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

Aditya Lahiri1, Sangeeta Shukla1, Ben Stear1, Taha Mohseni Ahooyi1, Deanne Taylor1,2

**Affiliations:**

1. The Department of Biomedical and Health Informatics, The Children’s Hospital of Philadelphia, Philadelphia PA
2. 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:**

Advancements in medicine and technology have propelled the 5-year survivorship of childhood cancer in the US from 58% in the 1970s to 85% in 2023. However, childhood cancers remain the leading cause of disease-related deaths in the US among children, adolescents, and young adults. Our overarching goal is to build a computational tool that integrates and analyzes various public biomedical databases and provides physicians and scientists with the landscape of targets and available therapeutic agents tested in childhood cancers. However, integrating and analyzing data from various public databases is non-trivial.

One such public database is the Clinical Trials Registry (ClinicalTrials.gov), which was launched by the National Institutes of Health (NIH) and contains 482,529 research studies from all 50 states in the US and 223 countries. While guidelines are in place to ensure data integrity, our initial inspection of the disease data revealed various inconsistencies in the form of extraneous information, typographical errors, missing values, etc. Furthermore, we observed that the tumor names do not necessarily follow standardized tumor names in the World Health Organization (WHO) Tumor Classification system or National Cancer Institute Terms (NCIT) and need to be identified from other diseases in the dataset.

Considering these factors, we designed a computational pipeline to identify and standardize tumors from the Clinical Trials Registry. The pipeline extracted over 107K diseases and identified 13,230 unique tumors, of which 6934 were pediatric. The pipeline employed different standardization techniques based on edit distances, text-embedding, and clustering, and compared their accuracy on a subset of tumor names from the registry. We found that text-embedding methods produced significantly more accurate standardization of the tumors in the registry compared to other methods. The pipeline standardizes all the 13,230 tumors, which are manually validated and provided as a data table for further downstream data integration and analysis.

*Objective:*

*Materials and Methods:*

*Results:*

*Discussion:*

*Conclusion:*

1. **Background and Significance:**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 (CT) \cite{CT.gov}. This eventually led to the creation of the ClinicalTrials.gov registry, which was publicly launched on February 29, 2000. Since then, the registry has amassed over 482,529 research studies across all 50 states in the US and 223 countries \cite{CT.gov}. Each record within the CT registry is self-reported by the trial sponsor through the web-based data entry platform known as the Protocol Registration and Results System (PRS) {\cite.gov}. The registry requires sponsors to enter basic details regarding the trial, such as purpose, design, patient eligibility criteria, and other critical information about the study [(Zarin et al. 2011)](https://paperpile.com/c/NPPxEM/Hrsy). The CT registry requires by law that data be entered in a tabular format and that an individual 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 CT registry. Following the submission of a record, the record is reviewed internally by CT registry staff before it is posted publicly on ClinicalTrials.gov.

Even with the established protocols and guidelines in the submission process, the registry data may contain various inconsistencies in the form of extraneous information, typographical errors, missing values, 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 "conditions" data file in the CT registry. The conditions data includes the names of the diseases or conditions that are the subject of the trial. Among the various diseases present in the CT registry, we focused on cancers as they are a leading cause of death in the US and the world [(Siegel et al. 2023)](https://paperpile.com/c/NPPxEM/GzPE). Thus, standardizing cancer names in the CT registry , i.e. by mapping each tumor name in the CT registry to their equivalent standardized names in the World Health Organization's tumor classification system or the National Cancer Institute terms (NCIT) will allow us to integrate tumor data from the CT registry with other databases such as Open Targets or Illuminating the Druggable Genome will allow us to deeply understand the landscape of tumors, targets, and drugs.

It must be noted that the CT registry 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 has been mapped to MeSH within the Unified Medical Language System (UMLS) metathesaurus for each of the conditions \cite{CT.gov}. While this recommendation adds a level of standardization to the disease/condition names present in CT registry, 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 CT registry, 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. In table 1, we show examples of CT records with their disease/condition names and associated mesh terms. Table 1 was created by performing a full join on the files “conditions.txt” and “browse\_conditions.txt”, the files were joined on CT ID.

| CT 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 1: Conditions data with MeSH Terms

In Table 1, for the condition triple negative breast cancer with CT ID: NCT05082610, the most appropriate MeSH terms is “triple negative breast neoplasms”, however, there are other associated MeSH terms such as “carcinoma, non-small-cell lung” and “respiratory tract diseases” which do not describe the condition of triple negative breast cancer. Since there is no metric in the CT registry by which we can computationally determine the most appropriate MeSH terms for a given condition. 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. Consider studies with CT ID: NCT05082610 and NCT04254107, in both the studies the condition names are “triple negative breast cancer”, however, the MeSH terms are not identical. For instance, in the study with CT ID: 04254107 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, these terms are not contained in the study with CT ID: NCT05082610. We can also see in study NCT04294784 , for the condition “gastroesophageal cancer” , there are no MeSH terms , however, for the same condition with a different CT ID NCT02669914 has multiple associated MeSH terms. Due to these inconsistencies, and a lack of metric to determine the most accurate MeSH term for a given condition, we can determine that using the MeSH terms to describe the conditions is not reliable. Therefore, we decided to use the conditions filed in the CT registry to extract the disease/conditions that are the subject of a CT. This study focuses on extracting tumors from the conditions field in the CT registry and standardizing them by matching them to their closest matching terms in the WHO tumor classification system or NCIT.

1. **Materials and Methods:**

***2.0 Data Availability :***The data used in this paper is obtained from the Clinical Trials registry (<https://clinicaltrials.gov/>). 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 31, 2023.

***2.1 Data Extraction Pipeline***

The clinical trials database contains information about every aspect of the study such as outcomes, drugs used, conditions (diseases) studied, design of experiments, sponsors of the studies etc in individual text files. Each text file contains the National Clinical Trial Identification Number (NCTID) which allows one to relate information contained in one text file to another. The NCTID serves as a foreign key for the clinical trial database. For the purpose of this study, we select only the conditions and interventions text files, which contain information regarding diseases and drugs used in a Clinical Trial respectively. The conditions file contains 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 that helps connect this record to related information contained in other files in the database. The “name” and “downcase\_name” fields contain names of the diseases studied in a specific clinical trial study , the only difference is that the “downcase\_name” contains the disease name in uncapitalized format. The disease names do not have classification such as tumors, viral diseases, sexually transmitted diseases, blood borne diseases etc. Thus the data needs to be filtered to identify tumors. Furthermore, for our analysis in this paper we only considered diseases that had a corresponding intervention belonging to the categories of “*Drug*”, “*Biological*”, “*Combination Product*”, “*Genetic*” in the interventions text file. The rationale behind limiting our diseases to these intervention types was to ensure that there is a corresponding targeted or chemotherapy, immunotherapy-based treatment option for the tumors from the clinical trials database. The intervention files list the “id”, “nct\_Id”, “intervention type”, “name”, and “description” for every study registered in the clinical trials database. There are 11 distinct “intervention types”: "*Drug","Biological","Radiation","Device","Behavioral", "Other","Genetic","Procedure" ,"Combination Product" ,"Dietary Supplement",* and *"Diagnostic Test*". The conditions file initially contained 801,197 diseases, after filtering with the intervention requirements we obtained 50,410 unique diseases. These diseases required further processing and filtering to identify the tumors from them.

***2.2 Disease Data Anomalies and Tumor Identification Process***

The disease data extracted contained names of different types of diseases and our next step in this pipeline was to extract the tumors from the rest of the diseases. A cursory analysis of the disease data revealed that it suffered from extensive noises such as typographical errors, extraneous information, missing values, drug names entered instead of disease name, etc. Furthermore, we also noticed that many tumor names in the disease file were not necessarily following standardized tumor names from the WHO Tumor Classification database (referred to as WHO database in the rest of the article) or the National Cancer Institute Thesaurus (NCIT). Table 2 shows some of the discrepancies associated with the disease data.

| NCT ID | Disease Name | Issue | Standardized WHO Tumor Name |
| --- | --- | --- | --- |
| NCT01782235 | primary sj√∂gren's syndrome (pss) | Special characters present in disease name. | NA |
| 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. |  |
| NCT01291602 | healthy male and female japanese volunteers | Description of study participant is provided instead of diseases. |  |
| 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. |  |
| 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) |

The dataset suffers from various sources of discrepancies some of which are outlined in Table 2. Furthermore, the tumors within the dataset are not standardized according to either the WHO database or NCIT, thus making it difficult to relate the tumors in the clinical trials database to tumors in other databases such as the Illuminating the Druggable Genome (IDG) or Open Targets (OT). Once the clinical trials tumors are standardized they can be related to external databases to draw further insights about the tumors such as potential drug targets and currently available FDA approved drugs.

Thus to extract the tumor names from the dataset, we devised a two step protocol to detect if a disease name is a tumor indeed. The first step consisted of checking if each disease name contained a tumor key word. The tumor key words are listed below:

"cancer|carcinoma|adenocarcinoma|tumor|lymphoma|blast|myeloma|melanoma|leukemia|astrocytoma|malignant|neoplasm|neoplasia|mesothelioma|ependymoma|glioma|thymoma|waldenstrom macroglobulinemia|myelodysplastic syndrome|polycythemia vera|myelofibrosis

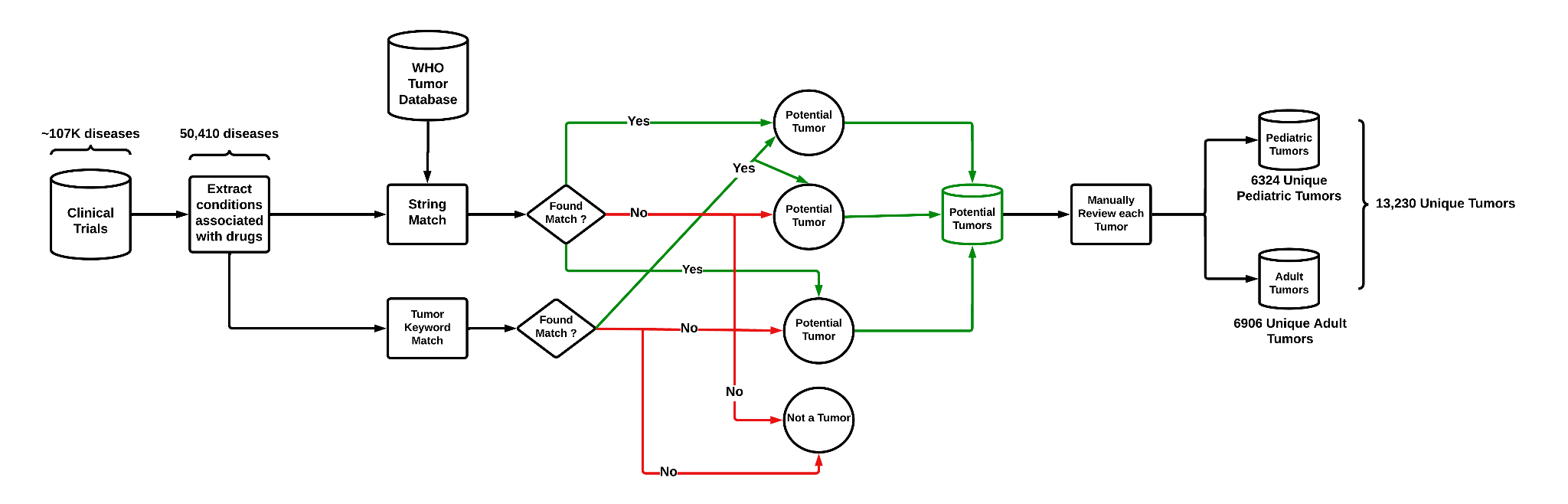
|myeloproliferative|sarcoma|gist-plus syndrome|macroglobulinemia|mycosis fungoides|sezary's disease|plasmacytoma"

If the disease name contained one of the above tumor key words, we flagged that disease as a potential tumor. In our second step in detecting tumors, we considered the tumor names contained in the WHO database (5 edition) and developed a fuzzy string match program to match disease names to tumor names in the WHO database. 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 trial disease to each WHO database term. If the clinical trial disease name was within a certain maximum distance threshold (defined in algorithm 1) any WHO tumor name then it was flagged as a potential tumor. Once every disease in clinical trials was flagged as a potential tumor using steps 1 and 2 , then they were manually validated as tumors. Furthermore during the validation process we also annotated whether the disease was a pediatric tumor and added a field containing a citation that suggested that the tumor was a pediatric tumor. The annotation steps of clinical trials diseases as cancers are described in detail in Algorithm 1.

***2.3 Standardardization Pipelines***

Upon processing the 50,410 unique diseases from the Clinical Trials database, we identified 13,230 diseases to be tumors. Among these tumors, 6324 were identified also to be pediatric tumors. There were 144 tumor names in the Clinical Trials database that could not be evaluated if they were pediatric tumors because the tumor names contained discrepancies that included multiple tumor names, vague tumor names, or tumor names containing treatment information. These terms were not removed as they contained information on tumors , however in the field designating them as pediatric tumor (“PedCanTumor) in the data file “tumor\_annotated\_adult\_ped.csv” , we provide the annotation of DA (Do not Annotate). Table 3 provides examples from a few of the 144 tumor terms which were designated as DA.

| Clinical Trial Tumor Name | Issue |
| --- | --- |
| locally advanced/metastatic or recurrent ovarian cancer, fallopian tube cancer, | Multiple tumor names |
| gastrointestinal neoplasms; esophageal, stomach, pancreas, colon neoplasms; malignant tumors of digestive organ; advanced gastrointestinal malignancies | Multiple tumor names |
| tumor vaccines | Vague description of tumor intervention |
| solid tumors and non-hodgkin's lymphoma | Solid tumors is a vague term whereas non-Hodgkin’s lymphoma is specific and a childhood tumor. |
| 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. |

Due to various sources of discrepancies in the tumor names in the clinical trials database, there is a need to standardize them according to an accepted standard. The tumor names in the WHO database is considered as a gold standard among oncologists and scientists studying cancers, thus we use the tumor names from the 5th, 4th, and 3rd edition of the WHO database as a reference for standardizing the tumor names in the clinical trials database [ADD CITATION] and provide the mapping between tumors names in clinical trials to the terms in the WHO database. In addition, we also considered the NCIT tumor as a reference for standardizing, and ran our pipeline to provide mappings between tumor names in clinical trials to the terms in the NCIT database. In the following subsections we discuss the various methods used in standardizing the clinical trials tumors. 

***2.3.1 Compute pairwise edit distance between CT tumors and WHO and NCIT Tumors***

To begin the process of standardization of tumor names in clinical trials databases, we started by comparing each tumor name in clinical trials to standardized terms in the WHO and NCIT databases using an edit distance measure. Edit distances are a measure of divergence between two given strings, it is a metric that specifies the minimum number of operations such as deletions, substitutions, insertions etc that are required to transform one string to another. A higher magnitude of edit distance between two strings, signifies a higher level of dissimilarity between the strings. 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:

| 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 3 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. There are different methods for calculating edit distances, in our analysis we have used Normalized Levenshtein , Jarro-Winkler, and cosine distance to calculate the edit distances between tumor names in clinical trials and those in the WHO and NCIT databases. 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 , thus 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 as follows:

In the above equation , |s1| and |s2| represent the respective lengths of strings s1 and s2 between which we are comparing the Levenshtein similarity. We calculate Levenshtein distance using the stringdist library in the R programming language\cite{}. 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 S\_1 and S\_2 respectively:

Where s1 and s2 are lengths of the strings S\_1 and S\_2 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 S\_1 and S\_2 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 in place, the Jarro-Winkler distance is simply defined as follows:

We calculate the Jarro-Winkler distance using the stringdist package in the R-programming language \cite{}.

**Cosine Distance:**

***2.3.2.1 Nearest match for edit distance***

***2.3.2.2 Affinity Propagation Clustering based on edit distance and standardization to WHO Terms***

***2.3.3 Extract tumor name embeddings from ADA 2.0 and V-3 Large Text Embeddings***

***2.3.3.1 Nearest match for ADA 2.0 and V-3 Large***

***2.3.3.2 Clustering with KMeans and Standardization to WHO Terms for ADA2.0 and V-3 Large Embeddings***

***2.3.3.3 Clustering with Affinity Propagation Clustering and Standardization to WHO Terms for ADA2.0 and V-3 Large Embeddings***

1. **Results:**
2. **Discussion:**
3. **Conclusion:**

**References**

**https://journal.r-project.org/archive/2014/RJ-2014-011/index.html**