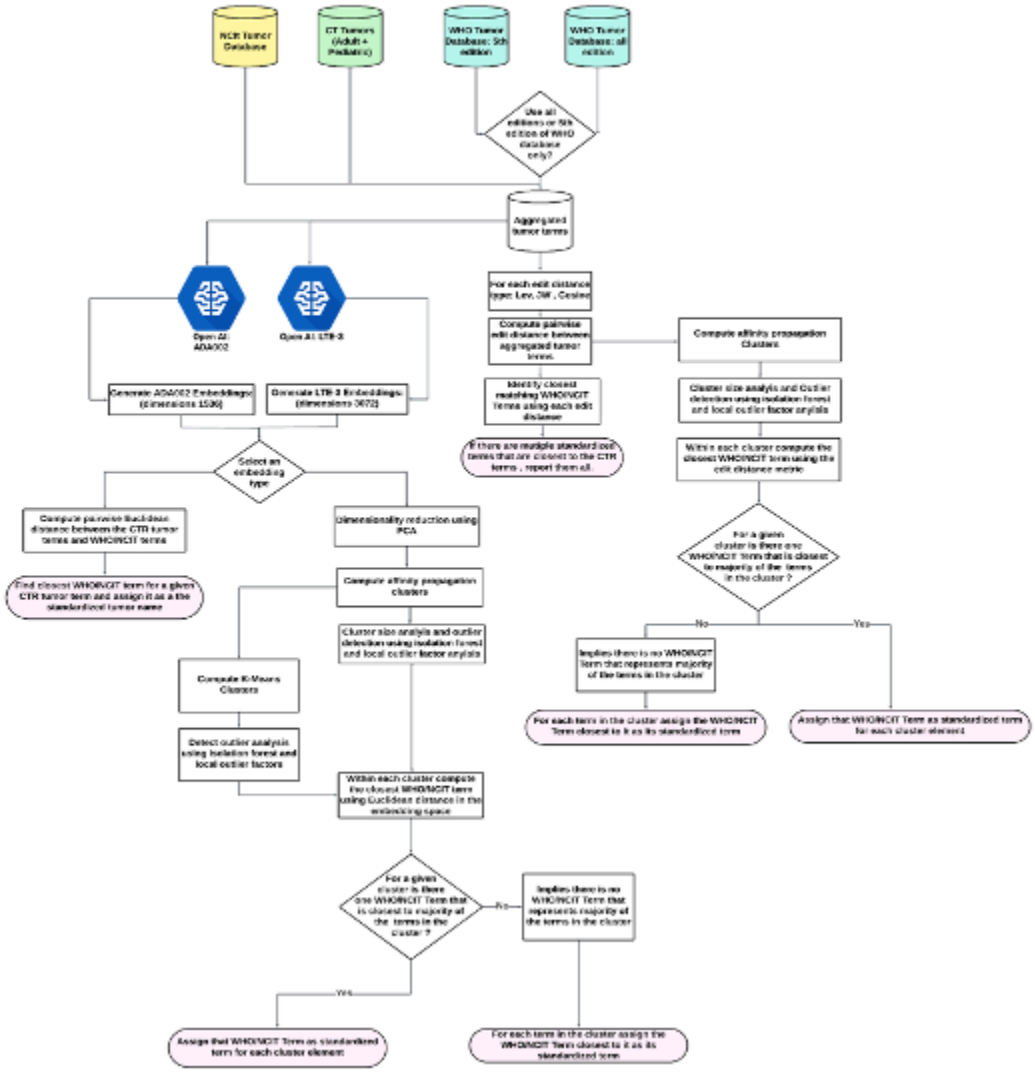
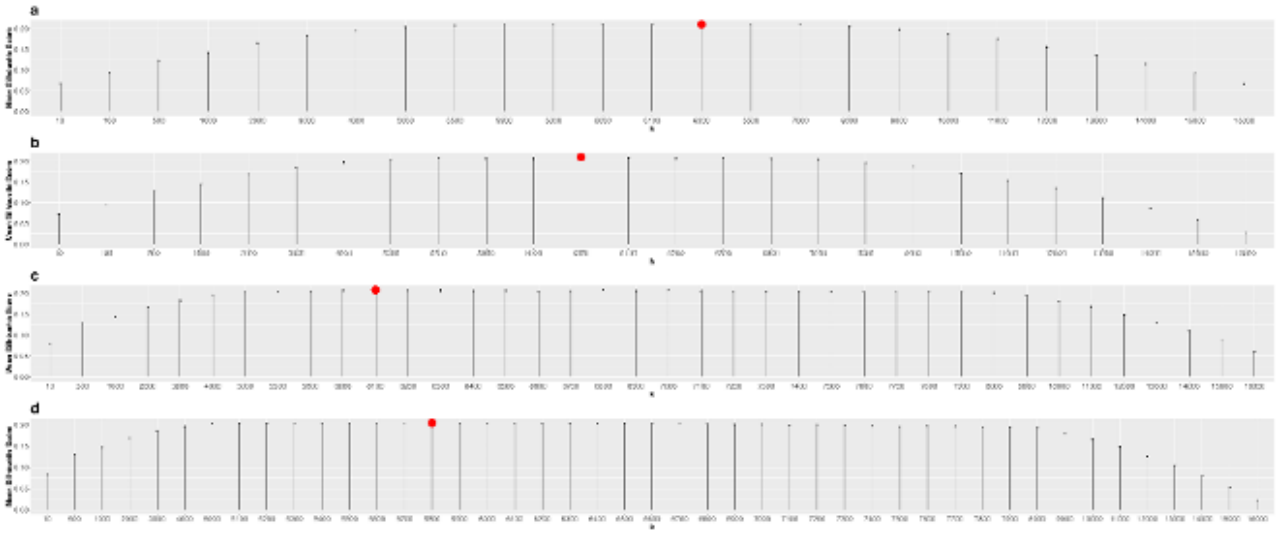


**Figure 3: Average Silhouette Score vs Number of clusters (K) used in K-means. a.** ADA002 when WHO database all editions terms are used. **b.** ADA002 when WHO database 5th edition terms are used **c.** LTE-3 when WHO database all editions terms are used **d.** LTE-3 when WHO database 5th editions terms are used.



**Supplemental information:**

**Figures**

**Supplementary Figure SF1:** CANTOS_Stanadardization_Workflow_Edit_Distances.pdf. This file contains a detailed view of the workflow employed by CANTOS to standardize the CTR tumor names using methods based on text-matching (edit distances).

**Supplementary Figure SF2:** CANTOS_Stanadardization_Workflow_Embedding.pdf. This file contains a detailed view of the workflow employed by CANTOS to standardize CTR tumor names using methods based on text-embedding.

**Tables**

**Supplementary Table ST1:** WHO_Tumor_all_edition.xlsx. This file contains the standardized tumor names from the WHO Tumor Classification System 3rd, 4th, and 5th editions.

**Supplementary Table ST2:** NCIT_Neoplasm_Core_terms.csv. This file contains the standardized tumor names from the NCIt database.

**Supplementary Table ST3:** who_cancer_key_words_general_5th_edition.xlsx. This file contains the standardized tumor names from the WHO Tumor Classification System 5th edition only.

**Supplementary Table ST4:** tumor_sample_df_gt_annotated_all.csv. This file contains the ground truth annotation with respect to all editions (3rd, 4th, and 5th) of WHO Tumor Classification System for the 1600 tumor names randomly sampled from the CTR. Furthermore, this file also contains the standardization results obtained from the 12 standardization methods implemented by CANTOS.

**Supplementary Table ST5:** tumor_sample_df_gt_annotated_5th.csv. This file contains the ground truth annotation with respect to the 5th edition of WHO Tumor Classification System for the 1600 tumor names randomly sampled from the CTR. Furthermore, this file also contains the standardization results obtained from the 12 standardization methods implemented by CANTOS.

**Supplementary Table ST6:** WHO_Results_all.csv. This file contains the standardized tumor names for each CTR tumor identified by CANTOS. The standardization was done with respect to all editions (3rd, 4th, and 5th) of

WHO Tumor Classification System. This result was obtained when all editions (3rd, 4th, and 5th) of WHO Tumor Classification System was in the CANTOS pipeline.

**Supplementary Table ST7:** WHO_Results_5thed.csv. This file contains the standardized tumor names for each CTR tumor identified by CANTOS. The standardization was done with respect to the 5th edition of WHO Tumor Classification System. This result was obtained when 5th edition of WHO Tumor Classification System was in the CANTOS pipeline.

**Supplementary Table ST8:** NCIT_Results_all.csv. This file contains the standardized tumor names for each CTR tumor identified by CANTOS. The standardization was done with respect to the NCIt database. This result was obtained when all editions (3rd, 4th, and 5th) of WHO Tumor Classification System was in the CANTOS pipeline.

**Supplementary Table ST9:** NCIT_Results_5thed.csv. This file contains the standardized tumor names for each CTR tumor identified by CANTOS. The standardization was done with respect to the NCIt database. This result was obtained when 5th edition of WHO Tumor Classification System was in the CANTOS pipeline.

**Supplementary Table ST10:** tumor_annotated_adult_ped.csv. This file contains the list of tumor identified using CANTOS from the conditions file in CTR. For each condition name, the file informs the user if it is a tumor and if it is also a pediatric tumor. For each pediatric tumor, the file also provides a literature citation confirming that it is a pediatric tumor.

**Supplementary Table ST11:** who_cancer_key_words_paediatric_5th_edition.xlsx. This file contains the standardized pediatric tumor names from the WHO Tumor Classification System 5th edition only.

**Documents**

**Supplementary Document SD1:** Condition_MeSH_Term_Comparison.pdf. This file compares the MeSH terms and conditions terms in the CTR and establishes that even though MeSH terms are standardized, they are not an appropriate representation of the condition names.

**Supplementary Document SD2:** Intervention_Type_NIH_CTR.pdf. This file lists the different types of intervention types found in the CTR.

**Supplementary Document SD3:** Tumor_Key_Words.pdf . This file contains the tumor key words that were used to detect tumor from CTR, along with standardized terms in ST3 and ST11.

**Supplementary Document SD4:** NIH_CTR_Conditions_Discrepancies.pdf . This file displays examples of condition names from the CTR and the various discrepancies associated with them and if there are any standardized term from the WHO Tumor Classification System.

**Supplementary Document SD5:** Example_Edit_Distance.pdf. This file demonstrates an example of how edit distance can be used to transform one string to another.

**Supplementary Document SD6:** data_download_instructions.pdf. This file shows the steps involved in downloading the CTR database.