Proceedings of Machine Learning Research 219:1–18, 2023 Machine Learning for Healthcare

Scaling Clinical Trial Matching Using Large Language Models: A Case Study in Oncology

Cliff Wong CLIFF.WONG@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

Sheng Zhang ZHANG.SHENG@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

Yu Gu AIDEN.GU@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

Christine Moung CHRISTINE.MOUNG@PROVIDENCE.ORG
Providence Health & Services Molecular Genomics Laboratory
Portland, OR, USA

Jacob Abel JACOB.ABEL@PROVIDENCE.ORG
Providence Health & Services Molecular Genomics Laboratory
Portland, OR, USA

Naoto Usuyama NAOTOUS@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

Roshanthi Weerasinghe ROSHANTHI.WEERASINGHE@PROVIDENCE.ORG
Clinical Research Analytics, Providence Health & Services
Portland, OR, USA

Brian Piening BRIAN.PIENING@PROVIDENCE.ORG
Earle A. Chiles Research Institute, Providence Cancer Institute
Portland, OR, USA

Tristan Naumann TRISTAN@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

Carlo Bifulco CARLO.BIFULCO@PROVIDENCE.ORG
Earle A. Chiles Research Institute, Providence Cancer Institute
Portland, OR, USA

Hoifung Poon HOIFUNG@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

Abstract

Clinical trial matching is a key process in health delivery and discovery. In practice, it is plagued by overwhelming unstructured data and unscalable manual processing. In this paper, we conduct a systematic study on scaling clinical trial matching using large language models (LLMs), with oncology as the focus area. Our study is grounded in a clinical trial matching system currently in test deployment at a large U.S. health network. Initial findings are promising: out of box, cutting-edge LLMs, such as GPT-4, can already structure elaborate eligibility criteria of clinical trials and extract complex matching logic (e.g., nested AND/OR/NOT). While still far from perfect, LLMs substantially outperform prior strong baselines and may serve as a preliminary solution to help triage patient-trial candidates with humans in the loop. Our study also reveals a few significant growth areas for applying LLMs to end-to-end clinical trial matching, such as context limitation and accuracy, especially in structuring patient information from longitudinal medical records.

1. Introduction

Clinical trial matching identifies eligible patients to be considered for clinical trial enrollment, which is essential for clinical discovery and drug development. For diseases such as cancers where standard of care often fails, clinical trials are also a last hope and represent an important treatment option. The standard process for clinical trial matching, however, is extremely inefficient. Most information needed for accurate matching is scattered in vast amount of unstructured text, such as elaborate eligibility criteria of a clinical trial or patient information in longitudinal electronic medical records. Curation and matching are currently done predominantly by manual effort, which is difficult to scale. For example, in the United States (U.S.) alone, there are two million new cancer patients every year and, at any given moment, there may be hundreds of active oncology trials (Siegel and Jemal, 2022; U.S. National Library of Medicine, 2023). Manually evaluating all oncology trial-patient candidates is infeasible. Consequently, less than 3% of patients were able to participate in a trial (Unger et al., 2019), whereas 38% of trial failure stemmed from insufficient recruitment (Stensland et al., 2014).

Natural language processing (NLP) has emerged as a promising direction for accelerating clinical trial matching. Prior work explores rule-based methods and supervised machine learning for structuring trial eligibility criteria (Weng et al., 2011; Kang et al., 2017; Yuan et al., 2019; Nye et al., 2020) or matching clinical trials against structured patient records (Zhang et al., 2020; Gao et al., 2020). While promising, these methods still require extensive manual effort for rule development and example annotation. Recently, large language models (LLMs), such as GPT-4, have demonstrated impressive zero-shot and few-shot capabilities in both general domains (OpenAI, 2023; Bubeck et al., 2023) and health applications (Lee et al., 2023a,b; Nori et al., 2023). In this paper, we conduct a systematic study on scaling clinical trial matching with LLMs by leveraging emergent in-context learning capabilities (Brown et al., 2020).

We choose oncology as the focus of our study because it incurs a high death burden and represents a large proportion of clinical trials (over 20% in 2022[1]). Oncology trials also tend

[1] As of April 2023, `ClinicalTrials.gov` has 32,341 trials with start date 01/01/2022–12/31/2022; further filtering for ‘cancer’ as disease yields 6,704 results. This reflects general trend with 449,665 available without start date restriction, of which 96,481 results include ‘cancer’.

to contain elaborate eligibility criteria, such as complex combination logic of disease histology and genomic biomarkers, with the latter (genomic biomarkers) typically out of scope for prior state of the art (Yuan et al., 2019; Zhang et al., 2020; Gao et al., 2020). We grounded our study in a real-world clinical trial matching system currently in test deployment at a large U.S. health integrated delivery network (IDN), which comprises three key components: structuring clinical trial eligibility criteria, structuring patient information from electronic medical records, and matching (Figure 2). The patient structuring component leverages state-of-the-art self-supervised deep learning (Preston et al., 2023). The trial-structuring and matching components are rule-based expert systems that require over 450 expert hours to develop.

For consistency and direct comparison with prior methods, we focus our study of LLMs primarily on structuring trial eligibility, but we also conduct a preliminary exploration on applying LLMs to end-to-end clinical trial matching.

Initial results are promising: using merely up to three examples, state-of-the-art LLMs, such as GPT-4, can already structure elaborate trial eligibility criteria and extract complex matching logic of disease histology and genomic biomarkers. We conduct both intrinsic evaluation on trial structuring and end-to-end matching evaluation, using expert-annotated test sets and legacy enrollment data, respectively[2]. While still far from perfect, LLMs substantially outperform prior strong baselines such as Criteria2Query (Yuan et al., 2019) and demonstrate competitiveness even against the oncology-specific expert system that requires many expert hours to develop and tailor for this domain.

Our study also reveals significant growth areas for applying LLMs to end-to-end clinical trial matching, especially in structuring patient information from electronic medical records. A cancer patient may have hundreds of notes, with key information such as tumor histology and biomarkers scattered across multiple notes (Preston et al., 2023) as shown in Figure 1. Naively concatenating all potentially relevant notes will almost always exceed even the largest context size available for GPT-4: 32K tokens.[3] It might also risk overwhelming the LLM with too much irrelevant information. In this paper, we thus resort to using structured patient information extracted by the state-of-the-art self-supervised deep-learning systems (Preston et al., 2023) and evaluate the LLM's capabilities in matching such information against the trial eligibility criteria. Preliminary results are promising and we leave more in-depth exploration to future work.

Generalizable Insights about Machine Learning in the Context of Healthcare

- We present the first systematic study on using LLMs to scale clinical trial matching. While we demonstrate this approach on a case study in oncology, our findings motivate the exploration of this approach across other areas.

- Our findings indicate that out of box, state-of-the-art LLMs such as GPT-4 can already handle complex eligibility criteria of clinical trials and extract matching logic.

[2] Our code is available at https://aka.ms/ctm-llm

[3] While GPT-4 has a context length of 8,192 tokens, there is limited access to GPT-4-32K which has a context length of 32,768–context (about 50 pages of text): https://openai.com/research/gpt-4.

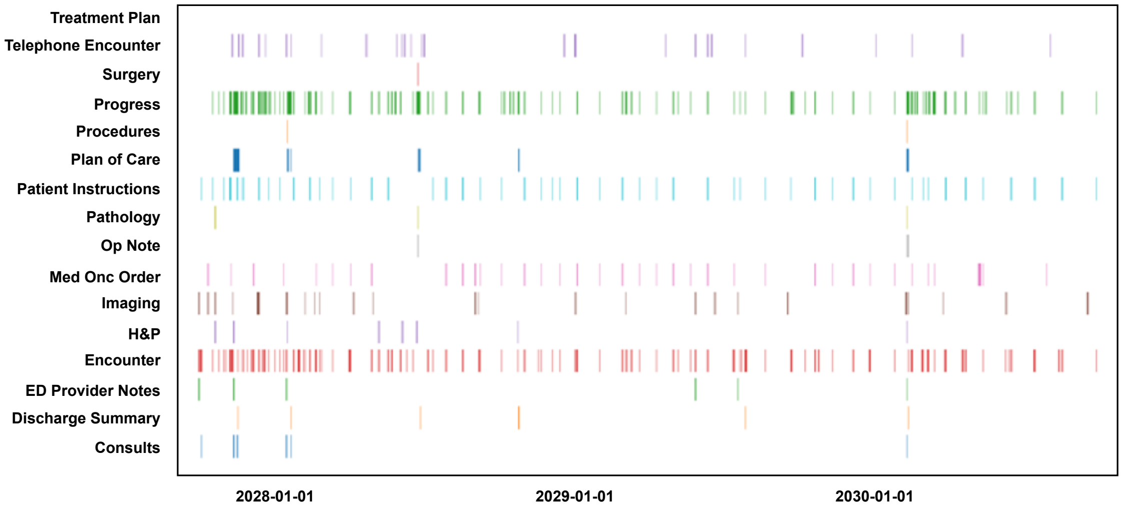


Figure 1: Patient de-identified timeline with various notes.

- We also identify several key growth areas for enhancing LLMs in end-to-end clinical trial matching, such as addressing the context limitation and accuracy issues, especially in extracting and structuring patient data from electronic medical records.

2. Related Work

The existing approaches for clinical trial matching can be divided into two categories depending on how matching is done.

Structure-then-Match Such systems first structure trial eligibility criteria by extracting key entities and relations for eligibility and then matching against structured patient information using manually crafted normalization rules to resolve superficial variations. Prior systems typically require extensive human annotations (Weng et al., 2011; Kang et al., 2017), supervised learning models (for rule extraction) (Bustos and Pertusa, 2018), or a combination of supervised learning and rules (e.g., Criteria2Query (Yuan et al., 2019)). Due to lexical variations and inadequate rule coverage, these systems often suffer from lower recall and generalizability. However, they can produce structured forms for eligibility criteria as intermediate results, which offer better interpretability and facilitate human-in-the-loop verification.

End-to-End These systems learn to encode patient and criteria for end-to-end matching via supervised learning from patient-trial or patient-criterion matching labels. For example, DeepEnroll (Zhang et al., 2020) jointly encodes patient records and trial eligibility criteria in the same embedding space, and then aligns them using attentive inference by learning from example patient-criteria matching data. COMPOSE (Gao et al., 2020) represents the state-of-the-art embedding-based model that uses hierarchical memory network to encode concepts at different granularity and differentiates between inclusion and exclusion criteria.

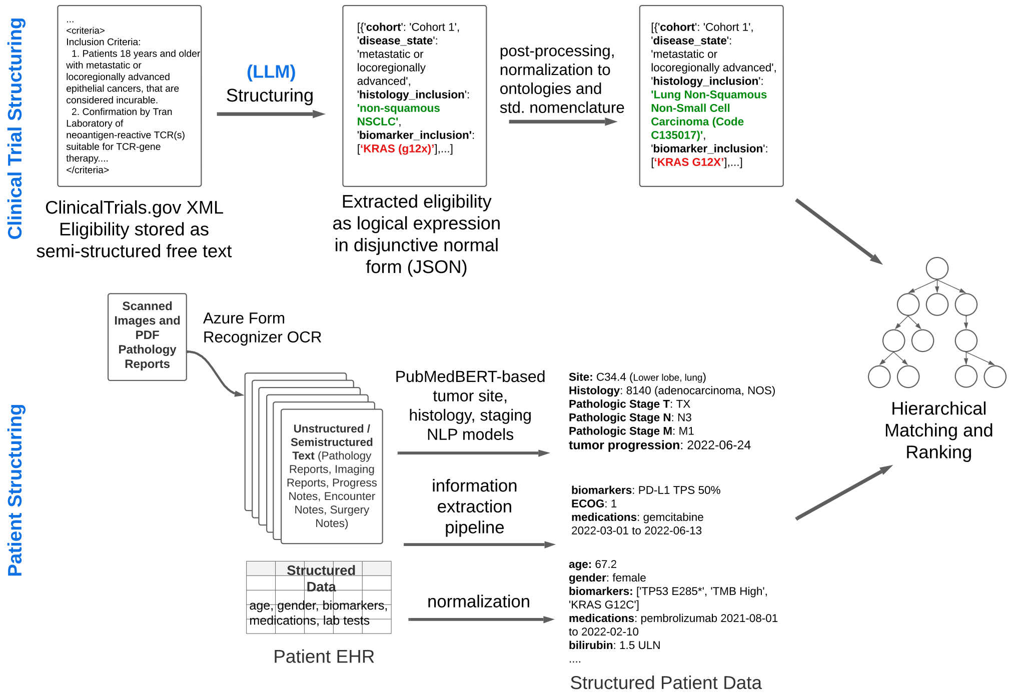

Clinical Trial Structuring
...
<criteria>
Inclusion Criteria:
1. Patients 18 years and older with metastatic or locoregionally advanced epithelial cancers, that are considered incurable.
2. Confirmation by Tran Laboratory of neoantigen-reactive TCR(s) suitable for TCR-gene therapy....
</criteria>
ClinicalTrials.gov XML Eligibility stored as semi-structured free text
(LLM) Structuring
[{'cohort': 'Cohort 1', 'disease_state': 'metastatic or locoregionally advanced', 'histology_inclusion': 'non-squamous NSCLC', 'biomarker_inclusion': ['KRAS (g12x)'],...]
Extracted eligibility as logical expression in disjunctive normal form (JSON)
post-processing, normalization to ontologies and std. nomenclature
[{'cohort': 'Cohort 1', 'disease_state': 'metastatic or locoregionally advanced', 'histology_inclusion': 'Lung Non-Squamous Non-Small Cell Carcinoma (Code C135017)', 'biomarker_inclusion': ['KRAS G12X'],...]
Hierarchical Matching and Ranking
Patient Structuring
Scanned Images and PDF Pathology Reports
Azure Form Recognizer OCR
Unstructured / Semistructured Text (Pathology Reports, Imaging Reports, Progress Notes, Encounter Notes, Surgery Notes)
Structured Data
age, gender, biomarkers, medications, lab tests
Patient EHR
PubMedBERT-based tumor site, histology, staging NLP models
information extraction pipeline
normalization
Site: C34.4 (Lower lobe, lung)
Histology: 8140 (adenocarcinoma, NOS)
Pathologic Stage T: TX
Pathologic Stage N: N3
Pathologic Stage M: M1
tumor progression: 2022-06-24
biomarkers: PD-L1 TPS 50%
ECOG: 1
medications: gemcitabine 2022-03-01 to 2022-06-13
age: 67.2
gender: female
biomarkers: ['TP53 E285*', 'TMB High', 'KRAS G12C']
medications: pembrolizumab 2021-08-01 to 2022-02-10
bilirubin: 1.5 ULN
....
Structured Patient Data

Figure 2: Overall schematic of matching.

Yuan et al. (2023) follows COMPOSE but uses LLMs to generate semantically equivalent criteria for data augmentation.

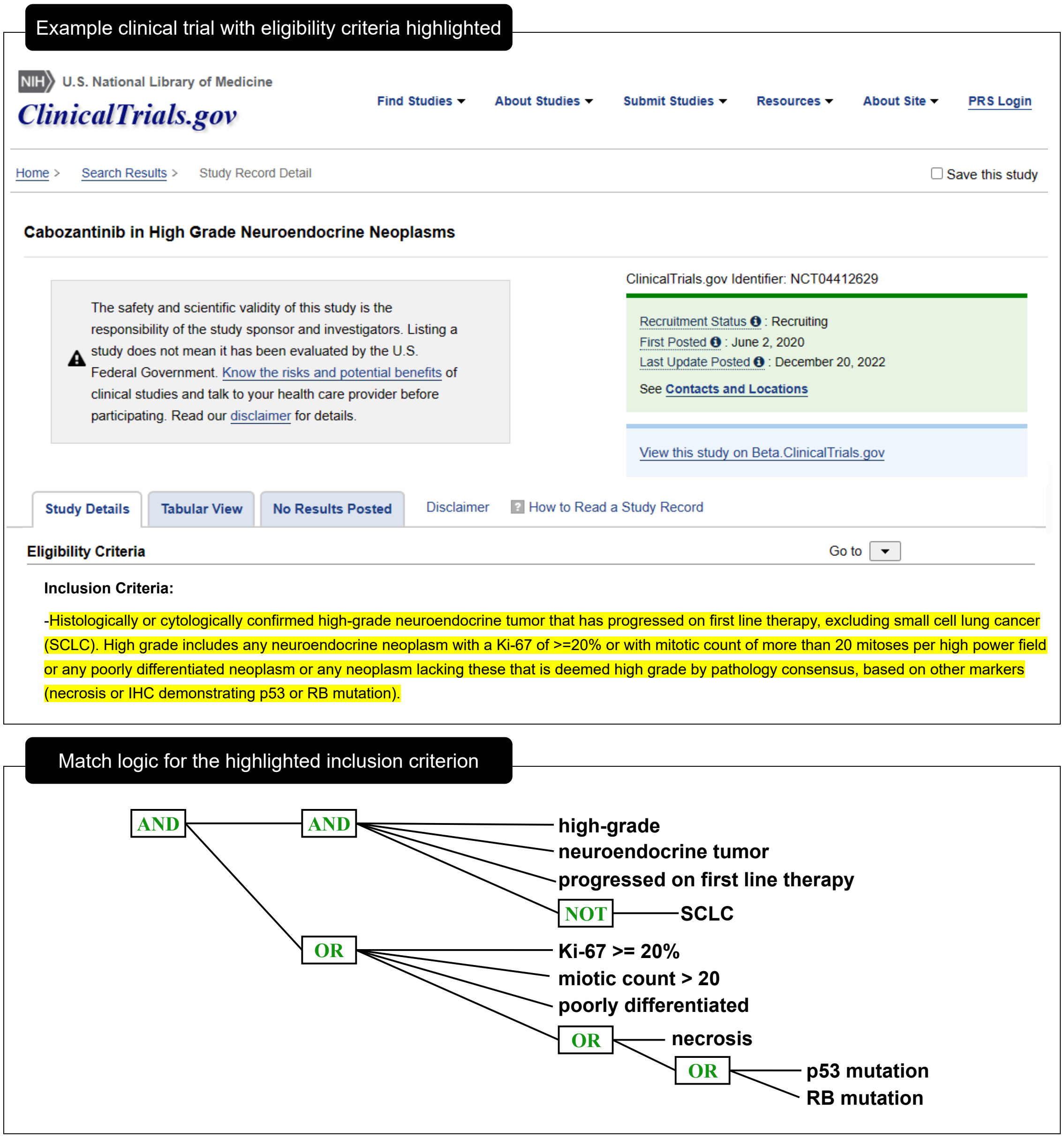
To the best of our knowledge, we are the first to explore using the emergent in-context learning capability of LLMs (Brown et al., 2020) for clinical trial matching. Out of box and with no more than three examples, cutting-edge LLMs, such as GPT-4, can already structure trial eligibility criteria with reasonable performance, outperforming strong baselines from prior systems such as Criteria2Query. We also show preliminary results in applying LLMs to end-to-end clinical trial matching.

3. Methods

In this section, we introduce the problem formulation for clinical trial matching and then describe the relevant systems and evaluation.

3.1. Problem Formulation

Given a clinical trial T and a patient P, clinical trial matching determines if P satisfies all eligibility criteria specified by T. The specification $T_{\texttt{spec}}$ generally comprises semi-structured text (see Figure 3 "Match logic"). Similarly, the patient information from electronic medical records (EMRs) $P_{\texttt{EMR}}$ also contains many free-text clinical documents, such as pathology



Example clinical trial with eligibility criteria highlighted

Cabozantinib in High Grade Neuroendocrine Neoplasms

The safety and scientific validity of this study is the responsibility of the study sponsor and investigators. Listing a study does not mean it has been evaluated by the U.S. Federal Government. Know the risks and potential benefits of clinical studies and talk to your health care provider before participating. Read our disclaimer for details.

ClinicalTrials.gov Identifier: NCT04412629

Recruitment Status : Recruiting
First Posted : June 2, 2020
Last Update Posted : December 20, 2022

See Contacts and Locations

View this study on Beta.ClinicalTrials.gov

Eligibility Criteria

Inclusion Criteria:

-Histologically or cytologically confirmed high-grade neuroendocrine tumor that has progressed on first line therapy, excluding small cell lung cancer (SCLC). High grade includes any neuroendocrine neoplasm with a Ki-67 of >=20% or with mitotic count of more than 20 mitoses per high power field or any poorly differentiated neoplasm or any neoplasm lacking these that is deemed high grade by pathology consensus, based on other markers (necrosis or IHC demonstrating p53 or RB mutation).

Match logic for the highlighted inclusion criterion

Figure 3: Example clinical trial (NCT04412629) with eligibility criteria highlighted and the match logic for the inclusion criterion.

reports, radiology reports, and encounter notes. To determine the match, there are two approaches:

- Structure-then-match: first structure both trial specification and patient information into common ontologies by extracting the relevant information, then execute the matching logic;

- End-to-end: learn to encode the trial and patient information directly and determine match by computing the similarity of embeddings.

DeepEnroll (Zhang et al., 2020) and COMPOSE (Gao et al., 2020) are representative examples of the end-to-end approach. Learning an accurate encoding is particularly challenging for patient, given the vast amount of potentially relevant medical records. Consequently, prior work focuses on a more limited setting where the patient records are already structured. Still, such methods require large amount of training data of known patient-trial matches (or patient-criteria matches) and their generalizability to unseen disease areas is underexplored. It is also hard for human experts to interpret or verify such end-to-end results.

In this paper, we instead focus our study on "structure-then-match". Let $T_{\texttt{structured}}$ denote the structured representation of the trial eligibility criteria and $P_{\texttt{structured}}$ denote that of a patient. Then clinical trial matching reduces to three subtasks:

- Structuring trial eligibility criteria: $T_{\texttt{spec}} \rightarrow T_{\texttt{structured}}$;
- Structuring patient records: $P_{\texttt{EMR}} \rightarrow P_{\texttt{structured}}$;
- Matching: $\texttt{Match}(T_{\texttt{structured}}, P_{\texttt{structured}}) \rightarrow \{\texttt{yes}, \texttt{no}\}$;

Matching is relatively straightforward, assuming that structuring has been done well, so we will focus on the structuring tasks. At the time of this study, it is not yet possible to applying cutting-edge LLMs, such as GPT-4, to protected health information (PHI) data in the environment available. Therefore, in the remainder of the paper, we focus on the structuring task for clinical trial eligibility criteria ($T_{\texttt{spec}} \rightarrow T_{\texttt{structured}}$).

3.2. Structuring Clinical Trial Eligibility Criteria

Eligibility criteria comprise inclusion and exclusion criteria. There is an implicit `AND` logic over components in the inclusion criteria, and an implicit `NOT AND` logic over components in the exclusion criteria. Additionally, there may be complex nested logic on combinations of patient attributes.

For oncology, a salient example is the interplay between disease histology (fine-grained subtypes of disease states) and genomic biomarkers (e.g., genetic point mutation). See Figure 3 for an example criterion. By contrast, standard demographic attributes (e.g., age, gender) are straightforward. We omit them for simplicity and focus our evaluation on disease histology and biomarkers.

3.3. Systems

Our main focus of the paper is to explore applying cutting-edge LLMs, such as GPT-3.5 and GPT-4, to the structuring tasks. For head-to-head comparison, we also consider a strong baseline using state-of-the-art biomedical entity extraction systems, prior state of the art Criteria2Query (Yuan et al., 2019), and an expert system in test deployment at a large health network.

LLM We use GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023) through Azure OpenAI Service,[4] which allows us to interact with the models efficiently and in HIPAA-compliant way. Through the API, we submit prompts and receive generated responses, which we postprocess to the requisite structured format.

Biomedical Entity Extraction In the absence of a general model such as LLMs, we can combine state-of-the-art biomedical entity extraction systems to assemble a strong baseline. Specifically, we consider SciSpaCy (Neumann et al., 2020) for extracting disease states, GNormPlus (Wei et al., 2015) for extracting gene entities, and tmVar (Li et al., 2013) for extracting genetic variants.

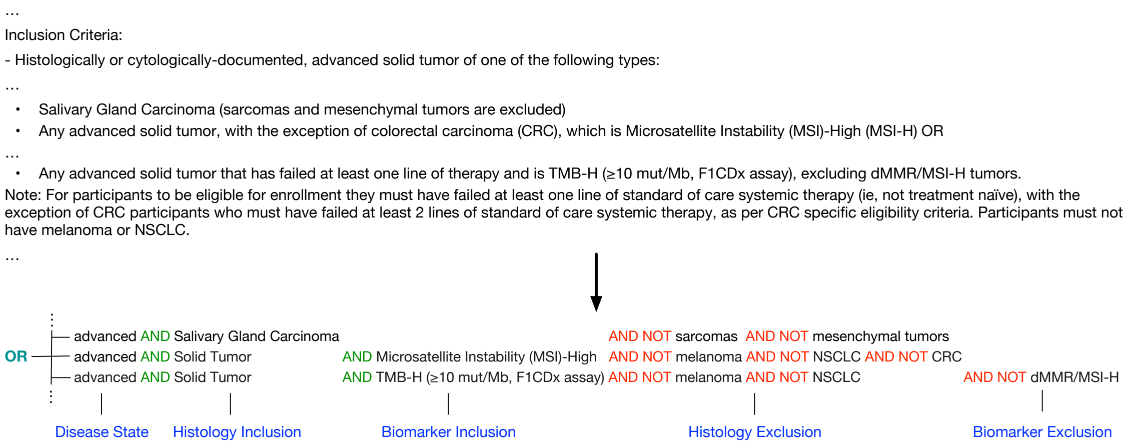
Criteria2Query Yuan et al. (2019) provide a standard baseline for structuring eligibility criteria. Criteria2Query combines supervised learning and rule-based heuristics and prior work shows state-of-the-art results in extracting disease states. We found that it can also extract biomarker information, although the capabilities are limited.

Expert System Some authors of this paper have previously developed an expert system for clinical trial matching, in collaboration with a large health network, where this system is currently in test deployment. This system comprises three key components: structuring trial eligibility criteria, matching against structured patient information, and a clinical trial matching application with a human-in-the-loop assisted clinical trial triaging user interface (UI). The structuring component encapsulates extensive heuristics for extracting and normalizing biomedical entities such as disease histology, genomic biomarkers, etc., as well as heuristics for processing semi-structured information (e.g., sections) and extracting matching directives (`AND`/`OR`/`NOT`). In internal evaluation, this system has demonstrated state-of-the-art performance and has been well received in test deployment by molecular tumor boards and trial coordinators. However, as common for rule-based approaches, this system has taken over 450 expert hours to develop and is specifically tailored for oncology (e.g., by focusing its structuring on disease histology and biomarkers). Exploring a more scalable approach thus represents an attractive direction for expanding to more fine-grained details and generalizing to other disease areas.

3.4. Evaluation

Intrinsic Evaluation For structuring trial eligibility criteria, we consider two settings. In the basic setting, we evaluate on entity extraction only, focusing on histology and biomarkers. Specifically, we compare system output against the gold entity set and report precision (i.e., positive predictive value), recall (i.e., sensitivity), and F1 (harmonic mean of precision and recall). This is similar to the approach taken by prior work. In the more advanced setting, we evaluate on extraction of the complete matching logic. Essentially, we regard structuring as a semantic parsing problem, where the structuring output is a logic form as in Figure 3 (see "Match logic"). To facilitate evaluation, we normalize the logic form to disjunctive normal form (DNF) (i.e., `OR` of `AND`s) and report precision/recall/F1 of the disjunctions. See Figure 4 for an example.

[4] https://azure.microsoft.com/en-us/products/cognitive-services/openai-service/



...

Inclusion Criteria:

- Histologically or cytologically-documented, advanced solid tumor of one of the following types:

...

- Salivary Gland Carcinoma (sarcomas and mesenchymal tumors are excluded)
- Any advanced solid tumor, with the exception of colorectal carcinoma (CRC), which is Microsatellite Instability (MSI)-High (MSI-H) OR

...

- Any advanced solid tumor that has failed at least one line of therapy and is TMB-H (≥10 mut/Mb, F1CDx assay), excluding dMMR/MSI-H tumors.

Note: For participants to be eligible for enrollment they must have failed at least one line of standard of care systemic therapy (ie, not treatment naïve), with the exception of CRC participants who must have failed at least 2 lines of standard of care systemic therapy, as per CRC specific eligibility criteria. Participants must not have melanoma or NSCLC.

...

OR
advanced AND Salivary Gland Carcinoma AND NOT sarcomas AND NOT mesenchymal tumors
advanced AND Solid Tumor AND Microsatellite Instability (MSI)-High AND NOT melanoma AND NOT NSCLC AND NOT CRC
advanced AND Solid Tumor AND TMB-H (≥10 mut/Mb, F1CDx assay) AND NOT melanoma AND NOT NSCLC AND NOT dMMR/MSI-H
Disease State
Histology Inclusion
Biomarker Inclusion
Histology Exclusion
Biomarker Exclusion

Figure 4: Example of disjunctive normal form (DNF). In DNF formulas must be a disjunction (OR) of conjunction (AND).

End-to-End Evaluation Given the structured trial information, we can simulate an end-to-end evaluation by triangulating it with structured patient information via a common matching component to determine end-to-end match results. We can then compare these results against a gold match dataset.

4. Experiments

4.1. Datasets

Clinical Trial We downloaded all the clinical trials in XML format from ClinicalTrials.gov (U.S. National Library of Medicine, 2023). In this work, we focus on oncology trials with "Treatment" as the primary purpose and "Interventional" as the study type. We use a list of oncology-related keywords to filter out non-oncology trials. For each trial, we extract the following XML elements: `brief_title`, `official_title`, `brief_summary`, `arm_group`, and `criteria`. We truncate the criteria element to the first 40 lines and apply heuristics to remove lines that are not related to oncology-specific criteria (disease histology and biomarkers).

Structured Eligibility Criteria To create the gold dataset for evaluating the structured criteria, three molecular pathologists at our collaborating health network manually structured 53 clinical trials as the test set. These 53 clinical trials were randomly selected from a larger set of trials that had significant enrollment at the health network and were more likely to contain complex criteria. The output organizes the criteria into five categories: disease state, histology inclusion, biomarker inclusion, histology exclusion, and biomarker exclusion. The criteria logical expression is in disjunctive normal form (DNF), which consists of a disjunction of conjunctive clauses (OR of ANDs). See Figure 4. Given the extracted histology terms and biomarkers, we further normalize them into the NCI Thesaurus ontology[5] and HGVS nomenclature[6].

Structured Patient Information To facilitate simulated end-to-end matching evaluation, we leverage the component for structuring patient information in the expert system at our collaborating health network (see Section 3.3). Briefly, this component system uses the Read OCR model in Azure Form Recognizer[7] to convert any scanned images and PDF documents into free text, and then appends such text to digitized medical records. It then applies a series of state-of-the-art biomedical NLP models to extract relevant patient attributes from the medical records. For example, it uses self-supervised PubMedBERT models (Preston et al., 2023) to extract the tumor site, histology, and staging information, as well as additional information extraction modules to extract other attributes such as health status, PD-L1 IHC result, and medications. Information available in structured EMRs is added directly, such as date of birth and gender. Other structured data fields such as lab tests and medications are normalized to the same units and NCI Thesaurus, respectively.

Historical Trial Enrollment Data Given a known patient enrollment into a trial, we can treat it as a gold label and test if a clinical trial matching system can correctly flag it as a match. In this way, we can estimate the recall for end-to-end matching. For this purpose, we use a dataset containing 523 patient-trial enrollment pairs at our collaborating health system. This is a subset of all historical enrollment data, after filtering out confidential trials and patients who do not have sufficient structured information.

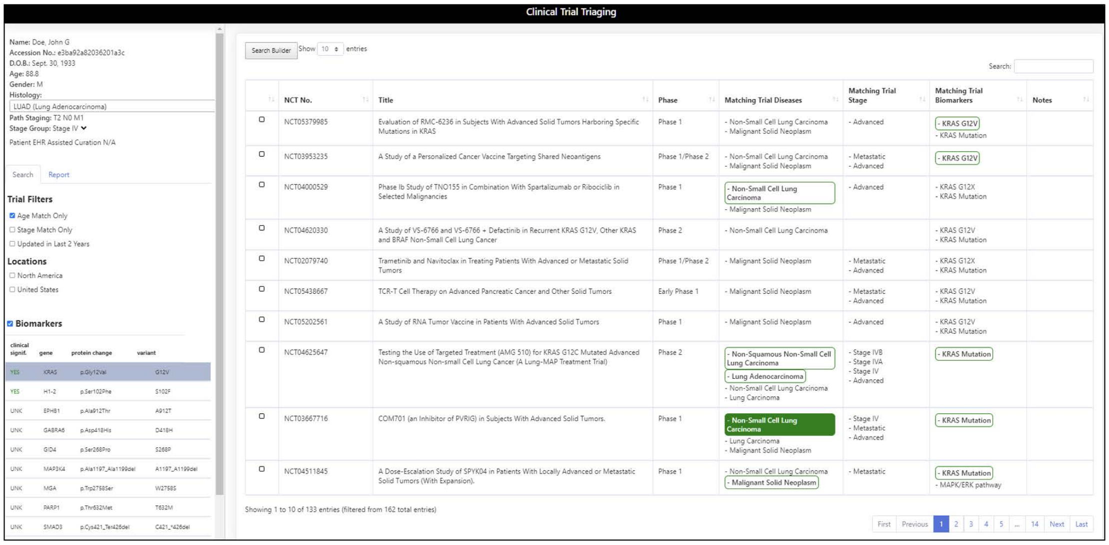
Clinical Trial Matching System Feedback Data As mentioned in Section 3.3, a clinical trial matching application is in test deployment at our collaborating health network (see Figure 5), which allows molecular pathologists to inspect patient-trial match candidates and select the most suitable trials for follow-up. We used molecular pathologists' selection as implicit feedback and consider selected candidates as positive and non-selected ones as negative. If no trials were selected for a patient, we skipped all candidates for this patient, as the molecular pathologists may not have inspected them yet. This yielded a dataset with 68,485 candidates with 84% positive.

Human Subjects, IRB, Data Security and Patient Privacy This work was performed under the auspices of an institutional review board (IRB)-approved research protocol (Providence protocol ID 2019000204) and was conducted in compliance with Human Subjects research, clinical data management procedures, as well as cloud information security policies and controls. All study data involving PHI were integrated, managed and analyzed exclusively and solely within our collaborating health network. All study personnel completed and were credentialed in training modules covering Human Subjects research, use of clinical data in research, and appropriate use of IT resources and IRB-approved data assets.

[5] https://ncithesaurus.nci.nih.gov/ncitbrowser/

[6] https://varnomen.hgvs.org/

[7] https://azure.microsoft.com/en-us/products/form-recognizer/



Clinical Trial Triaging

Name: Doe, John G
Accession No.: e3ba92a82036201a3c
D.O.B.: Sept. 30, 1933
Age: 88.8
Gender: M
Histology: LUAD (Lung Adenocarcinoma)
Path Staging: T2 N0 M1
Stage Group: Stage IV
Patient EHR Assisted Curation N/A

Search Report

Trial Filters

☑ Age Match Only
□ Stage Match Only
□ Updated in Last 2 Years

Locations

□ North America
□ United States

☑ **Biomarkers**

clinical signif.	gene	protein change	variant
YES	KRAS	p.Gly12Val	G12V
YES	H1-2	p.Ser102Phe	S102F
UNK	EPHB1	p.Ala912Thr	A912T
UNK	GABRA6	p.Asp418His	D418H
UNK	GID4	p.Ser268Pro	S268P
UNK	MAP3K4	p.Ala1197_Ala1199del	A1197_A1199del
UNK	MGA	p.Trp2758Ser	W2758S
UNK	PARP1	p.Thr632Met	T632M
UNK	SMAD3	p.Cys421_Ter426del	C421_*426del

Search Builder Show 10 entries

Search:

	NCT No.	Title	Phase	Matching Trial Diseases	Matching Trial Stage	Matching Trial Biomarkers	Notes
□	NCT05379985	Evaluation of RMC-6236 in Subjects With Advanced Solid Tumors Harboring Specific Mutations in KRAS	Phase 1	- Non-Small Cell Lung Carcinoma - Malignant Solid Neoplasm	- Advanced	- KRAS G12V - KRAS Mutation	
□	NCT03953235	A Study of a Personalized Cancer Vaccine Targeting Shared Neoantigens	Phase 1/Phase 2	- Non-Small Cell Lung Carcinoma - Malignant Solid Neoplasm	- Metastatic - Advanced	- KRAS G12V	
□	NCT04000529	Phase Ib Study of TNO155 in Combination With Spartalizumab or Ribociclib in Selected Malignancies	Phase 1	- Non-Small Cell Lung Carcinoma - Malignant Solid Neoplasm	- Advanced	- KRAS G12X - KRAS Mutation	
□	NCT04620330	A Study of VS-6766 and VS-6766 + Defactinib in Recurrent KRAS G12V, Other KRAS and BRAF Non-Small Cell Lung Cancer	Phase 2	- Non-Small Cell Lung Carcinoma		- KRAS G12V - KRAS Mutation	
□	NCT02079740	Trametinib and Navitoclax in Treating Patients With Advanced or Metastatic Solid Tumors	Phase 1/Phase 2	- Malignant Solid Neoplasm	- Metastatic - Advanced	- KRAS G12X - KRAS Mutation	
□	NCT05438667	TCR-T Cell Therapy on Advanced Pancreatic Cancer and Other Solid Tumors	Early Phase 1	- Malignant Solid Neoplasm	- Metastatic - Advanced	- KRAS G12V - KRAS Mutation	
□	NCT05202561	A Study of RNA Tumor Vaccine in Patients With Advanced Solid Tumors	Phase 1	- Malignant Solid Neoplasm	- Advanced	- KRAS G12V - KRAS Mutation	
□	NCT04625647	Testing the Use of Targeted Treatment (AMG 510) for KRAS G12C Mutated Advanced Non-squamous Non-small Cell Lung Cancer (A Lung-MAP Treatment Trial)	Phase 2	- Non-Squamous Non-Small Cell Lung Carcinoma - Lung Adenocarcinoma - Non-Small Cell Lung Carcinoma - Lung Carcinoma	- Stage IVB - Stage IVA - Stage IV - Advanced	- KRAS Mutation	
□	NCT03667716	COM701 (an Inhibitor of PVRIG) in Subjects With Advanced Solid Tumors.	Phase 1	- Non-Small Cell Lung Carcinoma - Lung Carcinoma - Malignant Solid Neoplasm	- Stage IV - Metastatic - Advanced	- KRAS Mutation	
□	NCT04511845	A Dose-Escalation Study of SPYK04 in Patients With Locally Advanced or Metastatic Solid Tumors (With Expansion).	Phase 1	- Non-Small Cell Lung Carcinoma - Malignant Solid Neoplasm	- Metastatic	- KRAS Mutation - MAPK/ERK pathway	

Showing 1 to 10 of 133 entries (filtered from 162 total entries)

Figure 5: Example user interface for clinical trial matching. NOTE: The data used in this example are synthetic and do not contain PHI.

4.2. System Details: Applying LLMs to Structuring Eligibility Criteria

To transform the trial XML into a structured representation, we employ a prompt template that guides GPT-4 (same for GPT-3.5) to extract and structure relevant criteria from each trial. Specifically, we focus on four types of criteria: trial cohort, disease state, tumor histology, and biomarkers. The prompt directs GPT-4 to output the structured representation in JSON format and provides instructions on how to handle unclear or missing information. In addition, the prompt may include few-shot example (input, output) pairs for in-context learning (Brown et al., 2020). Our prompt template can be found in Figures 6 to 9 in Appendix. For inference, we replace the placeholder {`input_trial`} in the prompt template with the input trial text and prompt GPT-4 to output the structured representation in JSON format. As shown in Figure 9, the output organizes the criteria into five categories: disease state, histology inclusion, biomarker inclusion, histology exclusion, and biomarker exclusion. The criteria logical expression is in disjunctive normal form (DNF). Our prompt instructs GPT-4 to assign a cohort name for each clause if possible.

4.3. Results: Structuring Trial Eligibility Criteria

Table 1 shows the test results on structuring oncology trial eligibility criteria in the basic setting, where we evaluate on entity extraction only and ignore complex match logic. We can't get GPT-3.5 to work in the 3-shot setting as adding the examples will render the prompt to exceed the context limit. Standard biomedical entity extraction tools such as GNormPlus (Wei et al., 2015), tmVar (Li et al., 2013), and SciSpaCy (Neumann et al., 2020) did not perform very well for one or both of histology and biomarkers. Criteria2Query (Yuan et al., 2019) also delivered subpar performance on oncology trial entities, revealing limitation in its generalizability. Remarkably, out of box, GPT-4 delivers strong performance in

Table 1: Comparison of test results on structuring oncology trial eligibility criteria. Gold labels provided by three molecular pathologists at the collaborating health network. Evaluation on inclusion and exclusion criteria entity extraction only, with complex match logic ignored. GPT-3.5 can't do 3-shot here due to limited context size.

	Histology			Biomarker		
	Precision	Recall	F1	Precision	Recall	F1
GNormPlus	-	-	-	6.8	19.6	10.2
SciSpaCy	34.2	70.2	46.0	58.3	6.9	12.3
Criteria2Query	29.6	40.2	32.8	68.3	27.5	39.2
GPT-3.5 (zero-shot)	35.1	31.6	34.2	61.2	29.4	39.7
GPT-4 (zero-shot)	62.1	69.0	**65.4**	75.3	59.8	66.7
GPT-4 (3-shot)	57.8	73.7	64.8	72.5	72.5	**72.5**

Table 2: Comparison of test results on structuring oncology trial eligibility criteria, evaluated on complete match logic in DNF form.

	Histology			Biomarker			Histology+Biomarker		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
SciSpaCy	9.2	16.8	11.8	36.4	5.3	9.3	7.4	10.8	8.8
Criteria2Query	23.2	34.8	27.8	22.0	12.0	15.5	9.6	17.4	12.4
GPT-3.5 (zero-shot)	36.0	20.6	26.2	48.3	18.7	26.9	30.7	14.6	19.7
GPT-4 (zero-shot)	43.0	41.3	42.1	31.0	17.3	22.2	27.8	21.1	24.0
GPT-4 (3-shot)	42.7	54.8	**48.0**	39.4	37.3	**38.4**	27.3	32.4	**29.6**

extracting both histology and biomarker information, despite requiring no extensive customization for this domain and task, outperforming Criteria2Query by over 30 absolute points in F1. In-context learning helped substantially for biomarker extraction, but didn't matter much for histology. GPT-4 also clearly outperforms GPT-3.5, which performs on par with Criteria2Query in the zero-shot setting.

Table 2 shows the test results on structuring oncology trial eligibility criteria in the full setting, where we evaluate on the complete match logic in DNF form. Note that this is a very stringent evaluation as each conjunction (`AND`s) needs to contain the exactly combination of atomic attributes to be considered correct. When there are complex nested logical forms, the number of conjunctions may become very large. Still, this evaluation offers a way to objectively compare system's capability in extracting complex match logic. Here, GPT-4 similarly dominates all other systems. Moreover, GPT-4 (3-shot) outperforms GPT-4 (zero-shot) in all scenarios, indicating that in this more challenging setting, in-context learning indeed plays a positive role.

4.4. Results: End-to-End Clinical Trial Matching

Given structured trial eligibility criteria and structured patient information, matching can be done by evaluating for each criterion component, where the patient information is compatible (thus eligible). This check is more elaborate than it may appear at first, since a criterion may specify a set of values (e.g., EGFR mutations without specifying which ones) or a more abstract concept than that in the patient information. Therefore, matching relies

Table 3: Comparison of recall results on historical patient-trial enrollment data. The expert system takes extensive expert manual effort to develop and has been designed to favor recall and ignore exclusion criteria, so it is expected to have high recall. The publicly available Criteria2Query instance runs slowly and we can only get 157 pairs evaluated in time (out of 466 total).

	Recall
SciSpaCy	50.0
Criteria2Query*	37.0
GPT-3.5 (zero-shot)	23.4
GPT-4 (zero-shot)	64.2
GPT-4 (3-shot)	76.8
Expert System	93.6

Table 4: Comparison of test results using feedback data from the clinical trial matching application in test deployment at our collaborating health network. The expert system is expected to perform well here, given that the users mainly evaluated its proposed candidates.

	Precision	Recall	F1
GPT-3.5 (zero-shot)	88.5	11.6	20.6
GPT-4 (zero-shot)	86.7	46.8	60.8
GPT-4 (3-shot)	87.6	67.3	76.1
Expert System	84.3	97.4	90.4

on a hierarchical representation of medical terms to check if a criterion entity subsumes the corresponding patient attribute. For histology matching, we convert both clinical trial criteria entities and patient tumor site/histology into OncoTree codes (Kundra et al., 2021), which are a standard hierarchical representation for cancer types.

For biomarker matching, we use a hierarchy that captures biomarkers at different levels of specificity: amino acid level, exon level, chromosomal level, gene level, or pathway level. A patient's tumor sequencing results typically specify a specific amino acid or DNA coding variant, whereas a clinical trial may specify a biomarker criterion at any of these levels. Matching needs to account for this hierarchy and the implied subsumption relations.

Table 3 shows the recall results for the historical patient-trial enrollment dataset. As expected, the expert system performs very well, given that it has been developed with extensive manual effort for the oncology domain. Remarkably, out of box, GPT-4 can already performs competitively, recovering 76.8% of gold patient-trial pairs.

Table 4 shows the test results using the feedback data from the test deployment of the expert system. Again, the expert system is expected to perform well here. E.g., the recall is expected to be close to 100%, given that the users mainly evaluated candidates proposed by the expert system and only occasionally added a trial they found by themselves. Out of box, GPT-4 already attains higher precision than the expert system. As observed in the historical enrollment evaluation, a major growth area lies in boosting the recall but GPT-4 already performs quite competitively overall.

We also explored applying GPT-4 directly to conduct end-to-end matching. Due to the context limit, we can't fit the entire medical record for most patients into the prompt. Therefore, it is only feasible to consider direct matching against patient information that has already been structured, and currently it is only feasible for the zero-shot setting. Due to current restriction on applying Azure OpenAI services to PHI, we only test this using a de-id example adapted from the public TCGA dataset. The results can be found in Figures 10 to 12. Remarkably, GPT-4 provides a cogent narrative of its matching decision, supported by detailed analysis of individual criteria. It is difficult to draw a generalizable conclusion from such an anecdotal example, but in combination with other experimental results, it certainly illustrates the potential in harnessing the general cognitive capabilities of large language models to scale clinical trial matching.

5. Discussion

In this work, we present the first systematic study on using large language models (LLMs) to scale clinical trial matching, using oncology as a case study to ground our exploration. Our findings foremost suggest that out-of-the-box, LLMs such as GPT-4 can already handle complex eligibility criteria of clinical trials and extract complex matching logic. In this regard, LLM provide a strong, practical baseline. And its general nature bodes well for the potential to apply to other disease areas.

Our study also reveals several key growth areas for enhancing LLMs in end-to-end clinical trial matching, such as addressing context length limitation and accuracy issues, especially in extracting and structuring patient data from EMRs. In future work, we plan to implement more sophisticated prompt engineering techniques and LLM fine-tuning.

While there are implicit limitations in the use of logical expressions, our system enables a triaging pipeline that can reduce the clinical trial candidates to a small number. This enables human-in-the-loop participation wherein a human can manually go through the small number of trials to verify all the criteria are met. Human-in-the-loop may be preferable given the safety-critical nature of healthcare. Nevertheless, we experimented with further accelerating this process by providing GPT-4 structured patient information, and asking GPT-4 to output all the matching and non-matching conditions. This may not be efficient to do for all possible patient-trial pairs but we can reserve this as a more expensive but higher-quality reranking step on a small number of candidates produced by a more efficient but less powerful system.

Limitations Clinical trial matching by structuring all criteria into a logical form (i.e. "structure-then-match") carries implicit limitations. In particular, it is not always possible to map an extracted criteria into an existing concept or ontology; and indeed, we found quite a few examples in our study where this was not possible. There are also various subtleties in the criteria language that are difficult to capture completely into a logical formula. Due to current restriction on applying GPT-4 to identifiable patient records, we were not able to explore LLMs in structuring patient information.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4, 2023.

Aurelia Bustos and Antonio Pertusa. Learning eligibility in cancer clinical trials using deep neural networks. *Applied Sciences*, 8(7):1206, 2018.

Junyi Gao, Cao Xiao, Lucas M. Glass, and Jimeng Sun. Compose: Cross-modal pseudo-siamese network for patient trial matching. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 803–812, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403123. URL https://doi.org/10.1145/3394486.3403123.

Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071, 04 2017. ISSN 1067-5027. doi: 10.1093/jamia/ocx019. URL https://doi.org/10.1093/jamia/ocx019.

Ritika Kundra, Hongxin Zhang, Robert Sheridan, Sahussapont Joseph Sirintrapun, Avery Wang, Angelica Ochoa, Manda Wilson, Benjamin Gross, Yichao Sun, Ramyasree Madupuri, Baby A. Satravada, Dalicia Reales, Efsevia Vakiani, Hikmat A. Al-Ahmadie, Ahmet Dogan, Maria Arcila, Ahmet Zehir, Steven Maron, Michael F. Berger, Cristina Viaplana, Katherine Janeway, Matthew Ducar, Lynette Sholl, Snjezana Dogan, Philippe Bedard, Lea F. Surrey, Iker Huerga Sanchez, Aijaz Syed, Anoop Balakrishnan Rema, Debyani Chakravarty, Sarah Suehnholz, Moriah Nissan, Gopakumar V. Iyer, Rajmohan Murali, Nancy Bouvier, Robert A. Soslow, David Hyman, Anas Younes, Andrew Intlekofer, James J. Harding, Richard D. Carvajal, Paul J. Sabbatini, Ghassan K. Abou-Alfa, Luc Morris, Yelena Y. Janjigian, Meighan M. Gallagher, Tara A. Soumerai, Ingo K. Mellinghoff, Abraham A. Hakimi, Matthew Fury, Jason T. Huse, Aditya Bagrodia, Meera Hameed, Stacy Thomas, Stuart Gardos, Ethan Cerami, Tali Mazor, Priti Kumari, Pichai Raman, Priyanka Shivdasani, Suzanne MacFarland, Scott Newman, Angela Waanders, Jianjiong Gao, David Solit, and Nikolaus Schultz. Oncotree: A cancer classification system for precision oncology. *JCO Clinical Cancer Informatics*, (5):221–230, 2021. doi: 10.1200/CCI.20.00108. URL https://doi.org/10.1200/CCI.20.00108. PMID: 33625877.

Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, limits, and risks of GPT-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023a. doi: 10.1056/NEJMsr2214184. URL https://doi.org/10.1056/NEJMsr2214184. PMID: 36988602.

Peter Lee, Carey Goldberg, and Isaac Kohane. *The AI Revolution in Medicine: GPT-4 and Beyond*. Pearson Education, 2023b. ISBN 9780138200138. URL https://books.google.com/books?id=AY-5zwEACAAJ.

Miao Li, Dongqing Zhang, Zhihao Yang, Weiwei Li, Xiaobo Liu, and Jian Wang. tmvar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, 29(11):1433–1439, 2013. doi: 10.1093/bioinformatics/btt162.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. scispacy: Fast and robust models for biomedical natural language processing. *Bioinformatics*, 36(4):1235–1241, 2020. doi: 10.1093/bioinformatics/btz682.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems, 2023.

Benjamin E. Nye, Ani Nenkova, Iain J. Marshall, and Byron C. Wallace. Trialstreamer: Mapping and browsing medical evidence in real-time. 2020.

OpenAI. GPT-4 technical report, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Sam Preston, Mu Wei, Rajesh Rao, Robert Tinn, Naoto Usuyama, Michael Lucas, Yu Gu, Roshanthi Weerasinghe, Soohee Lee, Brian Piening, Paul Tittel, Naveen Valluri, Tristan Naumann, Carlo Bifulco, and Hoifung Poon. Toward structuring real-world data: Deep learning for extracting oncology information from clinical text with patient-level supervision. *Patterns*, 4(4), 2023/04/14 2023. doi: 10.1016/j.patter.2023.100726. URL https://doi.org/10.1016/j.patter.2023.100726.

Rebecca L. Siegel and Ahmedin Jemal. Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(1):7–33, Jan 2022. doi: 10.3322/caac.21708. URL https://doi.org/10.3322/caac.21708.

Kristian D. Stensland, Russell B. McBride, Asma Latif, Juan Wisnivesky, Ryan Hendricks, Nitin Roper, Paolo Boffetta, Simon J. Hall, William K. Oh, and Matthew D. Galsky. Adult Cancer Clinical Trials That Fail to Complete: An Epidemic? *JNCI: Journal of* *the National Cancer Institute*, 106(9):dju229, 09 2014. ISSN 0027-8874. doi: 10.1093/jnci/dju229. URL https://doi.org/10.1093/jnci/dju229.

Joseph M Unger, Riha Vaidya, Dawn L Hershman, Lori M Minasian, and Mark E Fleury. Systematic review and meta-analysis of the magnitude of structural, clinical, and physician and patient barriers to cancer clinical trial participation. *J. Natl. Cancer Inst.*, 111 (3):245–255, March 2019.

U.S. National Library of Medicine. ClinicalTrials.gov. Accessed April, 2023, 2023. URL https://clinicaltrials.gov/.

Chih-Hsuan Wei, Chun-Nan Hsu, Wen-Lian Hsu, Yung-Chuan Liu, and Hong-Jie Dai. Gnormplus: An integrative approach for tagging genes, gene families, and protein domains. *BioMed Research International*, 2015:918710, 2015. doi: 10.1155/2015/918710.

Chunhua Weng, Xiaoying Wu, Zhihui Luo, Mary Regina Boland, Dimitri Theodoratos, and Stephen B Johnson. EliXR: an approach to eligibility criteria extraction and representation. *Journal of the American Medical Informatics Association*, 18(Supplement_1): i116–i124, 07 2011. ISSN 1067-5027. doi: 10.1136/amiajnl-2011-000321. URL https://doi.org/10.1136/amiajnl-2011-000321.

Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, and Chunhua Weng. Criteria2query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26(4):294–305, 02 2019. ISSN 1527-974X. doi: 10.1093/jamia/ocy178. URL https://doi.org/10.1093/jamia/ocy178.

Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability, 2023.

Xingyao Zhang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. Deepenroll: Patient-trial matching with deep embedding and entailment prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 1029–1037, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380181. URL https://doi.org/10.1145/3366423.3380181.

Appendix A. Prompt Templates

Prompt Template for Clinical Trial Structuring

(System Message) I am an intelligent and thorough agent designed is to extract the given clinical trial criteria input into a valid JSON logic structure. My response should be written in the language of JSON and should accurately capture the meaning of the input. Please note that my response should accurately reflect the logical connectives of the criteria (and, or, not). I will try my best attempt to normalize the logic formula into disjunctive normal form. In disjunctive normal form, each clause connected by a conjunction, or 'and', must be either a literal or contain a conjunction, or 'and' operator. Please keep in mind that my response should be flexible enough to allow for various relevant and creative solutions. I should also focus on providing an accurate and well-structured solution that can be easily understood by others. I should be sure to specify solid tumor and tumor staging when applicable. I only extract the most important tumor biomarkers, tumor histology or disease, and tumor staging information. I ignore criteria regarding life expectancy, written consent, pregnancy, contraception and adequate bone marrow, liver, and renal function. I ignore the medication treatment and dosages for trial arms. I only extract about the inclusion and exclusion criteria for the trial. I must be thorough and include all criteria related to biomarkers, histologies, disease state, and staging.

—

(Instructions) I will use the following output format:

[{“cohort”: (String) Cohort Name, “disease_state”:(String) Disease State Criteria, “histology_inclusion”: (String) Histology Inclusion Criteria, “biomarker_inclusion”: (List) Biomarker Inclusion Criteria, “histology_exclusion”: (List) Histology Exclusion Criteria, “biomarker_exclusion”: (List) Biomarker Exclusion Criteria}, ...]

This is a representation of the logical disjunctive normal form where each conjunctive (AND) clause is represented as a JSON and the outer list is a disjunction (logical OR) of those clauses or JSONs. I will try to separate the biomarkers and histologies.

Below are descriptions of those keys. Only the following keys are allowed. DO NOT include any other criteria keys:
cohort: a string that represents the name of the cohort for these criteria
disease_state: a string that describes the disease state (e.g, staging (stage I, II, IIIA, IV), refractory, relapsed, advanced, metastatic, etc.)
histology_inclusion: a string that represents the histology criteria that if satisfied, contributes to the expression evaluating to True. There should only be one histology per JSON item.

Figure 6: GPT-4 prompt template for clinical trial structuring.

Prompt Template for Clinical Trial Structuring (cont’d)

biomarker_inclusion: a list of biomarker inclusion criteria that if all elemnts of the list are satisfied, contributes to the expression evaluating to True.
histology_exclusion: a list of histology exclusion criteria if any elements of the list are satisfied, contributes to the expression evaluating to False.
biomarker_exclusion: a list of biomarker exclusion criteria that if any elements of the list are satisfied, contributes to the expression evaluating to False.

To match one of the clauses or JSON items in the list, it means all field values in each item are satisfied (”histology_inclusion” and ”biomarker_inclusion” and not any(”histology_exclusion”) and not any(”biomarker_exclusion”)). Each clause/JSON should have at least a value for histology_inclusion. If all the values for a JSON are empty or null, do you include in the list. I do not care about criteria that do not have at least a value histology_inclusion. If a field doesn’t have any value, set the value to ”” or []. Any of the JSON in the list matching means a patient is eligible for the trial. For criteria specific to a cohort, add a new item to the list. Skip biomarker or histology criteria that are inside if-then conditions unless the if condition is a cohort type. Skip prior treatment type criteria. If the extracted criteria only contains an abbreviation, I will find the long form of that abbreviation from the rest of the document and include that in the field value.

I recognize the following as biomarkers and will include them in the biomarker_inclusion or biomarker_exclusion fields if they are mentioned as criteria as defined above: gene mutations, gene expressions, TMB (tumor molecular burden, TMB-H or TMB-L), MSI (microsatellite instability, MSI-H, MSI-L, MSS, microsatellite stable) status, gene pathway alterations such as dMMR (deficient Mismatch Repair Pathway) or pMMR (proficient Mismatch Repair), PD-L1 expression levels determined by Combined Positive Score (CPS) or tumor proportion score (TPS), positive or negative status for breast cancer biomarkers such as HR, ER, PR and HER2.

When a criteria specifies a histology and all the subtypes of the histology, I will include all those subtypes as additional clauses (JSON) in the list of disjunctions.

When a trial exclusion includes primary brain (Central Nervous System, CNS) tumor (not just brain metastasis), also include brain / CNS tumor mention in the histology_exclusion list.

Do not include criteria about prior therapies or prior treatment. They are not considered biomarkers criteria and should not be included in the inclusion or exclusion criteria. Do not include any expression prior therapies, treatments, or therapies.

Figure 7: GPT-4 prompt template for clinical trial structuring (cont’d).

Prompt Template for Clinical Trial Structuring (cont'd)

If the inclusion criteria is for either of multiple biomarkers (and/or), list each of those biomarkers in a separate clause or JSON item because either of them being satisfied should contribute to the entire expression being True. I want to have the least restrictive accurate matching criteria output. Only list multiple biomarkers in biomarker_inclusion list for one clause JSON item if the trial absolutely require all of those biomarkers.
Do not include mentioned biomarkers if the presence or absense of the biomarker does not affect eligibility. And do not include biomarkers in the output that if the biomarker is present, then additional criteria is needed. I only extract if the criteria itself determines eligibility.

When there is a list of biomarker criteria in the same sentence and it's not clear whether the biomarkers have an AND or OR relationship, assume it's an OR criteria relationship between the biomarkers, which means assign each biomarker to a new clause. If the biomarkers are on separate main bullet points in criteria section, assume those are AND criteria relationship and all the biomarkers should be in the same clause.

Do not include cohort or arms that closed.

For multiple exclusion biomarker or histology criteria, those should be present in every clause JSON item.
DO NOT repeat the same JSON multiple times. Be succinct.
—

(Demonstrations)
Input:

```
<brief_title>A Phase 1/2 Study of DCC-3116 in Patients With MAPK Pathway
    Mutant Solid Tumors</brief_title>
<official_title>A Phase 1/2, First-in-Human Study of DCC-3116 as Monotherapy
    and in Combination With RAS/MAPK Pathway Inhibitors in Patients With
    Advanced or Metastatic Solid Tumors With RAS/MAPK Pathway Mutations</
    official_title>
<brief_summary>This is a Phase 1/2, multicenter, open label, first in human (
    FIH) study of DCC-3116 as monotherapy, and in combination with trametinib,
     binimetinib, or sotorasib in patients with advanced or metastatic solid
    tumors with RAS/MAPK pathway mutation. The study consists of 2 parts, a
    dose-escalation phase, and an expansion phase.</brief_summary>
<condition>Pancreatic Ductal Adenocarcinoma</condition>
<condition>Non-Small Cell Lung Cancer</condition>
<condition>Colorectal Cancer</condition>
<condition>Advanced Solid Tumor</condition>
<condition>Metastatic Solid Tumor</condition>
```

Figure 8: GPT-4 prompt template for clinical trial structuring (cont'd).

Prompt Template for Clinical Trial Structuring (cont'd)

```
<arm_group>
<arm_group_label>Dose Escalation (Part 1, Cohort A Monotherapy)</
    arm_group_label>
<arm_group_type>Experimental</arm_group_type>
<description>DCC-3116 tablets in escalating dose cohorts given orally twice
    daily (BID) in 28-day cycles as monotherapy (single agent). If no DLT in 3
     participants or 1 DLT/6 participants is observed, dose escalation may
    continue to the next planned dose cohort.</description>
</arm_group>
```

(... more arm groups are omitted for space saving)

```
<criteria>
Inclusion Criteria:
  1. Male or female participants >=18 years of age
  2. Dose Escalation Phase (Part 1):
    1. Participants must have a pathologically confirmed diagnosis of an
        advanced or metastatic solid tumor with a documented RAS, NF1, or RAF
        mutations. A molecular pathology report documenting mutational status
        of RAS, NF1, or RAF must be available.
    2. Progressed despite standard therapies, and received at least 1 prior
        line of anticancer therapy.
       - Participants with a documented mutation in BRAF V600E or V600K must
           have received approved treatments known to provide clinical benefit
            prior to study entry.
```

(... more criteria are omitted for space saving)

```
</criteria>
```

Eligibility Criteria Output:

```
[
    {"cohort": "Dose Escalation Phase (Part 1)", disease_state": "advanced or
        metastatic", "histology_inclusion": "Solid Tumor",  "
        biomarker_inclusion": ["RAS mutation"], "histology_exclusion": [], "
        biomarker_exclusion": []},
    {"cohort": "Dose Escalation Phase (Part 1)", "disease_state": "advanced or
         metastatic", "histology_inclusion": "Solid Tumor",  "
        biomarker_inclusion": ["NF1 mutation"], "histology_exclusion": [], "
        biomarker_exclusion": []}, {"cohort": "Dose Escalation Phase (Part 1)"
        , "disease_state": "advanced or metastatic", "histology_inclusion": "
        Solid Tumor",  "biomarker_inclusion": ["RAF mutation"], "
        histology_exclusion": [], "biomarker_exclusion": []},
```

(... more cohorts are omitted for space saving)

```
]
```

(User Message) Input:
{input_trial}

Eligibility Criteria Output:

Figure 9: GPT-4 prompt template for clinical trial structuring (cont'd).

Prompt Template for Direct Trial-Patient Matching

(System Message) As a helpful agent, my task is to analyze both the structured data and free-text pathology report of a patient and determine if they meet the clinical trial criteria text for eligibility. Please provide a clear and concise response that lists all the reasons why I believe the patient is eligible or not eligible. My analysis should take into account various factors such as age, medical history, current medications, histology, staging, biomarkers and any other relevant information provided in the report and structured data. I should also be able to handle complex scenarios with multiple variables and provide accurate reasoning behind my determination. My response should be flexible enough to allow for various relevant and creative reasons why the patient may or may not be eligible. My goal is to provide an informative and detailed analysis that assists healthcare professionals in making informed decisions regarding patient eligibility.

Clinical Trial Study Design Detail:

```
<brief_title>Study of Zotiraciclib for Recurrent High-Grade Gliomas With
    Isocitrate Dehydrogenase 1 or 2 (IDH1 or IDH2)
Mutations</brief_title>
...
<criteria>
- INCLUSION CRITERIA:
- Participants must have diffuse glioma, WHO grades 2-4, histologically
    confirmed by Laboratory of Pathology, NCI
- IDH1 or IDH2 mutation status confirmed by TSO500 performed in LP, NCI
- Participants must have received prior treatment (e.g., radiation,
    conventional chemotherapy) prior to disease
progression
- Participants must have recurrent disease, proven histologically or by
    imaging studies
- Participants who have undergone prior surgical resection are eligible for
    enrollment to cohorts 1-4.
- Age &gt;=18 years
- Karnofsky &gt;=70%
- Participants must have recovered from the adverse effects of prior therapy
    to grade 2 or less
EXCLUSION CRITERIA:
More than one prior disease relapse (WHO grade 3-4) or more than two prior
    disease relapses (WHO grade 2)
- Prior therapy with:
- bevacizumab for tumor treatment. Note: participants who received bevacizumab
     for symptom management,
including but not limited to cerebral edema, or pseudo progression can be
    enrolled
- Prolonged QTc &gt;470ms as calculated by Fridericia s correction formula on
    screening electrocardiogram (ECG)
...
</criteria>
```

—

Figure 10: Directly asking GPT-4 to match clinical trial criteria and patient record. The patient record is de-ID clinical data from public TCGA dataset.

Prompt Template for Direct Trial-Patient Matching (cont'd)

Patient Metadata:

```
Male,
29 years old,
Glioma (Anaplastic Astrocytoma),
LIVING (63 months),
Recurred/Progressed (53 months)
Samples: 1
Attribute Value
....
International Classification of Diseases for Oncology, Third Edition ICD-O-3
    Histology Code 9401/3
International Classification of Diseases for Oncology, Third Edition ICD-O-3
    Site Code C71.0
Karnofsky Performance Score 90
...
Neoplasm Histologic Grade G3
Neoplasm Histologic Type Name Astrocytoma
...
Performance Status 0
Performance Status Assessment Timepoint Category Pre-Adjuvant Therapy
...
Related symptom first present Seizures
Seizure History YES
Sex Male
...
```

Patient Mutations:

```
Missense IDH1 R132H
```

Pathology Metadata:

```
Oncotree Code AASTR
...
```

Patient Pathology Report:

```
SPECIMEN(S) SUBMITTED/ PROCEDURES ORDERED:
CLINICAL HISTORY: male, here for left side craniotomy for tumor.
GROSS DESCRIPTION: Received the following specimens in the Department of
    Pathology, labeled with
the patient's name and
A. Brain tumor
B. Brain tumor
...
```

DIAGNOSIS:

```
A. "Brain tumor":
Anaplastic astrocytoma (WHO Grade III) (see comment)
B. "Brain tumor":
Anaplastic astrocytoma (WHO Grade III)
...
```

—

Figure 11: Directly asking GPT-4 to match clinical trial criteria and patient record. The patient record is de-ID clinical data from public TCGA dataset (cont'd).

Prompt Template for Direct Trial-Patient Matching (cont'd)

Based on the provided information, the patient appears to be eligible for the clinical trial. Here are the reasons for eligibility:
1. Age: The patient is 29 years old, which meets the age requirement of being 18 years or older.
2. Diagnosis: The patient has a diagnosis of Anaplastic Astrocytoma, which is a high-grade glioma (WHO Grade III). The clinical trial is for recurrent high-grade gliomas with IDH1 or IDH2 mutations.
3. IDH1 Mutation: The patient has an IDH1 R132H mutation, which confirms the presence of an IDH1 mutation as required by the clinical trial.
4. Recurrent Disease: The patient's disease has recurred/progressed after 53 months, which meets the requirement of having recurrent disease.
5. Karnofsky Performance Score: The patient has a Karnofsky Performance Score of 90, which is greater than or equal to the required 70%.
6. Prior Treatment: The patient has received prior treatment (radiation, conventional chemotherapy) before disease progression, as required by the clinical trial.
7. Recovery from Adverse Effects: Although not explicitly mentioned, there is no indication that the patient has not recovered from the adverse effects of prior therapy to grade 2 or less.
However, there are some limitations in the provided information:
1. IDH2 Mutation: The IDH2 mutation status is not provided. However, since the patient has an IDH1 mutation, this is not a concern for eligibility.
2. Prior Therapy Exclusions: The patient's prior therapy with bevacizumab is not mentioned. However, there is no indication that the patient has received bevacizumab for tumor treatment, which would exclude them from the trial.
3. QTc Interval: The patient's QTc interval is not provided. However, there is no indication that the patient has a prolonged QTc >470ms, which would exclude them from the trial.
4. Other Exclusion Criteria: There is no information provided about prior invasive malignancies, history of allergic reactions to compounds similar to zotiraciclib, or uncontrolled intercurrent illness or social situations. However, there is no indication that the patient meets any of these exclusion criteria.
In conclusion, based on the available information, the patient appears to be eligible for the clinical trial.

Figure 12: Directly asking GPT-4 to match clinical trial criteria and patient record. The patient record is de-ID clinical data from public TCGA dataset (cont'd).