Standardization of Tumor Names in NIH-Clinical Trials Registry using Large Language Model Embedding Analysis

Aditya Lahiri1, Sangeeta Shukla1, Ben Stear1, Taha Mohseni Ahooyi1, Katherine Beigel1, Elizabeth Margolskee2, Deanne Taylor1,3

**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

**Key Words:** Clinical Trials Registry, WHO Tumor Classification, Text Embedding, Clinical Text Standardization

**Word Count:**

**ABSTRACT**

**Objective:**

This study aimed to extract tumor names from the National Institute of Health's (NIH) clinical trials registry (ClinicalTrials.gov) and standardize them according to the corresponding tumor terminology established in the World Health Organization's (WHO) tumor classification system and the National Cancer Institute Thesaurus (NCIt).

**Materials and Methods:**

We developed a computational pipeline that loaded the conditions data file from NIH's clinical trials registry and identified tumors from the rest of the conditions. Following the tumor identification process, each tumor from the registry was mapped to a standardized tumor terminology from the WHO tumor classification system and NCIt using twelve text standardization methods based on text-similarity and text-embedding methods. We evaluated the accuracy of each of these methods in mapping tumor names to standardized tumor terminology in the WHO tumor classification system on a subset of tumor names derived from the clinical trials registry. We limited the accuracy evaluation to only the WHO tumor classification system as it is considered the gold standard for tumor nomenclature.

**Results:**

Our results revealed that embedding-based text standardization methods outperformed methods based on text-matching algorithms. We generated two different sets of embeddings from OpenAI’s large language models and observed that accuracy of methods improved with embeddings that had higher dimensions. In particular, the method that mapped a given tumor name in the registry to the nearest term from WHO tumor classification system using Euclidean distance in the embedding space outperformed other methods.

**Discussion and Conclusion:**

The tumor names in the NIH clinical trials registry are not standardized, making integrating this data with other biomedical databases challenging. Therefore, we developed a computational pipeline that identified tumors from the NIH clinical trials registry and mapped them to their standardized terms established in the WHO tumors classification system.

**INTRODUCTION**

Cancer is a major global health problem [[1]](https://paperpile.com/c/NPPxEM/Weij) and is the second largest cause of deaths in the United States [[2]](https://paperpile.com/c/NPPxEM/6kRx). In the United States it is estimated that there will be over 2 million cases of newly diagnosed cancer and 611,720 deaths in 2024 [[2]](https://paperpile.com/c/NPPxEM/6kRx). Among children (ages 0 to 14 years) and adolescents (ages 15 to 19 years), in the US, pediatric cancer persists to be the second and fourth leading cause of deaths [[2]](https://paperpile.com/c/NPPxEM/6kRx), despite the jump in 5-year survival rate to 80% in the last 5 decades [[3]](https://paperpile.com/c/NPPxEM/tUKD) [[4]](https://paperpile.com/c/NPPxEM/FRbF). The increased survival rate of pediatric cancer treatment can primarily be attributed to successes in the treatment of common childhood hematological malignancies such as acute lymphoblastic leukemia, where the survival rate has risen from 10% to 90% [[5]](https://paperpile.com/c/NPPxEM/68hp), unfortunately this success is not uniformly shared across all types of pediatric cancers [[6]](https://paperpile.com/c/NPPxEM/NRZI) especially cancers of the brain and nervous system [[2]](https://paperpile.com/c/NPPxEM/6kRx). Compared to adult cancers, pediatric cancers are rarer and with fewer available therapeutic agents that have been tested in clinical trials due to challenges associated with recruiting statistically significant and diverse pediatric populations to support the various phases of clinical trials, logistical issues related to clinical trial-site location and molecular heterogeneity of tumors [[7,8]](https://paperpile.com/c/NPPxEM/Lq9U+5cZt). Therefore to understand the therapeutic landscape associated with adult and pediatric tumors it is critical to extract and analyze data from various biomedical databases, especially the NIH’s Clinical Trials Registry (referred to as CTR in the rest of the manuscript).

Cancer is a significant global health issue and the second leading cause of death in the United States [[1]](https://paperpile.com/c/NPPxEM/Weij). In 2024, over 2 million new cancer cases and 611,720 deaths are expected in the U.S. alone [[2]](https://paperpile.com/c/NPPxEM/6kRx). While pediatric cancer survival rates have improved to 80% over the last five decades [[3]](https://paperpile.com/c/NPPxEM/tUKD) [[4]](https://paperpile.com/c/NPPxEM/FRbF), which can primarily be attributed to successes in the treatment of common childhood hematological malignancies such as acute lymphoblastic leukemia [[5]](https://paperpile.com/c/NPPxEM/68hp), certain cancers, particularly those of the brain and nervous system [[2]](https://paperpile.com/c/NPPxEM/6kRx), remain difficult to treat and unfortunately this success is not uniformly shared [[6]](https://paperpile.com/c/NPPxEM/NRZI) . 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]](https://paperpile.com/c/NPPxEM/Lq9U+5cZt). 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]](https://paperpile.com/c/NPPxEM/hYaH). However, the data in the CTR with respect to tumor names contain 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]](https://paperpile.com/c/NPPxEM/Hrsy), these terms often fail to capture the specific details of many tumor types.Thus there is a need to standardize the tumor names in the CTR with respect to standardized nomenclature from the World Health Organization (WHO) Classification of Tumors and National Cancer Institute (NCIt) databases.

To address these inconsistencies, we applied methods based on text-matching (based on edit distances) and text-embedding (derived from OpenAI’s Large Language Models) to develop a computational pipeline that standardizes tumor names in the CTR’s “conditions” file. This pipeline maps tumor names to their standardized counterparts in the WHO and NCIt databases. Standardizing these names enables integration with other biomedical databases, such as Open Targets and Illuminating the Druggable Genome, facilitating a more comprehensive understanding of tumor biology, targets, and drugs.

This study addresses these issues by mapping non-standardized tumor names in the CTR to standardized names using OpenAI’s embedding methods. This automated approach improves the accuracy of tumor classification, enhancing the reliability of clinical trial data for research and therapeutic development.

The Food and Drug Administration Modernization Act of 1997 (FDAMA) mandated the National Institutes of Health (NIH) to create a publicly available resource to disseminate information on the effectiveness of drugs in federally or privately funded clinical trials [[9]](https://paperpile.com/c/NPPxEM/hYaH). This eventually led to the creation of the CTR (ClinicalTrials.gov), which was publicly launched on February 29, 2000. Since then, the CTR has amassed over 482,529 research studies across all 50 states in the US and 223 countries [[9]](https://paperpile.com/c/NPPxEM/hYaH). Each record within the CTR is self-reported by the trial sponsor through the web-based data entry platform known as the Protocol Registration and Results System (PRS) [[9]](https://paperpile.com/c/NPPxEM/hYaH). The CTR requires sponsors to enter basic details regarding the trial, such as purpose, design, patient eligibility criteria, and other critical information about the study [[10]](https://paperpile.com/c/NPPxEM/Hrsy). The CTR requires by law that data be entered in a tabular format and that an expert with knowledge of study design and data analysis be involved in the submission process to ensure that results are appropriately summarized and the data submission is consistent with the review criteria of the CTR. Following the submission of a record, the record is reviewed internally by the CTR staff before it is published publicly on ClinicalTrials.gov.

Users can download CTR data directly from the CTR website (link). The process requires that the user perform a faceted search and the resulting results are delivered as a suite of files through the web interface. The files include <file types> and a ‘conditions’ file which includes details on the conditions studied in each trial.

Thus by standardizing tumor names in the CTR, i.e. by mapping each tumor name in the CTR to their equivalent standardized names in the World Health Organization (WHO) Classification of Tumors or the National Cancer Institute terms (NCIt) will allow us to integrate tumor data from the CTR with other biomedical databases such as Open Targets or Illuminating the Druggable Genome will allow us to deeply understand the landscape of tumors, targets, and drugs.

Even with the established protocols and guidelines in the submission process, the CTR data may contain various inconsistencies in the form of extraneous information, typographical errors, missing values, non-standard nomenclature, etc. Such discrepancies must be addressed or filtered out before the data can be used for further downstream analysis. In this study, we developed a computational pipeline to standardize the tumor names contained in the output CTR "conditions" data file. The “conditions” data is a tab-delimited table that includes the names of the diseases or conditions that are the subject of the trial.

It must be noted that the CTR recommends adding relevant Medical Subject Headings (MeSH) terms or terms from another controlled vocabulary, such as the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), that have been mapped to MeSH within the Unified Medical Language System (UMLS) metathesaurus for each of the conditions. While this recommendation adds a level of standardization to the disease/condition names present in the CTR, the MeSH terms by themselves often fall short of describing the disease or condition accurately. Furthermore, for a given disease/condition there may be multiple associated MeSH terms, thereby leaving it to the user of the data to establish the most appropriate MeSH term for that disease/condition. Additionally, for certain records in the CTR, there might be no associated MeSH terms provided for a disease/condition name, therefore solely using the MeSH terms for analyzing the diseases or conditions that are the subject of a clinical trial record is not reliable. Therefore it is evident that the MeSH terms cannot be used to replace the terms enlisted in the conditions file, as the MeSH terms fail to accurately capture the specific information being conveyed by the conditions. In Supplementary Table S1, we provide examples of clinical trial records with their disease/condition names and associated mesh terms to illustrate the incompatibilities between condition names and MeSH terms.

For example, in Supplementary Table S1, the condition “triple negative breast cancer” is associated with two clinical trial studies with the identifiers (NCT IDs) NCT05082610 and NCT04254107. Each of these studies lists the various MeSH terms associated with “triple negative breast cancer” with the most appropriate MeSH term being “triple negative breast neoplasms”. However, there are other associated MeSH terms for each of these studies which are not appropriate: for study NCT05082610 there are MeSH terms such as “carcinoma, non-small-cell lung” and “respiratory tract diseases” while study NCT04254107 has associated MeSH terms such as “lymphoma” and “stomach neoplasms”, which do not describe the condition of “triple negative breast cancer”. Furthermore, the MeSH terms are not identical between studies where the condition names are the same, which adds to the inconsistencies between records with same condition names. For instance, NCT04254107 contains various MeSH terms associated with lymphomas such as “lymphoma, b-cell”, “large b-cell”, “lymphoma, large b-cell, diffuse”, “lymphoma, t-cell, peripheral”, etc., but these terms are not contained in the list of MeSH terms for NCT05082610. We can also see in study NCT04294784, for the condition “gastroesophageal cancer”, there are no MeSH terms, but for the same condition with a different NCT ID (NCT02669914), there are multiple associated MeSH terms.

We conclude from our analysis that the MeSH terms–even though they are internally standardized– are not suitable to be used to identify the exact term and condition name for the conditions they encode. Therefore, even though we observed that the condition names from the CTR contained syntactic and semantic inconsistencies, we decided to use the terms from the conditions file to extract tumor names and map them to their standardized nomenclature in 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 (“NCIt database”) (<https://ncithesaurus.nci.nih.gov/ncitbrowser/>).

**METHODS**

**Data Availability**

The data used in this paper is obtained from the CTR (<https://clinicaltrials.gov/>) [[9]](https://paperpile.com/c/NPPxEM/hYaH). The data can be publicly accessed via the Clinical Trials API or from the Aggregate Analysis of ClinicalTrials.gov-Clinical Trials Transformative Initiative (AACT-CTTI) website (<https://aact.ctti-clinicaltrials.org/download>). The ACCT-CTTI website is updated daily with contents from ClinicalTrials.gov and a static database is made available at the start of each month. The static database contains information about all the studies registered in ClinicalTrials.gov. We downloaded a copy of the database from the ACCT-CTTI website on August 22, 2023 which is available at the following at the ACCT-CTTI as a pipe-delimited file under the name of “20230822\_export.zip” in the section titled “Monthly Archive of Static Copies”.

**Data Extraction Pipeline**

The CTR contains information about various aspects of the study, such as outcomes, interventions used, conditions (diseases) studied, design of experiments, sponsors of the studies, etc. This information is presented to the public in individual text files. Each text file contains the National Clinical Trial Identification Number (NCTID), which serves as the unique identifier (foreign key) for that study. The NCTID allows one to reference a particular clinical trial and aggregate various information associated with it, which is stored in various text files in the database. In this study, we select the conditions ("conditions.txt" or conditions file) and interventions ("intervention.txt" or interventions file) files, which contain information regarding diseases and drugs used in each of the clinical trial studies, respectively. The conditions file contains 801,197 records and is annotated with the following fields: "*id*," "*nct\_id*," "*name*," and "*downcase\_name*." The "*id*" field represents the identification number for that record within the conditions file, whereas the "*nct\_id*" is the foreign key for that record. The "*name*" and "*downcase\_name*" fields contain names of the conditions (referred to as diseases in the rest of this manuscript) studied in a specific clinical trial study, with the "*downcase\_name*" containing the disease name in "*name*" in downcase format. We identified 105,483 unique diseases (by uniqueness of strings) in the conditions file.

The conditions file does not classify the disease based on any system. Therefore, we needed to develop a computational pipeline to extract tumor names (Phase 1), identify specific tumor names from the rest of the diseases (Phase 2), and identify and annotate unique tumors as pediatric or adult (Phase 3) (Figure 1). We focused on extracting tumors that are associated with a therapeutic agent (i.e., has a drug-target) registered in clinical trials. To achieve this, we subset the diseases in the conditions file with the types of drugs registered in the interventions file. The intervention file has 786,898 records and is annotated with the following fields: "*id*," "*nct\_id*," "*intervention\_type*", "*name*," and "*description*." The "*id*" field represents the identification number for that record within the conditions file, whereas the "*nct\_id*" is the foreign key for that record. The "*name*" and "*description*" fields provide names and details regarding the interventions used in that particular clinical trial, and the "*intervention\_type*" field classifies the interventions into one of the eleven distinct intervention types listed in Table 1.

**Table 1: Types of interventions listed in interventions file in the CTR**

| Index | Type of intervention | Included in analysis |
| --- | --- | --- |
| 1 | Drug | Yes |
| 2 | Biological | Yes |
| 3 | Radiation | No |
| 4 | Device | No |
| 5 | Behavioral | No |
| 6 | Other | No |
| 7 | Genetic | Yes |
| 8 | Procedure | No |
| 9 | Combination Product | Yes |
| 10 | Dietary Supplement | No |
| 11 | Diagnostic Test | No |

For extracting tumors with an associated therapeutic agent with a drug target, we first join the intervention file with the conditions file using the foreign key NCTID ("*nct\_id*") and then filter the file for interventions belonging to the following intervention type: "*Drug*", "*Biological*", "*Combination Product*", and "*Genetic*". We limited diseases to these intervention types to ensure that there is a corresponding targeted or chemotherapy, immunotherapy-based treatment option for the tumors from the CTR as these intervention types provide the most insight for researchers studying the therapeutic and drug-target landscape of a given tumor. After filtering based on the specified intervention types, we obtained 50,410 unique diseases for tumor identification. The tumor extraction process is visualized as Phase 1 of the pipeline in Figure 1.

**Figure 1**: **Tumor extraction and annotation pipeline**

Phase 2 of this pipeline differentiated tumors from the rest of the diseases (Figure 1, “Phase 2”). In Phase 2, there were two independent protocols to detect tumors from disease names. The first protocol consisted of checking if each disease name contained a tumor key word. The tumor keywords are listed in Supplemental Table S2. If the disease name contained a tumor keyword, that disease was flagged as a potential tumor.

The second protocol in Phase 2 matched each of the diseases in the CTR to the tumor names listed in the 5th edition (latest) of the WHO database using a fuzzy string match algorithm. If a disease from clinical trials exactly matched a term in the WHO database, it was flagged as tumor. If the disease did not match to any tumor within the WHO database, we performed a fuzzy (approximate) match with the disease name with each term in the WHO database. This was done by computing the generalized Levenshtein edit distance of the clinical trials disease to each WHO database term. The fuzzy matching using generalized Levenshtein edit distance was implemented using the agrepl function in R [[11]](https://paperpile.com/c/NPPxEM/rIfO) and the maximum distance was set to 0.2. If there was at least one WHO database term within this maximum distance, that clinical trial disease was flagged as a potential tumor. Once a clinical trial disease was flagged as a potential tumor by either of these two protocols, they were manually validated to confirm if they were indeed a tumor.

After the manual validation step in Phase 2, we identified a total of 13,230 unique (by string uniqueness) tumors in the clinical trial database. Out of these tumors, we developed phase 3 of the pipeline to identify which of the tumors were pediatric (Figure 1, phase 3”). In phase 3 we implemented a similar fuzzy string match algorithm as in phase 2, but instead compared the 13,230 tumors output from phase 2 to the pediatric tumors listed in the 5th edition of the WHO database (instead of the entire WHO database). Once the tumors were flagged as pediatric or adult by the pipeline, we manually validated the results. If the tumors were marked as pediatric by manual validation, we add a citation from peer reviewed literature, governmental websites or articles posted by research institution

where it states that the tumor in question is a pediatric tumor. If the tumor in question is found in the WHO database, we simply state “Listed in WHO Ped Tumor”. Out of the 50,410 diseases, the pipeline identified 13,230 tumors among which 6324 tumors were determined to be pediatric. These annotations are stored in the supplementary file S3 titled: “tumor\_annotated\_adult\_ped.csv”.

**Anomalies in Clinical Trials Tumor Names**

A cursory analysis of the tumor names data revealed various sources of discrepancy such as typographical errors, extraneous information, missing values, drug names entered instead of disease names, multiple tumor names etc. Furthermore, we also noticed many tumor names in the disease file did not follow standardized tumor names from the WHO or the NCIT databases. With the tumor names not being standardized, it becomes challenging to relate the tumors in the CTR to tumors in other biomedical databases, such as Illuminating the Druggable Genome (IDG) or Open Targets (OT). Some of these discrepancies also prevented us from annotating these tumor names as pediatric or adult cancers, and consequently 144 clinical trials tumors were annotated as “DA” (Do not Annotate) in the field designating them as pediatric tumor (“PedCanTumor) in the supplementary file S3. Table 2 outlines some of these common discrepancies associated with tumor names.

**Table 2: Discrepancies associated with Conditions Data**

| **NCT ID** | **Disease Name** | **Issue** | **Standardized WHO Tumor Name** |
| --- | --- | --- | --- |
| NCT02172768 | acute myeloid leucaemia | Leukemia is not spelled correctly. | acute myeloid leukaemia |
| NCT02658838 | lovastatin/ticagrelor [va drug interaction] | Drug names are presented instead of diseases. | Does not exist |
| NCT01291602 | healthy male and female japanese volunteers | Description of study participant is provided instead of diseases. | Does not exist |
| NCT04323956 | ann arbor stage iii follicular lymphoma | Extra information provided on staging of the tumor. | follicular lymphoma |
| NCT02637531 | adrenocortical carcinoma (part g) | Extra information presented inside parenthesis. | adrenal cortical carcinoma |
| NCT03712605 | pathologic stage i merkel cell carcinoma ajcc v8 | Extra information on staging of the tumor. | merkel cell carcinoma |
| NCT01780740 | disorder; heart, functional, postoperative, cardiac surgery | Multiple clinical terms entered in a comma separated format. | Does not exist |
| NCT01963481 | metastatic beast cancer | Breast cancer is spelled incorrectly. | invasive breast carcinoma of no special type (5th Edition WHO) ,  invasive breast carcinoma (3rd Edition WHO) |
| NCT01782235 | primary sj√∂gren's syndrome (pss) | Special characters present in disease name. | Does not exist |
| NCT01209195 | locally advanced/metastatic or recurrent ovarian cancer, fallopian tube cancer, | Multiple tumor names | Does not exist |
| NCT01863108 | tumor vaccines | Description of tumor intervention | Does not exist |
| NCT02018874 | solid tumors and non-hodgkin's lymphoma | Solid tumors is a vague term whereas non-Hodgkin’s lymphoma is specific and a childhood tumor. | Does not exist |
| NCT05050630 | at least one positive lesion according to the 2014 lugano criteria for hodgkin's and non-hodgkin's lymphoma | Vague tumor description. It is not clear whether the tumor is Hodkin’s or non-Hodgkin’s tumor. | Does not exist |

Due to various sources of discrepancies in the tumor names in the CTR, there is a need to standardize them. Once the clinical trials tumor names are standardized, they can be linked to external databases to draw further insights about the tumors, such as potential drug targets and currently available FDA-approved drugs.

To standardize the tumor names, we designed a computational pipeline to associate non-standardized tumor names identified from the CTR with the standardized tumor terms from the WHO and NCIT databases. The WHO database is considered the gold standard for tumor nomenclature and it has been updated over the years due to which there are multiple versions of this database. Thus, we considered the 5th, 4th, and 3rd editions of the WHO database which are publicly available online as a reference for standardization. We first aggregated the tumor names identified from the CTR, WHO database and NCIT database. We then ran the pipeline in two different iterations:, 1. with only the latest version of WHO database (5th edition) and 2. with all editions (3rd , 4th and 5th) of the WHO database. The purpose of this was to standardize the tumors from clinical trials with respect to two variations of the WHO database: the latest (referred to in this text as the WHO database 5th Edition) and the combined editions of the WHO database (referred to in this text as the WHO database All Editions). During each of these iterations of the pipeline, the tumor names from clinical trials were also standardized with respect to the NCIT database, however, we evaluate the performance of the various methods used in this pipeline only with respect to the WHO database as it is considered the gold standard for tumor nomenclature. In the following section we will discuss the various methods used to standardize the tumor names CTR.

**Standardization Pipeline**

We tested various methods to standardize the names of tumor names in the CTR. These methods are based either on text-matching (edit-distances) or text-embedding. We also developed methods that combined unsupervised clustering along with text-matching and text-embedding to standardize the terms in CTR, in total we implemented 12 methods and tested their performance accuracies which are discussed in the results sections.. In the following subsections, we will first discuss each of these text-matching and then text-embedding-based standardization methods.

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

When standardizing the tumor names from the CTR to their equivalent WHO or NCIT terms, we are essentially comparing pieces of text (also referred as strings in this manuscript) with the objective of mapping the terms in a way in which meaning conveyed by the CTR term is conserved by the WHO or NCIT term to which it is mapped. Edit distances offer a way to compare the similarity between two strings. Edit distances can be based on the minimum number of edit-operations (deletions, substitutions, insertions, etc), q-grams or heuristics that are required 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. There are several methods to compute edit distances between strings; in this paper, we use normalized Levenshtein distance, Jarro-Winkler distance, and cosine distance, which are commonly used edit distances. An example of how edit distances can be used to compare strings is discussed in supplementary file S4. Following are brief descriptions of each method.

*Normalized Levenshtein distance*: Levenshtein distance between two strings is defined as the minimum number of single character edits which include insertions, deletions or substitutions required to transform a string to its target string. The Levenshtein distance between two strings is not normalized. To normalize them, we divide the Levenshtein distance by the length of the longest string, so that we get a distance in the interval [0,1]. By normalizing the Levenshtein distance, we can compare the dissimilarity between a string and multiple target strings on the same scale ([0,1]). Furthermore, we can also define the similarity between two strings S1 and S2 as follows:

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 [[12]](https://paperpile.com/c/NPPxEM/MKyW). 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:

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 which estimating m that two characters from S1 and S2 are only considered to be matching if they are the same and are less than characters apart.

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:

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:

We calculate the Jarro-Winkler distance using the stringdist package in the R-programming language [[12]](https://paperpile.com/c/NPPxEM/MKyW).

*Cosine Distance***:** In order to define cosine distance , we first need to define 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:

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

We calculate the cosine distance using the stringdist package in the R-programming language [[12]](https://paperpile.com/c/NPPxEM/MKyW).

Based on the three edit distances, we 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 clinical trial tumor name, we select the nearest standardized terms under each edit distance. If more than one term qualified as the closest term, we reported them all by separating the terms with a semicolon.

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

We explored another set of standardization methods that are based on edit distances and clustering techniques. These methods consist of a clustering step and a mapping step. To form the clusters, we used affinity propagation (AP) clustering and applied the distance matrices computed by calculating the pairwise edit distances in the previous section as a divergence metric for AP. We selected the AP algorithm for cluster as it automatically determines the number of clusters instead of requiring the number cluster (which is not known to us) to be a user-defined hyperparameter. Unlike other clustering algorithms, AP is also not dependent on the initialization conditions and is deterministic [[13]](https://paperpile.com/c/NPPxEM/OxXw). 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 [[14]](https://paperpile.com/c/NPPxEM/ZrGQ). Furthermore, AP clustering methods have shown success in clustering textual data [[15,16]](https://paperpile.com/c/NPPxEM/vxqy+FX6U). Once the clusters were computed using AP, the median cluster size was determined and clusters larger than the median cluster size were identified. These clusters are designated as large clusters. We perform AP clustering within these clusters until their sizes drop below the previously determined median cluster size of the AP algorithm converges, and no more clustering can be performed.

We then check for outliers within each cluster using isolation forest and local outlier factors (LOF). If a data point within a cluster is determined to be an outlier using either of the methods, it is removed from that cluster and labeled as a new cluster with just that data point. We implemented isolation forest using the 'isolation.forest' package in the R programming language [[17]](https://paperpile.com/c/NPPxEM/5rKu). The number of trees (ntrees argument) is set to 100 as recommended by Lie et al.2008 [[18]](https://paperpile.com/c/NPPxEM/vpqB) in their original introduction of the isolation forest algorithm and the dims argument to 3, as suggested for numeric datasets in the package documentation. The isolation scores are calculated for each data point within a cluster, and if the isolation score is above 0.5, that data point is deemed an outlier. Similarly, for LOF, we use the lof function within the "dbscan" package in the R programming language [[19]](https://paperpile.com/c/NPPxEM/ibnD). To calculate the LOF value of this function, we needed to specify the number of nearest neighbors used to define the local neighborhood of a data point ("minPts"). We compute the LOF values for "minPts" ranging from 2 (clusters need to have more than one element to have an outlier) to . We compute the median LOF for each data point in the cluster, and if this value is above 1, that point is deemed an outlier.

Upon completing the outlier analysis, we implement the pipeline's second step, which maps each cluster member to a standardized term. To achieve this, we iterate through each cluster and identify 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 draw, 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 figure 2. The following section will discuss the pipeline that standardizes the clinical trial tumors based on text-embeddings.

**Fig 2: Text Match Pipeline using Edit Distances**

***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 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 [[20–23]](https://paperpile.com/c/NPPxEM/SePY+AIyW+WbHn+bKvK). Text-embeddings have been used in various applications such as developing search engines [[24,25]](https://paperpile.com/c/NPPxEM/fu7E+8uDZ), text clustering [[26]](https://paperpile.com/c/NPPxEM/NP5Q) and classification [[27]](https://paperpile.com/c/NPPxEM/73kP), recommender systems [[28]](https://paperpile.com/c/NPPxEM/1mjP), and anomaly detection [[29]](https://paperpile.com/c/NPPxEM/ZV6P). 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 [[23]](https://paperpile.com/c/NPPxEM/bKvK). In this paper, we generate text-embeddings from the following two embedding models offered by Open AI: text-embedding-ada-002 (referred as ADA002 in rest of the text) and text-embedding-3-large (referred as LTE-3 in rest of the text) [[30]](https://paperpile.com/c/NPPxEM/l3Uz). LTE-3 is a more recent and improved model and generates embeddings that have 3072 dimensions, whereas ADA002 generates embeddings consisting of 1536 dimensions.

We used both LTE-3 and ADA002 to generate embeddings for all the tumors identified in clinical trials, each term in every edition of the WHO database, and all the terms in the NCIT database. Once these sets of embeddings were generated, the tumor names from the CTR were standardized by calculating their Euclidean distances from each standardized tumor term in the WHO and NCIT database and identifying the nearest matching term. We did this using each set of embeddings and standardized the clinical trials tumor with respect to the 5th edition of WHO database, combined editions (3rd-5th) (referred to as “all editions”) of the WHO database, and the NCIT database.

***Text Embedding Analysis: Embeddings and Clustering*** Similar to how we used edit distances as a divergence metric for AP clustering and performed standardization on each cluster, we considered using the Euclidean distance between terms in the embedding space as a divergence metric in AP clustering. However, calculating the pairwise Euclidean distance in the embedding space for all the tumor terms in clinical trials, WHO, and NCIT database is computationally expensive and time consuming as the embedding space has a large number of dimensions for both ADA002 (1536 dimensions) and LTE-3 (3072 dimensions) embeddings. Instead of computing all pairwise Euclidean distances on the raw embeddings, we performed principal component analysis (PCA) on each set of embeddings (ADA002 and LTE-3) for each of the pipeline iterations (WHO 5th Edition and WHO All Editions). Thus, we generated four sets of PCA transformed embeddings, and their dimensions are listed in Table 4. For each case, we only retain the minimum number of principal components that explain 80% of the variance in the data.

**Table 4: PCA dimensions for each embeddings 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 |

With the four sets of PCA-transformed embeddings, we calculated the pairwise Euclidean distance that was used for AP clustering. Once AP clustering is completed, we identify the large clusters by computing the z-scores for each cluster based on the number of cluster members. We set the z-score threshold at 2.5, meaning that any clusters with a z-score greater than 2.5 were designated as ‘large’ clusters. Using a z-score threshold of 2.5, we are able to determine the maximum number of members per cluster. For each cluster we designated as ‘large’, we performed AP within these clusters until either their sizes were below or equal to the maximum number of cluster members or the AP clustering algorithm converged. In this embedding-based AP clustering, we did not use median cluster size as the threshold for determining the large clusters as we did in edit-distance-based AP clustering. This is because we observed that the median cluster size in embedding-based AP clusters was lower than the maximum cluster size established by z-score based clustering. Due to this several clusters which had fairly homogeneous cluster members were determined to be “large” and required further sub clustering. The maximum cluster size determined using z-score was larger and determined fewer such clusters to be large, thus we selected it to be the threshold for determining large clusters. It should be noted that CTR has a diverse range of tumor names, and thus clustering these tumor names will produce clusters of non uniform cluster size, hence it is not trivial to estimate a maximum cluster size when we consider 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 the clustering was completed, we performed isolation forest and LOF analysis to determine outliers within each cluster just as we did in the case for edit-distance based AP clustering. The hyperparameters for both isolation forest and LOF analysis were kept the same as they were for edit-distance based AP clustering. Following the outlier detection step, we iterate through each cluster and determine the standardized term (WHO database 5th Edition, WHO database All Edition, or the NCIT database) that is closest to each cluster element 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 tumor name. The final standardization process is similar to the standardization process discussed in the edit-distance based AP clustering, however, instead of edit distances, we used the Euclidean distance in the embedding space to determine the closest standardized terms. In Table 5, we compare the number of AP clusters when we use text-embeddings and text-matching (edit distances).

**Table 5: 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, All Editions WHO Terms | Number of Clusters for CT Terms, NCIT Terms, 5th Edition WHO |
| --- | --- | --- | --- |
| 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 |

In addition to AP clustering, we also implemented K-means clustering on the PCA transformed embeddings to standardize the tumor names in clinical trials. Unlike AP clustering, the K-means algorithm requires the user to define the numbers of clusters to be formed, “K”, as a hyperparameter [[31]](https://paperpile.com/c/NPPxEM/wp0i). Since we do not have any a priori information on the types of tumors that are present in the CTR and have no means to classify the tumors based on their tissue or molecular subtypes, we needed to use other computational methods to decide on a value for the number of clusters to provide as an input to the K-means algorithm. To determine the number of clusters, we computed a commonly used cluster performance metric known as the silhouette coefficient [[32]](https://paperpile.com/c/NPPxEM/rUdk). After the clustering is completed, the silhouette coefficient is computed for each data point. A silhouette coefficient ranges from -1 to 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 [[33]](https://paperpile.com/c/NPPxEM/HATf). 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 [[33]](https://paperpile.com/c/NPPxEM/HATf). 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.

For several values of “K”, we perform K-means clustering and calculate the silhouette coefficient for each data point. For each cluster, we calculate the average silhouette coefficient which indicates the performance for that cluster. We also 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, we 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.

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

Once the clustering is completed using K-means, we follow the exact same steps for outlier detection (isolation forest and LOF analysis) and standardization (compute euclidean distances with standardized term and identify the closest match) as we did for AP clusters,maintaining the exact hyperparameter configurations. The entire text-embedding based standardization pipeline is displayed in Figure 4.

**Figure 4: Text Match Pipeline using ADA002 and LTE-3 Embeddings**

**RESULTS**

In this paper, we implemented 12 methods to standardize the tumor terms from the CTR. These methods were based on text-matching algorithms (edit-distances) and text embedding. We identified 13,230 tumors in the CTR and standardized them using these methods by considering only the 5th edition of WHO database, then all editions of WHO database, and finally the NCIT database. To evaluate the performance of each of these 12 methods, we needed to know the ground truth or the appropriate standardized tumor names for each of the 13,230 clinical trials tumors. Since this information is not available and manual annotation of all tumors is not feasible, we arbitrarily sampled 1,600 tumors from the 13,230 CTR tumors for ground truth annotation and to evaluate the performance of our methods.

For these 1,600 tumors, we annotated the ground truths with respect to both the WHO database 5th Edition and the WHO database All Editions We did not evaluate the performance on the NCIT tumor terms, as the WHO database is considered as the gold standard for tumor nomenclature. During the ground truth annotation process, we noticed that several CTR tumors either had multiple ground truths associated with them or did not have ground truths available from the WHO database.

Tumors from the CTR that could not be manually assigned a ground truth from the WHO database were excluded when evaluating the accuracy of each method. When only the 5th edition of the WHO database (WHO database 5th Edition) was used for standardization, we identified 567 CTR tumors that did not have a ground truth. When we considered all editions of the WHO database (WHO database All Editions), we identified 482 CTR tumors that did not have a ground truth. This is a consistent finding as all editions of the WHO database have a greater number of standardized tumor terms (compared to only the 5th edition, which contains only the currently accepted terms but not all terms that may have been used in the past) to which the clinical trials tumors can be mapped. Therefore, out of the 1,600 terms chosen for ground truth annotation, we evaluated the accuracy for 1,033 terms when we considered only the 5th edition of the WHO database for standardization and 1,118 terms when we considered all editions of the WHO database. To compute the accuracy of each method, we simply count the instances in which a given method identified at least one ground truth associated with the tumor name drawn from the CTR and divide it by total number of terms (1033 for 5th Edition of WHO database or 1118 for All Editions of WHO database). We report the accuracy of each method for the WHO database All Editions (Table 6) and the WHO database 5th Edition (Table 7).

Table 6: Accuracies for Standardization Methods when all Editions of WHO terms are used

| **Ranking** | **Basis** | **Methods** | **Accuracy All Editions WHO** |
| --- | --- | --- | --- |
| 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 7: Accuracies for Standardization Methods when only 5th Editions of WHO terms are used

| **Ranking** | **Basis** | **Methods** | **Accuracy 5th Edition WHO** |
| --- | --- | --- | --- |
| 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 6 and Table 7 show that text-embedding-based methods outperform text-match-based methods irrespective of which variation of the WHO database was used for standardization. Typically, methods utilizing LTE-3 embeddings performed better than ADA002 embeddings, with method LTE-3 + K-means as the exception, which performed marginally worse than ADA002+AP and ADA002+Euclidean+Dist when we standardized against all editions of the WHO database (Table 6). However, it should be noted that LTE-3+K-means ranked higher than ADA002+K-means when standardized against either variation of the WHO database. We attribute the better performance of LTE-3-based methods to the fact that LTE-3 embeddings have twice the number of dimensions as ADA002 and are able to better capture the complexity in the input data.

Among the edit distances, Levenshtein distance performed better than Jarro-Winkler distance and cosine distance or any of their implementations involving AP clustering. The Jarro-Winkler distance is useful when there are minor discrepancies between the texts being compared and if there are common prefixes between the texts, whereas cosine distance is based on the frequency of occurrence of each word (“bag of words”) in a text and does not take into account the order of words. Comparatively, Levenshtein distance only counts the number of edit operations needed to transform one text to another, does not factor the prefix similarity, and maintains the order of words/alphabet. The ground truth annotations of the 1,600 clinical trials tumors along with their standardization results are available in Supplementary File S5 (for WHO database All Editions) and Supplementary File S6 (for WHO database 5th Edition). For each iteration of the pipeline, we report the WHO and NCIT standardized terms for each tumor identified in the CTR in the following supplementary files: Supplementary File S7-S10.

**Discussion:**

In this paper, our objective was to develop a computational pipeline to extract tumors from the CTR and standardize them with respect to the WHO and NCIT databases. To this end, we built a tumor extraction pipeline to extract tumors from the CTR and then annotated the tumors as adult and pediatric tumors. We manually validated the diseases that were identified as adult tumors and pediatric tumors, and we assigned a field with a citation that confirmed that the tumor in question is indeed pediatric (Supplementary File S3). In total we identified 13,230 tumors, among which 6,324 were pediatric tumors.

We standardized the tumors identified from the CTR with respect to both the WHO and NCIT databases. Because the WHO database is considered the gold standard for tumor nomenclature, we evaluated our standardization methods against only the WHO database. The standardization methods we used were based on text-matching (edit distances) techniques and text-embedding techniques. To evaluate the performance of these techniques, we drew 1,600 tumor terms arbitrarily from the tumors listed in the CTR and standardized them according to the variation of the WHO database that was in that iteration of the pipeline (either the “WHO database 5th Edition” which included only the 5th edition of the database, or the “WHO database All Editions” which included the 3rd, 4th, and 5th editions of the database). We observed that when we used all the editions of the WHO database, we were able to manually annotate (ground truth) more of the tumor terms in the CTR because the 5th edition of the WHO database has fewer terms compared to all three editions of the WHO database combined. In general, the text-embedding based methods were more accurate than text-matching based methods, which we attribute to the fact that text-matching techniques are based on edit-distances and can only evaluate syntactical differences between text, whereas text-embeddings are able to capture the semantic and contextual meaning of text and map similar texts close to each other in the embedding space.

We generated the text embeddings from OpenAI’s LLM:text-embedding-ada-002 (ADA002) and text-embedding-3-large (LTE-3). We observed that the LTE-3 based methods performed better than the ADA002 methods, with the exception of LTE-3+K-means, which slightly underperformed compared to ADA002+Euclidean Dist and ADA002+AP when standardized against all the editions of the WHO database. Irrespective of the variation of the WHO database that was considered in the pipeline, LTE-3+Euclidean Dist achieved the highest accuracy followed by LTE-3 + AP. LTE-3 + Euclidean Dist standardizes the clinical trials tumor by identifying and assigning the WHO database term that is closest in terms of Euclidean distance in the LTE-3 embedding space. Compared to the LTE-3 +AP method, the LTE-3 +Euclidean Dist is a simpler and faster technique that does not require additional steps such as clustering, cluster size analysis and outlier detection.

As more samples from the CTR are annotated with their respective ground truth, the performance accuracies of each method will merge towards their true accuracies.

While we expect the accuracy to change, it is clear from our ground truth annotation for the 1,600 clinical tumors that the text-embedding based methods will very likely outperform the text-matching (edit distance) based methods. Expert annotation of the tumors in clinical trials is essential for accurate performance evaluation of these methods and is a limitation in our study. Furthermore, as the CTR is updated, new tumor names need to be detected and potentially standardized, which will require rerunning the pipeline, identifying the WHO database tumor which is at the closest Euclidean distance (LTE-3 + Euclidean Dist method), and annotating the ground truths for each new tumor entry. In addition to requiring expert annotation of ground truths, there is a computational cost associated with running the pipeline and storing the data, which can become expensive in the long run. Another limiting factor in this study are the embeddings that were generated by OpenAI. If the OpenAI models are updated or discontinued, we will need to switch to other available LLMs that generate embeddings. Furthermore, the LLMs that generate the embeddings for OpenAI are not specifically trained on a medical or tumor corpus, and an LLM which is trained on such a corpus will likely have better performance and will be able to differentiate between the tumors more precisely.

The methods explored here provide a basis to extract and standardize the tumors from the CTR. While there are guidelines in place for submitting data to the CTR to maintain basic data integrity, there are no enforced protocols to standardize tumor names that are entered as conditions data in the CTR. While studies have called for standardization of clinical trials with respect to study design [[34,35]](https://paperpile.com/c/NPPxEM/9u0o+9AUN) evidence reproducibility [[36]](https://paperpile.com/c/NPPxEM/MrPy), there have been no studies to our knowledge that have attempted to standardize tumors or any other diseases in the CTR. Standardizing the tumor names in the CTR will make these tumor names searchable in other biomedical databases, which will allow researchers to quickly develop an expansive overview of the associated targets, drugs, clinical outcomes for a given tumor.

**Conclusion:**

The CTR records information on various aspects of a clinical trial study which includes the conditions being studied. The information is disseminated to the public in text file format. However, the CTR lacks neither provides a mechanism to identify tumor names from other conditions nor are the tumor names standardized according to the WHO tumor classification system or the National Cancer Institute thesaurus, making it challenging to extract and link these tumors with other biomedical databases for integrative analysis. In this paper, we designed a pipeline to extract tumors from the CTR and annotated them as adult or pediatric tumors. Furthermore, we tested 12 methods to standardize the identified tumor names using text-matching and text-embedding based methods. We observed that embedding based methods performed better than text-matching methods, and in particular the LTE-3 + Euclidean Dist method had the highest accuracy in standardizing the clinical trials tumors. We standardize all the 13,230 tumors in the CTR with respect to both the WHO and NCIT database using all the methods and report them in our results.

**Author Contribution:**

**Supplementary Material:**

**Funding:**

**Conflict of Interest:**

**Data availability:**

**References**

1 [Bray F, Laversanne M, Sung H, *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74:229–63.](http://paperpile.com/b/NPPxEM/Weij)

2 [Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. 2024;74:12–49.](http://paperpile.com/b/NPPxEM/6kRx)

3 [Matt GY, Sioson E, Shelton K, *et al.* St. Jude Survivorship Portal: Sharing and Analyzing Large Clinical and Genomic Datasets from Pediatric Cancer Survivors. *Cancer Discov*. 2024;14:1403–17.](http://paperpile.com/b/NPPxEM/tUKD)

4 [Aristizabal P, Winestone LE, Umaretiya P, *et al.* Disparities in Pediatric Oncology: The 21st Century Opportunity to Improve Outcomes for Children and Adolescents With Cancer. *Am Soc Clin Oncol Educ Book*. 2021;41:e315–26.](http://paperpile.com/b/NPPxEM/FRbF)

5 [Hunger Stephen P., Mullighan Charles G. Acute Lymphoblastic Leukemia in Children. *N Engl J Med*. ;373:1541–52.](http://paperpile.com/b/NPPxEM/68hp)

6 [Laetsch TW, DuBois SG, Bender JG, *et al.* Opportunities and Challenges in Drug Development for Pediatric Cancers. *Cancer Discov*. 2021;11:545–59.](http://paperpile.com/b/NPPxEM/NRZI)

7 [Renfro LA, Ji L, Piao J, *et al.* Trial Design Challenges and Approaches for Precision Oncology in Rare Tumors: Experiences of the Children’s Oncology Group. *JCO Precis Oncol*. 2019;3. doi:](http://paperpile.com/b/NPPxEM/Lq9U) [10.1200/PO.19.00060](http://dx.doi.org/10.1200/PO.19.00060)

8 [Rivers Z, Hyde B, Ronski K, *et al.* Exploring Barriers to Pediatric Cancer Clinical Trials: The Role of a Networked, Just-in-Time Study Program. *Clin Ther*. 2023;45:1148–50.](http://paperpile.com/b/NPPxEM/5cZt)

9 [National Institutes of Health Clinical Trials Registry. ClinicalTrials.gov.](http://paperpile.com/b/NPPxEM/hYaH) <https://clinicaltrials.gov/> [(accessed 26 August 2024)](http://paperpile.com/b/NPPxEM/hYaH)

10 [Zarin DA, Tse T, Williams RJ, *et al.* The ClinicalTrials.gov Results Database — Update and Key Issues. *N Engl J Med*. 2011;364:852–60.](http://paperpile.com/b/NPPxEM/Hrsy)

11 [Snášel V, Keprt A, Abraham A, *et al.* Approximate String Matching by Fuzzy Automata. *Man-Machine Interactions*. Springer Berlin Heidelberg 2009:281–90.](http://paperpile.com/b/NPPxEM/rIfO)

12 [van der Loo M. Stringdist: Approximate string matching, fuzzy text search, and string distance functions. CRAN: Contributed Packages. 2013.](http://paperpile.com/b/NPPxEM/MKyW)

13 [Frey BJ, Dueck D. Clustering by passing messages between data points. *Science*. 2007;315:972–6.](http://paperpile.com/b/NPPxEM/OxXw)

14 [Kitahara YFGI. Fast Algorithm for Affinity Propagation. In: Walsh T, ed. *Twenty-Second International Joint Conference on Artificial Intelligence*. IJCAI/AAAI :2238–43.](http://paperpile.com/b/NPPxEM/ZrGQ)

15 [Shi XH, Guan RC, Wang LP, *et al.* An incremental affinity propagation algorithm and its applications for text clustering. *2009 International Joint Conference on Neural Networks*. IEEE 2009.](http://paperpile.com/b/NPPxEM/vxqy)

16 [Shailendra Kumar Shrivastava, J.L.Rana, R.C.Jain. Text document clustering based on phrase similarity using affinity propagation. *International Journal of Computer Applications*. 2013;61. doi:](http://paperpile.com/b/NPPxEM/FX6U) [10.5120/10032-5077](http://dx.doi.org/10.5120/10032-5077)

17 [Cortes D. isotree: Isolation-Based Outlier Detection. CRAN: Contributed Packages. 2019.](http://paperpile.com/b/NPPxEM/5rKu)

18 [Liu FT, Ting KM, Zhou Z-H. Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*. IEEE 2008:413–22.](http://paperpile.com/b/NPPxEM/vpqB)

19 [Hahsler M, Piekenbrock M. Dbscan: Density-based spatial clustering of applications with noise (DBSCAN) and related algorithms. CRAN: Contributed Packages. 2015.](http://paperpile.com/b/NPPxEM/ibnD)

20 [Morris J, Kuleshov V, Shmatikov V, *et al.* Text embeddings reveal (almost) as much as text. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics 2023.](http://paperpile.com/b/NPPxEM/SePY)

21 [Mikolov T. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:13013781*. Published Online First: 2013.](http://paperpile.com/b/NPPxEM/AIyW)

22 [Incitti F, Urli F, Snidaro L. Beyond word embeddings: A survey. *Inf Fusion*. 2023;89:418–36.](http://paperpile.com/b/NPPxEM/WbHn)

23 [Khattak FK, Jeblee S, Pou-Prom C, *et al.* A survey of word embeddings for clinical text. *J Biomed Inform*. 2019;100S:100057.](http://paperpile.com/b/NPPxEM/bKvK)

24 [Gökçe O, Prada J, Nikolov NI, *et al.* Embedding-based scientific literature discovery in a text editor application. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics 2020.](http://paperpile.com/b/NPPxEM/fu7E)

25 [Mai G, Janowicz K, Yan B. Combining text embedding and knowledge graph embedding techniques for academic search engines. In: Key-Sun Choi, Luis Espinosa Anke, Thierry Declerck, Dagmar Gromann, Jin-Dong Kim, Axel-Cyrille Ngonga Ngomo, Muhammad Saleem, Ricardo Usbeck, ed. *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 2018:77–88.](http://paperpile.com/b/NPPxEM/8uDZ)

26 [Mehta V, Bawa S, Singh J. WEClustering: word embeddings based text clustering technique for large datasets. *Complex Intell Systems*. 2021;7:3211–24.](http://paperpile.com/b/NPPxEM/NP5Q)

27 [Stein RA, Jaques PA, Valiati JF. An analysis of hierarchical text classification using word embeddings. *Inf Sci* . 2019;471:216–32.](http://paperpile.com/b/NPPxEM/73kP)

28 [Musto C, Semeraro G, de Gemmis M, *et al.* Learning word embeddings from Wikipedia for content-based recommender systems. *Lecture Notes in Computer Science*. Cham: Springer International Publishing 2016:729–34.](http://paperpile.com/b/NPPxEM/1mjP)

29 [Pande A, Ahuja V. WEAC: Word embeddings for anomaly classification from event logs. *2017 IEEE International Conference on Big Data (Big Data)*. IEEE 2017:1095–100.](http://paperpile.com/b/NPPxEM/ZV6P)

30 [New embedding models and API updates.](http://paperpile.com/b/NPPxEM/l3Uz) <https://openai.com/index/new-embedding-models-and-api-updates/> [(accessed 4 September 2024)](http://paperpile.com/b/NPPxEM/l3Uz)

31 [Wu J. Cluster Analysis and K-means Clustering: An Introduction. In: Wu J, ed. *Advances in K-means Clustering: A Data Mining Thinking*. Berlin, Heidelberg: Springer Berlin Heidelberg 2012:1–16.](http://paperpile.com/b/NPPxEM/wp0i)

32 [Shahapure KR, Nicholas C. Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE 2020:747–8.](http://paperpile.com/b/NPPxEM/rUdk)

33 [Shutaywi M, Kachouie NN. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy* . 2021;23. doi:](http://paperpile.com/b/NPPxEM/HATf) [10.3390/e23060759](http://dx.doi.org/10.3390/e23060759)

34 [Canonica GW, Baena-Cagnani CE, Bousquet J, *et al.* Recommendations for standardization of clinical trials with Allergen Specific Immunotherapy for respiratory allergy. A statement of a World Allergy Organization (WAO) taskforce. *Allergy*. 2007;62:317–24.](http://paperpile.com/b/NPPxEM/9u0o)

35 [Katz MHG, Marsh R, Herman JM, *et al.* Borderline resectable pancreatic cancer: need for standardization and methods for optimal clinical trial design. *Ann Surg Oncol*. 2013;20:2787–95.](http://paperpile.com/b/NPPxEM/9AUN)

36 [Dickersin K, Mayo-Wilson E. Standards for design and measurement would make clinical research reproducible and usable. *Proc Natl Acad Sci U S A*. 2018;115:2590–4.](http://paperpile.com/b/NPPxEM/MrPy)

**FIGURES**

**Figure 1**: **Tumor extraction and annotation pipeline**

****

**Figure 2: Text Match Pipeline using Edit Distances**

****

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

****

**Figure 4: Text Match Pipeline using ADA002 and LTE-3 Embeddings**

****

**SUPPLEMENTARY DATA:**

**Table S1: Conditions data with MeSH Terms:** This table was created by performing a full join on the files “conditions.txt” and “browse\_conditions.txt”, the files were joined on clinical trials ID.

| NCT ID | Condition name | MeSH term |
| --- | --- | --- |
| NCT05082610 | triple negative breast cancer | neoplasms,triple negative breast neoplasms,carcinoma, non-small-cell lung,breast neoplasms,neoplasms by site,breast diseases,skin diseases,carcinoma, bronchogenic,bronchial neoplasms,lung neoplasms,respiratory tract neoplasms,thoracic neoplasms,lung diseases,respiratory tract diseases |
| NCT04254107 | triple negative breast cancer | lymphoma,carcinoma,lymphoma, t-cell, peripheral,lymphoma, large b-cell, diffuse,triple negative breast neoplasms,squamous cell carcinoma of head and neck,stomach neoplasms,neoplasms by histologic type,neoplasms,lymphoproliferative disorders,lymphatic diseases,immunoproliferative disorders,immune system diseases,neoplasms, glandular and epithelial,neoplasms by site,carcinoma, squamous cell,lymphoma, b-cell,lymphoma, non-hodgkin,lymphoma, t-cell,breast neoplasms,breast diseases,skin diseases,head and neck neoplasms,gastrointestinal neoplasms,digestive system neoplasms,digestive system diseases,gastrointestinal diseases,stomach diseases |
| NCT01590680 | neuroblastoma | neuroblastoma,pheochromocytoma,paraganglioma,neuroectodermal tumors, primitive, peripheral,neuroectodermal tumors, primitive,neoplasms, neuroepithelial,neuroectodermal tumors,neoplasms, germ cell and embryonal,neoplasms by histologic type,neoplasms,neoplasms, glandular and epithelial,neoplasms, nerve tissue,neuroendocrine tumors |
| NCT04081701 | medulloblastoma | adenoma,meningioma,medulloblastoma,paraganglioma,pituitary neoplasms,esthesioneuroblastoma, olfactory,central nervous system neoplasms,hemangioblastoma,neoplasms, glandular and epithelial,neoplasms by histologic type,neoplasms,pituitary diseases,hypothalamic diseases,brain diseases,central nervous system diseases,nervous system diseases,endocrine system diseases,neoplasms, nerve tissue,neoplasms, vascular tissue,meningeal neoplasms,nervous system neoplasms,neoplasms by site,glioma,neoplasms, neuroepithelial,neuroectodermal tumors,neoplasms, germ cell and embryonal,neuroectodermal tumors, primitive,neuroendocrine tumors,endocrine gland neoplasms,hypothalamic neoplasms,supratentorial neoplasms,brain neoplasms,neuroblastoma,neuroectodermal tumors, primitive, peripheral,olfactory nerve diseases,cranial nerve diseases,hemangioma, capillary,hemangioma |
| NCT04294784 | gastroesophageal cancer | NA |
| NCT02669914 | gastroesophageal cancer | lung neoplasms,carcinoma, non-small-cell lung,colorectal neoplasms,pancreatic neoplasms,ovarian neoplasms,brain neoplasms,kidney neoplasms,carcinoma, renal cell,breast neoplasms,respiratory tract neoplasms,thoracic neoplasms,neoplasms by site,neoplasms,lung diseases,respiratory tract diseases,carcinoma, bronchogenic,bronchial neoplasms,intestinal neoplasms,gastrointestinal neoplasms,digestive system neoplasms,digestive system diseases,gastrointestinal diseases,colonic diseases,intestinal diseases,rectal diseases,endocrine gland neoplasms,pancreatic diseases,endocrine system diseases,ovarian diseases,adnexal diseases,genital diseases, female,female urogenital diseases,female urogenital diseases and pregnancy complications,urogenital diseases,genital neoplasms, female,urogenital neoplasms,genital diseases,gonadal disorders,central nervous system neoplasms,nervous system neoplasms,brain diseases,central nervous system diseases,nervous system diseases,urologic neoplasms,kidney diseases,urologic diseases,male urogenital diseases,adenocarcinoma,carcinoma,neoplasms, glandular and epithelial,neoplasms by histologic type,breast diseases,skin diseases |

**Table S2: contains the key words used for detecting tumors**

| # | Tumor Key Words |
| --- | --- |
| 1 | cancer |
| 2 | carcinoma |
| 3 | adenocarcinoma |
| 4 | tumor |
| 5 | lymphoma |
| 6 | blast |
| 7 | myeloma |
| 8 | melanoma |
| 9 | leukemia |
| 10 | astrocytoma |
| 11 | malignant |
| 12 | neoplasm |
| 13 | neoplasia |
| 14 | mesothelioma |
| 15 | ependymoma |
| 16 | glioma |
| 17 | thymoma |
| 18 | waldenstrom macroglobulinemia |
| 19 | myelodysplastic syndrome |
| 20 | polycythemia vera |
| 21 | myelofibrosis |
| 22 | myeloproliferative |
| 23 | sarcoma |
| 24 | gist-plus syndrome |
| 25 | macroglobulinemia |
| 26 | mycosis fungoides |
| 27 | sezary's disease |
| 28 | plasmacytoma |

**S3:** tumor\_annotated\_adult\_ped.csv

**S4: Demonstration of comparing strings using edit distances**

In order to illustrate the concept of how strings can be compared using edit distances, consider the following two strings

String 1: Breast Cancer

String 2: Brain Cancer

Let us now suppose, we can make the following operations: deletions, substitution, and insertion to transform string 1 to string 2. Then we can proceed to transform String 1 to String 2 in the following ways:  
  
Table A1: step by step demonstration of comparing strings using edit distances.

| Method 1 | Method 1 operation | Method 2 | Method 2 operation |
| --- | --- | --- | --- |
| Breast Cancer |  | Breast Cancer |  |
| Brast Cancer | Delete ‘e’ | Braast Cancer | Substitute ‘e’ with ‘a’ |
| Bras Cancer | Delete ‘t’ | Braist Cancer | Substitute ‘a’ with ‘i’ |
| Bra Cancer | Delete ‘s’ | Braint Cancer | Substitute ‘s’ with ‘n’ |
| Brai Cancer | Insert ‘i’ | Brain Cancer | Delete ‘t’ |
| Brain Cancer | Insert ‘n’ |  |  |

We can observe from Table A1 that the two methods transform string 1 to string 2, but method 2 employs fewer steps to achieve this objective. Furthermore, with the given set of operations the minimum number of steps required to transform string 1 to string 2 is four. This is achieved by method 2 , thus the edit distance between string 1 and string 2 is four. The set of operations that were used to transform the strings and method with which we calculated the edit distance is also known as the Levenshtein distance.

**S5:** tumor\_sample\_df\_gt\_annotated\_all\_sep11.csv

**S6:**tumor\_sample\_df\_gt\_annotated\_5th\_sep11.csv

**S7:** WHO\_Results\_all\_10sep.csv

**S8:** WHO\_Results\_5thed\_9sep.csv

**S9:** NCIT\_Results\_all\_10sep.csv

**S10:**NCIT\_Results\_5thed\_9sep.csv