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

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

[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

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 loads the conditions data file from NIH's clinical trials registry and identifies tumors from the rest of the conditions. Following the tumor identification, each tumor from the registry is 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 limit 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 identifies tumors from the NIH clinical trials registry and maps them to their standardized terms established in the WHO tumors classification system.

Background and Methods

Cancer is a major global health problem [1] and is the second-largest cause of death in the United States [2]. 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 death [2], despite the jump in 5-year survival rate to 80% in the last five decades [3-4]. 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]. 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 National Institutes of Health's (NIH) Clinical Trials Registry (CT Registry). The CT registry stores data about various aspects of a clinical trial in separate text files in its database. One such file is the conditions file, which informs the users about the conditions/diseases being studied in a particular clinical trial. Thus, the condition file is the key to extracting cancer information from the CT registry.

Even though there are established protocols for submitting data into the CT registry to ensure data integrity, our analysis of the conditions file revealed that the conditions data contain various inconsistencies in the form of extraneous information, typographical errors, missing values, etc. Furthermore, the tumor names need to be extracted from the rest of the conditions, and the tumor names are not necessarily standardized with respect to the World Health Organization's tumor classification system (WHO database) or the National Cancer Institute Thesaurus (NCIt database). These discrepancies must be addressed before the data can be used for further downstream analysis or data integration.

To this end, we developed a computational pipeline that:

1. **Extracts tumors from the rest of the conditions in the CT registry.**
2. **Annotates extracted tumors as pediatric or adult tumors.**
3. **Standardizes tumors with respect to the WHO and NCIt databases.**

We manually validated each of the CT registry condition terms that were identified as tumor by the pipeline. For each tumor that was determined to be a pediatric tumor, we also manually added a citation from peer-reviewed literature, governmental websites, or articles published by research institutions stating that the tumor in question is a pediatric tumor. We implemented several methods based on text-matching and text-embedding to standardize the tumor terms in the CT registry with respect to the WHO and NCIt databases. Text matching methods are based on edit distances, which are a class of metric that quantifies the syntactical differences between strings to measure text similarity. 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. Text embeddings, on the other hand, 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 space, texts with similar meanings should have embeddings close to each other, and texts that differ in their meaning should be further apart [9–12].

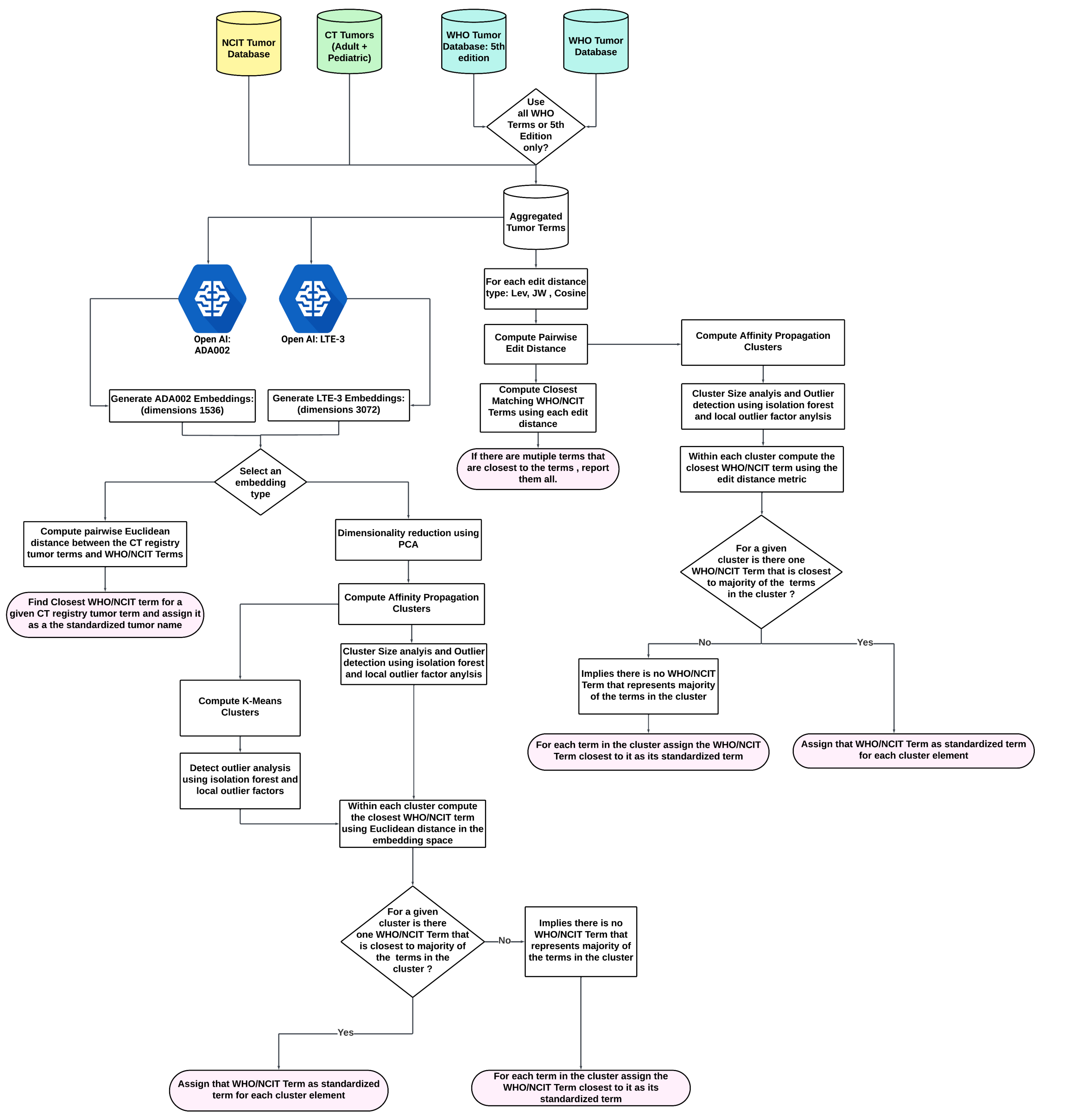
Databases

The Clinical Trials (CT) registry contains information regarding various aspects of a clinical trial study and is stored in the database in the form text files. Using the conditions file in the CT registry, we extracted 801,197 records of conditions. Our pipeline extracted 105,483 unique diseases of which 13,230 were tumors. Among the 13,230 tumors, 6324 tumors were pediatric tumors.

The WHO Tumor Classification database consists standardized terms for tumor names. This database is considered the gold standard for tumor nomenclature and is used for standardization of the tumors identified from the CT registry. We considered 5th, 4th and 3rd editions of the WHO Tumor Classification database for standardizing the tumor names in the CT registry.

NCI Thesaurus (NCIt) provides reference terminology for many NCI and other systems. It covers vocabulary for clinical care, translational and basic research, and public information and administrative activities. The NCIt database also provides standardized terms for tumor names. We used this database to standardize the tumor names from the CT registry.

Standardization Pipeline


NCIT Tumor Database
CT Tumors (Adult + Pediatric)
WHO Tumor Database: 5th edition
WHO Tumor Database
Use all WHO Terms or 5th Edition only?
Aggregated Tumor Terms
Open AI: ADA002
Open AI: LTE-3
Generate ADA002 Embeddings: (dimensions 1536)
Generate LTE-3 Embeddings: (dimensions 3072)
Select an embedding type
Compute pairwise Euclidean distance between the CT registry tumor terms and WHO/NCIT Terms
Find Closest WHO/NCIT term for a given CT registry tumor term and assign it as a the standardized tumor name
Dimensionality reduction using PCA
Compute Affinity Propagation Clusters
Cluster Size analyis and Outlier detection using isolation forest and local outlier factor anylsis
Compute K-Means Clusters
Detect outlier analysis using isolation forest and local outlier factors
Within each cluster compute the closest WHO/NCIT term using Euclidean distance in the embedding space
For a given cluster is there one WHO/NCIT Term that is closest to majority of the terms in the cluster ?
No
Implies there is no WHO/NCIT Term that represents majority of the terms in the cluster
For each term in the cluster assign the WHO/NCIT Term closest to it as its standardized term
Yes
Assign that WHO/NCIT Term as standardized term for each cluster element
For each edit distance type: Lev, JW , Cosine
Compute Pairwise Edit Distance
Compute Closest Matching WHO/NCIT Terms using each edit distance
If there are mutiple terms that are closest to the terms , report them all.
Compute Affinity Propagation Clusters
Cluster Size analyis and Outlier detection using isolation forest and local outlier factor anylsis
Within each cluster compute the closest WHO/NCIT term using the edit distance metric
For a given cluster is there one WHO/NCIT Term that is closest to majority of the terms in the cluster ?
No
Implies there is no WHO/NCIT Term that represents majority of the terms in the cluster
For each term in the cluster assign the WHO/NCIT Term closest to it as its standardized term
Yes
Assign that WHO/NCIT Term as standardized term for each cluster element

Text matching methods: We used the following three edit distances to measure the differences between the tumor terms: Normalized Levenshtein, Jarro-Winkler, and cosine distance.

Closest match using edit distance: In this method, we computed the edit distance between each tumor term in the CT registry and the terms in the WHO and NCIt databases. We then identified the WHO and NCIt terms closest to each tumor term in the CT registry and mapped them as the standardized term.

Edit distance and Affinity Propagation Clustering: In this method, we computed the pairwise edit distance between each tumor term in the CT registry, WHO, and NCIt databases and used it as a divergence metric for affinity propagation clustering. For each cluster, we evaluated if they were large and performed nested clustering if necessary. In the following step we performed outlier analysis using isolation forest and local outlier factors on each cluster. Finally, for each cluster, we assigned a standardized cluster label. This was done by identifying the WHO and NCIt terms closest to each cluster member. If there was a WHO or NCIt term closest to most of the cluster members (majority), then that term was assigned as the standardized tumor name for each cluster member; otherwise, each cluster member was assigned to its nearest (in terms of edit distances) matching WHO and NCIt terms.

Text Embedding methods: Embeddings were generated for all tumor terms in the CT Registry, WHO, and NCIt databases using Open AI's text embedding models: text-embedding-ada-002 (ADA002) and text-embedding-3-large (LTE-3).

Closest match using Embeddings: We computed the Euclidean distance in the embedding space (LTE-3 or ADA002) between each tumor term in the CT registry and the terms in the WHO and NCIt databases. We then identified the WHO and NCIt terms closest to each tumor term in the CT registry and assigned them as the standardized term.

Embeddings and Clustering: In this method, we first performed principal component analysis to reduce the dimensionality of the embedding space. We then computed pairwise Euclidean distance between each tumor term in the CT registry, WHO, and NCIt databases in the PCA transformed embedding space and used it as a divergence metric for affinity propagation and K-Means clustering. For clusters formed using affinity propagation only, we evaluated if they were large and performed nested clustering if required. Then, for both clustering methods, we carried out outlier analysis and cluster label assignment as we did in edit distance based affinity propagation clustering.

Results

To evaluate each method's performance accuracies, we needed to annotate the ground truth, i.e., the appropriate standardized tumor nomenclature for each tumor name from the CT registry. These annotations are not available in the CT registry, therefore need to be manually annotated. Since it is not feasible to manually annotate all the 13,230 tumors, we arbitrarily sampled 1600 tumors from the CT registry for manual annotation, so that the accuracies of each standardization method could be estimated. However, we limited the ground truth annotation and thereby the accuracy evaluation to the 5th edition and the combined editions (3rd, 4th and 5th) of the WHO database. This is because the WHO database is considered the gold standard for tumor nomenclature. However, we provided the WHO and NCIT standardized terms for each tumor term in the CT registry as supplemental files.

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

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

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.
2. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. 2024;74:12–49.
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.
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.
5. Hunger Stephen P., Mullighan Charles G. Acute Lymphoblastic Leukemia in Children. *N Engl J Med*. ;373:1541–52.
6. Laetsch TW, DuBois SG, Bender JG, *et al.* Opportunities and Challenges in Drug Development for Pediatric Cancers. *Cancer Discov*. 2021;11:545–59.
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: 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.
9. 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.
10. Mikolov T. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:13013781*. Published Online First: 2013.
11. Incitti F, Urli F, Snidaro L. Beyond word embeddings: A survey. *Inf Fusion*. 2023;89:418–36.
12. Khattak FK, Jeblee S, Pou-Prom C, *et al.* A survey of word embeddings for clinical text. *J Biomed Inform*. 2019;100S:100057.