

AutoCriteria:

a generalizable clinical trial eligibility criteria extraction system powered by large language models

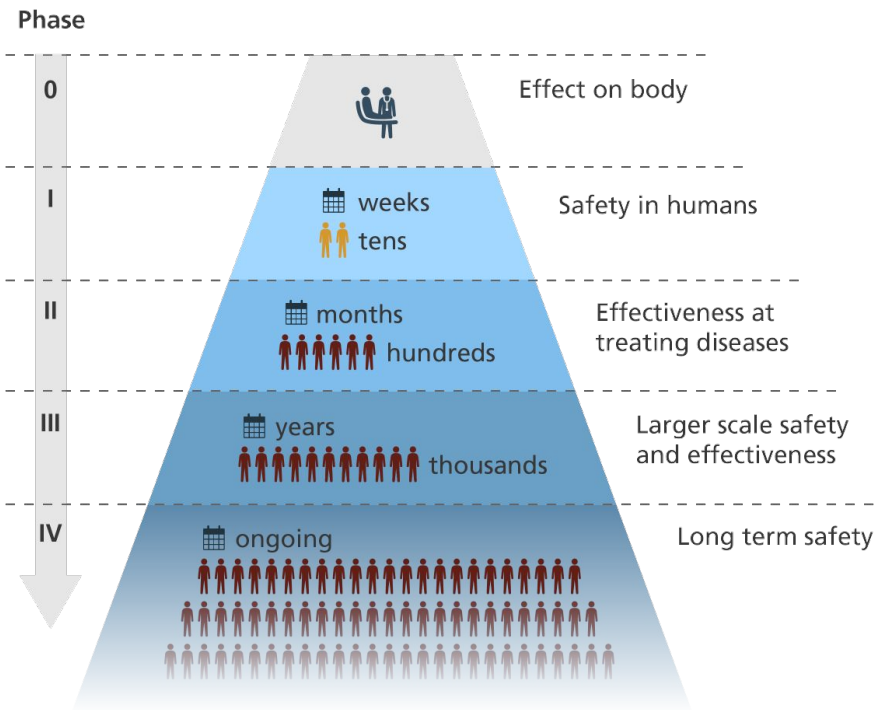
Nat Tangchitnob & Pongsakorn Tanupatrasakul

What are clinical trials?

“A research study in which one or more human subjects are prospectively assigned to one or more interventions (which may include placebo or other control) to evaluate the effects of those interventions on health-related biomedical or behavioral outcomes.”

Recruitment challenges

- Insufficient participants
- Long recruitment timings
- Participant dropouts



Generative Pre-trained Transformer (GPT-4)

Large Language Models (LLMs) hold potential in applying to medicine.

With the introduction of GPT-3 model, researchers have begun exploring the clinical use cases such as

- Building domain lexicons

- Automatic diagnosis and triage

Recently, as the more powerful ChatGPT and GPT-4 models became available, studies are evaluating their medical applications such as

- Licensing examinations

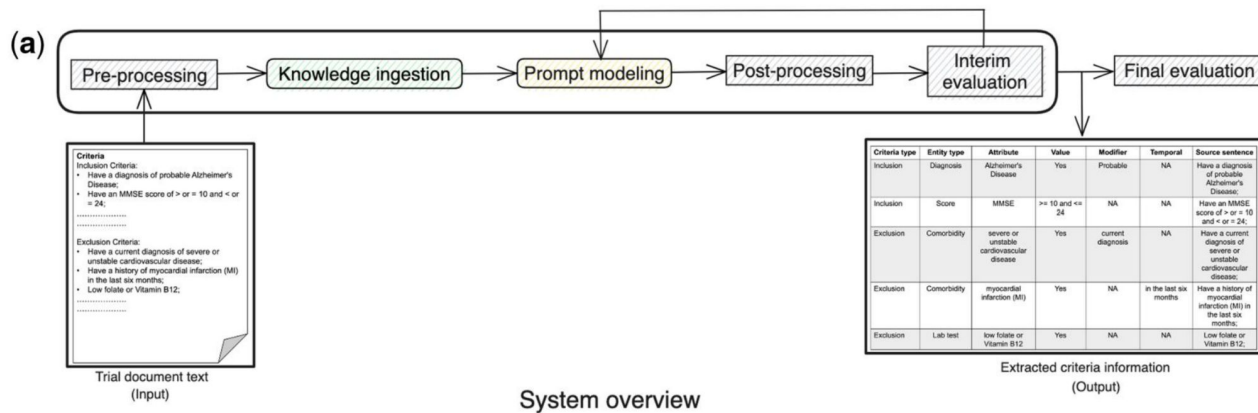
- Question answering

- Medical education

- Creating medical domain-specific datasets to for medical chatbot applications.

AutoCriteria

- Extract granular eligibility criteria information
- Based on the GPT-4 model from OpenAI
- Easily adaptable and knowledge-driven prompt
- Post-processing improves the medical grounding and consistency of the extracted information
- Generalizability with minimum prompt adjustment was evaluated on clinical trial text belonging to a wide variety of diseases



Main contributions of this work

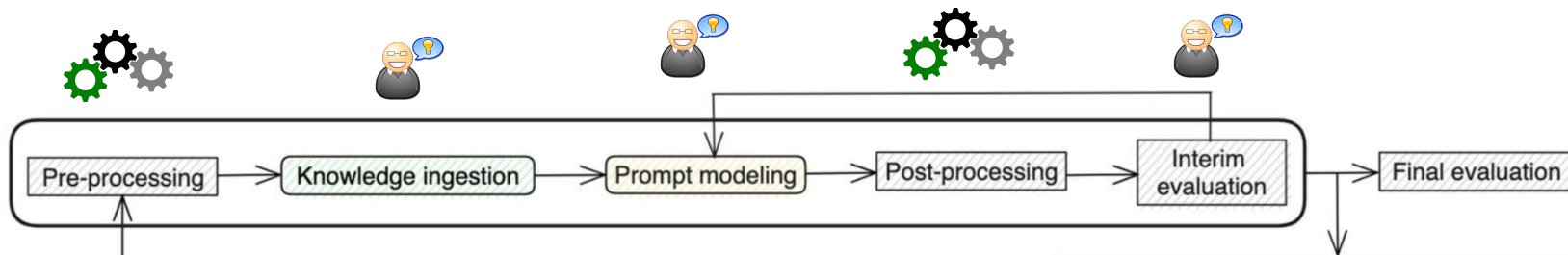
1. A **zero-shot** GPT-based system to extract granular eligibility criteria including identifying temporal information and criteria conditions from clinical trial documents **without requiring manual annotations**.
2. **Adaptable prompts** for the system that are flexible and generalizable across **different disease domains** and can be easily extended to **new diseases** (without manual annotation and training).

Data

- Clinical trial document for the 9 diseases were collected from ClinicalTrials.gov
- For each disease, 28 trials were retrieved.
 - 3 trials for prompt design
 - 5 trials for prompt calibration
 - 20 trials randomly selected for evaluation
- Prompt design refers to initial version of the prompt
- Prompt calibration refers to iteratively improving the prompt based on expert guided evaluation

Preprocessing

1. Split the raw criteria text in trial document into inclusion and exclusion by case-insensitive regular expression
2. Split each of the part into chunk of 200 words while preserving sentence boundaries



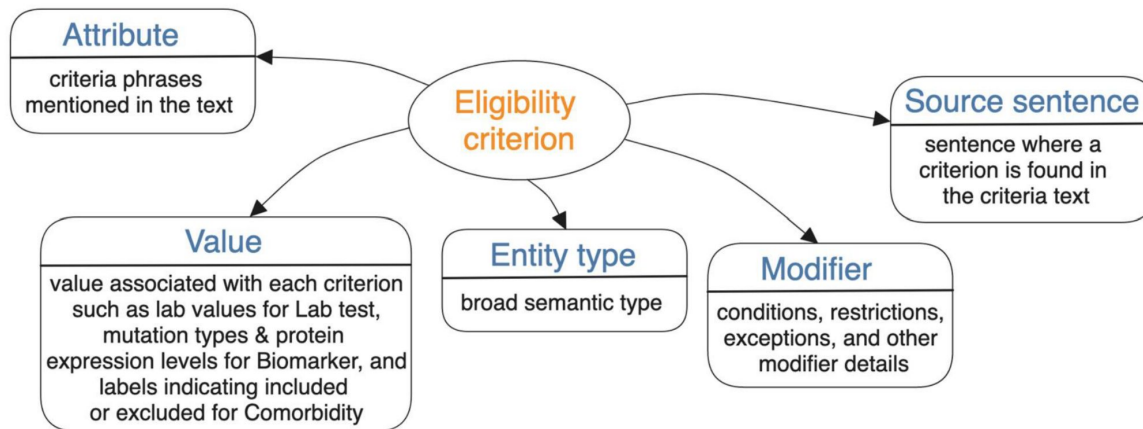
<p>Raw text</p>	<p>Inclusion Criteria:</p> <p>Males, and females of at least 50 years old with a primary caregiver</p> <p>Probable Alzheimer's disease</p> <p>Mini-Mental State Examination (MMSE) score of ≥ 10 and ≤ 26</p> <p>Patients initiating therapy for the first time with a Cholinesterase (ChE) inhibitor (patients prescribed both rivastigmine and memantine are allowed) or patients who failed to benefit from previous ChE inhibitor treatment</p> <p>Residing with someone in the community throughout the study or, if living alone, in contact with the responsible caregiver everyday</p> <p>Exclusion Criteria:</p> <p>Patients not treated according to the product monograph for capsules</p> <p>Current diagnosis of an active skin lesion/disorder that would prevent accurate assessment of the adhesion and potential skin irritation of the patch (e.g., atopic dermatitis, wounded or scratched skin in the area of the patch application)</p> <p>History of allergy to topical products containing any of the constituents of the patches</p> <p>Other protocol-defined inclusion/exclusion criteria may apply.</p>	
<p>Preprocessed text</p>	<ul style="list-style-type: none"> - Males, and females of at least 50 years old with a primary caregiver - Probable Alzheimer's disease - Mini-Mental State Examination (MMSE) score of ≥ 10 and ≤ 26 - Patients initiating therapy for the first time with a Cholinesterase (ChE) inhibitor (patients prescribed both rivastigmine and memantine are allowed) or patients who failed to benefit from previous ChE inhibitor treatment - Residing with someone in the community throughout the study or, if living alone, in contact with the responsible caregiver everyday 	<ul style="list-style-type: none"> - Patients not treated according to the product monograph for capsules - Current diagnosis of an active skin lesion/disorder that would prevent accurate assessment of the adhesion and potential skin irritation of the patch (e.g., atopic dermatitis, wounded or scratched skin in the area of the patch application) - History of allergy to topical products containing any of the constituents of the patches <p>Other protocol-defined inclusion/exclusion criteria may apply. 8</p>

Knowledge ingestion

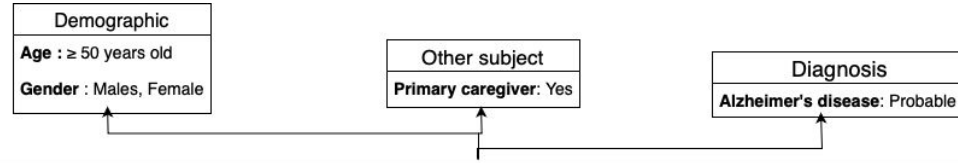


Key criteria entities and attribute for each disease were identified by knowledge experts

The knowledge is also leveraged in the prompt modeling component

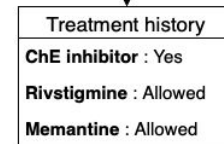
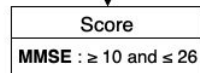


Example



[Inclusion Criteria Text]:

- Males, and females of at least 50 years old with a primary caregiver
- Probable Alzheimer's disease
- Mini-Mental State Examination (MMSE) score of ≥ 10 and ≤ 26
- Patients initiating therapy for the first time with a Cholinesterase (ChE) inhibitor (patients prescribed both rivastigmine and memantine are allowed) or patients who failed to benefit from previous ChE inhibitor treatment
- Residing with someone in the community throughout the study or, if living alone, in contact with the responsible caregiver everyday



Cancer trials

Include Breast Cancer (BC) and Multiple Myeloma (MM) for cancer related trial

10 criteria entity types:

1. Demographic
2. Tumor characteristic
3. Score (performance score such as ECOG and Karnofsky)
4. Biomarker
5. Diagnosis (cancer diagnosis criteria and specific cases of cancer)
6. Comorbidity
7. Treatment history (all treatments, therapies, medications, drugs, and procedures)
8. Lab tests
9. Contraception related
10. Survival (life expectancy)

Alzheimer's Disease (AD) trial

Additional **entity types** for this trial:

- Symptom
- Cognitive requirement
- Other subject requirement

Score for Alzheimer are classified by

- Dementia scores (Clinical dementia rating and mini-mental state examination)
- Depression scales and examinations

Biomarker

- Imaging biomarker
- Treatment history
- rain imaging scan findings

Additional of education were added as part of demographic entity

Nonalcoholic steatohepatitis

Additional criterion for this trials:

NASH included vitals such as BMI and weight.

Biomarker entity include: Liver fat content on MRI proton density fat fraction

Treatment include imaging scans

For this trials, **score** refer to

- NASH, nonalcoholic fatty liver disease activity score
- Lobular inflammation and ballooning degeneration scores

Inflammatory bowel disease (IBD) trial

2 Inflammatory bowel diseases were considered in this trial:
Crohn's disease (CD) and ulcerative colitis (UC)

Extracted IBD-specific criteria:

- CD activity index
- Harvey-Bradshaw Index
- Simple endoscopic score for CD
- Mayo endoscopic subscore for UC
- Complete mayo scores
- Vital

Rare disease trial

3 Rare diseases were included in the study:

Sickle cell disease (SCD)

Homozygous familial hypercholesterolaemia (HoFh)

Heritable pulmonary arterial hypertension (HpAH)

Vitals were extracted for all 3 diseases:

- WHO functional class criteria
- Pulmonary arterial hypertension (PAH)

Prompt modeling

2 comprehensive prompts for inclusion and exclusion criteria

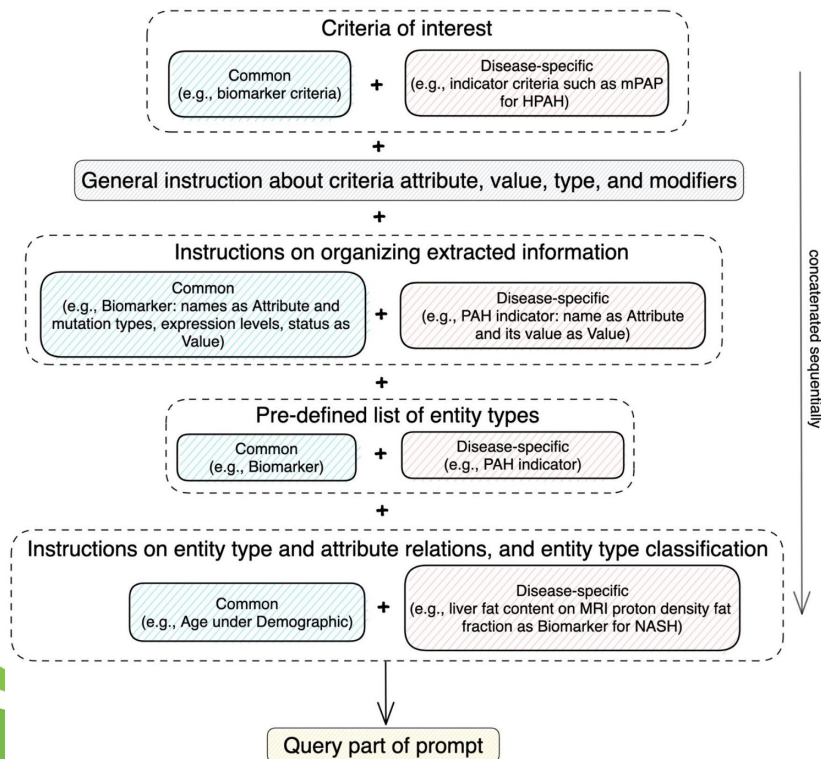
Each prompt has 3 main components

1. **General instruction** to always extract text spans from the given criteria text
2. Inclusion or Exclusion criteria **text**
3. **Query**
 - Criteria attributes (phrases)
 - Values
 - Modifier information
 - Entity types
 - Source sentences as evidence
 - Desired tabular response format

Prompt composition
1. General instruction 2. [Inclusion Criteria Text]: < criteria text > 3. Query part for Inclusion
Sample output
Entity type: Lab test Attribute: Hemoglobin Value: ≥ 10.0 g/dL Modifier: NA Source Sentence: Hemoglobin greater than 10.0 g/dL.

Figure 4. Sample prompt and output.

Prompt modeling



concatenated sequentially

From this text, identify the age, gender, education, dementia scores (such as MMSE and CDR), depression scales, mental or nervous system examinations, dementia symptoms (like memory loss and cognitive symptoms), cognitive requirements, other subject requirement, biomarkers (including all imaging biomarkers and brain imaging scan findings), lab tests, contraception-related criteria, prior treatments or therapies, medications or drugs, procedures (including imaging scans), cognitive decline or alzheimer's disease diagnosis, all other diseases or comorbidities or any health conditions or issues or complications (including other dementias and brain disease), and life expectancy along with any conditions. Extract all specific medications that are used to fight alzheimer's as well as medications that are generic (e.g., brain medications) or those for other diseases. Extract all treatments and therapies including chemotherapy, immunotherapy, targeted therapy, inhibitors, or antibodies against molecules, and interventional studies. Extract all comorbidities including all disease names (both hypernyms and their hyponyms), any health conditions, mental issues, complication, syndromes, disorder, symptoms, abnormalities, allergy, hypersensitivities, contraindication, adverse events, or side effects.

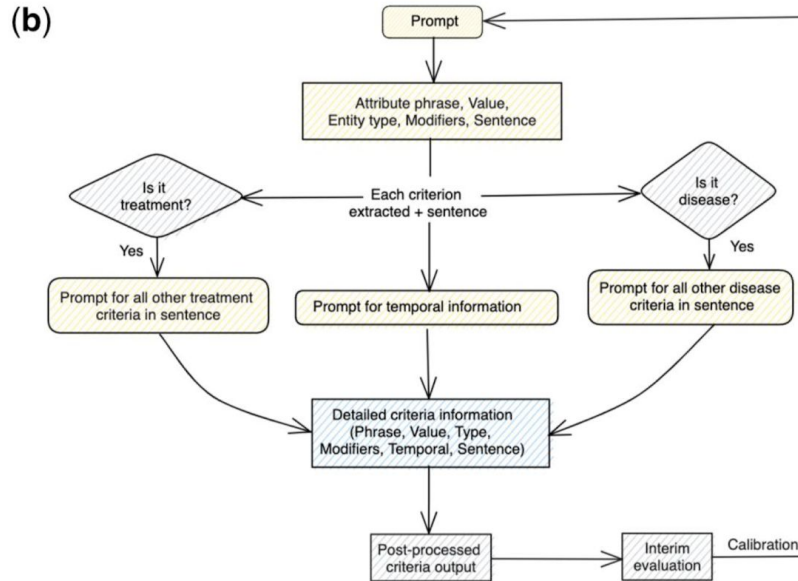
Show everything in a table with 5 column headers Entity, Attribute, Value, Condition, and Sentence. The table must contain 5 columns. Include any other condition/restriction/exception, other details, or specific conditions for specific patient groups under the "Condition" column of each treatment or therapy, medication, biomarker, lab test, and disease. Please think deeply to extract this information. Also include the corresponding sentence or phrase in the text where the entity was found under "Sentence" column. The phrase under "Attribute" column must be present in the phrase under "Sentence" column. The "Condition" should show temporal/condition information very precisely. Show only the lab tests and entries that are mentioned in the above text. Put lab test names under "Attribute" column and their numeric values under "Value" column. Only show the lab test names that are mentioned. In case only adequate organ function is mentioned, specify that as a different entity under "Attribute" column. Also create different lab test entities when the conditions are different (e.g., varying disease conditions).

For Biomarkers, put all gene, gene product names, and imaging biomarker names under "Attribute" column, and their mutation types or expression level or status under "Value" column and any specific condition under "Condition" column. If mutation type or expression level is not mentioned, put "Yes" for required inclusion, "Allowed" when they are allowed or eligible under certain specific conditions or time frames, and "No" for not included cases under the "Value" column. Also put the specific condition under the "Condition" column. Put each disease or health condition or comorbidity name in the text under "Attribute" column and include either the abbreviated or the full form of the disease. Put any specific condition or the status (e.g., active) described for each disease under "Condition" column. For each disease, previous treatment or therapy, procedure, and medication, put either "Yes" for required inclusion, "Allowed" when they are allowed or eligible under certain conditions or time frames, or "No" for not included cases under the "Value" column. Also put the specific condition under the "Condition" column. Any tests related to infections/diseases and any imaging exam for any disease should go under "Condition" column. Put the value for Age by including the number (usually with greater than or less than symbols) as is present in the given original criteria text. For dementia scores, depression scales, and mental or nervous system examinations, put "Score" under "Entity" column, put the scoring names (such as CDR and MMSE), scaling system names, and examination names under "Attribute" column, and their corresponding values under "Value" column. For cognitive requirements, put the cognitive category (e.g., visual) under the "Attribute" and the requirement (e.g., adequate) under "Value" column. For other subject requirement, put the specific subject (e.g., informant) under "Attribute" column. For dementia or alzheimer's-related diagnosis criteria, put "Diagnosis" under "Entity" column, put the disease or condition name under "Attribute" and put either "Yes" for required inclusion, "Allowed" for allowed criteria, or "No" for not included cases under "Value". Put each entity in each row. Each unique disease or health condition or comorbidity entity must be in different rows.

Classify each entity into one of the following classes - Demographic, Score, Symptom, Cognitive Requirement, Other Subject, Contraceptive, Biomarker, Diagnosis, Comorbidity, Previous Treatment, Lab test, and Survival. And put that class under the "Entity" column. Every row should have unique phrase under "Attribute" column. Entity column can contain Demographic, Score, Symptom, Cognitive Requirement, Other Subject, Contraceptive, Biomarker, Diagnosis, Comorbidity, Previous Treatment, Lab test, and Survival.

Note that the "Comorbidity" class includes all disease or comorbidity terms, any health condition or issue or complication terms, whereas the "Previous Treatment" class includes all treatments (including antibody or inhibitor treatments), medications, therapies, drugs, and procedures. Attribute column will include "Age" for age criteria, "Gender" for gender criteria, and "Education" for education criteria. "Age", "Gender", and "Education" falls under "Demographic" entity. "Attribute" column will include "Life Expectancy" for the life expectancy criteria. If it is mentioned in the text, put the life expectancy time period under the "Value" column, otherwise put "NA". "Life Expectancy" falls under "Survival" entity.

Prompt modeling



Prompt modeling

Temporal information

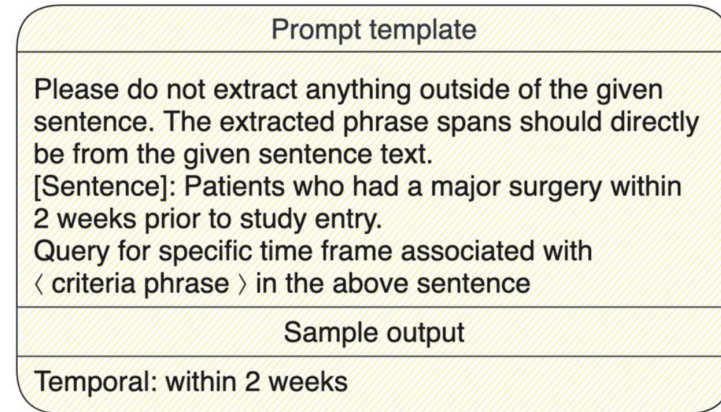
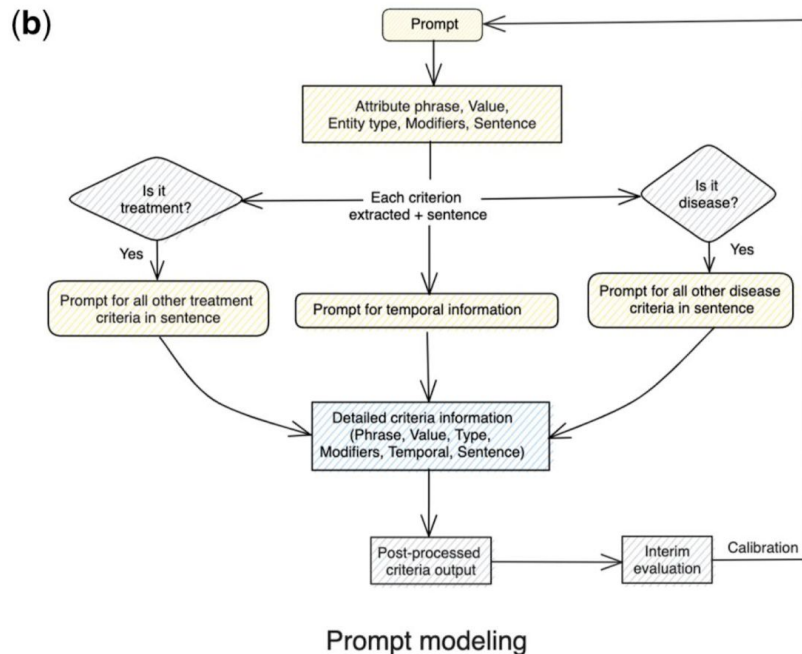


Figure 5. Temporal prompt and output.

Prompt modeling



Handling exceptions

For criteria entities that do not have an associated quantitative value, the assigned values are:

- Yes
- No
- Allowed

Example

“Patients **may** have previously received chemotherapy in the adjuvant/neoadjuvant setting, though this is **not required**,”

Chemotherapy is an allowed inclusion criteria

Post processing

Processing the responses from the GPT model to

1. Handle any inconsistencies in the model output

- Lack of specific information such as phrase like “other disease” will be removed

2. Incorporate medical knowledge through simple rules

- Sometimes phrase like “antibodies against” which refers to a treatment is classified by the model as biomarker

3. Additional response cleaning

- Removing “-” and numbers such as “1” and “3” that appear at the beginning of text

Evaluation

Interim evaluation: the prompts are evaluated manually and are calibrated **iteratively using expert feedback** for every disease

Final system assessment: we evaluated both quantitatively and qualitatively

Quantitative evaluation

- 180 trial documents (20 for each disease), manually annotated by knowledge experts
- Report the precision, recall, and F1 scores
- True positives at the level of entity (e.g., “*type 2 diabetes mellitus*” = “*type 2 diabetes mellitus*”)
- Accuracy measures for 4 combinations
 - attribute + value
 - attribute + value + entity type
 - attribute + value + entity type + temporal
 - attribute + value + entity type + temporal + modifier
- Must exactly match the gold standard value

Qualitative evaluation

- Thematic analysis of the missing and incorrect entities on the 180 trials

Evaluation

Type	Entity	Attribute	Value	Temporal	Modifier	Sentence
Inclusion	Demographic	Age	≥ 50 years old	NA	NA	Males, and females of at least 50 years old with a primary caregiver
Inclusion	Score	MMSE	≥ 10 and ≤ 26	NA	NA	Mini-Mental State Examination (MMSE) score of ≥ 10 and ≤ 26
Inclusion	Treatment History	ChE inhibitor	Yes	NA	Initiating therapy for the first time	Patients initiating therapy for the first time with a Cholinesterase (ChE) inhibitor
Inclusion	Treatment History	Rivastigmine	Allowed	NA	NA	Patients prescribed both rivastigmine and memantine are allowed
Inclusion	Diagnosis	Alzheimer's disease	Probable	NA	NA	Probable Alzheimer's disease
Inclusion	Other Subject	Primary caregiver	Yes	NA	Residing with someone in the community throughout the study	Residing with someone in the community throughout the study or, if living alone, in contact with the responsible caregiver everyday

Results

Table 1. Results of AutoCriteria on 180 clinical trial documents (20 for each disease) in extracting eligibility criteria phrases.

Disease	Precision (%)	Recall (%)	F1
Breast cancer	87.26	81.17	84.10
Multiple myeloma	85.53	86.83	86.18
Alzheimer's	94.54	92.38	93.45
NASH	95.08	95.81	95.44
Crohn's	87.21	88.30	87.75
Ulcerative colitis	91.85	92.99	92.42
SCD	90.46	90.15	90.30
HPAH	87.39	90.23	88.79
HoFH	90.38	88.01	89.18
All	89.62	89.23	89.42

HoFH, Homozygous familial hypercholesterolaemia; HPAH, Heritable pulmonary arterial hypertension; NASH, Nonalcoholic steatohepatitis; SCD, Sickle cell disease.

Results

Table 2. Accuracy (%) of AutoCriteria on 180 clinical trial documents (20 for each disease) in extracting a combination of criteria information.

Disease	Attribute + value	Entity type + attribute + value	Entity type + attribute + value + temporal	Entity type + attribute + value + temporal + modifier
Breast cancer	76.86	74.09	71.51	67.21
Multiple myeloma	81.68	73.96	68.74	66.54
Alzheimer's	91.05	90.67	88.00	85.14
NASH	91.87	91.61	88.56	86.28
Crohn's	85.10	83.43	81.20	78.83
Ulcerative colitis	88.01	86.76	86.29	83.64
SCD	87.10	85.91	85.91	83.36
HPAH	90.00	89.42	89.19	86.05
HoFH	86.70	84.83	84.27	82.21
All	85.90	83.37	81.09	78.95

HoFH, Homozygous familial hypercholesterolaemia; HPAH, Heritable pulmonary arterial hypertension; NASH, Nonalcoholic steatohepatitis; SCD, Sickle cell disease.

Results: Error Analysis

Among the **correctly extracted entity**,

- 38% (73/189) of the errors related to values (assigned “*allowed*” instead of “*yes*”)

Among the **incorrect entity type** classifications,

- 33% (27/82) are related to the system predicting a “*treatment history*” criteria as a “*comorbidity*.”

Among the **incorrect temporal** predictions,

- 46% (12/26) are associated with multiple temporality conditions in a sentence
- (extracts “*within the past 6 months, within the last 3 months*” instead of “*within the last 3 months*”)
- 23% (6/26) are related to capturing partial information
- (extracts “*within 6 months*” instead of “*during the trial or within 6 months after the last infusion*”)

For the **incorrect modifier** predictions,

- 87% (228/260) related to extracting partial information
- (extracts “*≥15 cm from anal verge*” instead of “*to the rectum, ≥15 cm from anal verge*”)

For cancer trials, the accuracy scores (especially considering modifiers) are low as there are more complex and long criteria conditions

Results

Table 3. Gold standard statistics: frequency of eligibility criteria entities.

Criteria type	BC	MM	AD	NASH	CD	UC	SCD	HPAH	HoFH
Biomarker	50	82	19	40	21	2	18	25	33
Comorbidity	390	524	292	356	335	246	241	334	226
Contraceptive	60	78	9	20	14	18	35	41	23
Demographic	42	44	44	45	42	43	48	53	47
Diagnosis	36	78	29	21	55	59	27	90	21
Lab test	85	165	7	68	38	25	52	61	56
Treatment history	307	299	67	174	181	211	146	186	100
Score	20	19	30	34	27	34	16	6	6
Survival	7	5	2	2	–	–	–	5	3
Cognitive Requirement	–	–	4	–	–	–	–	–	–
Other subject	–	–	17	–	–	–	–	–	–
Symptom	–	–	5	–	–	–	–	–	–
Tumor characteristics	49	20	–	–	–	–	–	–	–
Vital	–	–	–	27	3	4	2	9	17
PAH indicator	–	–	–	–	–	–	–	37	–
WHO class	–	–	–	–	–	–	–	13	–
Others	–	7	–	–	2	–	4	–	2
All	1046	1321	525	788	718	642	589	860	534

AD, Alzheimer's disease; BC, Breast cancer; CD, Crohn's disease; HoFH, Homozygous familial hypercholesterolaemia; HPAH, Heritable pulmonary arterial hypertension; MM, Multiple myeloma; NASH, Nonalcoholic steatohepatitis; SCD, Sickle cell disease; UC, Ulcerative colitis.

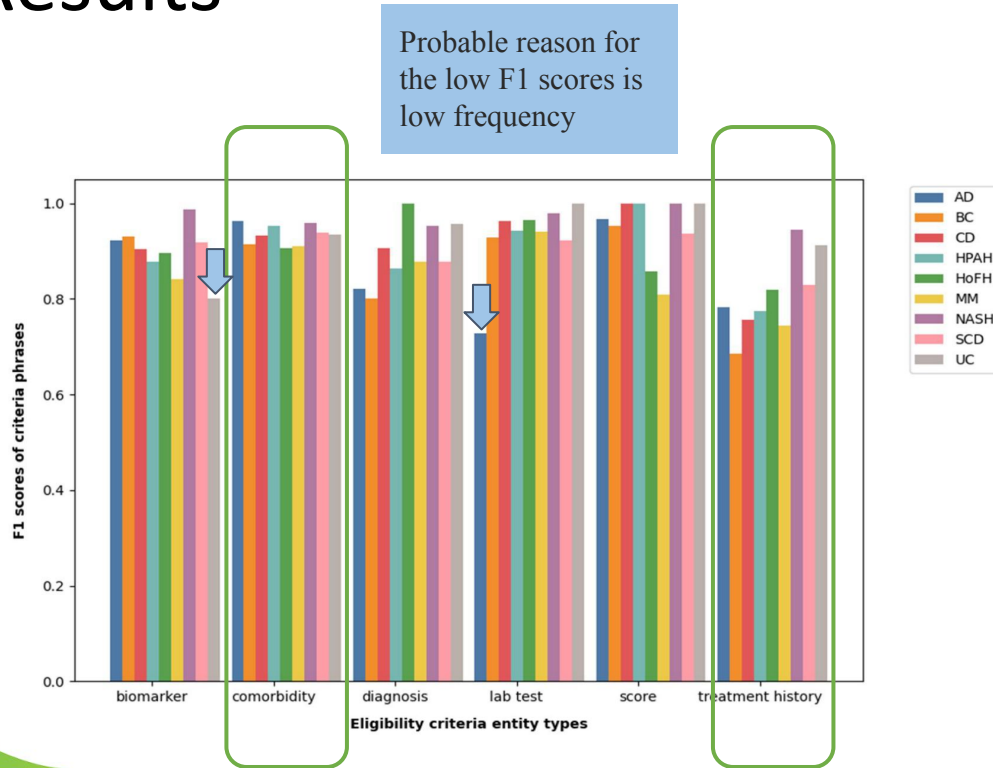
Results

Table 4. Gold standard statistics: maximum, minimum, mean, and median number of words in the criteria text for each disease.

Item	BC	MM	AD	NASH	CD	UC	SCD	HPAH	HoFH
Max # words (inclusion)	893	716	250	532	613	894	519	519	345
Min # words (inclusion)	95	83	43	27	35	23	37	48	22
Mean # words (inclusion)	303.0	472.9	117.85	157.85	195.15	195.1	222.25	265.05	118.5
Median # words (inclusion)	237.0	492.0	101.0	154.0	141.0	140.5	193.0	268.0	90.0
Max # words (exclusion)	1078	918	593	1109	1976	786	745	978	690
Min # words (exclusion)	100	184	41	37	36	36	44	80	23
Mean # words (exclusion)	413.5	499.6	188.65	301.0	309.85	189.8	263.05	390.0	225.35
Median # words (exclusion)	405.5	518.0	149.0	191.0	130.0	130.5	213.5	282.5	225.35

AD, Alzheimer's disease; BC, Breast cancer; CD, Crohn's disease; HoFH, Homozygous familial hypercholesterolaemia; HPAH, Heritable pulmonary arterial hypertension; MM, Multiple myeloma; NASH, Nonalcoholic steatohepatitis; SCD, Sickle cell disease; UC, Ulcerative colitis.

Results



Entity types	Average f1 score
demographic	92.16
contraceptive	85.71
survival	96.97
vital	91.11
Cognitive requirement	100
Other subject requirement	97.14
symptom	100
Tumor characteristics	87.30
PAH indicator	90.91
WHO functional class	88.89

Discussion

AutoCriteria : GPT-based system for eligibility criteria extraction from clinical trial text

- Negation
- “may be” included cases
- Temporal information
- Other modifier information
- Entity type of those criteria

Multiple NLP tasks all performed at once through efficient prompt design

- Entity recognition
- Relation extraction
- Classification

Comprehensive prompt developed for 1 disease can be easily transformed to identify information from a diverse set of diseases

Discussion

The role of knowledge experts in this study included:

1. Identifying the important **inclusion/exclusion criteria**
2. Framing the criteria **descriptions** in the prompts
3. Verifying the **clinical correctness** of prompts
4. Providing **feedback** for prompt adjustments/calibration
5. Validating system **output**

AutoCriteria understands eligibility criteria conditions including

- Exceptions
- Context
- Cohort information

Prior studies for eligibility criteria extraction using pretrained language models

- Not evaluate criteria relations
- Focused on quantitative results for relations

Discussion: Strength

Table 5. Areas of strengths and shortcomings of AutoCriteria.

Identifier	Theme	Examples	
Strengths			
S1	Accurate interpretation of logic in criteria	<p>Patients with type 2 diabetes mellitus may be included if they fulfill the following criteria;</p> <ol style="list-style-type: none"> Stable therapeutic regimen as defined by no changes in oral agents or dose for at least 3 months before screening and the stable dose can be maintained throughout the study. HbA1c \leq 9.5%. 	<p>“Allowed”</p> <p>Modifiers</p>
S2	Context-aware comprehension	<p>Histological confirmation of steatohepatitis on a diagnostic liver biopsy by central reading of the slides (biopsy obtained within 6 months prior to randomization or during the screening period) with at least 1 in each component of the NAS score (steatosis scored 0-3, ballooning degeneration scored 0-2, and lobular inflammation scored 0-3).</p>	<p>Different values from different scores</p>
S3	Precise recognition of cohort-specific attributes	<ol style="list-style-type: none"> For both Cohorts 1 and 2, Subjects must have estrogen (ER) receptor and progesterone (PR) receptor staining <10% and be human epidermal growth factor receptor 2 (HER2) negative defined as immunohistochemistry (IHC) 0 to 1+ (Inclusion criteria) For Cohort 2 only: Subject has severe hypersensitivity (\geqGrade 3) to nab-paclitaxel (Abraxane). (Exclusion criteria) 	<p>Different criteria for different cohorts</p>
S4	Different modifiers and values associated with the same criteria phrase	<ol style="list-style-type: none"> Part A and B patients: presence of NASH by histological evidence (liver biopsy obtained 2 years or less prior to randomization) with fibrosis level of F1, F2 or F3 (fibrosis in the absence of cirrhosis) Part C patients: presence of NASH by histological evidence (liver biopsy obtained during the Screening period or 6 months or less prior to randomization) with fibrosis level of F2 or F3 	<p>Different values from same entity</p>

Discussion: Limitation

Extract "Pain" and "Sickle cell disease"
Instead of "Pain not sickle cell disease related"

Extract "Liver disorder" and "NASH"
Instead of "history of a liver disorder"

Shortcomings

- SC1 Lacks precision in identifying crucial details
- SC2 Overlooking main criterion
- SC3 Differentiating the main criteria entities from modifier entities

- 1) Predominate cause of pain is **not sickle cell disease related**.
 - 2) History of a liver disorder **other than NASH**.
- Had a ventilation-perfusion (V/Q) lung scan, spiral/helical/electron beam computed tomography (CT), or pulmonary angiogram prior to Day 1 that shows **no evidence of thromboembolic disease** (ie, should note normal or low probability for pulmonary embolism).
- 1) Medical history or ongoing gastrointestinal disorders potentially affecting the **absorption of SAR439859 or letrozole**.
 - 2) Unable to complete surgery with curative intent after conclusion of **neoadjuvant systemic therapy**

Should extract
"thromboembolic disease"
with value "no"

Extract "absorption of letrozole"
Instead of "medical history or ongoing
gastrointestinal disorders"

The important terms or phrases in the examples are bolded for reference.

Discussion: Limitation

1. AutoCriteria does not currently handle certain sophisticated criteria conditions including “at least one of these criteria,” and “one or more of following criteria.”
2. Relatively low performance for BC and MM reflects the complexity of information and language in cancer trial text compared to other diseases
3. The system suffers from low recall for the “Treatment history” criteria type for BC as it fails to capture a few prior therapy entities. These are mostly broad therapy terms such as:
 - Cancer therapy (investigational, approved)
 - Chemotherapy (any previous, neo/adjuvant, cytotoxic)
 - Corticosteroid therapy
 - Endocrine therapy (adjuvant)
 - Systemic therapy
 - Anti-cancer therapy, and radiotherapy







Future work

1. Normalizing the extracted criteria entities into **standard terminologies** (such as ICD codes for disease criteria) to facilitate real-world applications
2. **Larger sample** of trial documents and explore using GPT-4 to extract information in a **multi-turn question-answering** fashion
3. **Other large language models** such as Davinci-003 was also considered, however, manual review on a subset of trials suggested their **underperformance** compared to GPT-4
4. Our main focus was on investigating its zero-shot generalizability across diverse diseases rather than comparing different model variants
5. Another interesting future direction could be the **generated large-scale data** for fine-tuning available pretrained language models

Conclusion

- A generalizable GPT-based system that can identify granular eligibility criteria information from clinical trial documents across a variety of disease domains
- Query part of the prompts are separated into different components that can be easily modified while extending to new diseases
- The disease-specific components in the query can be filled in through expert input
- While the traditional deep learning methods usually rely on manually annotated data and retrain the models, our proposed approach generalizes well across disease domains without requiring annotation or retraining.
- Such a generalizable and scalable criteria extraction system could significantly streamline the patient recruitment process and expedite the construction of criteria knowledge base

Why We Support and Encourage the Use of Large Language Models in *NEJM AI* Submissions

Authors: Daphne Koller, Ph.D. , Andrew Beam, Ph.D. , Arjun Manrai, Ph.D. , Euan Ashley, M.B., Ch.B., D.Phil. , Xiaoxuan Liu, M.B.Ch.B., Ph.D. , Judy Gichoya, M.B.Ch.B., M.S. , Chris Holmes, Ph.D. ,  for the editors and editorial board of *NEJM AI** [Author Info & Affiliations](#)

Published December 11, 2023 | *NEJM AI* 2023;1(1) | DOI: 10.1056/AIe2300128 | [VOL. 1 NO. 1](#)

LLMs may enable and accelerate behaviors both good and bad. As is the case for many technologies — ranging from nuclear power and computers to stem cell research and genetic engineering to cryptography.

The better the tools that we provide to scientists,
the greater their ability to produce robust, novel scientific findings and disseminate them broadly.