IMPROVING DATA QUALITY IN AN INSTITUTIONAL CLINICAL TRIAL DATA REPOSITORY TO SUPPORT PATIENT-TRIAL MATCHING

by
Jessica A. Patricoski

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, Maryland
March 2022

Abstract

Institutional clinical data repositories often suffer from poor data quality, creating demand for advanced natural language processing (NLP) tools to support secondary use tasks. The need to address specific data quality issues and link unstructured aggregated institutional clinical trial summaries with their ClinicalTrials.gov records inspired the work in this project. A modern language representation model, the Bidirectional Encoder Representations from Transformers (BERT) model, has shown promise in many NLP tasks and been the basis for other BERT-based models pre-trained with domain-specific resources. My thesis aimed to evaluate the abilities of biomedical-domain-specific BERT models to discriminate between pairs of clinical trial texts belonging to the same trial ("matches") and those belonging to different trials ("mismatches"), using trial titles and eligibility criteria (EC).

Trials records from an institutional repository were paired with trial records from the Database for Aggregate Analysis of ClinicalTrials.gov. Next, BERT and six biomedical-domain-specific BERT models computed semantic similarity scores between the trial titles and trial EC for each trial pairing.

I evaluated the models using the difference in median similarity scores between matched and mismatched pairs. I also examined model performance by analyzing the overlap between matched and mismatched pairs' kernel density estimate (KDE) plots. Lastly, I conducted exploratory analyses using different similarity score thresholds to convert score outputs into binary match/mismatch classifications and evaluated model performance using the standard metrics of recall and precision; the true negative rate and accuracy were also calculated.

SciBERT was the only domain-specific model to demonstrate a greater difference in median similarity between matched and mismatched pairs (0.153; 0.061) than BERT (0.098; 0.051).

BlueBERT had the smallest KDE overlap between matched and mismatched titles (0.057) followed by Bio+Clinical BERT (0.061) and PubMedBERT (tied with CODER; 0.066), while PubMedBERT had the smallest KDE overlap between matched and mismatched EC (0.110) followed by CODER (0.111) and BioBERT (0.122).

Bio+Clinical BERT and PubMedBERT had the best title classification performance, while Bio+Clinical BERT and CODER had the best EC classification performance.

Domain-specific models outperformed BERT in all evaluation methods used, but larger studies with more balanced datasets are required to determine the generalizability of this claim.

Primary Reader and Thesis Advisor: Dr. Taxiarchis Botsis

Secondary Reader: Dr. Harold Lehmann

Acknowledgments

I want to acknowledge and thank everyone who has supported the research and writing that comprise this thesis, including my research mentor Dr. Taxiarchis Botsis and degree advisor Dr. Harold Lehmann, and all others in the Johns Hopkins University School of Medicine Department of Biomedical Informatics and Data Science.

I also want to thank everyone at the Sidney Kimmel Comprehensive Cancer Center who supports and contributes to the Johns Hopkins Molecular Tumor Board, including Dr. Valsamo Anagnostou, Dr. Jessica Tao, Archana Balan, Kory Kreimeyer, Jonathan Spiker, Ken Hardart, and Dr. Geoffrey Zhang.

Lastly, I want to express my gratitude to my parents, friends, and fiancé, whose support and encouragement played an indispensable role in my academic achievements.

Table of Contents

List of Tables

List of Figures

Chapter 1 – Introduction

The Johns Hopkins (JH) Molecular Tumor Board (MTB) is a multidisciplinary team of experts who review tumor and liquid biopsy profiling reports to provide personalized recommendations tailored to the genetic footprint of cancers. This expert review requires access to high-quality data from the JH's clinical data warehouse and external sources, including but not limited to biomedical literature and clinical trial data. Improving the data quality of the aggregated institutional summary of clinical trials, integrating necessary trial data from external clinical trial registries and databases, and developing models to match MTB patients with institutional clinical trials are among MTB's top priorities. I focused on one of these priorities and explored information retrieval techniques to enrich the institutional clinical trial data.

Institutional clinical data repositories often suffer from incompleteness, inconsistency, and inaccuracy, which may be attributed to multiple factors, such as errors in data entry and lack of quality controls. These challenges, specifically data incompleteness and inconsistency, are common obstacles to the secondary use of biomedical data for health services and research.[1] Secondary use of biomedical and health data can support clinical research, public health initiatives, scientific discovery, and essential patient-side tasks like precision-targeted therapy matching. Improving data quality is a complex task, requiring an ecology of tools and methods. Therefore, it is essential to prioritize the development of storage technologies, natural language processing (NLP) techniques, mining tools, and other methods for addressing and reducing the amount of incomputable, unstructured, poor-quality data.[1]

The work discussed herein took place in the context of a larger project, whose goal was to link patients to relevant clinical trials at the Johns Hopkins Sidney Kimmel Comprehensive

Cancer Center (SKCCC) for which they were eligible. To do so, the MTB needed to retrieve structured trial data for the studies from ClinicalTrials.gov. Because these studies are routinely registered at ClinicalTrials.gov, it should be simple to look at the institutional summary of the study, read off the National Clinical Trial identifier (NCT ID) from within that record, and use that ID to search ClinicalTrials.gov. However, the above data quality issues were all observed in the institutional summary of clinical trial data I was provided. For example, the NCT ID field often contained incomplete and inaccurate values or was empty, making it impossible to match the SKCCC clinical trial entries with their public ClinicalTrials.gov records.

This thesis was inspired by the need to overcome this barrier of linking our local data to the national database. To accomplish that, I investigated modern language representation models that would enable me to compare the semantic similarity of texts (record elements) between the SKCCC records and ClinicalTrials.gov records.

Exploring various language models for this purpose introduced me to the Bidirectional Encoder Representations from Transformers (BERT) family, which has shown promise in solving multiple NLP problems and has a wide range of models pre-trained on various biomedical corpora. These corpora include PubMed abstracts, PMC full-text articles, clinical notes, and synthetic vocabularies.[2-11] Language model pre-training has improved NLP in the past, but the effect of *domain-specific* pre-training for semantic similarity is still relatively unexplored.

This thesis aims to evaluate and compare the abilities of biomedical-domain-specific BERT models to discriminate between clinical trial texts using official trial titles and trial eligibility criteria (EC). The following thesis sections introduce the architecture of BERT

models, outline the methods used to create the dataset and calculate semantic similarity, break down the results, and discuss the findings.

Chapter 2 – Background

2.1 BERT

Unlike other language representation models, BERT is the first unsupervised deeply bidirectional language representation model pre-trained on a plain text corpus. Most importantly, it is a contextual model, meaning it generates word representations based on the other words in a given sentence, both before and after the word of interest, hence *bidirectionally*.[2,3] The original BERT model comes in two sizes: $\text{BERT}_{\text{BASE}}$ (12 encoders) and $\text{BERT}_{\text{LARGE}}$ (24 encoders)[*]. Both configurations are pre-trained on a document-level corpus comprised of the BooksCorpus (800M words) and English Wikipedia (2.5B words).[2]

The BERT architecture is comprised of 12 (or 24) stacked encoder blocks. Encoder blocks or layers iteratively processes word vectors (i.e., vectors showing a word's position within a given sentence) and assign weights representing the word's relevance to the other words in the sentence. Each encoder learns from the previous layers and repeats the process to create a contextualized vector of relationships, relevance, and semantic meaning (called a word embedding) for each word. A final sentence embedding (i.e., a vector of semantic meaning for the entire sentence) is created by averaging the last output vector for each word.[12]

[*] Encoders are responsible for building relationships among the words in an input sentence, thus a greater number of encoder layers results in a more complex semantical representation of the sentence overall.

2.2 BERT-Based Language Models for Biomedical and Clinical Data

The success of BERT and the overall generalizability of transformer models inspired the development of many biomedical-domain-specific variations of BERT, as broad domain texts lack a robust set of medical and clinical terminology. This work evaluated and compared six such variations against the $BERT_{BASE}$ model.

BioBERT has a $BERT_{BASE}$ architecture and is initialized with the same pre-trained weights but uses two additional datasets in pre-training: PubMed abstracts (4.5B words) and PubMed Central full-text articles (13.5B words).[4] BlueBERT follows the same process as BioBERT but pre-trains on MIMIC-III clinical notes (500M words) instead of PubMed Central articles.[5,6] Unlike BioBERT and BlueBERT, Bio+Clinical BERT initializes on BioBERT and pre-trains on all notes in the MIMIC-III database (880M words).[6-8] Medical Knowledge Embedded Term Representation (CODER) initializes on PubMedBERT but, unlike the others, adds only one additional dataset to the original pre-training: the Unified Medical Language System (UMLS) Metathesaurus.[9]

All BERT-based models detailed above utilize the same vocabulary (and thus, the same weights) as $BERT_{BASE}$, BaseVocab (30K words).[7,10] SciBERT utilizes the same architecture as $BERT_{BASE}$ but uses a different vocabulary, SciVocab (capped at 30K words to match BaseVocab), constructed using SentencePiece from 1.14M Semantic Scholar full-text papers (18% computer science, 82% biomedical; 3.2B words), and pre-trains from scratch.[7,10] PubMedBERT also has a domain-specific vocabulary, but with fewer irrelevant words; the one used in this study has a vocabulary built from 30M PubMed abstracts (3.1B words) and is trained

from scratch using PubMed abstracts only.[11] Table 2.1 presents a side-by-side comparison of the pre-training details for each BERT model used in this study.

Table 2.1. Overview of the pre-training details for the seven BERT models compared in this study. Note that corpus sizes are approximate and may fluctuate between models due to varying extraction methods and pre-processing.

Model	Vocabulary	Pretraining	Corpus	Text Size
BERT	Wiki + BooksCorpus (BaseVocab)		Wiki + BooksCorpus	3.3B
BioBERT		Continual ($BERT_{BASE}$)	PubMed abstracts + PMC full-text articles	18B
BlueBERT		Continual ($BERT_{BASE}$)	PubMed abstracts + MIMIC-III (clinical notes)	4.5B
Bio+Clinical BERT		Continual (BioBERT)	MIMIC-III (all note types)	0.9B
CODER		Continual (PubMedBERT)	UMLS	15.5M
SciBERT	PMC + CS (SciVocab)	From scratch	PMC full-text articles (biomedical + CS)	3.2B
PubMedBERT	PubMed	From scratch	PubMed abstracts	3.1B

These models are often used for many NLP tasks: named entity recognition (concept /entity extraction, and other; NER), de-identification, text inference, normalization, relation extraction, relation classification, document classification, language inference, sentence similarity (SS), and question answering. Use cases include recognizing drug names/gene names/chemicals/diseases, predicting disease, identifying high-risk patients, classifying phenotypes, finding gene-disease associations, predicting outcomes, and more.[7,13] Table 2.2 details BERT model performance on multiple of the tasks listed above. The models' high performance on the SS task suggests a high capacity for semantic similarity challenges, such as discriminating between biomedical texts.

Table 2.2. Performance outcomes for some of the biomedical and clinical NLP tasks used to evaluate BERT models in previous studies. [8,9,11,14]

Study	Task	Dataset	BERT	BioBERT	SciBERT	Blue BERT	PubMed BERT	Bio+Clinical BERT	CODER
Alsentzer et al.[8]	TI	MedNLI [a]	77.6%	80.8%	-	-	-	82.7%	-
	RE	i2b2 [b]	*79.7	*82.7	-	-	-	*83.1	-
	De-ID	i2bw [b]	**93.4	**93.9	-	-	-	**93.6	-
Peng et al.[14]	RE	i2b2 [b]	-	72.2	-	76.4	-	-	-
	TI	MedNLI [a]	-	80.5%	-	84.0%	-	-	-
	NER	ShARe/ CLEFE [b]	-	72.8	-	77.1	-	-	-
	NER	BC5CDR Disease [b]	-	85.9	-	85.4	-	-	-
Gu et al.[11]	NER	BC5CDR Disease [b]	81.4	84.7	84.5	83.7	85.6	83.0	-
	SS	BIOSSES [c]	82.7	89.5	86.3	85.4	90.4	91.2	-
Yuan et al.[9]	No	Cadec [d]	25.7	17.6	18.0	20.6	15.9	18.4	76.2
	No	PsyTar [d]	19.8	11.0	15.1	18.1	13.8	15.5	71.1

*Average performance on the i2b2 2010 concept extraction and i2b2 2012 entity extraction challenges
**Average performance on the i2b2 2006 1B and i2b2 2014 7A de-identification challenges
Metrics: a (accuracy), b (F1), c (Pearson), d (top-3 accuracy)

TI = Text Inference
RE = Relation Extraction
De-ID = De-identification
NER = Named Entity Recognition
SS = Sentence Similarity
No = Normalization

Chapter 3 – Methods

My goal was to determine which BERT models could best discriminate between pairs of clinical trial records describing the same trial vs. pairs describing different trials. I anticipated that this strategy would help me identify the highest-performing model(s) that might efficiently link SKCCC and ClinicalTrials.gov trials through textual elements, such as trial titles, eliminating the need for NCT IDs.

3.1 Clinical Trial Data Sources

There were two sources of data in this study: clinical trial records from the Sidney Kimmel Comprehensive Cancer Center and clinical trial records from the Database for Aggregate Analysis of ClinicalTrials.gov.[15,16]

3.1.1 Source 1: Sidney Kimmel Comprehensive Cancer Center

As outlined in Chapter 1, the first source of clinical trial records was an unstructured aggregated summary of SKCCC clinical trials retrieved from an institutional repository. The corpus included all clinical trials registered as of June 1, 2021 (N=2,118) with missing NCT IDs, incorrectly formatted IDs, or unrecognized IDs. Trials with valid NCT IDs or empty EC fields were excluded from the final SKCCC set of trials.

3.1.1.1 Data Quality Improvement Efforts

Before using the SKCCC trial data for BERT analysis, I pre-processed the provided data to improve their quality by correcting spelling and grammatical errors and removing non-ASCII characters.

Each trial record was expected to have 16 data fields (study number, IRB number, study status, study domain, department, sponsor, NCT ID, primary investigator, coordinators, nurses, sites, title, purpose, EC, treatment information, and keywords). However, records had their information stored in as few as 1 or as many as 89 fields due to formatting issues. Additionally, 16 trials had study locations in the title field. Most of the inconsistencies came following the "purpose" and within the "EC" and "treatment information" fields that contained the pipe symbol, generally used as the delimiter in the SKCCC data export. I, therefore, elected to combine all fields after the "purpose" field (fields 13-73) and removed the keywords, which were preceded with "|" delimiters. This method was not foolproof, as some trials had their EC information in the purpose field, but it maximized the number of trials whose EC could be captured while minimizing the amount of noise; both were paramount in this exploration.

3.1.2 Source 2: Database for Aggregate Analysis of ClinicalTrials.gov

The second set of clinical trial records originated from the Clinical Trials Transformation Initiative (CTTI)'s Database for Aggregate Analysis of ClinicalTrials.gov (AACT).[15,16] The AACT is publicly available and refreshed daily to represent up-to-date information for every clinical trial registered with ClinicalTrials.gov, a registry of human clinical research studies operated by the National Library of Medicine at the National Institutes of Health.[16] For this project, I downloaded a static copy of the database (387,486 trials) as a pipe-delimited flat file on August 23, 2021, and extracted the NCT IDs, official study titles, and EC. All static copies of the live AACT database are available in the CTTI AACT Archive.[17]

3.2 Creating the Clinical Trial Pairs

This step aimed to create a dataset of trial pairings for the seven BERT models to assess. The idea was to find the closest match for an SKCCC entry from the AACT trials using a generic model and then manually review the pairs to determine whether they matched. Once there was a collection of matching and mismatching pairs, the models could compute their level of similarity and support the evaluation of their performance based on scores they assigned to matches vs. scores they assigned to mismatches (next step).

To create pairs of clinical trial text between the clinical trials in our registry (Source 1) and those registered with ClinicalTrials.gov (Source 2), I conducted a "semantic search," which is an NLP method for identifying which sentence in a large corpus is most similar to a given querying sentence.[18] In this case, the SKCCC trial titles were the querying sentences, and the collection of AACT trial titles was the large corpus. Note that only trial titles (no EC) were used to create pairs.

After basic pre-processing for all trial titles (from both sources), I embedded all titles from the AACT corpus into a vector space using SentenceTransformers and a pre-trained sentence embedding model. [9,18,19] In other words, a sentence embedding (vector of the title's semantic meaning) was generated for each title and placed in a shared vector space.[20] I chose SentenceTransformers, a Python framework for sentence and text embedding that facilitates semantic text comparison, which was selected for this task because of its simple implementation and high-performance, low-runtime general-purpose pre-trained models.[18,21] At the date of download (August 23, 2021), the selected model (paraphrase-mpnet-base-v2) was the top-performing SentenceTransformers model recommended for symmetric semantic searches.

Once the corpus was established, each SKCCC title was independently embedded into the corpus's vector space and compared to the surrounding embeddings. Distances between embeddings were then calculated using cosine similarity, and the SKCCC trials were matched to the AACT trial associated with the closest corpus embedding. A visual representation of this process is presented in Figure 3.2. After manually reviewing the 689 pairs, I labeled each pair as either belonging to the same trial (hereafter, a "match") or different trials (hereafter, a "mismatch"). A brief review of the terminology is as follows:

Trial pair – a set of two trial records (one from SKCCC and one from AACT) and their trial data (official title and eligibility criteria).

Match **–** a trial pair where both trial records represent data from the same trial.

Mismatch **–** a trial pair where the trial records represent data from different trials.

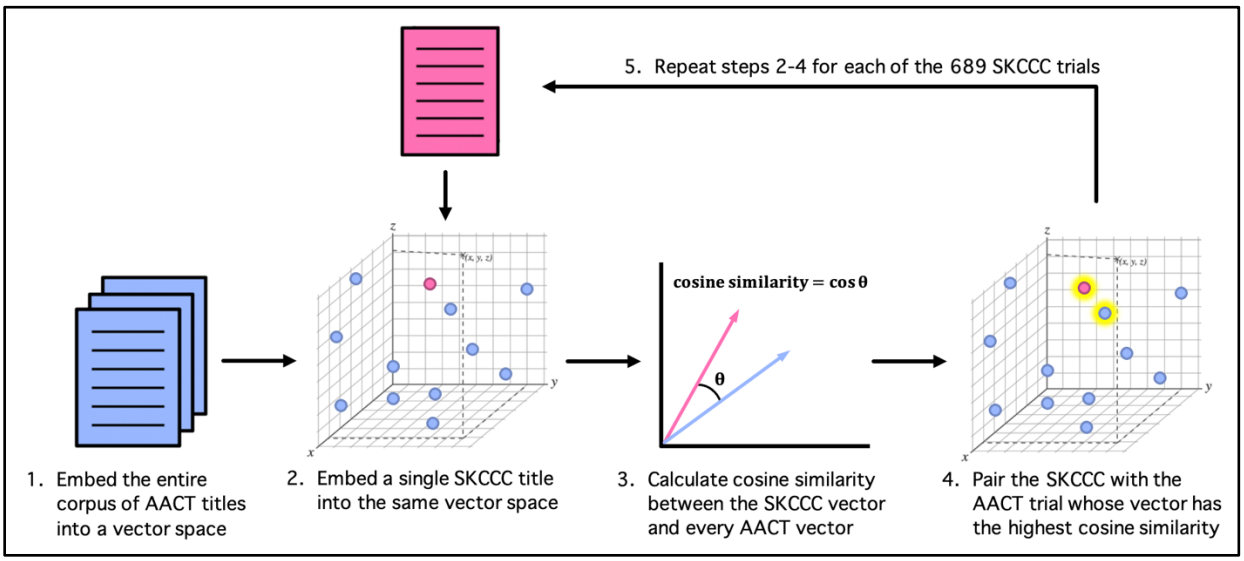

Figure 3.2. A visual representation of the pair creation process used to create the dataset for this study.

3.3 Calculating Semantic Similarity

Once the trials were paired, the BERT models assessed semantic similarity between each pair by computing embeddings for the raw title and EC texts (by averaging the output of each word's last layer) and then calculating cosine similarity between the embeddings. All seven models were sourced from the publicly available Hugging Face Model Hub and deployed using the open-source NLP library, Transformers.[22-29]

3.4 Evaluation

Comparing BERT models could mean finding the model that gave the highest similarity score or discriminated the most between matching and non-matching. The discrimination measures included the difference in median similarity scores for matched and mismatched pairs and the overlap in score distribution between matched and mismatched pairs as determined by kernel density estimate (KDE) plots. It is important to note that when extracting SKCCC trial data from the institutional database, EC could not be separated from treatment information, and as a result, the EC field contained noise that might have affected similarity calculations.

3.4.1 Evaluation Using Binary Classification

Another method I used to assess model performance was converting each model's similarity score outputs into binary classifications (0 = different trials, 1 = same trial) and using the predicted labels to calculate recall (true positive rate; TPR), precision (positive predictive value; PPV), true negative rate (TNR), and accuracy.

I used this method twice. For the first iteration, I considered each model's median similarity value (for all pairs) to be the match/mismatch cutoff point because the wide variation

among similarity ranges would have made applying a universal arbitrary threshold (e.g., 0.5) to all models unreliable. For the second iteration, I found and used the optimal cutoff value for each model. I defined the "optimal" similarity score threshold as the point at which Youden's J statistic was maximized. In other words, I calculated Youden's J statistic at every similarity score threshold between 0.000 and 1.000 and chose the cutoff that yielded the largest J statistic value.

Chapter 4 – Results

The following section discusses the characteristics of the dataset, the similarity scores calculated by each model, and the results of the binary classification exploratory analysis.

4.1 Dataset Outcomes

Of the 2,118 clinical trials in the SKCCC institutional database, 23 (1.1%) trials had null EC fields, and 1,406 (66.4%) trials had valid NCT IDs, resulting in 712 trials requiring the BERT-based effort. Subcategory distributions for the 689 included SKCCC trials are available in Figure 4.1. Some of the most misspelled words in the SKCCC texts were trial, platelet, aggressive, colorectal, steroid, interstitial, approximately, stabilize, androgen, discretion, measurable, receive, separate, persistent, and life-threatening.

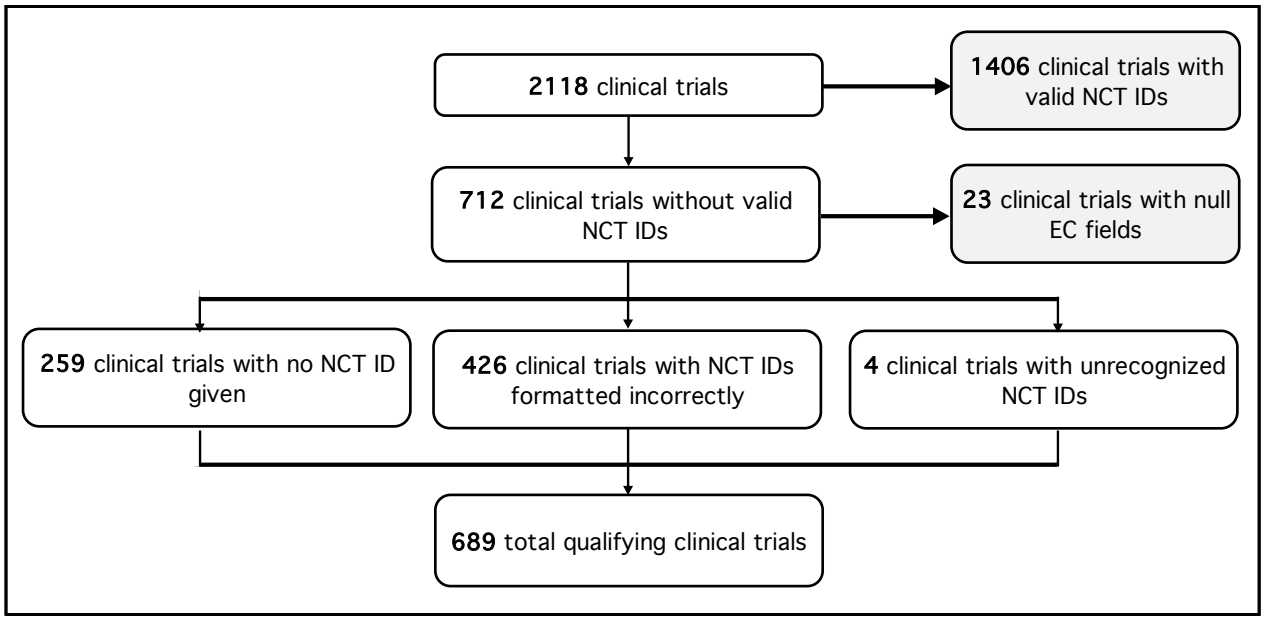

Figure 4.1. A visual breakdown of the SKCCC (Source 1) clinical trials. Grey boxes indicate excluded trials.

Following pair creation, a manual review revealed that of the 689 SKCCC-AACT clinical trial pairings, 603 (87.5%) were matches, and 86 (12.5%) were mismatches. See Table 4.1 for various examples of pair titles, including matches with dissimilar titles and mismatches with similar titles. An equivalent table for pair EC is available in Appendix A, Table A1.

Table 4.1. Examples of official trial titles from the clinical trials of matching pairs (titles belong to the same trial) and mismatching pairs (titles belong to different trials).

Matching Pairs ("Matches")		
	SKCCC	**AACT**
Visually Similar	"Phase 2 study of CGC-11047 in patients with metastatic hormone refractory prostate cancer"	"A Phase II Study of CGC-11047 in Patients With Metastatic Hormone Refractory Prostate Cancer (47-02-001)"
Visually Dissimilar	"Artesunate intravaginal inserts or ointment for the treatment of high grade dysplasia in patients with persistent recurrence of HPV disease of the lower ano-genital tract"	"A Phase I Proof-of-Concept Study of Artesunate Ointment for the Treatment of Patients With High-Grade Vulvar Intraepithelial Neoplasia (HSIL VIN 2/3)"
	"Childrens Brain Tumor Tissue Consortium (CBTTC) Collection Protocol (09-007316)"	"A Children's Oncology Group Protocol for Collecting and Banking Pediatric Brain Tumor Research Specimens"
Mismatching Pairs ("Mismatches")		
Visually Similar	"A Ph. II Evaluation of Bevacizumab in the Treatment of Persistent or Recurrent Epithelial Ovarian or Primary Peritoneal Carcinoma"	"An Indian Multicentric, Open Label, Prospective Phase 4 Study of Bevacizumab in the Front Line Management of Advanced/Metastatic Epithelial Ovarian Cancer, Fallopian Tube Cancer or Primary Peritoneal Cancer in Real-life Clinical Practice"
	"EAY131-Z1C Phase II Study of Palbociclib (PD-0332991) in Patients with Tumors with CDK4 or CDK6 Amplification (NCI-MATCH sub-protocol)"	"MATCH Treatment Subprotocol Z1B: Phase II Study of Palbociclib (PD-0332991) in Patients With Tumors With CCND1, 2, 3 Amplification"
Visually Dissimilar	"Donor Lymphocyte Infusions (DLI) plus Rapamycin to Decrease Toxicity Associated with DLI"	"Rapamycin in Relapsed Acute Lymphoblastic Leukemia"

4.2 Semantic Similarity Scores

Visual representations of the distribution of similarity scores calculated for titles and EC by the BERT-based models are shown in Figure 4.2 and Figure 4.3, respectively. Similarity-score statistics grouped by model, match status, and data type are available in Appendix A, Table A2.

SciBERT, while ranking lowest in mean similarity of matching pairs, demonstrated the largest overall difference in median similarity between matched and mismatched pairs (0.153 for titles and 0.061 for EC) and assigned lower values for mismatched pairs than all other models. The differences in the median similarity between matched and mismatched pairs, along with KDE plots and the area of overlap between curves, for all models are provided in Figure 4.4.

Outside of SciBERT, no other domain-specific model had a larger difference in median similarity than $BERT_{BASE}$, which had an overall difference in mean similarity of 0.102 for titles and 0.054 for EC. As expected, the median similarity scores for trial EC were generally lower than those for title scores, likely due to the noise present in the EC field of the SKCCC trials.

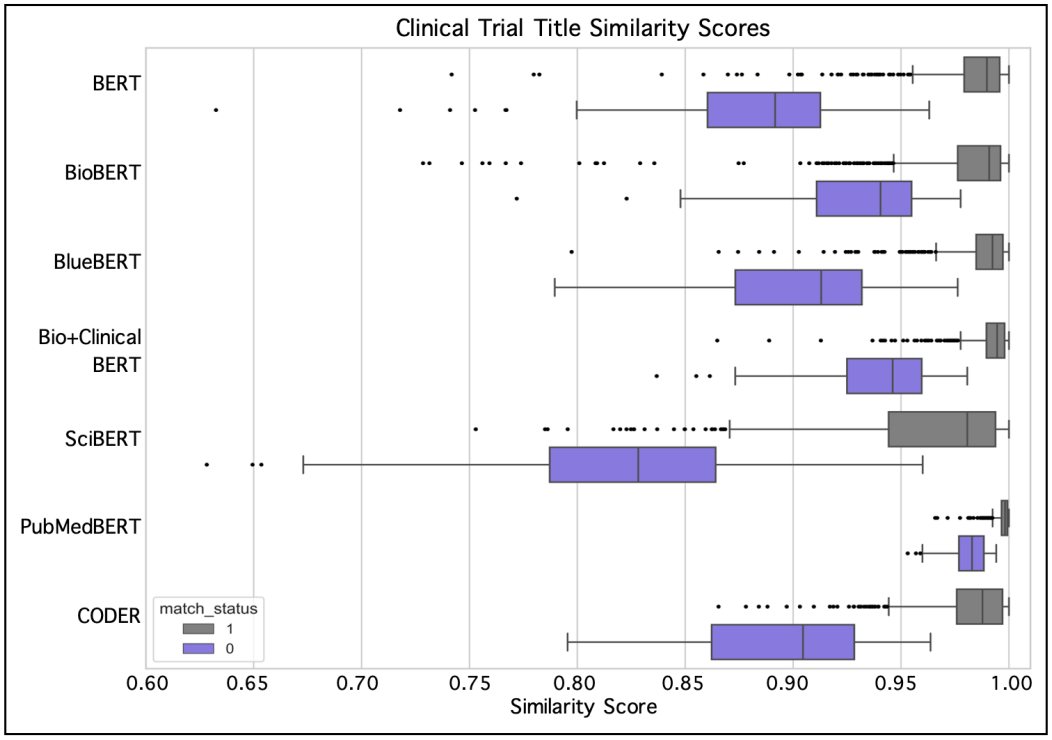

Figure 4.2. Boxplots showing the spread of similarity scores for paired clinical trial titles, grouped by match status (1 indicates the titles belonged to the same clinical trial, 0 indicates the titles belonged to different trials).

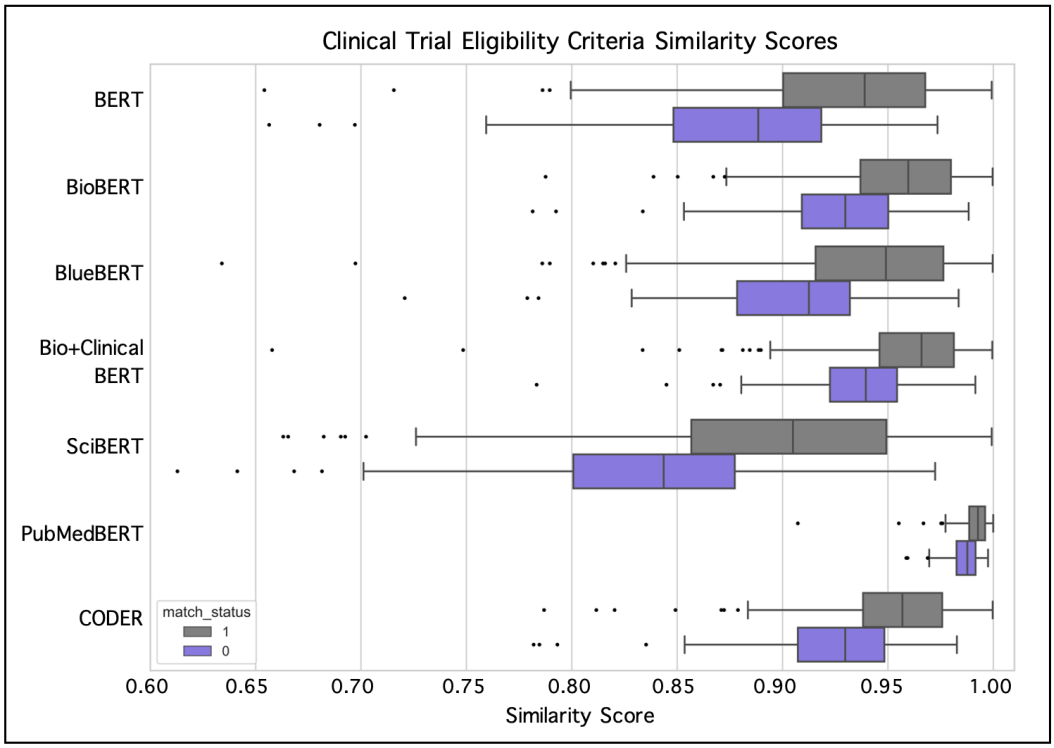

Figure 4.3. Boxplots showing the spread of similarity scores for paired clinical trial EC, grouped by match status (1 indicates the EC belonged to the same clinical trial, 0 indicates the EC belonged to different trials).

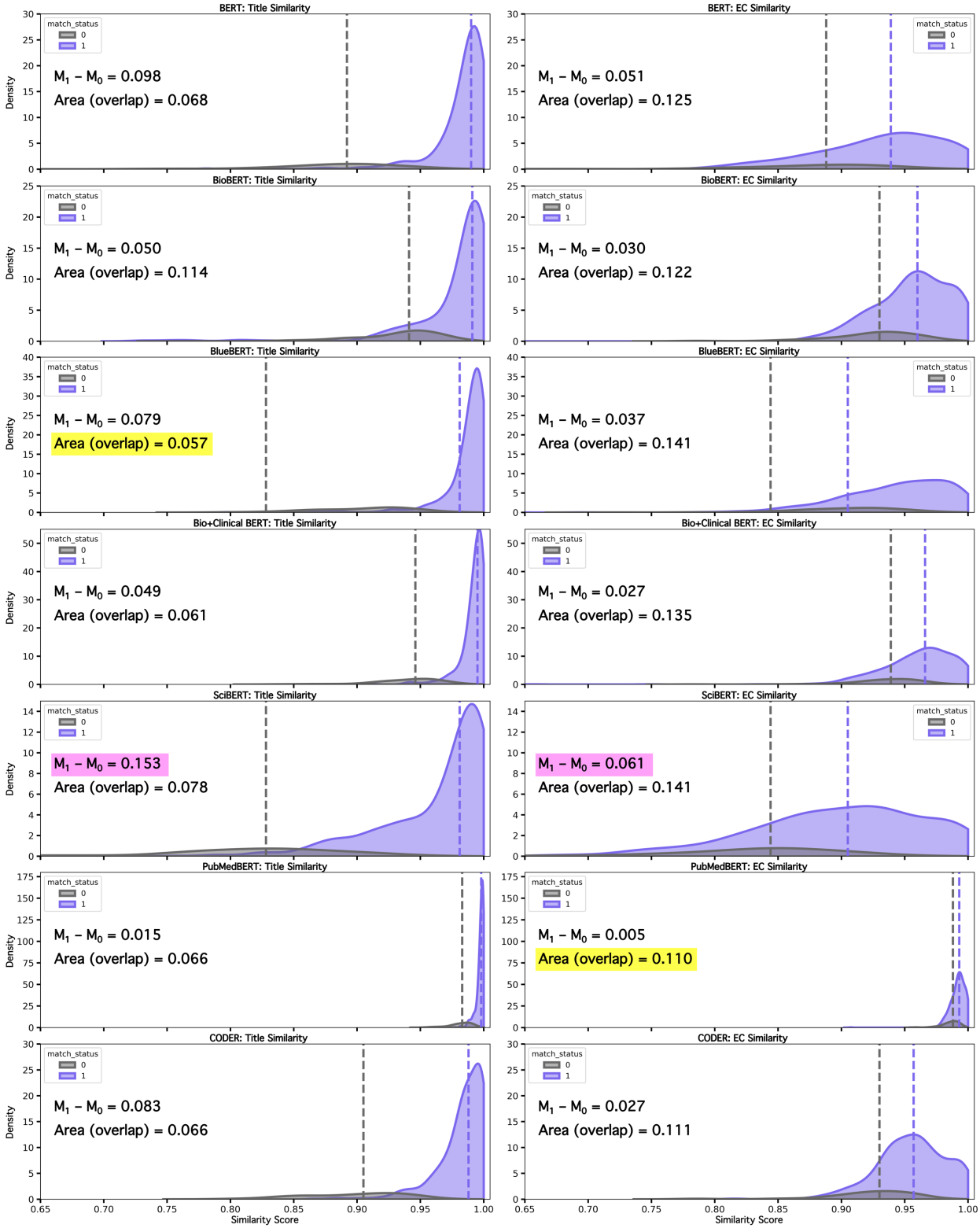

Figure 4.4. KDE plots for each model's similarity score distributions, grouped by match status and type (title plots are on the left, EC plots are on the right). Dashed lines represent the models' median scores for matches (M_1) and mismatches (M_0). The difference in median similarity and area of KDE curve overlap are listed on each plot.

Though not immediately apparent in the boxplots from Figures 4.2 and 4.3, Figure 4.4 shows that while SciBERT had the greatest median distance between similarity scores for matches and mismatches, it also had significant overlap between scores for the two categories. From this perspective, BlueBERT had the greatest title similarity performance, and PubMedBERT had the greatest EC similarity performance.

4.3 Exploratory Analysis Using Binary Classification

Using the median-cutoff technique for binary classification (see Section 3.4.1), I found SciBERT performed best overall in title classification, while BlueBERT performed best overall in EC classification. For title classification, SciBERT had the highest recall and accuracy. Interestingly, all seven models had TNRs and precisions of 1.000. For EC classification, BlueBERT had the highest recall (tied with BERT), highest TNR (tied with Bio+Clinical BERT), highest accuracy, and highest precision.

The patterns visible in these metrics (available in Table 4.2), specifically the perfect TNRs, perfect precisions, and low recalls for title classification, indicate that the median values are not the optimal cutoff points for classification. It also suggests that SciBERT has only a marginally higher probability of identifying matching titles than the other models, and all models had equal success identifying mismatching titles.

Table 4.2. Evaluation of the BERT models for match/mismatch classification, using overall median similarity scores as the threshold to create binary outputs. If the model assigned a similarity score for a given text pairing as greater than the overall median (listed in the Thresh column), it was classified as a match (1), otherwise a mismatch (0).

	Title Classification Using Median Thresholds				
	Thresh	**TPR**	**TNR**	**Accuracy**	**PPV**
BERT	0.988	0.547	1.000	0.604	1.000
BioBERT	0.974	0.556	1.000	0.611	1.000
BlueBERT	0.991	0.544	1.000	0.601	1.000
Bio+Clinical BERT	0.994	0.504	1.000	0.566	1.000
SciBERT	0.977	**0.562**	1.000	**0.617**	1.000
PubMedBERT	0.998	0.456	1.000	0.524	1.000
CODER	0.985	0.556	1.000	0.611	1.000
	EC Classification Using Median Thresholds				
	Thresh	**TPR**	**TNR**	**Accuracy**	**PPV**
BERT	0.933	**0.547**	0.860	0.586	0.965
BioBERT	0.956	0.542	0.826	0.578	0.956
BlueBERT	0.945	**0.547**	**0.872**	**0.588**	**0.968**
Bio+Clinical BERT	0.962	0.544	**0.872**	0.585	0.968
SciBERT	0.896	0.546	0.849	0.583	0.962
PubMedBERT	0.992	0.516	0.849	0.557	0.960
CODER	0.954	0.539	0.837	0.576	0.959

TPR = True Positive Rate (Recall) TNR = True Negative Rate PPV = Positive Predictive Value (Precision)

Using the optimal-cutoff technique, I found that there was no one clear top-performing model and the performance metrics were generally split between two models, reflecting similar findings to the KDE plots. The optimal thresholds and re-calculated metrics are listed in Table 4.3.

For title classification, Bio+Clinical BERT had the highest recall and accuracy while PubMedBERT had the highest TNR and precision. Unlike before, SciBERT was not the highest performer in any category. For EC classification, Bio+Clinical BERT had the highest TNR (tied with BlueBERT) and precision, while CODER had the highest recall and accuracy.

Table 4.3. Evaluation of the BERT models for match/mismatch classification, using optimal similarity score thresholds to create binary outputs. If the model assigned a similarity score for a given pairing as greater than the optimal similarity threshold (listed in the Thresh column), it was classified as a match (1), otherwise a mismatch (0).

	Title Classification Using Optimal Thresholds				
	Thresh	**TPR**	**TNR**	**Accuracy**	**PPV**
BERT	0.959	0.907	0.988	0.917	0.998
BioBERT	0.974	0.773	0.977	0.798	0.996
BlueBERT	0.958	0.940	0.977	0.945	0.996
Bio+Clinical BERT	0.972	**0.947**	0.953	**0.948**	0.993
SciBERT	0.902	0.900	0.895	0.900	0.984
PubMedBERT	0.994	0.867	**1.000**	0.884	**1.000**
CODER	0.951	0.934	0.965	0.938	0.995
	EC Classification Using Optimal Thresholds				
	Thresh	**TPR**	**TNR**	**Accuracy**	**PPV**
BERT	0.932	0.554	0.860	0.592	0.965
BioBERT	0.950	0.637	0.767	0.653	0.950
BlueBERT	0.944	0.556	**0.872**	0.595	0.968
Bio+Clinical BERT	0.959	0.584	**0.872**	0.620	**0.970**
SciBERT	0.871	0.677	0.733	0.684	0.947
PubMedBERT	0.990	0.678	0.698	0.681	0.940
CODER	0.935	**0.798**	0.616	**0.775**	0.936

TPR = True Positive Rate (Recall) TNR = True Negative Rate PPV = Positive Predictive Value (Precision)

Chapter 5 – Discussion

In trying to match imperfect institutional information with data posted to its official destination, I found that SciBERT had the highest difference in median similarity between sets of biomedical texts belonging to the same clinical trial and sets belonging to different clinical trials. However, when used for classification, it did not perform as well as its counterparts. SciBERT's poor classifier performance might be attributed to the density of its score distribution. Looking at the KDE plots, SciBERT had the greatest median distance but also a wide spread of scores, while the other models' scores were more concentrated around their medians, causing SciBERT to have the greatest area of intersect between KDE curves and the largest overlap between scores for matching and mismatching pairs (after BioBERT).

This study has three limitations. First, I found a class imbalance between the number of matching pairs (603) and mismatching pairs (86), which applies to multiple text classification tasks. Second, due to the data quality challenges of the aggregated institutional summary extraction, treatment information could not be separated from the EC, resulting in an intangible level of interference in EC similarity assessment. However, the handling of noisy information is a common challenge in free-text processing and comparison that was successfully handled by some of the selected models. Third, the BERT architecture had a maximum word length of 512, and as a result, most EC texts were not compared in full. I acknowledge that this limitation may have introduced a bias to the similarity calculations but did apply to all models that used the same word length and may have helped reduce the amount of treatment information assessed.

One important observation worth investigating in the future is the role individual vocabularies played in this study. For example, BioBERT (BaseVocab) is trained on PubMed

abstracts and PMC full-text articles, while PubMedBERT (unique vocab) is only trained on PubMed abstracts. Yet, PubMedBERT substantially outperformed BioBERT in KDE curve overlap, median difference in semantic similarity, and multiple metrics for binary classification using optimal thresholds. It is also possible that SciBERT's performance varied from the other domain-specific models due to its unique vocabulary, SciVocab, which includes computer science terminology and overlaps with BaseVocab by only 43%.[10] It is also worth considering the impact of MIMIC-III notes in pre-training, as both Bio+Clinical BERT (continuation of BioBERT plus pre-training on MIMIC-III notes) and BlueBERT (continuation of $BERT_{BASE}$ plus pre-training on MIMIC-III clinical notes) outperformed BioBERT in title classification and title KDE curve overlap.

Based on the above findings, the conclusion that domain-specific BERT models outperform $BERT_{BASE}$ in all evaluation methods used is not generalizable outside of this study. However, the strategy I took of assembling candidate pairs using different BERT models, calculating their similarity, and assessing the performance of that similarity certainly is. Further research on larger more-balanced datasets is necessary to determine the full influence of pre-training for comparing clinical trial text semantic similarity. Future research would also benefit from knowing the semantic similarity between the texts compared, as this would enable more precise model evaluation.

The next steps include assessing the feasibility of integrating semantic similarity techniques into patient-matching workflows and using advanced text mining tools to extract matching criteria, such as gene names, alterations, biomarker requirements, and cancer types.

References

1. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit Transl Bioinform*. 2010;2010:1-5. Published 2010 Mar 1.
2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of NAACL-HLT; 2019 Jun 2-7; Minneapolis, MN. Stroudsburg (PA): ACL; c2019. p. 4171-4186.
3. Delvin J. Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. *Google AI Blog*. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html. Published November 2, 2018. Accessed February 15, 2022.
4. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019 Sep;36(4):1234-40.
5. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In: Demner-Fushman D, Cohen KB, Ananiadou S, Tsujii J, editors. Proceedings of the BioNLP workshop; 2019 Aug 1; Florence, IT. Stroudsburg (PA): ACL; c2019. p. 58-65.
6. Johnson A, Pollard T, Shen L et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):160035. doi:10.1038/sdata.2016.35
7. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. Journal of Biomedical Informatics. 2019;100:100057.

8. Alsentzer E, Murphy J, Boag W et al. Publicly Available Clinical BERT Embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, MN: Association for Computational Linguistics; 2019:72-78. doi:10.18653/v1/w19-1909
9. Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: Knowledge infused cross-lingual medical term embedding for term normalization. arXiv: 2011.02947 [Preprint]. 2017 [cited 2021 Aug 20]: [11 p.]. Available from: https://arxiv.org/pdf/2011.02947.pdf
10. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2019 Nov 3-7; Hong Kong, CN. Stroudsburg (PA): ACL; c2019. p. 3615-33620
11. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. arXiv: 2007.15779v5 [Preprint]. 2020 [cited 2021 Aug 20]: [24 p.]. Available from: https://arxiv.org/abs/2007.15779
12. Fridman L. *MIT Deep Learning: State of The Art*; 2019. Available at: https://www.youtube.com/watch?v=53YvP6gdD7U&t=432s.
13. Kanakarajan K, Kundumani B, Sankarasubbu M. BioELECTRA: Pretrained Biomedical text Encoder using Discriminators. *Proceedings of the 20th Workshop on Biomedical Language Processing*. 2021:143-154. doi:10.18653/v1/2021.bionlp-1.16
14. Peng Y, Chen Q, Lu Z. An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. 2020:205-214. doi:10.18653/v1/2020.bionlp-1.22

15. Database for Aggregate Analysis of ClinicalTrials.gov (AACT). United States: Clinical Trials Transformation Initiative; 2016. https://aact.ctti-clinicaltrials.org/. Updated daily. Accessed August 23, 2021.
16. Tasneem A, Aberle L, Ananth H, Chakraborty S, Chiswell K, McCourt BJ, Pietrobon R. The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. PLoS One. 2012;7(3):e33677.
17. AACT Archive (before November 15, 2021). Clinical Trials Transformation Initiative. https://aact.ctti-clinicaltrials.org/archive. Updated November 15, 2021.
18. Reimers J, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2019 Nov 3-7; Hong Kong, CN. Stroudsburg (PA): ACL; c2019. p. 3892-3992.
19. Reimers J, Gurevych I (2021) paraphrase-mpnet-base-v2 (Version 2.0) [Source code]. https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2
20. Briggs J. Dense Vectors: Capturing Meaning with Code. Natural Language Processing (NLP) for Semantic Search. https://www.pinecone.io/learn/nlp. Published 2022. Accessed February 28, 2022.
21. Piao G. Scholarly Text Classification with Sentence BERT and Entity Embeddings. In: Gupta M, Ramakrishnan G, editors. Trends and Applications in Knowledge Discovery and Data Mining-Lecture Notes in Computer Science; 2021 May 11-14; Virtual. Switzerland (AG): Springer, Cham;12705:79-87 doi: 10.1007/978-3-030-75015-2_8
22. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2019) BioBERT: biobert-v1.1 (Version 1.1) [Source code]. https://huggingface.co/dmis-lab/biobert-v1.1/tree/main

23. Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: bert-base-uncased [Source code]. https://huggingface.co/bert-base-uncased
24. Peng Y, Yan S, Lu Z (2019) BlueBERT: BlueBert-Base, Uncased, PubMed and MIMIC-III [Source code]. https://huggingface.co/bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12
25. Alsentzer E, Murphy J, Boag W et al. (2019) Bio_ClinicalBERT: Bio+ClinicalBERT [Source code]. https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT
26. Beltagy I, Lo K, Cohan A (2019) SciBERT: scibert_scivocab_uncased [Source code]. https://huggingface.co/allenai/scibert_scivocab_uncased
27. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. (2020) PubMedBERT (abstracts only): BiomedNLP-PubMedBERT-base-uncased-abstract [Source code]. https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract
28. Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S (2017) CODER: UMLSBert_ENG [Source code]. https://huggingface.co/GanjinZero/UMLSBert_ENG
29. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, et al. Transformers: State-of-the-art Natural Language Processing. In: *Proceedings of the 2020 EMNLP*. Brooklyn, NY: Association for Computational Linguistics; 2020:38-45. Available from: https://aclanthology.org/2020.emnlp-demos.6.pdf

Appendix A

Table A1. Examples of EC from the clinical trials of matching pairs (titles belong to the same trial) and mismatching pairs (titles belong to different trials).

Matching Pairs ("Matches")		
Visually Similar	**SKCCC**	"Inclusion Criteria: - Histologic or cytologic confirmation of a solid tumor that is advanced (metastatic, recurrent and/or unresectable) with measurable disease per RECIST v1.1 - At least 1 lesion accessible for biopsy - Eastern Cooperative Oncology Group Performance Status of 0 or 1 Exclusion Criteria: - Participants with primary central nervous system (CNS) tumors, or with CNS metastases as the only site of active disease (Participants with controlled brain metastases; however, will be allowed to enroll) - Participants with active, known or suspected autoimmune disease - Participants with conditions requiring systemic treatment with either corticosteroids (> 10mg prednisone equivalents) or other immunosuppressive medications within 14 days of study treatment administration - Participants with a known history of testing positive for Human Immunodeficiency Virus (HIV) or known Acquired Immunodeficiency Syndrome (AIDS) - Cytotoxic agents, unless at least 4 weeks have elapsed from last dose of prior anti-cancer therapy and initiation of study therapy Other protocol defined inclusion/exclusion criteria could apply Study drug BMS-986253 will be administered in combination with Nivolumab at specified doses and specified intervals depending on which part of the study you are enrolled."
	AACT	"For more information regarding Bristol-Myers Squibb Clinical Trial participation, please visit www. BMSStudyConnect. com Inclusion Criteria: - Histologic or cytologic confirmation of a solid tumor that is advanced (metastatic, recurrent and/or unresectable) with measurable disease per RECIST v1.1 - At least 1 lesion accessible for biopsy - Eastern Cooperative Oncology Group Performance Status of 0 or 1 Exclusion Criteria: - Participants with CNS metastases as the only site of active disease (Participants with controlled brain metastases; however, will be allowed to enroll) - Participants with active, known or suspected autoimmune disease - Participants with conditions requiring systemic treatment with either corticosteroids (> 10mg prednisone equivalents) or other immunosuppressive medications within 14 days of study treatment administration - Participants with a known history of testing positive for Human Immunodeficiency Virus (HIV) or known Acquired Immunodeficiency Syndrome (AIDS) - Cytotoxic agents, unless at least 4 weeks have elapsed from last dose of prior anti-cancer therapy and initiation of study therapy Other protocol defined inclusion/exclusion criteria could apply"
Visually Dissimilar	**SKCCC**	"***Subjects who have existing specimens:1) Males or females of any age.2) Patient diagnosis of any nervous system tumor.3) Nervous system tumor specimens for which one of the following four conditions apply: a) The specimens are identifiable, were collected under an IRB-approved protocol and the subject consented to permit storage of specimens and data for future use consistent with the objectives of this protocol; b) The specimens are identifiable and the IRB has issued a waiver of informed consent and authorization for release and use of the specimens and data for the CBTTC Collection Protocol; c) The specimens and data are de-identified (contain no PHI); ORd) The subject provides informed consent for the use of his/her specimens and data for the CBTTC Collection Protocol and CBTTC Repository. ****Prospectively enrolled subjects:1) Males or females of any age.2) Diagnosis of any nervous system tumor including metastatic to the brain.3) Undergoing, or underwent, clinical surgery for the tumor(s) or deceased.4) Parental/guardian permission

		(informed consent) and written HIPAA authorization and if appropriate, child assent. Specimen collection will occur at time of surgery/autopsy and clinical data updates will occur per CNS Tumor diagnosis at least every 6 months until 60m post-surgery then once every 5 years thereafter per the standard of care of the specific brain tumor diagnosis.”
	AACT	“Inclusion Criteria: - Diagnosis of brain tumor - Previously treated at a Children?s Oncology Group (COG) institution - Patients are eligible at time of diagnosis, second-look surgery, recurrence, or development of a second malignant neoplasm - Must have brain tumor biological specimens derived from primary tumors of the CNS available for submission”
Mismatching Pairs (“Mismatches”)		
Visually Similar	**SKCCC**	“Patients must fulfill all eligibility criteria outlined in Section 3.1 ofMATCH Master Protocol; Patients must have CDK4 amplification or CDK6 amplification, or another aberration, as determined via the MATCH Master Protocol; Patients must have an electrocardiogram (ECG) within 8 weeks priorto treatment assignment and must have no clinically important abnormalities in rhythm, conduction or morphology of resting ECG; Patients must not have breast cancer, mantle cell lymphoma, myeloma, or liposarcoma; Patients must not have known hypersensitivity to palbociclib or compounds of similar chemical or biologic composition; Patients with known or symptoms of left ventricular dysfunction will be excluded; Patients must not have received prior therapy with a CDK4 or CDK6 inhibitor (including but not limited to palbociclib, abemaciclib, or ribociclib); Patients must not be using drugs or foods that are known potent CYP3A4 inhibitors or inducers, or are CYP3A substrates with narrow therapeutic indices. All study participants will get the same study intervention which consists of the study drug palbociclib. You will take palbociclib by mouth daily in the morning for three weeks, followed by a week off. Each 4 week time period is considered a cycle. Palbociclib should be taken with meals, and should be swallowed whole.”
	AACT	“Inclusion Criteria: - Patients must have met applicable eligibility criteria in the Master MATCH Protocol prior to registration to treatment subprotocol - Patients must have amplification of CCND1, 2, or 3, or another aberration, as determined via the MATCH Master Protocol - Patients must have an electrocardiogram (ECG) within 8 weeks prior to treatment assignment and must have no clinically important abnormalities in rhythm, conduction or morphology of resting ECG (e. g. complete left bundle branch block, third degree heart block) Exclusion Criteria: - Patients must not have known hypersensitivity to palbociclib or compounds of similar chemical or biologic composition - Patients must not have breast cancer, mantle cell lymphoma or myeloma - Patients with known or symptoms of left ventricular dysfunction will be excluded - Patients must not have had prior treatment with palbociclib, ribociclib, abemaciclib or any other CDK4/6 inhibitors - Patients must not be using drugs or foods that are known potent CYP3A4 inhibitors or inducers, or are CYP3A4 substrates with narrow therapeutic indices”
Visually Dissimilar	**SKCCC**	“About 26 patients with proven, localized adenocarcinoma of the prostate, who are eligible for 3D-CRT and who are at intermediate-risk for biochemical (PSA) failure following irradiation. There will be two treatment groups in this trial. Participants in the first treatment group will receive CG7870 on Day 1 and 3D-CRT will begin on Day 4. 3D-CRT will be administered at a daily dose of 180 cGy, five days a week, for 41 treatments, for a total dose of 7,380 cGy. During the treatment period, the patients will be seen daily for 5 days and weekly (1 day) for 9 weeks. An additional 180 cGy fraction may be added (total of 7560 cGy), if determined to be clinically indicated and safe by the investigator. If an additional fraction is given, then patients will be treated with 180 cGy daily, five days a week, for 42 treatments. Participants in the second treatment group will receive CG7870 on Day 1 and Day 22. 3D-CRT will begin on Day 4. 3D-CRT will be administered at a daily dose of 180

		cGy, five days a week, for 41 treatments, for a total dose of 7,380 cGy. During the treatment period, the patients will be seen daily for 5 days and weekly (1 day) for 9 weeks. An additional 180 cGy fraction may be added (total of 7560 cGy), if determined to be clinically indicated and safe by the investigator. If an additional fraction is given, then patients will be treated with 180 cGy daily, five days a week, for 42 treatments"
	AACT	"Inclusion Criteria: - locally confined adenocarcinoma of the prostate - all T-stages with a PSA < 60ng/ml, except any T1a tumor and well-differentiated (or Gleason score < 5) T1b-c tumors with PSA-levels ≤ 4 ng/ml - Karnofsky Performance Status of 80 or more Exclusion Criteria: - distant metastases - positive regional lymph nodes proven by surgical or cytological sampling - on anticoagulants - previous prostatectomy - previous pelvic irradiation"

Table A2. Similarity score statistics grouped by model, match status, and data type.

	Trial Matches (N=603)		**Trial Mismatches (N=86)**	
Measure	**Title**	**EC**	**Title**	**EC**
BERT				
Mean ± STD	0.983 ± 0.026	0.929 ± 0.058	0.881 ± 0.057	0.875 ± 0.069
Median (IQR)	0.990 (0.017)	0.939 (0.068)	0.892 (0.053)	0.888 (0.070)
Range [min, max]	[0.742, 1.000]	[0.253, 0.999]	[0.633, 0.963]	[0.582, 0.973]
BioBERT				
Mean ± STD	0.978 ± 0.037	0.955 ± 0.034	0.930 ± 0.037	0.924 ± 0.038
Median (IQR)	0.991 (0.020)	0.960 (0.043)	0.941 (0.044)	0.930 (0.041)
Range [min, max]	[0.728, 1.000]	[0.588, 1.000]	[0.772, 0.978]	[0.781, 0.988]
BlueBERT				
Mean ± STD	0.987 ± 0.019	0.942 ± 0.044	0.904 ± 0.039	0.903 ± 0.044
Median (IQR)	0.992 (0.012)	0.949 (0.060)	0.913 (0.059)	0.912 (0.054)
Range [min, max]	[0.797, 1.000]	[0.634, 1.000]	[0.789, 0.976]	[0.721, 0.983]
Bio+Clinical BERT				
Mean ± STD	0.991± 0.012	0.960 ± 0.031	0.939 ± 0.028	0.935 ± 0.030
Median (IQR)	0.995 (0.008)	0.966 (0.035)	0.946 (0.035)	0.939 (0.032)
Range [min, max]	[0.865, 1.000]	[0.658, 0.999]	[0.837, 0.981]	[0.783, 0.991]
SciBERT				
Mean ± STD	0.964 ± 0.043	0.897 ± 0.072	0.823 ± 0.067	0.833 ± 0.068
Median (IQR)	0.981 (0.050)	0.905 (0.092)	0.828 (0.077)	0.844 (0.076)
Range [min, max]	[0.753, 0.994]	[0.468, 0.999]	[0.628, 0.960]	[0.613, 0.972]
PubMedBERT				
Mean ± STD	0.997 ± 0.004	0.992 ± 0.007	0.982 ± 0.009	0.986 ± 0.007
Median (IQR)	0.998 (0.002)	0.993 (0.007)	0.983 (0.011)	0.988 (0.009)
Range [min, max]	[0.966, 1.000]	[0.907, 1.000]	[0.953, 0.994]	[0.959, 0.997]
CODER				
Mean ± STD	0.983 ± 0.019	0.955 ± 0.029	0.898 ± 0.040	0.922 ± 0.038
Median (IQR)	0.988 (0.021)	0.957 (0.038)	0.905 (0.066)	0.930 (0.041)
Range [min, max]	[0.865, 1.000]	[0.787, 1.000]	[0.795, 0.964]	[0.782, 0.983]