Granularly, Precisely, and Timely: Leveraging Large Language Models for Safety and Efficacy Extraction in Oncology Clinical Trial Abstracts (SEETrials)

Hunki Paek, PhD[1]*, Kyeryoung Lee, PhD[1] *, Surabhi Datta, PhD[1], Liang-Chin Huang, PhD[1], Josh Higashi[1], Nneka Ofoegbu, PharmD[1], Jingqi Wang, PhD[1], Frank J. Manion, PhD[1], Jeremy L. Warner, MD[2], Hua Xu, PhD[3], Xiaoyan Wang, PhD[1]

[1] IMO health, Rosemont, IL 60018 USA
[2] Brown University, Providence, RI 02912 USA
[3] Yale University, New Haven, CT 06510 USA

*** Equal contribution**

Corresponding author

Xiaoyan Wang, PhD
IMO health,
9600 West Bryn Mawr Avenue
Suite 100, Rosemont, IL 60018 USA
Phone: 8476136655
Email: xw108@caa.columbia.edu

KEYWORDS

Cancer Clinical Trial Outcome Extraction
Conference Abstracts
GPT-4
Zero-shot Prompting
Large Language Model

ABSTRACT

Objective

Conference abstracts provide preliminary evidence for clinical trial outcomes. This study aims to develop an automatic extraction system to precisely extract and convert granular safety and efficacy information from abstracts into a computable format for timely downstream analysis.

Materials and Methods

We collected multiple myeloma clinical trial abstracts from ASCO, ASH, and PubMed (2012-2023) to develop SEETrials, a GPT-4 based system. Qualitative and quantitative evaluations were conducted. Descriptive data analysis on efficacy and safety entities was performed. The generalizability of the system was tested in three other cancer trial studies.

Results

SEETrials achieved precision, recall, and F1 of 95.5%, 93.9%, and 94.7% across 70 data elements. Analysis of 245 multiple myeloma clinical trial abstracts revealed variations in safety and efficacy entity distribution across different modalities and phases. Application to other cancer trial studies demonstrated consistent performance with precision, recall, and F1 of 96.9%, 95.4%, and 96.1%, respectively.

Discussion

Qualitative error analysis identified a common source of inaccuracy, notably in cohort identification and categorization. Phase1/2 studies prioritize early indications of treatment efficacy, while phase 2/3 studies emphasize prolonged effects. Adverse event distribution is consistent across phases, with noticeable increases in fatal events in later phases, suggesting longer follow-up or more patients may reveal severe adverse events missed in phase 1 safety studies.

Conclusion

SEETrials displayed high accuracy and generalizability to diverse drug modalities and disease domains. Its capacity to streamline large-scale dataset analysis is crucial for advancing clinical trial research, ensuring timely and accurate data extraction, and facilitating efficient dissemination.

INTRODUCTION

Effective clinical decision-making relies profoundly on the continuous flow of data and outcomes derived from clinical trial studies [1]. The timely dissemination of clinical trial outcomes, including risks and benefits from early-phase trials, holds importance in facilitating well-informed decision-making in patient care, ultimately influencing patient treatment outcomes [2]. Swift access to data also plays a critical role in optimizing patient recruitment for clinical trials, a pivotal factor in expediting the development of novel treatments, particularly in high-need medical areas such as oncology [3].

Conference abstracts often serve as crucial sources of preliminary findings, unveiling clinical trial results and providing insights into the efficacy and safety of investigational treatments. However, the retrieval of these abstracts poses a challenge, and trial results initially presented in the conference abstracts may not progress to subsequent journal publication [4]. This discrepancy is significant, with more than 50% of outcomes from abstracts and nearly one-third of randomized trial results, initially introduced in conference abstracts, failing to advance to full publication, and this trend has not shown improvement over time [5]. Moreover, despite the recommendation to include conference abstracts in systematic literature reviews by both Cochrane and the United States National Academy of Sciences [6], concerns persist regarding the unclear impact of unpublished conference abstracts on the final conclusions [7] and inconsistency between conference abstracts and papers in reporting results [8–10]. Given these various benefits and concerns, a complete and thorough systematic review requires clinical outcomes to be extracted not only from published articles but also from all conference abstracts. This approach allows for the generation of a comprehensive clinical evidence base and for gaining insights into the consistency between results from the conference and full paper abstracts. Nevertheless, extracting granular information from these sources with sufficient precision, particularly within the context of large-scale datasets, in a timely manner, poses an additional challenge.

Natural Language Processing (NLP) has proven to be a transformative technology in the medical and clinical research domains, offering a significant advantage in extracting clinical data elements from unstructured free-text documents [11–13]. Recently, large language models (LLMs) have gained tremendous attention within the NLP domain, exhibiting remarkable capabilities in understanding context, reasoning, and processing vast amounts of textual information [14–16]. The introduction of more potent generative pre-trained transformer (GPT) models such as ChatGPT and GPT-4, have prompted a thorough investigation into their effectiveness across diverse medical applications.

Within the realm of information extraction from clinical documents, various LLMs have been deployed in diverse applications, such as analyzing radiology reports [17–19] and electronic

health records [15]. In the clinical trial domain, LLMs play crucial roles in trial information retrieval [20], criteria text generation [21], and clinical trial eligibility criteria analysis [22]. Despite recognized limitations, ChatGPT has also demonstrated value in supporting systematic literature reviews [23]. Notably, prior studies have not explored the application of these models for extracting clinical trial outcomes from conference abstracts, especially in scenarios involving dense information presented in both text and tables.

In this study, we present "SEETrials," an innovative LLM pipeline designed to automatically extract safety and efficacy information from a substantial volume of conference and published journal abstracts related to multiple myeloma clinical trial studies. The significance of this study lies in addressing the challenge of extracting comprehensive information, especially from conference abstracts, which often contain condensed tables with crucial contents briefly described or left unspecified due to word limitations. Furthermore, we evaluated the generalizability of our system by applying it to three additional types of cancer clinical trial studies, assessing its effectiveness across diverse cancer research contexts. The overarching objective is to enhance the accessibility of timely and precise clinical data in a computable format for downstream analysis and synthesizing clinical evidence.

BACKGROUND

Conference abstracts play a crucial role in generating evidence for clinical decision-making across diverse domains, but their dissemination is limited. Time lags from conference abstracts to full paper publication [4], and a low conversion rate from abstract to journal paper [9,24,25] contribute to this disparity. Various factors, including time constraints [26] hinder the publication of abstracts in full journals. Publication bias further compounds this issue, resulting in the underrepresentation of "not positive or negative" results or "positive but not statistically significant" results in publications [5,27,28] despite the significance of all study results in influencing clinical decision-making and patient care. Difficulties in extracting information from free-text formats also pose additional challenges.

We focused on multiple myeloma clinical trial studies as a use case for development of our LLM model, recognizing the pressing need for timely and precise clinical data in this clinical context. Despite significant advancements in treatment options and survival rates over the last 15 years, multiple myeloma remains an incurable malignancy, marked by elevated rates of relapse and therapy resistance [29]. Furthermore, the introduction of numerous new therapies has created significant uncertainty regarding therapy selection and sequencing, given the limited availability of consensus guidelines and recent systematic literature reviews [30,31]. Our analysis concentrated on the most recent mechanism of action drugs developed in the last decade to explore the feasibility of ensuring timely access to accurate clinical outcomes from both annual conference abstracts and published journal abstracts. Leveraging the resulting LLM pipeline should enhance clinical decision-making and facilitate patient recruitment in clinical trials.

MATERIALS AND METHODS

Data

For the model development and evaluation, we collected 100 clinical trial abstracts related to multiple myeloma from various drug groups, focusing on five different modalities. The sources included the American Society of Clinical Oncology (ASCO: https://ascopubs.org/jco/meeting), American Society of Hematology conferences (ASH: https://ashpublications.org/blood/issue/142/Supplement%201), and PubMed (https://pubmed.ncbi.nlm.nih.gov/) websites. The selection was based on keywords ("clinical trial" AND "each modality"). The modalities covered are 1) chimeric antigen receptor T cell therapy (CAR-T therapy) ; idecabtagene vicleucel and ciltacabtagene autoleucel, 2) Bispecific Antibodies (BsAbs)/Bispecific T cell engager (BiTE) therapy; talquetamab-tgvs, teclistamab-cqyv, and elranatamab-bcmm, 3) Antibody-drug conjugates (ADC) therapy; Belantamab-mafodotin, and 4) Cereblon E3 ligase modulator therapy (CELMoD) ; Iberdomide and Mezigdomide 5) Others abstracts including Histone deacetylase inhibitor (HDACi) and immune checkpoint inhibitors (ICI); pembrolizumab.

For the prompt design phase, we used 20 abstracts, followed by 30 abstracts for prompt calibration. Subsequently, another set of 50 abstracts was employed for the system evaluation. "Prompt design" involved creating initial prompts through manual scrutiny of abstracts, while "prompt calibration" refined prompts iteratively based on expert evaluation of the system output, both guided by expert judgment (HP and KL).

After establishing the system, we implemented the system on an additional 145 multiple myeloma clinical trial study abstracts for comprehensive data analysis, resulting in a total of 245 abstracts used for the data analysis. To assess the generalizability of our system, we included 18 additional clinical trial abstracts (6 related to breast cancer, 7 to non-small cell lung cancer, and 5 to leukemia) sourced from PubMed.

SEETrials

We developed SEETrials to extract crucial details from clinical trial studies presented at annual conferences and published journal abstracts, leveraging the capabilities of GPT-4. The subsequent sections outline the components of the system and Figure 1 provides a schematic overview.

Pre-processing steps for abstracts collection and table conversion

We collected abstract texts from ASCO, ASH, and PubMed websites. Abstract texts from PubMed were directly saved in text format, while those from ASCO and ASH websites were

initially captured in a Word file and later converted to a text file (UTF-8) to facilitate the efficient loading of a substantial number of abstracts for automated data extraction. Notably, abstracts from annual conferences often contain tabular data structures, and the word to text file conversion process may lead to formatting issues or broken tables, creating challenges for GPT comprehension. To address potential formatting issues, we developed an "Automatic Table Organizer" module leveraging GPT, effectively realigning columns and rows. For ASH conference abstracts featuring tables in picture format, we applied Optical Character Recognition (OCR) to convert picture files into text files before employing the "Automatic Table Organizer" for further processing. Two input variations for abstracts with tables were then generated, one with the table included and one without.

Knowledge ingestion

SEETrials operates on a knowledge-driven framework and require foundational comprehension and knowledge of oncology and clinical trials. We did not presume that this knowledge was inherent in the GPT models we used. Rather we designed prompts to incorporate foundational comprehension and knowledge of oncology clinical trials. We also incorporated an understanding of their primary and secondary endpoints for efficacy and safety outcome measurement.

Prompt modeling for data extraction and merging outcomes

We directed our GPT system to extract and align information related to cohort composition. To enhance GPT performance and output accuracy, we employed two distinct GPT modules for each of the comprehensive prompts: one for "Individual cohorts/arms and entire group" and another for " Individual cohorts/arms and subgroup group".

Each prompt included three main components: a general instruction to extract study details, clinical safety and efficacy outcomes; the abstract text; and a query. The query explicitly requested information on entities and values for specific tables, with detailed instructions for formatting.

The outputs from both modules were combined in the subsequent step, where we merged outputs from the "Individual cohorts/arms and entire group" and " Individual cohorts/arms and subgroup group" modules. This merging process employed a prompt to accurately combine extracted entities and values and eliminate any overlaps.

To finalize the output format, we transferred the results to csv file. In this file, each entity was represented in individual rows, and associated values for each cohort, entire group, or subgroup were organized in columns to facilitate further downstream analysis.

Postprocessing

We implemented a post-processing module to refine outputs from the data extracting module by addressing two key formatting challenges: (1) addressing undesired allocations of entities and their corresponding values in the outputs and (2) ensuring that each cell within the value's column contained a singular value, optimizing the presentation for computational analysis.

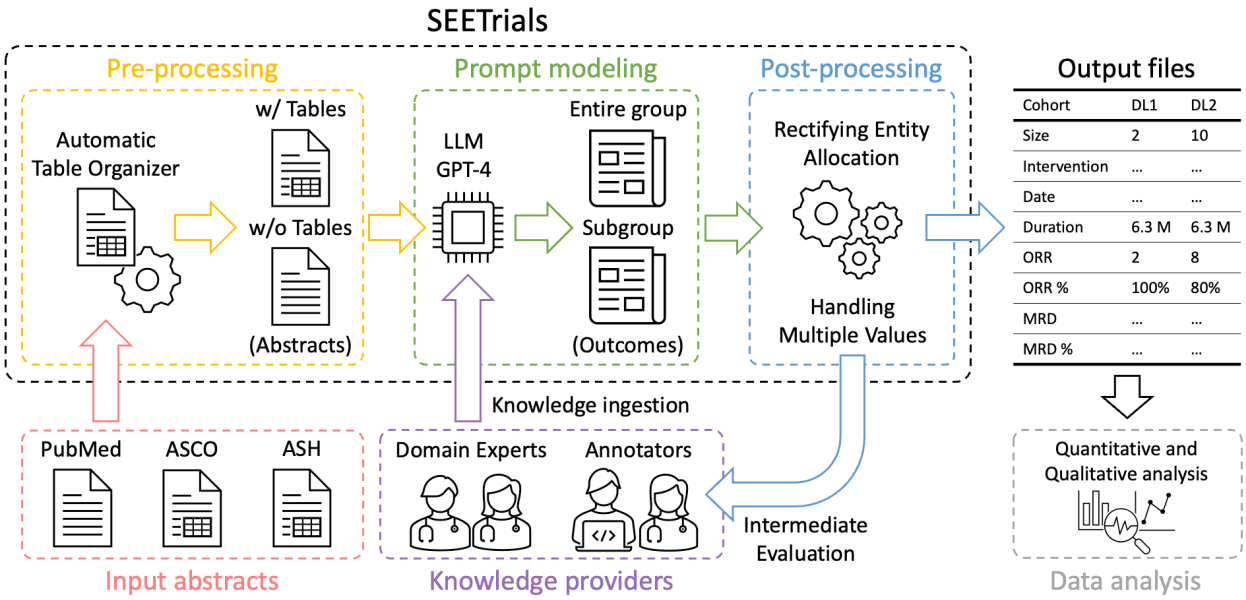
Evaluation

For the system assessment, we used 50 randomly selected abstracts, manually annotated by knowledge experts (KL, NO, and HP). The evaluation was conducted in two ways: a strict approach and a relaxed approach. In the strict evaluation, entities and their values, whether in a single arm or multiple arms, were treated as a unified entity. Any inaccuracies or omissions in values across different arms, were considered as false positives or false negatives or both for each entity (each row in the output file). In the relaxed evaluation, entities and values for each arm were assessed independently (each cell in the output file). We presented the precision, recall, and F1 scores of SEETrials for each abstract, considering both strict and relaxed evaluations.

Descriptive Data Analysis

The Descriptive Data Analysis initiated with the normalization of extracted entities, converting terms such as "OSR" to "Overall Survival Rate (OSR)" for standardization. Subsequently, the total occurrences of entities across diverse modalities were tabulated, leading to the creation of a master data file. We examined the distribution and a quantitative overview of entities, specifically emphasizing treatment efficacy and safety, in conjunction with trial phases.

Figure 1. SEETrials System Overview.
SEETrials is an autonomous system designed to extract critical details from clinical trial studies presented at annual conferences and published journal abstracts. Utilizing the capabilities of GPT-4, the system is delineated in subsequent sections, with a schematic overview of its components.



ASCO, American Society of Clinical Oncology; ASH, American Society of Hematology; LLM, Large language model; GPT, Generative Pre-trained Transformer; ORR, overall response rate; MRD, minimum residual disease.

RESULTS

Establishment and Evaluation of the LLM System, SEETrials

Recognizing potential formatting challenges in abstracts with tables, we developed an "Automatic Table Organizer" GPT module. This module effectively realigned columns and rows in the text file, rectifying any formatting discrepancies. Following the application of the Automatic Table Organizer, we observed improved table information extraction but noted a decrease in performance for extracting main body text information (data not shown). To address this, we input two abstract sets into our GPT system: one with organized tables and the other without tables. The outcomes from two different input sets were combined in the Merging module.

The performance evaluation of SEETrials, conducted on a subset of 50 selected abstracts, is delineated in Table 1. The system demonstrated a high level of precision, recall, and F1 score, with an overall performance of 95.5%, 93.9%, and 94.7%, respectively, across all Mechanisms of Action (MOAs) in strict evaluation. In the relaxed evaluation, the system exhibited an even higher performance, achieving an overall precision of 98.1%, a recall of 98.9%, and an F1 score of 98.5%. Comprehensive details, including the total number of entities, and values counted, as well as true positive, false positive, and false negative numbers, are provided in the supplementary Table 1.

Table 1. Performance Metrics of the SEETrials System in Multiple Myeloma Clinical Trial Study Abstracts.

	STRICT			**RELAXED**		
Abstract No.	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Average	**0.955**	**0.939**	**0.947**	**0.979**	**0.987**	**0.983**
1	1	1	1	1	1	1
2	1	1	1	1	1	1
3	0.965	0.982	0.973	0.996	0.982	0.989
4	1	1	1	1	1	1
5	0.927	0.864	0.894	0.960	0.984	0.972
6	0.957	0.943	0.950	0.994	0.991	0.993
7	1	1	1	1	1	1
8	1	0.892	0.943	0.925	1	0.961
9	1	0.985	0.992	0.990	1	0.995
10	1	1	1	1	1	1
11	1	1	1	1	1	1

12	0.917	0.917	0.917	0.975	0.975	0.975
13	1	0.975	0.987	0.980	1	0.990
14	0.960	0.960	0.960	1	0.990	0.995
15	0.824	0.824	0.824	0.949	0.949	0.949
16	1	1	1	1	1	1
17	1	0.980	0.990	0.992	1	0.996
18	1	0.941	0.970	0.970	1	0.985
19	0.962	0.962	0.962	0.990	0.990	0.990
20	0.635	0.635	0.635	0.878	0.878	0.878
21	0.923	0.947	0.935	0.982	0.966	0.974
22	1	1	1	1	1	1
23	1	1	1	1	1	1
24	0.931	0.915	0.923	0.972	0.979	0.976
25	1	0.891	0.942	0.924	1	0.961
26	0.984	0.984	0.984	0.996	0.996	0.996
27	1	0.962	0.980	0.974	1	0.987
28	1	0.944	0.971	0.977	1	0.988
29	1	1	1	1	1	1
30	1	1	1	1	1	1
31	1	0.983	0.992	0.994	1	0.997
32	1	1	1	1	1	1
33	0.900	0.900	0.900	0.990	0.985	0.987
34	1	0.979	0.989	0.986	1	0.993
35	0.914	0.914	0.914	0.983	1	0.991
36	0.816	0.816	0.816	0.969	0.987	0.978
37	0.767	0.742	0.754	0.934	0.988	0.961
38	1	1	1	1	1	1
39	1	1	1	1	1	1
40	1	1	1	1	1	1
41	1	1	1	1	1	1
42	1	1	1	1	1	1
43	1	1	1	1	1	1
44	0.974	0.974	0.974	0.995	0.995	0.995
45	1	1	1	1	1	1
46	0.792	0.764	0.778	0.939	0.939	0.939
47	0.972	0.921	0.946	0.949	0.978	0.964
48	0.895	0.819	0.855	0.914	0.952	0.933
49	1	1	1	1	1	1
50	0.760	0.760	0.760	0.982	0.949	0.965

Qualitative Error Analysis

We conducted a qualitative error analysis on individual documents, presenting our findings from the abstracts with F1 scores lower than 80% (4/50 abstracts, # 20, 37, 46, and 50) in Figure 2.

- Case 1 (#20) is a phase 1 dose-escalation study determining the recommended phase 2 dosage (RP2D). It included information on the entire cohort and intervention dosage. However, data for treatment outcomes, including safety and efficacy, were derived from a small portion of the RP2D group. The values were redundantly presented for both the entire group and the RP2D cohort (Figure 2A).

- Case 2 (#37) represents a dose escalation study (trial phase information was not specified). The study comprises four distinct arms, each associated with a different dosage. Despite the abstract's mention of these four arms and the investigational drug dosage for each, the cohort size and study findings outlined in the abstract pertain to the combined cohorts 1 and 2, specifically the lowest two doses. Notably, SEETrials erroneously presented identical information for both cohort 1 and cohort 2, resulting in a uniform cohort size of 6 for each (Figure 2B).

- Case 3 (#46) introduces a phase 1 dose-escalation study featuring two arms, each linked to a different dosage. The abstract provides information on cohort size and treatment outcomes for the overall group, as well as individual details for Arm 1 and Arm 2. Furthermore, the abstract introduces specific information about dose escalation subgroups, encompassing cohort size and response rates. While SEETrials accurately predicted most entities and values for the entire cohort and Arms 1 and 2, it did not capture information about the dose escalation subgroups (Figure 2C).

- Case 4 (#50) is a phase 3 study with one control arm and three intervention arms with an investigated drug at three different dosages. SEETrials accurately extracted information for the three intervention arms but missed details on the control arm. This led to all clinical finding entities being considered false negatives in STRICT evaluation (F1 76%), though in RELAXED evaluation, the overall performance was higher (F1 96%) due to correct extraction for the three intervention arms (Figure 2D).

As described, a prevalent source of inaccuracy in the extraction process pertains to cohort identification and categorization. Abstracts frequently encompass diverse clinical information for multiple cohorts, including Dose-escalation, Dose-expansion, RP2D groups, entire groups, and subgroups. The lack of explicit descriptions in the text occasionally poses a challenge for the

system in determining the appropriate allocation of values. Additionally, our observations revealed the presence of other errors in the extraction process.

- Author's affiliation information in PubMed abstracts was extracted and used as the study location in cases where actual study location was not mentioned in the abstracts.

- Cohort size discrepancies arise when the initial abstract mentions a specific size (e.g., n=14) but describes outcomes for a different number (e.g., n=10). Lack of clarification about the missing patients in the cohort further complicates allocation, leading to occasional misplacement of entities and values.

- GPT occasionally captures only the percentage value (e.g., 25%) without extracting other relevant information (e.g., 14 patients). To ensure output consistency, improvement is needed for extracting all values and allocating them into separate cells.

Figure 2. Examples of the qualitative error analysis. A) Partial output from the abstract #20 ASCO_ADC. B) Partial output from abstract #37 PubMed_CAR T. C) Partial output from abstract #46 ASCO_CAR T. D) Partial output from abstract #50_ASCO_CELDoM. (Yellow Highlight: False Positive – Not annotated in Gold Standard but predicted by SEETrials; Red-Colored Characters: False Negative – Annotated in Gold Standard but not predicted by SEETrials)

A)

Clinical Findings	Entire Group	Subgroup 1 (RP2D)
---	---	---
Registry Number	NCT03399799	NCT03399799
Cohort Size	174	28
Intervention for Each Arm	Talquetamab intravenously (IV; range 0.5–180 µg/kg) or subcutaneously (SC; range 5.0–800 µg/kg) weekly or biweekly	Weekly SC 405 µg/kg talquetamab
Clinical Study Information	Not mentioned	
Overall Response Rate	63% (Not mentioned)	63%
Adverse Events - CRS	79% (Not mentioned)	79%

B)

Clinical Findings	Cohort 1,2 (Lowest 2 dose levels)	Cohort 2 (Lowest 2 dose levels)	Cohort 3 (Third dose level)	Cohort 4 (Fourth dose level)
---	---	---	---	---
Registry Number	NCT02215967	NCT02215967	NCT02215967	NCT02215967
Cohort Size	6	6	Not mentioned	2
Intervention for Each Arm	CAR-BCMA T cells (lowest 2 dose levels)	CAR-BCMA T cells (lowest 2 dose levels)	CAR-BCMA T cells (third dose level)	CAR-BCMA T cells (fourth dose level of 9 × 10(6) CAR(+) T cells/kg body weight)

C)

Clinical Findings	Arm1 (DL1)	Arm2 (DL2)	Entire Group	Dose escalation group
----------	----------	----------	----------	
Registry Number	NCT04155749	NCT04155749	NCT04155749	NCT04155749
Cohort Size	19	6	25	12
Intervention for Each Arm	CART-ddBCMA infusion at 100 x 106 CAR+T cells	CART-ddBCMA infusion at 300 x 106 CAR+T cells	CART-ddBCMA infusion	CART-ddBCMA dose escalation
Median Follow-up	9.8 months	9.8 months	9.8 months	
Follow-up Range	2-23.7 months	2-23.7 months	2-23.7 months	longest follow up
Overall Response Rate (ORR)	Not mentioned	Not mentioned	100%	100%
Stringent Complete Response (sCR)/Complete Respons	67%	Not mentioned	67%	75%

D)

Clinical Findings	Arm1 (or Cohort 1)	Arm2 (or Cohort 2)	Arm3 (or Cohort 3)	Arm4 (or Cohort 4)
---	---	---	---	
Registry Number	NCT05552976	NCT05552976	NCT05552976	NCT05552976
Cohort Size	Not specified	Not specified	Not specified	Not specified
Intervention for Each Arm	MeziKd (1.0 mg), carfilzomib, and dexamethasone	MeziKd (0.6 mg), carfilzomib, and dexamethasone	MeziKd (0.3 mg), carfilzomib, and dexamethasone	carfilzomib and dexamethasone

Analyzing Extracted Entities from Abstracts

We curated a comprehensive dataset comprising 245 abstracts, covering diverse therapeutic modalities. This includes 130 abstracts related to CAR-T cell therapies, 63 to BsAbs/BiTE therapies, 38 to ADC therapies, 6 to CELMoD therapies, 4 to HDACi therapies, and 4 to ICI therapies. Among these, there were 75, 47, 46, and 16 abstracts corresponding to phases 1, 1/2, 2, and 3. 61 abstracts did not have phase information. The distribution of clinical trial phases in each modality is detailed in Table 2.

Table 3 and Supplemental Table 2 present the descriptive statistics of extracted clinical outcomes, focusing on treatment effectiveness and safety across three distinct modalities. The complete list of extracted data elements was detailed in Figure 3. Notably, the overall response rate emerged as the most frequently reported entity related to treatment efficacy in clinical trial abstracts, with percentages of 23.9% in CAR T studies, 27.3% in BsAbs/BiTE studies, and 25% in ADC studies. Complete response rates were prominent, particularly in CAR T studies at 21.8% and in BsAbs//BiTE studies at 15.4%. Similarly, (very good) partial responses were noteworthy in ADC studies, accounting for 20% of the mentions.
The entity of minimal residual disease featured prominently in CAR T cell therapy clinical trial abstracts (8%), surpassing mentions in BsAbs/BiTE (3.5%) and ADC (2.5%) studies.

Examining safety-related entities, neutropenia emerged as the most frequently mentioned adverse event across all three modalities, with percentages of 13.4% in CAR T, 13.5% in BsAbs//BiTE, and 12.4% in ADC studies. Cytokine release syndrome was prevalent, with rates of 13.0% in CAR T, 9% in BsAbs//BiTE, and 8.6% in ADC studies. Thrombocytopenia and anemia were also commonly cited adverse events across modalities, showing variations in percentages. Infection occurrences were noted in 4.8% of CAR T studies, while higher percentages of 12% and 8.6% were observed in BsAbs/BiTE and ADC studies, respectively.

Figure 4 provides a visual summary of the percentages of 11 efficacy and safety-related entities across the three modalities, offering a comprehensive overview of the comparative landscape. These findings contribute to a deeper understanding of the distinct clinical profiles associated with CAR T cell therapy, BsAbs/BiTE, and ADC modalities, facilitating informed decision-making in therapeutic development.

Table 2. Characteristics overview of abstracts included.

	Total	phase 1	phase 1/2	phase 2	phase 3	Not mentioned
CART	130	40	16	26	5	43
BsAbs/BiTE	63	19	18	10	6	10
ADC	38	10	9	10	2	7
CELMoD&Others	14	6	4	0	3	1
Total	245	75	47	46	16	61

CAR T, chimeric antigen receptor T cell; BsAbs, Bispecific antibody; BiTE, Bispecific T cell engager; ADC, antibody drug conjugate; Cereblon E3 ligase modulator therapy, CELMoD.

Table 3. Descriptive statistics of extracted clinical outcomes on treatment efficacy.

Entities related to efficacy	CAR T (n)	CAR T (%)	BsAbs/BiTE (n)	BsAbs/BiTE (%)	ADC (n)	ADC (%)
ORR	69	23.9	39	27.3	20	25.0
CR	63	21.8	22	15.4	8	10.0
(VG)PR	43	14.9	19	13.3	16	20.0
PFS	32	11.1	13	9.1	7	8.8
MRD	23	8.0	5	3.5	2	2.5
OS	20	6.9	15	10.5	8	10.0
DoR	15	5.2	15	10.5	8	10.0
SD	8	2.8	2	1.4	4	5.0
PD	7	2.4	0	0.0	0	0.0
TTR	5	1.7	3	2.1	2	2.5
MR	4	1.4	3	2.1	1	1.3
TTP	0	0.0	3	2.1	0	0.0
TTTD	0	0.0	1	0.7	0	0.0
TTTF	0	0.0	0	0.0	0	0.0
TTNT	0	0.0	3	2.1	2	2.5
DCR	0	0.0	0	0.0	2	2.5

CAR T, chimeric antigen receptor T cell; BsAbs, Bispecific antibody; BiTE, Bispecific T cell engager; ADC, antibody drug conjugate; ORR, overall response rate; CR, complete response; (VG)PR, (very good) partial response; PFS, progression free survival; MRD, minimum residual disease; OS, overall survivor; DoR, duration of response; SD, stable disease; PD, progressive disease; TTR, time to response; MR, minimal response; TTP, time to progress; TTTD, time to treatment discontinuation; TTTF, time to treatment failure; TTNT, time to next treatment; DCR, disease control rate.

In Figures 5 and 6, we present a detailed breakdown of trial numbers and percentages stratified by clinical trial phases (phases 1, 1/2, 2, and 3) for efficacy and safety-related entities, respectively. Notably, response rates and status-related entities were frequently reported in phase 1 studies, indicating a focus on early indications of treatment efficacy. Conversely, survival or treatment duration-related entities were predominantly extracted in phase 2 studies, highlighting a shift towards assessing prolonged treatment effects and overall therapeutic impact.

The distribution of adverse event entities exhibited consistency across different phases of clinical trials. Only noticeable differences were observed in fatal adverse events which appear significantly more in later phase trials compared to phase 1 initial study. This observation suggests that longer follow-up times and or increased number of patients can show more severe grade of adverse events that were not detected in phase 1 safety study.

Figure 3. Comprehensive Overview of Extracted Granular Entities for Study details, Treatment Efficacy, and Safety. This illustrates a comprehensive breakdown of extracted granular entities, encompassing Study Details, Treatment Efficacy, and Treatment Safety parameters, with a total of 16, 18, and 36 entities respectively, offering a detailed overview of the clinical trial landscape.

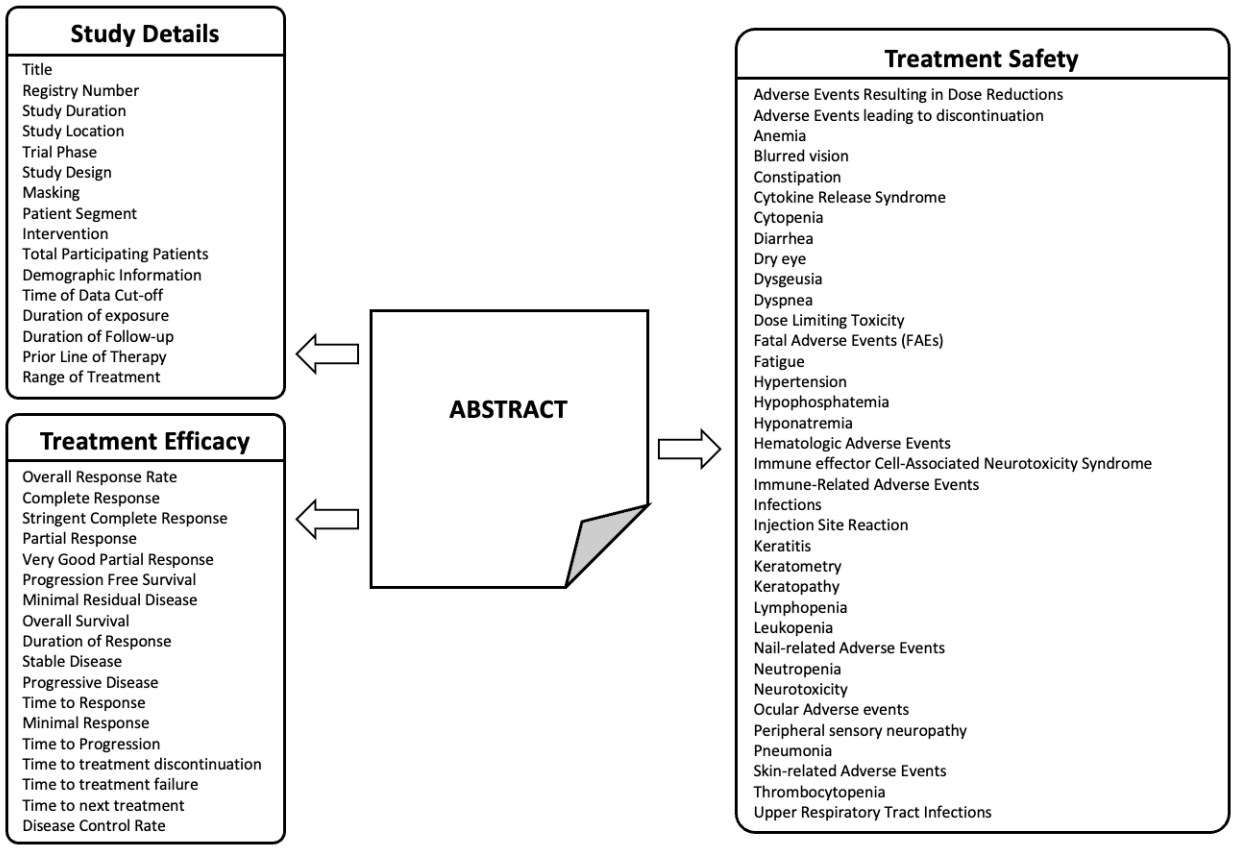
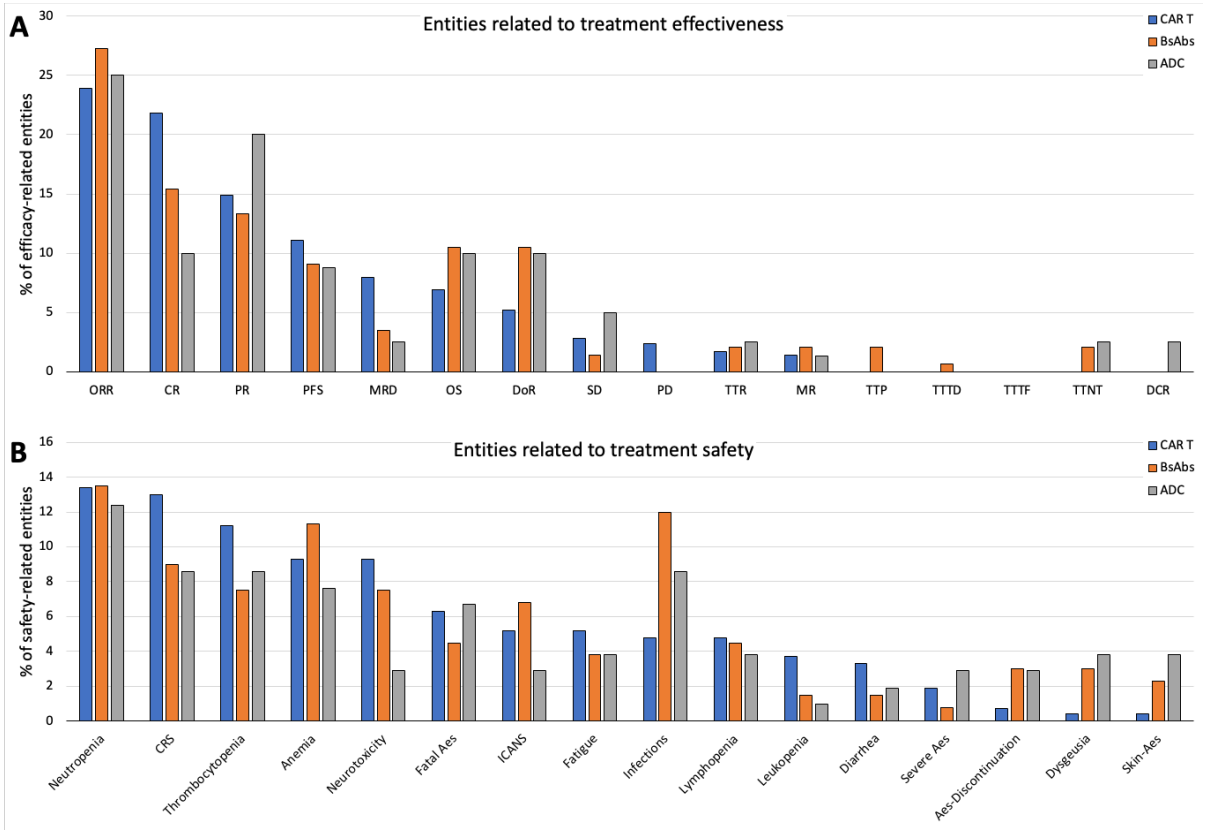
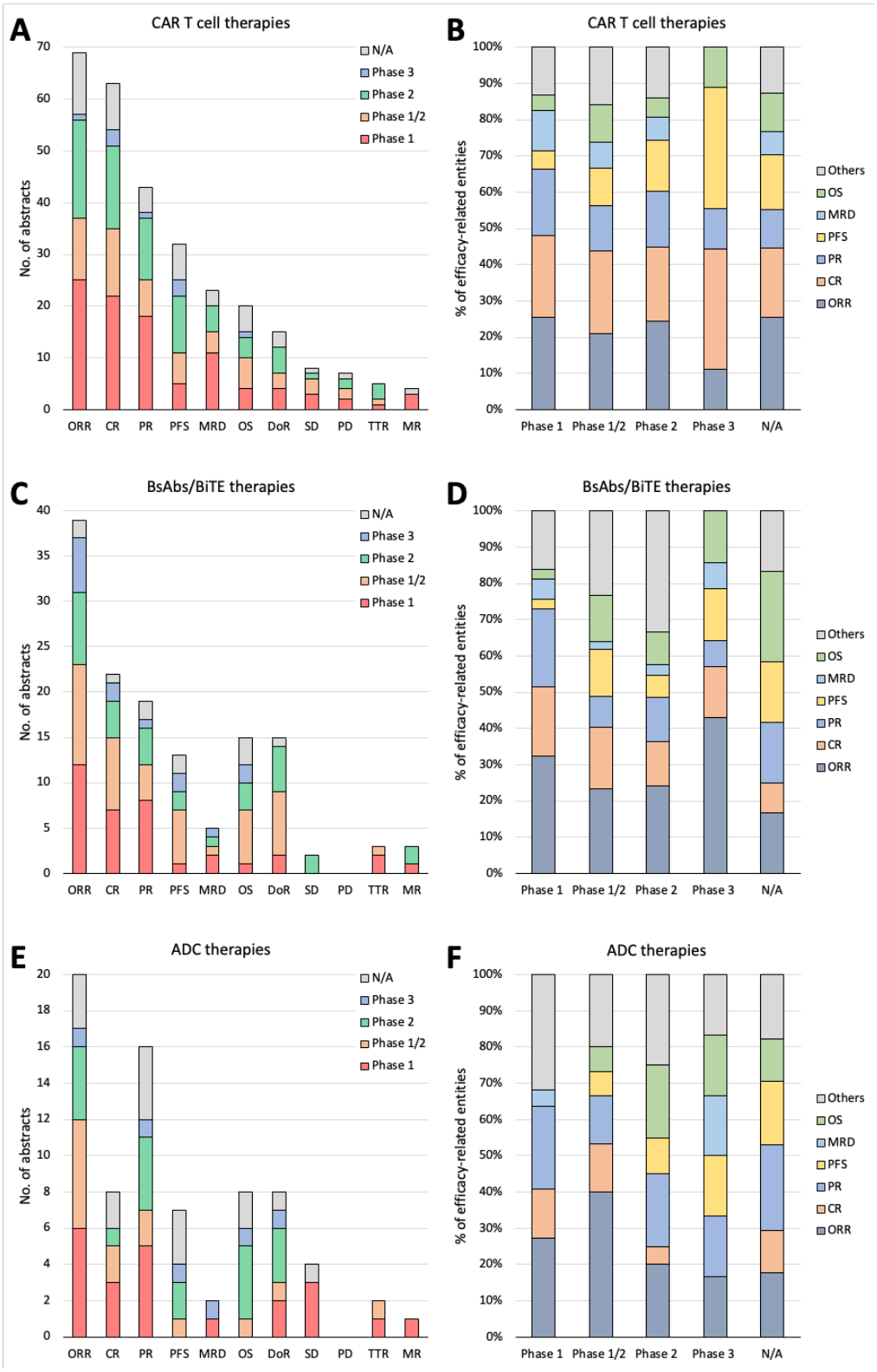


Figure 4. Comparative Landscape of Efficacy and Safety Entities Across CAR T, BsAbs/BiTE, and ADC Modalities. This visual summary illustrates the percentages of 11 efficacy and 13 safety-related entities across CAR T cell therapy, BsAbs/BiTE, and ADC modalities, providing a comprehensive overview of their comparative clinical profiles. A) Entities related to treatment effectiveness. B) Entities related to treatment safety.



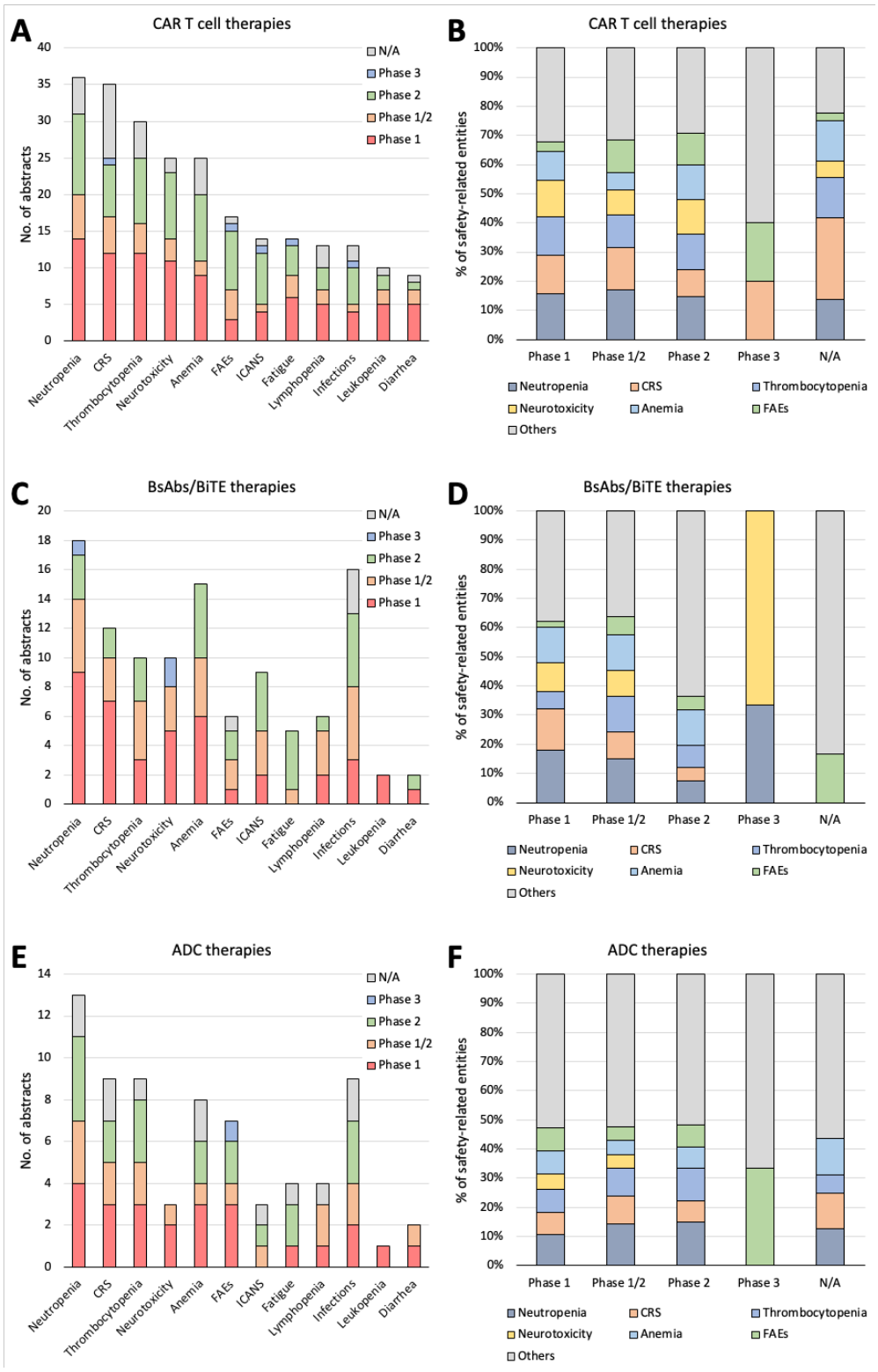
CAR T, chimeric antigen receptor T cell; BsAbs, Bispecific antibody; BiTE, Bispecific T cell engager; ADC, antibody drug conjugate; ORR, overall response rate; CR, complete response; (VG)PR, (very good) partial response; PFS, progression free survival; MRD, minimum residual disease; OS, overall survivor; DoR, duration of response; SD, stable disease; PD, progressive disease; TTR, time to response; MR, minimal response; TTP, time to progress; TTTD, time to treatment discontinuation; TTTF, time to treatment failure; TTNT, time to next treatment; DCR, disease control rate; CRS, cytokine release syndrome; Aes, adverse events; ICANS, immune effector cell associated neurotoxicity syndrome.

Figure 5. **Trial Numbers Stratified by Clinical Trial Phases for Efficacy-Related Entities.**
A detailed breakdown of trial numbers and percentages is presented, categorizing clinical trials into phases 1, 1/2, 2, and 3 for efficacy-related entities. A and B: CAR T cell therapies. C and D: BsAbs/BiTE therapies. E and F: ADC therapies.



CAR T, chimeric antigen receptor T cell; BsAbs, Bispecific antibody; BiTE, Bispecific T cell engager; ADC, antibody drug conjugate; ORR, overall response rate; CR, complete response; (VG)PR, (very good) partial response; PFS, progression free survival; MRD, minimum residual disease; OS, overall survivor; DoR, duration of response; SD, stable disease; PD, progressive disease; TTR, time to response; MR, minimal response.

Figure 6. Trial Numbers Stratified by Clinical Trial Phases for Safety-Related Entities.
This figure provides a detailed breakdown of trial numbers and percentages stratified by clinical trial phases (phases 1, 1/2, 2, and 3) for safety-related entities. A and B: CAR T cell therapies. C and D: BsAbs/BiTE therapies. E and F: ADC therapies.



CAR T, chimeric antigen receptor T cell; BsAbs, Bispecific antibody; BiTE, Bispecific T cell engager; ADC, antibody drug conjugate; CRS, cytokine release syndrome; FAEs, fatal adverse events; ICANS, immune effector cell associated neurotoxicity syndrome.

Evaluating the Generalizability of the System

We implemented our system on randomly chosen cancer clinical trial studies, including two solid cancers (breast cancer and lung cancer) and one hematologic cancer (leukemia). The performance scores across 18 abstracts are highly consistent with both each other and the earlier 50 multiple abstracts (refer to Table 4). The average precision, recall, and F1 scores in strict evaluation were 96.9%, 95.4%, and 96.1%, respectively. Comprehensive details, including the total number of entities, and values counted, as well as true positive, false positive, and false negative numbers, in these 18 abstract analyses, are provided in the supplementary Table 2.

Table 4. Performance Metrics of the SEETrials System in Various Cancer Clinical Trial Study Abstracts

		STRICT			RELAXED		
Abstract No.	**Disease type**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Average		**0.969**	**0.954**	**0.961**	**0.980**	**0.992**	**0.986**
1	Breast cancer	1	1	1	1	1	1
2	Breast cancer	0.969	0.969	0.969	0.992	0.992	0.992
3	Breast cancer	0.970	0.914	0.941	0.942	1	0.970
4	Breast cancer	0.861	0.861	0.861	0.965	0.965	0.965
5	Breast cancer	1	1	1	1	1	1
6	Breast cancer	0.963	1	0.981	1	0.972	0.986
7	Lung cancer	1	1	1	1	1	1
8	Lung cancer	1	0.955	0.977	0.977	1	0.988
9	Lung cancer	0.838	0.838	0.838	0.959	0.959	0.959
10	Lung cancer	0.967	0.967	0.967	0.983	0.983	0.983
11	Lung cancer	0.958	0.958	0.958	0.993	0.993	0.993
12	Lung cancer	1	0.956	0.977	0.966	1	0.983
13	Lung cancer	1	1	1	1	1	1
14	Leukemia	1	0.963	0.981	0.981	1	0.990
15	Leukemia	1	0.936	0.967	0.975	1	0.987
16	Leukemia	0.973	0.973	0.973	0.989	0.995	0.992
17	Leukemia	0.925	0.925	0.925	0.981	0.981	0.981
18	Leukemia	1	0.976	0.988	0.982	1	0.991

DISCUSSION

In the development of our knowledge-driven system, SEETrials, our primary aim was the systematic extraction of clinical trial study details and treatment outcomes and their corresponding values automatically, from abstracts of multiple myeloma clinical trial studies. This encompassed both conference abstracts and published article abstracts. Our study significantly contributes by strategically addressing the challenge of extracting comprehensive information, particularly from conference abstracts. These abstracts often present condensed tables, where crucial content is concisely summarized or left unspecified in the main text due to word limitations. Moreover, our study significantly improved the accessibility of timely and precise clinical data in a computable format for downstream data analysis.

The implementation of "SEETrials" demonstrated remarkable potential, evidenced by high precision, recall, and F1 scores of 95.5%, 93.9%, and 94.7%, respectively. These robust performance metrics serve as a validation of the efficacy of our pipeline in the accurate extraction of detailed information from multiple myeloma clinical trial studies. Additionally, this study systematically explored the zero-shot capabilities of GPT-4 [20,32], particularly in tasks such as interventions and the recognition of treatment outcome entities. These tasks included the association of outcomes with corresponding values in each cohort and the classification of cohorts, all seamlessly executed through the use of efficiently designed single-turn prompts. Notably, our findings underscored the adaptability of a comprehensive prompt initially tailored for a specific cancer clinical trial (e.g., Multiple Myeloma) to successfully extract trial information from diverse cancer types (e.g., breast cancer, lung cancer, and leukemia). This observation underscores the expansive applicability of SEETrials, ensuring the uniform analysis of clinical trial study outcomes across diverse disease domains. Consequently, this enhances analysis efficiency, scalability, and interoperability within the field. To our knowledge, this study marks a pioneering attempt to explore the zero-shot capabilities of a GPT-powered model for the comprehensive extraction of clinical trial study outcomes from various types of abstracts, in addition to our previously established system—a generalizable clinical trial eligibility criteria extraction system powered by a large language model [22].

In this study, we analyzed a range of drug modalities within the framework of multiple myeloma, with specific emphasis on CAR-T, BsAbs/BiTE, ADCs, and CELMoD therapies. The examination of treatment efficacy and safety outcome distributions across diverse drug modalities and trial phases yielded valuable insights. A thorough understanding of how these outcomes vary across different modalities and phases is crucial for contextualizing clinical trial results. Notably, phase 1 studies exhibit a prevalence of response rates and status-related entities, highlighting an emphasis on early indicators of treatment efficacy while phase 2 studies have more survival or treatment duration-related entities, indicating an evolving focus on prolonged treatment effects and overall therapeutic impact as trials progress. Our study aligns with existing

literature, revealing differences in infection occurrences among different modalities, reflecting consistent safety profiles in CAR T studies, BsAbs/BiTE studies, and ADC studies [33,34]. Interestingly, our analysis indicates notable differences in the incidence of fatal adverse events between later phase trials and the initial phase 1 study. The increased prevalence of fatal adverse events in later phases raises considerations regarding the dynamics of safety assessments over time. A plausible explanation is the potential manifestation of more severe grades of adverse events with longer follow-up times and an increased number of enrolled patients. This suggests that certain adverse events, particularly those of a more critical nature, may not be readily apparent during early clinical development, emphasizing the critical importance of extended follow-up periods and larger patient cohorts. While phase 1 studies provide crucial insights into initial safety profiles, the evolution of adverse events, particularly those with fatal outcomes, becomes more apparent in subsequent phases. Therefore, these observed differences underscore the need for sustained vigilance and thorough monitoring in later phases to capture and address adverse events that may emerge with prolonged exposure or in a broader patient population.

While this study did not place significant emphasis on certain aspects, future work could delve into critical details such as previous treatment history (e.g., the percentage of patients with a different prior line of therapy, specific numbers like "triple-class refractory" multiple myeloma), the percentage of high-risk cytogenetics or other high risk group patients, or the percentage of patients with specific biomarkers (e.g., PD-1 or mismatch repair status etc.) enrolled in the study for more in-depth analysis. We also would like to expand the adaptability in various conference abstracts in future study. Furthermore, the feasibility test comparing treatment outcomes in earlier conference abstracts and subsequent full articles leveraging SEETrials' potential can address community inquiries. Concerns about reporting accuracy in conference abstracts arise from the lack of elaborate peer-reviews unlike publications [7].

CONCLUSION

We developed SEETrials, a GPT-based system with broad applicability, capable of identifying granular safety and efficacy entities across various clinical trial studies. The system demonstrated high precision, recall, and F1, affirming its reliability and accuracy. SEETrials shows significant potential in ensuring timely and accurate data extraction and facilitating efficient dissemination of clinical trial outcomes.

FUNDING

This work did not receive any specific grant or funding support.

AUTHOR CONTRIBUTIONS

HP, KL, SD and XW conceptualized the study design, spearheaded the development and evaluation of the system. HP, KL, SD, and LH led the data extraction, prompt engineering and computational analysis, and drafted the initial manuscript. HP, KL, and NO oversaw the knowledge ingestion, gold standard and coordinated the evaluation of the system. HP, KL, SD, LH, FM, JW, HX and XW edited and revised the manuscript. LH, JH, and JW provided the infrastructure support for computation. All authors read and approved the final manuscript.

CONFLICT OF INTEREST STATEMENT

HP, KL, SD, LH, NO, FM, JW, and XW are currently employees of IMO health Inc. JW is currently employees of Brown university. HX is currently employees of Yale university. No other conflict of interest.

DATA AVAILABILITY

The SEETrials source code, prompts, and output are available at *[URL provided upon acceptance]*.

References

1 Subbiah V. The next generation of evidence-based medicine. *Nat Med*. 2023;29:49–58.

2 Chen EX, Tannock IF. Risks and Benefits of Phase 1 Clinical Trials Evaluating New Anticancer Agents: A Case for More Innovation. *JAMA*. 2004;292:2150.

3 Weber JS, Levit LA, Adamson PC, *et al.* American Society of Clinical Oncology Policy Statement Update: The Critical Role of Phase I Trials in Cancer Research and Treatment. *JCO*. 2015;33:278–84.

4 Wright EC, Kapuria D, Ben-Yakov G, *et al.* Time to Publication for Randomized Clinical Trials Presented as Abstracts at Three Gastroenterology and Hepatology Conferences in 2017. *Gastro Hep Adv*. 2023;2:370–9.

5 Scherer RW, Meerpohl JJ, Pfeifer N, *et al.* Full publication of results initially presented in abstracts. *Cochrane Database Syst Rev*. 2018;11:MR000005.

6 Scherer RW, Saldanha IJ. How should systematic reviewers handle conference abstracts? A view from the trenches. *Syst Rev*. 2019;8:264.

7 Schmucker CM, Blümle A, Schell LK, *et al.* Systematic review finds that study data not published in full text articles have unclear impact on meta-analyses results in medical research. *PLoS ONE*. 2017;12:e0176210.

8 Rowhani-Farid A, Hong K, Grewal M, *et al.* Consistency between trials presented at conferences, their subsequent publications and press releases. *BMJ Evid Based Med*. 2023;28:95–102.

9 Snedeker KG, Campbell M, Totton SC, *et al.* Comparison of outcomes and other variables between conference abstracts and subsequent peer-reviewed papers involving pre-harvest or abattoir-level interventions against foodborne pathogens. *Prev Vet Med*. 2010;97:67–76.

10 Li G, Abbade LPF, Nwosu I, *et al.* A scoping review of comparisons between abstracts and full reports in primary biomedical research. *BMC Med Res Methodol*. 2017;17:181.

11 Sheikhalishahi S, Miotto R, Dudley JT, *et al.* Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Med Inform*. 2019;7:e12239.

12 Cook MJ, Yao L, Wang X. Facilitating accurate health provider directories using natural language processing. *BMC Med Inform Decis Mak*. 2019;19:80. doi: 10.1186/s12911-019-0788-x

13 Dave AD, Ruano G, Kost J, *et al.* Automated Extraction of Pain Symptoms: A Natural Language Approach using Electronic Health Records. *Pain Physician*. 2022;25:E245–54.

14 Singhal K, Azizi S, Tu T, *et al.* Large Language Models Encode Clinical Knowledge. Published Online First: 2022. doi: 10.48550/ARXIV.2212.13138

15 Yang X, Chen A, PourNejatian N, *et al.* A large language model for electronic health records. *npj Digit Med*. 2022;5:194.

16 Arsenyan V, Shahnazaryan D. Large Language Models for Biomedical Causal Graph Construction. Published Online First: 2023. doi: 10.48550/ARXIV.2301.12473

17 Reichenpfader D, Müller H, Denecke K. Large language model-based information extraction from free-text radiology reports: a scoping review protocol. Health Informatics 2023. https://doi.org/10.1101/2023.07.28.23292031

18 Fink HA, Linskens EJ, Silverman PC, *et al.* Accuracy of Biomarker Testing for Neuropathologically Defined Alzheimer Disease in Older Adults With Dementia. *Ann Intern Med*. 2020;172:669–77.

19 Hasani AM, Singh S, Zahergivar A, *et al.* Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports. *Eur Radiol*. Published Online First: 8 November 2023. doi: 10.1007/s00330-023-10384-x

20 Hu Y, Ameer I, Zuo X, *et al.* Zero-shot Clinical Entity Recognition using ChatGPT. Published Online First: 2023. doi: 10.48550/ARXIV.2303.16416

21 Peikos G, Symeonidis S, Kasela P, *et al.* Utilizing ChatGPT to Enhance Clinical Trial Enrollment. Published Online First: 2023. doi: 10.48550/ARXIV.2306.02077

22 Datta S, Lee K, Paek H, *et al.* AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *J Am Med Inform Assoc*. 2023;ocad218.

23 Anghelescu A, Firan FC, Onose G, *et al.* PRISMA Systematic Literature Review, including with Meta-Analysis vs. Chatbot/GPT (AI) regarding Current Scientific Data on the Main Effects of the Calf Blood Deproteinized Hemoderivative Medicine (Actovegin) in Ischemic Stroke. *Biomedicines*. 2023;11:1623.

24 Collier T, Roadley-Battin M, Darlow C, *et al.* Analysis of conference abstract-to-publication rate in UK orthopaedic research. *BMJ Evid Based Med*. 2018;23:7–11.

25 Coppens DG, Gardarsdottir H, van den Bogert CA, *et al.* Publication rates and reported results in a cohort of gene- and cell-based therapy trials. *Regen Med*. 2020;15:1215–27.

26 Scherer RW, Ugarte-Gil C, Schmucker C, *et al.* Authors report lack of time as main reason for unpublished research presented at biomedical conferences: a systematic review. *Journal of Clinical Epidemiology*. 2015;68:803–10.

27 Ioannidis JPA. Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials. *JAMA*. 1998;279:281.

28 Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*. 1997;315:640–5.

29 Cowan AJ, Green DJ, Kwok M, *et al.* Diagnosis and Management of Multiple Myeloma: A Review. *JAMA*. 2022;327:464.

30 Van Nieuwenhuijzen N, Frunt R, May AM, *et al.* Therapeutic outcome of early-phase clinical trials in multiple myeloma: a meta-analysis. *Blood Cancer J*. 2021;11:44.

31 Tanenbaum B, Miett T, Patel SA. The emerging therapeutic landscape of relapsed/refractory multiple myeloma. *Ann Hematol*. 2023;102:1–11.

32 Kartchner D, Ramalingam S, Al-Hussaini I, *et al.* Zero-Shot Information Extraction for Clinical Meta-Analysis using Large Language Models. *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Toronto, Canada: Association for Computational Linguistics 2023:396–405. https://doi.org/10.18653/v1/2023.bionlp-1.37

33 Swan D, Murphy P, Glavey S, *et al.* Bispecific Antibodies in Multiple Myeloma: Opportunities to Enhance Efficacy and Improve Safety. *Cancers (Basel)*. 2023;15:1819.

34 Raje N, Anderson K, Einsele H, *et al.* Monitoring, prophylaxis, and treatment of infections in patients with MM receiving bispecific antibody therapy: consensus recommendations from an expert panel. *Blood Cancer J*. 2023;13:116.

Supplemental Table 1. Comprehensive summary of entities and values counted, true positive, false positive, and false negative numbers from multiple myeloma studies

Note: Row means entities and their values, whether in a single arm or multiple arms, as a unified entity. Cell means entities and values for each arm, which were assessed independently.

Abstract No.	Total Rows Compared	True Positives	False Positives	False Negatives	Total Cells Compared	True Positives	False Positives	False Negatives
1	53	46	0	0	138	138	0	0
2	59	55	0	0	165	165	0	0
3	61	55	2	1	285	279	1	5
4	34	31	0	0	93	93	0	0
5	48	38	3	6	264	242	10	4
6	76	66	3	4	350	342	2	3
7	39	34	0	0	204	204	0	0
8	40	33	0	4	111	99	8	0
9	70	66	0	1	201	198	2	0
10	61	58	0	0	290	290	0	0
11	50	43	0	0	172	172	0	0
12	29	22	2	2	120	117	3	3
13	46	39	0	1	200	195	4	0
14	29	24	1	1	100	99	0	1
15	38	28	6	6	272	258	14	14
16	36	33	0	0	132	132	0	0
17	55	48	0	1	245	240	2	0
18	38	32	0	2	136	128	4	0
19	30	25	1	1	104	103	1	1
20	68	40	23	23	189	166	23	23
21	42	36	3	2	117	112	2	4
22	28	24	0	0	144	144	0	0
23	34	30	0	0	240	240	0	0
24	66	54	4	5	300	281	8	6
25	68	57	0	7	192	171	14	0
26	67	60	1	1	549	547	2	2
27	56	50	0	2	156	150	4	0
28	42	34	0	2	180	170	4	0
29	33	29	0	0	87	87	0	0
30	45	42	0	0	126	126	0	0

31	67	59	0	1	360	354	2	0
32	71	63	0	0	252	252	0	0
33	46	36	4	4	200	196	2	3
34	51	46	0	1	141	138	2	0
35	40	32	3	3	175	172	3	0
36	43	31	7	7	228	221	7	3
37	37	23	7	8	186	171	12	2
38	29	26	0	0	78	78	0	0
39	37	33	0	0	99	99	0	0
40	45	40	0	0	120	120	0	0
41	45	41	0	0	205	205	0	0
42	46	41	0	0	123	123	0	0
43	52	46	0	0	230	230	0	0
44	44	38	1	1	195	194	1	1
45	46	42	0	0	126	126	0	0
46	61	42	11	13	220	199	13	13
47	42	35	1	3	342	317	17	7
48	89	68	8	15	415	360	34	18
49	65	61	0	0	183	183	0	0
50	30	19	6	6	175	166	3	9

Supplemental Table 2. Descriptive statistics of extracted clinical outcomes on treatment safety.

Entities related to safety	CAR T (n)	CAR T (%)	BsAbs/BiTE (n)	BsAbs/BiTE (%)	ADC (n)	ADC (%)
Neutropenia	36	13.4	18	13.5	13	12.4
CRS	35	13.0	12	9.0	9	8.6
Thrombocytopenia	30	11.2	10	7.5	9	8.6
Anemia	25	9.3	15	11.3	8	7.6
Neurotoxicity	25	9.3	10	7.5	3	2.9
Fatal AEs	17	6.3	6	4.5	7	6.7
Fatigue	14	5.2	5	3.8	4	3.8
ICANS	14	5.2	9	6.8	3	2.9
Infections	13	4.8	16	12.0	9	8.6
Lymphopenia	13	4.8	6	4.5	4	3.8
Leukopenia	10	3.7	2	1.5	1	1.0
Diarrhea	9	3.3	2	1.5	2	1.9
Severe AEs	5	1.9	1	0.8	3	2.9
Cytopenia	4	1.5	1	0.8	1	1.0
URT Infections	3	1.1	1	0.8	1	1.0
AEs Leading Discontinuation	2	0.7	4	3.0	3	2.9
Constipation	2	0.7	0	0.0	1	1.0
DLT	2	0.7	0	0.0	0	0.0
Hypertension	2	0.7	1	0.8	0	0.0
Hyponatremia	2	0.7	0	0.0	0	0.0
AEs Leading Dose Reductions	1	0.4	1	0.8	0	0.0
Dysgeusia	1	0.4	4	3.0	4	3.8
Dyspnea	1	0.4	1	0.8	0	0.0
Hematologic AEs	1	0.4	0	0.0	0	0.0
Nail- AEs	1	0.4	3	2.3	2	1.9
Skin-AEs	1	0.4	3	2.3	4	3.8
TRAEs (Gr≥3)	0	0.0	1	0.8	2	1.9
BCVA	0	0.0	0	0.0	0	0.0
Blurred vision	0	0.0	0	0.0	2	1.9
Dry eye	0	0.0	0	0.0	0	0.0
Hypophosphatemia	0	0.0	0	0.0	2	1.9
Immune-AEs	0	0.0	0	0.0	0	0.0
ISR	0	0.0	1	0.8	1	1.0
Keratitis	0	0.0	0	0.0	0	0.0

Keratometry	0	0.0	0	0.0	0	0.0
Keratopathy	0	0.0	0	0.0	3	2.9
Ocular-AEs	0	0.0	0	0.0	1	1.0
PSN	0	0.0	0	0.0	2	1.9
Pneumonia	0	0.0	0	0.0	0	0.0
TRAEs	0	0.0	0	0.0	1	1.0

CAR T, chimeric antigen receptor T cell; BsAbs, Bispecific antibody; BiTE, Bispecific T cell engager; ADC, antibody drug conjugate; CRS, cytokine release syndrome; AEs, adverse events; ICANS, immune effector cell associated neurotoxicity syndrome; URT, upper respiratory tract infections; TRAEs, treatments related adverse events; BCVA, best corrected visual acuity; ISR, injection site reaction; PSN, peripheral sensory neuropathy.

Supplemental Table 3. Comprehensive summary of entities and values counted, true positive, false positive, and false negative numbers from multiple myeloma studies

Note: Row means entities and their values, whether in a single arm or multiple arms, as a unified entity. Cell means entities and values for each arm, which were assessed independently.

Abstract No.	Disease type	Total Rows Compared	True Positives	False Positives	False Negatives	Total Cells Compared	True Positives	False Positives	False Negatives
1	Breast cancer	35	31	0	0	93	93	0	0
2	Breast cancer	38	31	1	1	128	127	1	1
3	Breast cancer	40	32	1	3	210	194	12	0
4	Breast cancer	42	31	5	5	144	139	5	5
5	Breast cancer	63	55	0	0	220	220	0	0
6	Breast cancer	30	26	1	0	108	104	0	3
7	Lung cancer	30	27	0	0	108	108	0	0
8	Lung cancer	52	42	0	2	176	168	4	0
9	Lung cancer	42	31	6	6	148	142	6	6
10	Lung cancer	33	29	1	1	120	118	2	2
11	Lung cancer	29	23	1	1	144	143	1	1
12	Lung cancer	51	43	0	2	180	172	6	0
13	Lung cancer	38	35	0	0	105	105	0	0
14	Leukemia	32	26	0	1	162	156	3	0
15	Leukemia	53	44	0	3	376	352	9	0
16	Leukemia	41	36	1	1	185	183	2	1
17	Leukemia	44	37	3	3	160	157	3	3
18	Leukemia	45	41	0	1	168	164	3	0