Evaluating ChatGPT in Information Extraction: A Case Study of Extracting Cognitive Exam Dates and Scores

Authors:

Neil Jethani, NYU Grossman School of Medicine
Simon Jones, NYU Grossman School of Medicine
Nicholas Genes, NYU Grossman School of Medicine
Vincent J. Major, NYU Grossman School of Medicine
Ian S. Jaffe, NYU Grossman School of Medicine
Anthony B. Cardillo, NYU Grossman School of Medicine
Noah Heilenbach, NYU Grossman School of Medicine
Nadia Fazal Ali, NYU Grossman School of Medicine
Luke J. Bonanni, NYU Grossman School of Medicine
Andrew J. Clayburn, NYU Grossman School of Medicine
Zain Khera, NYU Grossman School of Medicine
Erica C. Sadler, NYU Grossman School of Medicine
Jaideep Prasad, NYU Grossman School of Medicine
Jamie Schlacter, NYU Grossman School of Medicine
Kevin Liu, NYU Grossman School of Medicine
Benjamin Silva, NYU Grossman School of Medicine
Sophie Montgomery, NYU Grossman School of Medicine
Eric J. Kim, NYU Grossman School of Medicine
Jacob Lester, NYU Grossman School of Medicine
Theodore M. Hill, NYU Grossman School of Medicine
Alba Avoricani, NYU Grossman School of Medicine
Ethan Chervonski, NYU Grossman School of Medicine
James Davydov, NYU Grossman School of Medicine
William Small, NYU Grossman School of Medicine
Eesha Chakravartty, NYU Grossman School of Medicine
Himanshu Grover, NYU Grossman School of Medicine
John A. Dodson, NYU Grossman School of Medicine
Abraham A. Brody, NYU Rory Meyers College of Nursing, NYU Grossman School of Medicine
Yindalon Aphinyanaphongs, NYU Grossman School of Medicine
Narges Razavian*, NYU Grossman School of Medicine

** Correspondence should be addressed to:*
Narges Razavian, Assistant Professor, NYU Grossman School of Medicine
Ste 642, 227 East 30th Street, New York, NY 10016
narges.razavian@nyulangone.org

Abstract

Background: Large language models (LLMs) provide powerful natural language processing (NLP) capabilities in medical and clinical tasks. Evaluating LLM performance is crucial due to potential false results. In this study, we assessed ChatGPT, a state-of-the-art LLM, in extracting information from clinical notes, focusing on cognitive tests, specifically the Mini Mental State Exam (MMSE) and the Cognitive Dementia Rating (CDR). We tasked ChatGPT with extracting MMSE and CDR scores and corresponding dates from clinical notes.

Methods: Our cohort had 135,307 clinical notes (Jan 12th, 2010 to May 24th, 2023) mentioning MMSE, CDR, or Montreal Cognitive Assessment (MoCA). After applying inclusion criteria and excluding notes with only MoCA, 34,465 notes remained. Among them, 765 were randomly selected and underwent analysis by ChatGPT. 22 medically-trained experts reviewed ChatGPT's responses and provided ground truth. ChatGPT (GPT-4, version "2023-03-15-preview") was used on the 765 notes to extract MMSE and CDR instances with corresponding dates. Inference was successful for 742 notes. We used 20 notes for fine-tuning and training the reviewers. The remaining 722 were assigned to reviewers, with 309 assigned to two reviewers simultaneously. Inter-rater-agreement (Fleiss' Kappa), precision, recall, true/false negative rates, and accuracy were calculated.

Results: For MMSE information extraction, ChatGPT achieved 83% accuracy. It demonstrated high sensitivity with a Macro-recall of 89.7% and outstanding true-negative rates of 96%. The precision for MMSE was also high at 82.7%. In the case of CDR information extraction, ChatGPT achieved 89% accuracy. It showed excellent sensitivity with a Macro-recall of 91.3% and a perfect true-negative rate of 100%. However, the precision for CDR was lower at 57%. Analyzing the ground truth data, it was found that 89.1% of the notes included an MMSE documentation, whereas only 14.3% had a CDR documentation, which affected the precision of CDR extraction. Inter-rater-agreement was substantial, supporting the validity of our findings. Reviewers considered ChatGPT's responses correct (96% for MMSE, 98% for CDR) and complete (84% for MMSE, 83% for CDR).

Conclusion: ChatGPT demonstrates overall accuracy in extracting MMSE and CDR scores and dates, potentially benefiting dementia research and clinical care. Prior probability of the information appearing in the text impacted ChatGPT’s precision. Rigorous evaluation of LLMs for diverse medical tasks is crucial to understand their capabilities and limitations.

Introduction

Large-scale language models (LLMs) [1–4] have emerged as powerful tools in natural language processing (NLP), capable of performing diverse tasks when prompted [5] [6]. These models have demonstrated impressive clinical reasoning abilities [7], successfully passing medical licensing exams [8] [9] [10] and generating medical advice on distinct subjects, including cardiovascular disease [11], breast cancer [12], colonoscopy [13], and general health inquiries [14], [6], [15] [16]. These models can produce clinical notes [16] and assist in writing research articles [16]. Medical journals have begun developing policies around use of LLMs in writing [17] [18] [19] [20] [21] [22] and reviewing. Examples of such LLMs include ChatGPT [2] [1], Med-PALM-2 [3], LLaMA [4], and open-source models actively produced by the community [23].

In this study, we focus on evaluating *information extraction* abilities of ChatGPT (powered by GPT-4 [2]) from clinical notes. Information extraction involves the retrieval of specific bits of information from unstructured clinical notes, a task historically handled by rule-based systems [24,25] [26] [27] [28] [29] [30] or language models explicitly trained on datasets annotated by human experts [31] [32] [33] [34] [35] [36]. Rule-based systems lack a contextual understanding and struggle with complex sentence structures, ambiguous language, and long-distance dependencies, often leading to high false positive rates and low sensitivities [37] [38] [39] [40]. Additionally, training a new model for this task can be computationally demanding and require substantial human effort. In contrast, LLMs, such as ChatGPT, operate at “zero-shot” capacity [41] [42] [43], i.e., only requiring a prompt describing the desired information to be extracted.

Despite their promise, LLMs also have a potential limitation - the generation of factually incorrect yet highly convincing outputs, commonly known as “hallucination.” The massive architectures and complex training schemes of LLMs hamper “model explanation” and the ability

to intrinsically guarantee behavior. This issue has been extensively discussed in the literature, emphasizing the need for cautious interpretation and validation of information generated by LLMs [44] [2] [45].

One area where LLMs may greatly benefit healthcare is in the identification of memory problems and other symptoms indicative of Alzheimer's Disease and Alzheimer's Disease Related Dementias (AD/ADRD) within clinical notes. AD/ADRD is commonly underdiagnosed or diagnosed later in the disease trajectory, particularly in racial and ethnic minoritized groups [46] [47] [48] [49] [50] [51]. The precise extraction of cognitive test scores holds significant importance in the development and clinical validation of tools that can facilitate early detection [52] of AD/ADRD in the clinic. Earlier identification can lead to a host of benefits, including assisting with advanced care planning, performing secondary cardiovascular disease prevention, which may reduce worsening of cognitive impairment [53] [54], identification for serving in research trials [55] [56,57], and with the rapid advancement in biologic therapeutics, the opportunity to receive potentially disease modifying drugs [57] [58]. Accurately extracting cognitive exam scores (often buried in clinical notes and not documented in any structured field), enables validation, training and fine-tuning of models at a much larger scale in a clinical setting for a much more racial/ethnically diverse patient population set compared to current research cohorts.

The primary focus of this paper is therefore on the validation of a state-of-the-art LLM (ChatGPT powered by GPT-4), for information extraction related to cognitive tests, specifically the Mini-Mental State Examination (MMSE) [59] and Clinical Dementia Rating (CDR) [60], from clinical notes of a racially and ethnically diverse patient population. Our objective is to accurately extract both the exam score and the date when the exam was administered using ChatGPT.

This study represents a large-scale formal evaluation of ChatGPT's performance in information extraction from clinical notes. Going forward, we intend to employ this benchmark dataset to validate other (open or closed-source) LLMs. Furthermore, we plan to adopt a similar approach to validate LLMs for information extraction across various clinical use cases. By prioritizing prompt engineering with ChatGPT for extracting clinical information, this research aims to enhance our understanding of the potential of LLMs in healthcare and facilitate the development of reliable and robust clinical information extraction tools.

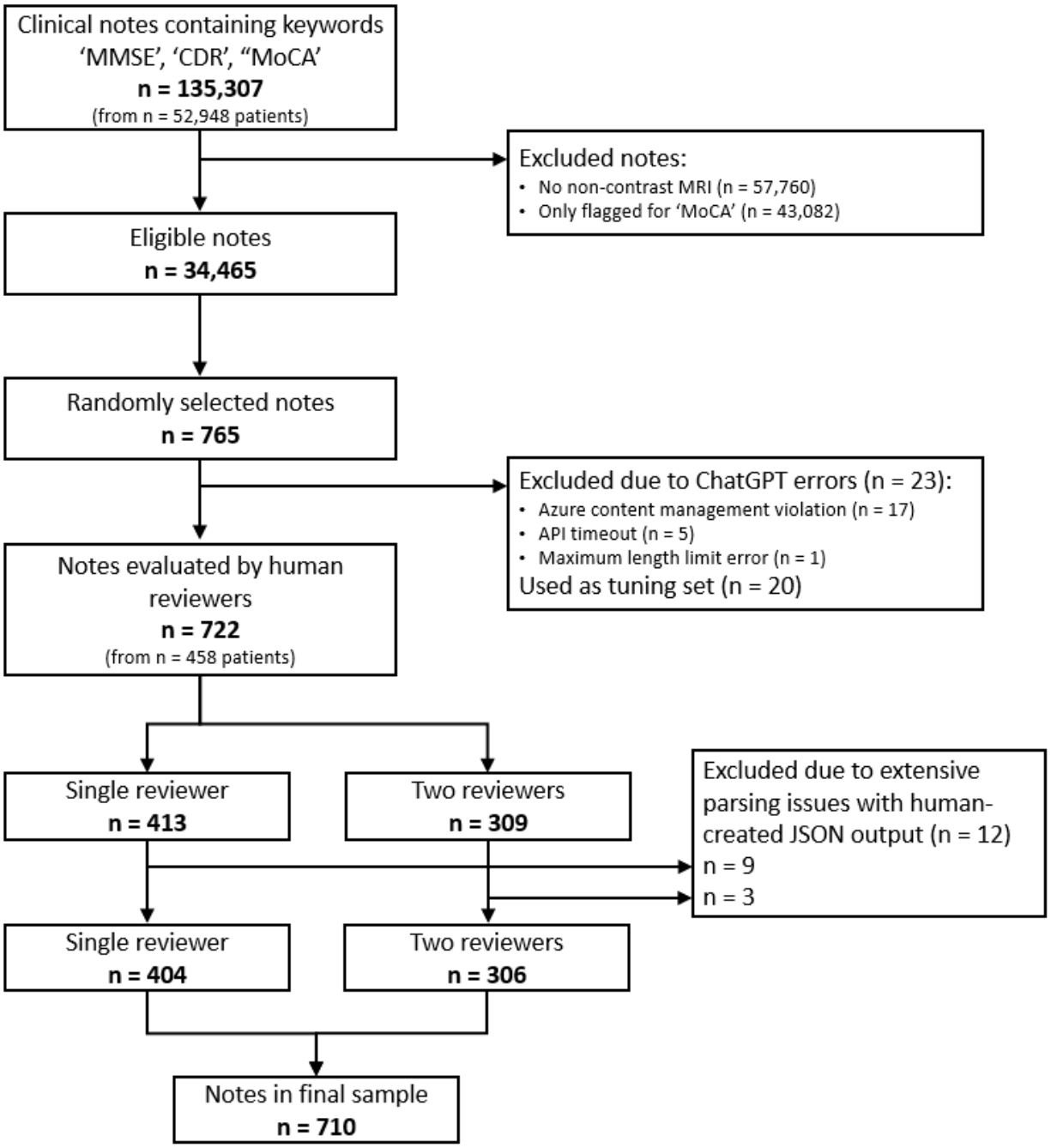
Methods

This study is approved under IRB i20-01095, “Understanding and predicting Alzheimer’s Disease.” NYU DataCore services were utilized to prepare the data as described below.

Dataset

An original cohort of 135,307 clinical notes corresponding to inpatient, outpatient, and emergency department visits between January 12th 2010 and May 24th 2023, which included any of the following keywords (‘MMSE’, ‘CDR,’ or ‘MoCA’ case-insensitive) were identified (see Figure 1). MMSE stands for Mini Mental State Exam, CDR stands for Cognitive Dementia Rating, and MoCa stands for Montreal Cognitive Assessment [61]. These notes belonged to 52,948 patients. From among these patients, 26,355 had a non-contrast brain Magnetic Resonance Imaging (MRI) in the system. Limiting the clinical notes to those who had an MRI in the system resulted in 77,547 notes. These notes were extracted. At this stage we further limited the notes to those including any mentions of MMSE and/or CDR (ignoring MoCA), which yielded 34,465 clinical notes for analysis.

Figure 1: Flowchart of clinical notes evaluated for inclusion in the final sample of GPT-analyzed notes



The choice for requiring patients to have a brain MRI as well as MMSE and/or CDR enables us to have a similar level of granularity as the Alzheimer's Disease Neuro-Imaging Initiative (ADNI) [62], which also uses MMSE and CDR for definition of mild cognitive impairment and dementia

stages. This further enables us to harmonize our clinical dataset with these large research cohorts. Similarly, the choice to ignore MoCA was due to the lack of inclusion of MoCA in standard definition for stages of cognitive impairment in ADNI. The mild cognitive impairment and (mild, moderate or severe) dementia definition criterias utilized in ADNI are included in Supplementary Table S1. Data harmonization is beyond the scope of this paper, although information extraction plays a substantial role in enabling it.

From among 34,465 notes that fit the inclusion criteria, a random selection of 765 notes was identified to undergo information extraction via ChatGPT and manual evaluation. 765 was the total number of the notes needed to satisfy two conditions: 1) Each reviewer not being assigned more than 50 notes to review, and 2) at least around 15 notes per reviewer being double-reviewed by another random reviewer. From among these 765 notes, ChatGPT encountered application programming interface (API) errors in 23 cases (3%). These errors arose from “Azure content management violations" [63] (17 cases), API timeouts (5 cases), and maximum length limit errors (1 case). Supplementary Table S2 includes a more detailed description of these errors. The remaining 742 were considered for assignment to domain expert reviewers.

Generative AI

ChatGPT (GPT-4, API version “2023-03-15-preview”) was used on these 765 notes to extract all instances of the cognitive tests—MMSE and CDR—along with the dates at which the tests were mentioned to have been administered. The inference was successful for 742 notes. The complete API call, along with the exact prompt, the temperature, and other hyper-parameters are included in Supplementary Table S3. The prompt included a request to return these results in a JSON format. ChatGPT’s response (full), as well as the JSON formatted dialogue response were recorded in one session on June 9th 2023. The notes sent to ChatGPT were text-only,

stripped of the rich-text formatting (RTF) native to our EHR system (Epic Systems, Verona, WI). This reduced token count by approximately ten-fold, enabling notes to fit into the GPT4-8K input window and removing a substantial source of confusion for the LLM in prompt tuning. The date that the encounter was recorded in Epic was appended at the beginning of the note, proceeding with a column (":") then the note text. See Supplementary Table S3 for the API request, including the prompt.

Hyper-parameter and Prompt Tuning

We assigned 20 notes out of the 742 as our hyper-parameter and prompt tuning set. An interactive cloud-based environment (i.e playground) was utilized initially to fine-tune the prompt. After initial exploratory analysis using these 20 notes, they were scored via the API using the best prompt and hyper-parameter found in the interactive mode. All human expert reviewers (detailed below) were instructed to first review the results of the 20 cases in a RedCap survey. The goal of this step was to train the reviewers, refine the information presented in RedCap, improve clarification of the questions, and potentially refine the prompt. These 20 notes were then excluded from any additional analysis.

Human Expert Reviewers

Our team included 22 medically trained expert reviewers who volunteered and were trained to review a (HTML formatted) note, provide ground truth, and judge the correctness and completeness of ChatGPT answers for each cognitive test. Fully (HTML) formatted notes were pulled using an Epic web service, and were fed into the RedCap survey. Redcap survey rendered the note's HTML formatting, to ensure notes could be displayed to users in the same format as the readers are accustomed to seeing them clinically, rather than the text-only, computer-friendly format provided to GPT. For 21 of these reviewers, each reviewer was assigned approximately 50 clinical notes to evaluate. From among each reviewer's 50 assigned

notes, about 15 notes were assigned to another random reviewer. The assignment algorithm randomly selected a pair of reviewers for each of our 309 double-reviewed notes and assigned the remaining notes to a randomly selected reviewer until each reviewer reached 50 notes or we fully assigned all notes. This random assignment was a necessary step for ensuring correctness of Fleiss' Kappa [64] metric for inter-rater-agreement. As a result, there was a slight variation in the total number of assigned notes for each reviewer.

Overall, 722 notes were assigned to these 21 reviewers, of which 309 were double-reviewed and 413 were solo-reviewed. The double-reviewed 309 notes were utilized in reporting inter-rater-agreement metrics. After the review, 69 out of 309 notes had at least one disagreement between the two reviewers based on one of the four questions: *Whether ChatGPT's response on MMSE was correct; whether ChatGPT's response on MMSE included all instances of MMSE found in the clinical note; whether ChatGPT's response on CDR was correct; and whether ChatGPT's response on CDR included all instances of CDR found in the clinical note*. A 22nd reviewer was then tasked to review these 69 notes again to provide a third review. Majority vote was then employed to identify the final answer and the ground truth provided by the reviewer whose answer was in the majority vote was used to calculate detailed precision/recall metrics. When both reviewers fully agreed and their JSON results were both valid for analysis, we randomly selected one to compute the precision and recall.

Parsing the JSON results

ChatGPT was instructed to provide answers in JSON format. The human expert reviewers were instructed to provide ground truth data also in JSON format. To facilitate the review process for expert reviewers and reduce errors due to word choices, human reviewers were recommended to start from ChatGPT's answer and correct the incorrect parts. Both results were then parsed automatically and turned into a list of (score, date) tuples, (four lists of tuples per note: 1)

ChatGPT's MMSE, 2) ChatGPT's CDR, 3) reviewer's MMSE, 4) reviewer's CDR.) MMSE and CDR scores were analyzed individually. To normalize the JSON entries, any entry corresponding to the "score" which had obvious wrong test types (i.e "MoCA", "GDS") was automatically excluded.

Statistical Approach

For double-reviewed notes, we reported Fleiss' Kappa [64] as a measure for inter-rater-agreement. We reported this metric for the four questions (Is MMSE complete/correct, and is CDR complete/correct). Additionally, for double-reviewed notes, we computed a 2-way Fleiss Kappa for MMSE and CDR lists of (outcome and date) tuples extracted from the JSON responses of expert reviewers, comparing them against each other, to derive inter-rater-agreement. Fleiss' Kappa is useful when the assignment of a note to reviewer pairs has been random (uniform), and each note has been reviewed by a subset of reviewers [65] [66]. We only considered exact matches (i.e MMSE-27/30 with MMSE-26/30 is just as bad as MMSE-5/30). Kappa can be interpreted as follows: 40%–59% would be *Weak,* 60%–79% would be *Moderate,* 80%–90% would be Strong, and Above 90% would be *Almost Perfect [65].* In addition to 2-way Kappa, we also report a 3-way Kappa on the entries of MMSE and CDR results extracted from the JSON results, computing the joint agreement between the results of ChatGPT and the results provided by two human reviewers.

We also report per test type (MMSE and CDR), Accuracy, True and False Negative Rates, Micro- and Macro-Precision and Micro- and Macro-Recall. Accuracy is defined as the percentage of correct results (at clinical note level), correct being defined as the list of (Value/Date) tuples in the JSON entries for ChatGPT and Ground Truth being fully identical. Macro-Precision for MMSE (or CDR) is the average (at the note level) of percentage of correct MMSE (or CDR) tuples extracted (correct both in date and score values compared to an entry

mentioned in the ground truth for MMSE (or CDR)). Macro-Recall for MMSE (or CDR) is the average (at the note level) of the percentage of the MMSE items in the ground truth that are extracted by the ChatGPT. *Micro*-precision is calculated as percentage of *correct* MMSE (or CDR) items extracted by ChatGPT, from among all extracted MMSE (or CDR) items by ChatGPT and is calculated as one number across all notes combining all notes' entries. Micro-recall is similarly calculated as the percentage of all MMSE (or CDR) items mentioned in the ground truth that were extracted by ChatGPT.

Results

ChatGPT analyzed 765 notes for extraction of Mini Mental Status Exam (MMSE) and Cognitive Dementia Rating (CDR) scores and exam dates. Of these, 23 encountered API error (3%), and 20 were used to fine-tune prompt and hyper-parameters. The remaining 722 notes were assigned to human expert reviewers who manually reviewed (and provided ground truth for) these notes. The characteristics of these 722 notes and associated patients are included in Table 1.

Table 1: Characteristics of 722 notes which are manually evaluated, and their corresponding patients

Feature	**All notes (N=722 notes from 458 patients)**	**Double reviewed notes (N=309 notes from 236 patients)**
Patient demographics		
Age at time of note (mean(sd))	72.64 (14.01)	73.68 (14.01)
Gender		
Female (%)	242 (52.84 %)	124 (52.54 %)
Male(%)	216 (47.16 %)	112 (47.46 %)
Race		
Asian	27 (5.90 %)	10 (4.24 %)

Black	39 (8.52 %)	17 (7.20 %)
White	334 (72.93 %)	178 (75.42 %)
American Indian	1 (0.22 %)	0 (0.00 %)
Unknown	57 (12.45 %)	31 (13.14 %)
Note characteristics		
Date ranges (min to max)	2011/11/21 to 2023/05/10	2011/11/21 to 2023/05/10
Length (in words) (mean (SD))	8428.2 (3822.3)	8306.2 (3851.1)
Open source BERT (prompt tokens)	2167.3 (1049.5)	2123.6 (1023.5)
ChatGPT (Prompt Tokens)	2212.93 (1002.9)	2174.9 (992.3)
ChatGPT (Completion Tokens)	64.3 (49.6)	64.2 (46.5)
ChatGPT (Total Tokens)	2277.3 (1017.9)	2239.1 (1005.0)

Of the double-reviewed 309 notes, 69 had at least one disagreement between the responses to the four questions (if ChatGPT's response for MMSE/CDR is correct/complete) and were assigned to a new reviewer for a third opinion. Among the responses with disagreement, 9 disagreed about correctness of MMSE answers, 40 disagreed about completeness of MMSE answers, 17 disagreed about correctness of CDR answers, and 22 disagreed about completeness of CDR answers. The average response (at the note level) by the included reviews for the four yes/no questions are included in Table 2. Overall reviewers considered ChatGPT's response to be 96.5% and 98% correct for MMSE and CDR respectively. The assessment for whether ChatGPT's answers are also complete (i.e. they do not miss anything) was slightly lower averaging about 84% and 83% for MMSE and CDR respectively.

Table 2: Average response (at the note level) of the responses of reviewers in judging if ChatGPT's answers for MMSE and CDR are correct and/or complete.

	All notes (N=722)	**Double reviewed notes (N=309)**
Is ChatGPT's answer for MMSE correct?	96.5% (sd 18.2)	96.4% (sd 18.5)
Is ChatGPT's answer for MMSE complete?	85.0% (sd 35.7)	84.7% (sd 36.0)
Is ChatGPT's answer for CDR correct?	98.0% (sd 13.7)	99.6% (sd 5.6)

Is ChatGPT's answer for CDR complete?	80.4% (sd 39.6)	83.4% (sd 37.1)

The inter-rater-agreements between reviewers were calculated based on Fleiss' Kappa and are summarized in Table 3. In addition to measuring Fleiss' Kappa between reviewers based on double-reviewed notes (reported as 2-way Fleiss' Kappa in Table 3), we also report agreement between ChatGPT, and the two human reviewers (reported as 3-way Fleiss' Kappa in Table 3). The 2-way agreement on the yes/no questions was high (94% agreement between reviewers for MMSE and 89% agreement for CDR). There was some disagreement in judging the completeness of the answer, leading to a Kappa value of 75% for MMSE (and 85% for CDR). More notably, when analyzing the elements of the ground truth JSON, the 2-way agreement was excellent both for scores (83% for MMSE and 80% for CDR) and for dates (93% for MMSE and 79% for CDR). When measuring the 3-way agreement, there was an increase in all the metrics except MMSE dates.

Table 3: Fleiss' kappa inter-rater-agreement metric between reviewers (2-way) and reviewers and ChatGPT (3-way) over the double-reviewed notes.

	2-way Fleiss' kappa (Among human reviewers) On N=309 double-reviewed notes, n=21 reviewers	3-way Fleiss' Kappa (between ChatGPT and two human reviewers) On N=309 double-reviewed notes, n=21 reviewers
Binary Questions		
Is MMSE list generated by ChatGPT correct?	94.2 %	NA
Is MMSE list generated by ChatGPT complete?	75.2 %	NA
Is CDR list generated by ChatGPT correct?	89.0 %	NA
Is CDR list generated by ChatGPT complete?	85.8 %	NA
Individual (value/date) tuples from ChatGPT and Ground-Truth JSON results.		

MMSE values (of the scores in the note)	83.6 %	93.7 %
MMSE dates (of the scores in the note)	93.3 %	87.2 %
CDR values (of the scores in the note)	80.5 %	87.0 %
CDR dates (of the scores in the note)	79.0 %	82.5 %

JSON responses

Two (out of 722) responses from ChatGPT were not in correct JSON format and failed automatic parsing. These two cases were both reporting “no MMSE or CDR.” Another 26 of the remaining 722 ChatGPT JSON results had “MMSE_Scores” and “CDR_Scores” as the JSON entry, while the rest had “MMSE” and “CDR” as the entries. The JSON results of 79 notes had additional entries, primarily including other cognitive tests (47 “MoCA” and 20 “GDS” being the most frequent wrong tests). These additional entries in JSON results of both ChatGPT and human reviewers were automatically excluded from analysis. Of all 1031 (309×2 + 413) human expert responses, 131 were not in correct JSON format and failed automatic parsing. Most, 92, of those were missing a single formatting character (e.g. comma, bracket) and were manually corrected. The remaining 39 were long entries not provided in JSON format that would have required substantial edits and were consequently excluded from analysis. As a result, 12 of the notes (3 coming from the double-reviewed notes) were excluded from further analysis. In the remaining 710 notes (306 of which were double-reviewed), all parsed ground truth and ChatGPT instances of MMSE and CDR tests were analyzed for computing accuracy, precision, and recall. ChatGPT’s (micro- and macro-) precision and recall for MMSE and CDR items are included in Table 4.

Table 4: Aggregate Accuracy, True Negative Rate, (Micro- and Macro-) Precision and Recall for MMSE and CDR scores extracted by ChatGPT.

	All notes with parsed JSON (N=710)	**Double-reviewed notes with parsed JSON (N=306)**
MMSE		
Total notes without any MMSE (in ground truth)	115	48
Total notes without any MMSE (in GPT results)	77	25
Total correctly predicted empty MMSEs	76	24
ChatGPT's True Negative Rate for MMSE	98.7%	96%
ChatGPT's False Negative Rate for MMSE	1.2%	4%
Remaining notes with un-empty GPT response undergone Precision/Recall calculation for MMSE	633	281
Total MMSE instances predicted	831	366
MMSE Macro Precision (mean % (sd %))	82.9% (sd 36.2)	82.7% (sd 36.8)
MMSE Macro Recall (mean % (sd %))	87.8% (sd 30.4)	89.7% (sd 28.3)
MMSE Micro Precision	83.8%	84.1%
MMSE Micro Recall	83.7%	87.5%
Total notes with any error MMSE result	121	52
Overall accuracy of MMSE	82.9%	83.0%
CDR		
Total notes without CDR (in ground truth)	608	260
Total notes without CDR (in GPT results)	533	233
Total correctly predicted empty CDR	532	233
CDR True Negative Rate	99.8%	100%
CDR False Negative Rate	0.2%	0%
Remaining notes with un-empty GPT response undergone Precision/Recall calculation for CDR	177	73
Total CDR instances predicted	256	92
CDR Macro Precision (mean % sd %)	48.3% (sd 49.9)	57.5% (sd 49.4)
CDR Macro Recall (mean % sd %)	84.3% (sd 36.3)	91.3% (sd 28.1)

CDR Micro Precision	36.3%	51.0%
CDR Micro Recall	85.3%	92.1%
Total notes with any error CDR result	91	31
Overall accuracy of CDR	87.1%	89.8%

ChatGPT had an excellent True Negative Rate—over 96% for MMSE and 100% for CDR in double-reviewed notes. Both results had high recall (sensitivity), reaching 89.7% for MMSE (macro-recall) and 91.3% for CDR (macro-recall). MMSE was more frequently mentioned in the notes and ChatGPT's macro precision (PPV) was 82.7%. CDR, on the other hand, was less frequent, and we observed that ChatGPT hallucinates (factitiously generates) results occasionally leading to a macro precision of only 57.5%. We analyzed 31 notes from the double-reviewed notes where ChatGPT's predictions for CDR had any incorrect elements in them. In 3 notes, there was no sign of a CDR score and the result was hallucinated without an obvious reason. In 4 cases, the date was wrong but the score was correct. The remaining 24 notes' errors were all due to GPT mistakenly reporting results of another test (e.g. GDS, CD4, and Cup/Disk Ratio, also abbreviated CDR) instead of CDR. We similarly analyzed the 52 notes which had any errors for the MMSE predicted list. These notes' errors included 19 with an error only in the reported date while the score was correct; 17 cases of wrong test scores (i.e mostly MoCA, only 2 MiniCog and 1 MSK cases) being picked instead of MMSE; 4 cases of inconsistency in the note itself; 2 typos (one case where the note said "AOx3" for the score, another where MMSE was written as "MMS E") according to the reviewer comments; 3 cases of hallucinations without specific reasons; 4 cases of non-numeric MMSE values (i.e. "deferred" or "preserved"); 1 case of reporting partial MMSE scores instead of the total MMSE score; 1 case was missed from a table that listed 7 MMSE score/dates that were easy to see in the HTML version of the note but confusing in text-only version; and finally, 1 case was actually correct and the error was due to not considering "not found"' equal to an empty list. Taking positive and

negative results into account, overall the MMSE and CDR results were 83% and 89% accurate according to the double-reviewed notes.

Discussion

In this study, our primary objective was to evaluate the performance of ChatGPT, a large-scale language model, in extracting information from clinical notes, specifically focusing on cognitive tests such as the Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR).
To achieve this, we tasked ChatGPT with extracting MMSE and CDR scores and their corresponding dates from clinical notes. We employed 22 human expert reviewers to assess the accuracy and completeness of ChatGPT's responses on 722 clinical notes. Through our evaluation, we measured accuracy, true negative rate and (micro- and macro-) precision and recall of ChatGPT's extraction results for MMSE and CDR items. We also analyzed the feedback provided by the clinical reviewers while considering inter-rater-agreement using statistical techniques.

Our results revealed that ChatGPT successfully extracted relevant information for MMSE and CDR scores, as well as their associated dates, with high recall, capturing nearly all of the pertinent details present in the clinical notes. The overall accuracy of the information extraction results for MMSE and CDR were 83% and 89% respectively. The extraction was highly sensitive (e.g. Macro-recall of 89.7% for MMSE and 91.3% for CDR) and had outstanding true-negative-rates (96% for MMSE and 100% for CDR). The precision of the extracted information was also high for MMSE (82.7%) although in the case of CDR, we observed that ChatGPT occasionally mistook other tests for CDR and therefore the precision was only 57%. Based on the ground-truth provided by our reviewers, 89.1% of the notes included an MMSE documentation instance, whereas only 14.3% of the notes included a CDR documentation

instance. This, combined with our analysis of the errors, explain lower precision in the CDR case, and suggest combining ChatGPT with basic NLP preprocessing may improve the LLM performance further. The substantial inter-rater-agreement among our expert reviewers further supported the robustness and validity of our findings, and the reviewers considered ChatGPT's responses correct (96% and 98% of reviewers judging ChatGPT's responses correct for MMSE and CDR respectively) and complete (84% and 83% of reviewers judging ChatGPT's responses complete for MMSE and CDR respectively.)

The findings of our study demonstrate that large language models, such as ChatGPT, offer a promising solution for extracting valuable clinical information from unstructured notes. This approach provides a more efficient and scalable approach compared to previous methods that either rely on rigid rule-based systems that struggle to capture the necessary contextual information or involve training models on text samples where humans have painstakingly extracted the relevant information. This capability is expected to further improve with OpenAI's introduction of "function calling" [67], which ensures the output of correct JSON from their LLM. LLMs can be effortlessly applied to enhance the value of clinical data for research, enable harmonization with disease registries and biobanks, improve outreach programs within health centers, and contribute to the advancement of precision medicine. Additionally, the availability of large labeled datasets resulting from this information extraction process can also enable AI models to be trained for a wide variety of tasks.

Furthermore, our findings have implications for future AD/ADRD research. Currently, the majority of research in scalable development and validation of AI tools for early AD/ADRD detection rely on research cohorts. These cohorts are overwhelmingly white (NACC cohort is 83% white [68] ADNI cohort is 92% white [62], and do not represent true at-risk populations who tend to have higher comorbid disease burden [50]. Due to late detection and diagnosis of

AD/ADRD [46] [47] [48] [49], clinical data often lacked the details necessary for accurate case identification (i.e. structured data such as ICD codes would yield low sensitivities). Using LLMs to extract data from clinical notes has the potential to improve the quality of clinical data, paving the way for clinical validation and development of clinically applicable novel AI tools and performing cognitive-health precision medicine at scale.

In a broader scope, validation and use of LLMs on unstructured Electronic Health Records (EHRs) have much larger implications. EHRs provide an untapped resource for data-driven knowledge discovery and biomedical research. Active research areas, from causal studies on observational data [69] [70] [71] [72], to disease prognosis [73] [74] [75] [76] [77] [78] [79] [80] [81] [82] [83], disease/genomics understanding [84] [85] [86] [87] [88], treatment response monitoring [89] [90] [91] [92], and fairness and bias in medicine [93] [94] [95] [96] have been using and deriving insight from EHR data. Yet, various studies show that the structured EHR data needs to be complemented by the information in unstructured clinical notes, and structured EHR data alone do not contain the complete healthcare data captured in clinics [97] [98] [99] [100] [37]. Linguistically motivated (often rule-based) information extraction approaches emerged to augment EHRs with information from free text [24,25] [26] [27] [28] [29] [30], but these approaches were not sensitive enough [37] [38] [39] [40]. Standard machine- and deep-learning-based approaches emerged to improve the sensitivity and precision of information extracted [31] [32] [33] [34] [35], but these approaches were task-specific and required substantial manual labor to develop and validate tools. With the introduction of LLMs which enable “zero-shot” learning (i.e. prompt-based inference on a massively trained generative AI model), the ability to extract diverse information without any task-specific model pre-training emerged. This creates immense potential to curate high-quality datasets from EHR data. However, many LLMs have not been developed for healthcare settings, and are known to have the ability to hallucinate. Accordingly, we need to rigorously evaluate them in various

clinical use cases before adopting them. This study addresses the validation need for two specific cognitive tests (namely MMSE and CDR) and is the beginning of a series of works in benchmarking LLMs for medical information extraction. We also acknowledge that given the variability in LLM sensitivity and specificity, further work is needed to combine these information with other clinical evidence and/or predictive models, to ensure and that LLMs don't exacerbate inequities, as racial and ethnic minoritized individuals are less likely to have the MMSE and CDR used here performed on them.

When considering the implications of this study, it is important to acknowledge its limitations. Our focus was on evaluating the information extraction capabilities of the current state of the art of LLMs, specifically ChatGPT powered by GPT-4, rather than comparing it to other LLMs or NLP methods. We believe that our results may be enhanced with better prompt engineering and combining LLMs with standard NLP. Additionally, we conducted a large-scale human evaluation for a single dementia use case, prioritizing result reliability over assessing various clinical scenarios. It is also important to note that our findings pertain specifically to information retrieval from clinical notes and do not predict how ChatGPT will perform on medical tasks requiring diagnosis, treatment recommendation, or summarization.

Overall our study findings indicate that large language models such as ChatGPT have the potential to enhance the extraction of valuable clinical information from unstructured notes, benefiting research, disease registries, biobanks, and precision medicine. The use of LLMs on unstructured electronic health records (EHRs) also holds broader implications for knowledge discovery, biomedical research, and addressing gaps in structured EHR data. Nevertheless, rigorous evaluation and consideration of limitations are crucial, and further work is needed to combine LLM information with other clinical evidence to avoid exacerbating inequities and ensure equitable healthcare delivery.

References

1. OpenAI. ChatGPT. 2023 [cited 3 Jul 2023]. Available: http://openai.com/chatgpt (accessed June 2023)

2. OpenAI. GPT-4 Technical Report. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2303.08774

3. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2305.09617

4. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2302.13971

5. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2303.12712

6. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2303.13375

7. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023;388: 1233–1239.

8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2: e0000198.

9. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. BMJ Neurology Open. 2023;5. doi:10.1136/bmjno-2023-000451

10. Matias Y. Our latest health AI research updates. In: Google [Internet]. 14 Mar 2023 [cited 3 Jul 2023]. Available: https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/

11. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA. 2023;329: 842–844.

12. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. Radiology. 2023;307: e230424.

13. Lee T-C, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B. ChatGPT Answers Common Patient Questions About Colonoscopy. Gastroenterology. 2023. doi:10.1053/j.gastro.2023.04.033

14. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media

forum. JAMA Intern Med. 2023;183: 589–596.

15. Dash D, Thapa R, Banda JM, Swaminathan A, Cheatham M, Kashyap M, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. arXiv [cs.AI]. 2023. Available: http://arxiv.org/abs/2304.13714

16. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023;47: 33.

17. Koo M. The Importance of Proper Use of ChatGPT in Medical Writing. Radiology. 2023;307: e230312.

18. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. In: Nature Publishing Group UK [Internet]. 18 Jan 2023 [cited 4 Jul 2023]. doi:10.1038/d41586-023-00107-z

19. Thorp HH. ChatGPT is fun, but not an author. Science. 2023;379: 313–313.

20. Nature. Authorship. In: Nature Authorship [Internet]. Springer Nature; 2023 [cited 4 Jul 2023]. Available: https://www.nature.com/nature/editorial-policies/authorship

21. JAMA. Instructions for Authors. In: JAMA Authorship Guidelines [Internet]. 4 Jul 2023 [cited 4 Jul 2023]. Available: https://jamanetwork.com/journals/jama/pages/instructions-for-authors

22. Hosseini M, Rasmussen LM, Resnik DB. Using AI to write scholarly publications. Account Res. 2023; 1–9.

23. Park D. Open LLM Leaderboard. In: Open LLM Leaderboard [Internet]. 4 Jul 2023 [cited 4 Jul 2023]. Available: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

24. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17: 229–236.

25. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP--a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc. 2018;25: 331–336.

26. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc. 2018;25: 530–537.

27. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17: 507–513.

28. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. J Biomed Inform. 2001;34: 301–310.

29. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest

x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. pp. 2097–2106.

30. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901 07031. 2019. Available: https://www.aaai.org/Papers/AAAI/2019/AAAI-IrvinJ.6537.pdf

31. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. arXiv [cs.CL]. 2020. Available: http://arxiv.org/abs/2004.09167

32. McDermott MBA, Hsu TMH, Weng W-H, Ghassemi M, Szolovits P. CheXpert++: Approximating the CheXpert labeler for Speed,Differentiability, and Probabilistic Output. arXiv [cs.LG]. 2020. Available: http://arxiv.org/abs/2006.15229

33. Le Glaz A, Haralambous Y, Kim-Dufor D-H, Lenca P, Billot R, Ryan TC, et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. J Med Internet Res. 2021;23: e15708.

34. Weng W-H, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak. 2017;17: 1–13.

35. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023; 1–6.

36. Leiter RE, Santus E, Jin Z, Lee KC, Yusufov M, Chien I, et al. Deep Natural Language Processing to Identify Symptom Documentation in Clinical Notes for Patients With Heart Failure Undergoing Cardiac Resynchronization Therapy. J Pain Symptom Manage. 2020;60: 948–958.e3.

37. Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc. 2015;23: e20–e27.

38. Taggart M, Chapman WW, Steinberg BA, Ruckel S, Pregenzer-Wenzler A, Du Y, et al. Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients. JAMA Netw Open. 2018;1: e183451–e183451.

39. Wu Y, Denny JC, Trent Rosenbloom S, Miller RA, Giuse DA, Xu H. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. AMIA Annu Symp Proc. 2012;2012: 997.

40. Fan Y, Wen A, Shen F, Sohn S, Liu H, Wang L. Evaluating the Impact of Dictionary Updates on Automatic Annotations Based on Clinical NLP Systems. AMIA Summits Transl Sci Proc. 2019;2019: 714.

41. Larochelle H, Erhan D, Bengio Y. Zero-data learning of new tasks. Proceedings of the 23rd national conference on Artificial intelligence - Volume 2. AAAI Press; 2008. pp. 646–651.

42. Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, et al. Finetuned language models are

zero-shot learners. arXiv [cs.CL]. 2021. Available: https://research.google/pubs/pub51119/

43. Rezaei M, Shahidi M. Zero-shot learning and its applications from autonomous vehicles to COVID-19 diagnosis: A review. Intelligence-Based Medicine. 2020;3-4: 100005.

44. Borji A. A Categorical Archive of ChatGPT Failures. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2302.03494

45. Maynez J, Narayan S, Bohnet B, McDonald R. On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. pp. 1906–1919.

46. Tsoy E, Kiekhofer RE, Guterman EL, Tee BL, Windon CC, Dorsman KA, et al. Assessment of Racial/Ethnic Disparities in Timeliness and Comprehensiveness of Dementia Diagnosis in California. JAMA Neurol. 2021;78: 657–665.

47. Lin P-J, Daly A, Olchanski N, Cohen JT, Neumann PJ, Faul JD, et al. Dementia diagnosis disparities by race and ethnicity. Alzheimers Dement. 2020;16. doi:10.1002/alz.043183

48. Saadi A, Himmelstein DU, Woolhandler S, Mejia NI. Racial disparities in neurologic health care access and utilization in the United States. Neurology. 2017;88: 2268–2275.

49. Drabo EF, Barthold D, Joyce G, Ferido P, Chang Chui H, Zissimopoulos J. Longitudinal analysis of dementia diagnosis and specialty care among racially diverse Medicare beneficiaries. Alzheimers Dement. 2019;15: 1402–1411.

50. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet. 2020;396: 413–446.

51. Harper LC. 2022 Alzheimer's Association Facts and Figures. https. Available: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/915A476B938D0AF39A218D34852AF645/9781009325189mem_205-207.pdf/resources.pdf

52. US Dept of Health and Human Services. National Plan to Address Alzheimer's Disease: 2020 Update. 2021 [cited 1 Nov 2021]. Available: https://aspe.hhs.gov/reports/national-plan-address-alzheimers-disease-2020-update-0

53. SPRINT MIND Investigators for the SPRINT Research Group, Williamson JD, Pajewski NM, Auchus AP, Bryan RN, Chelune G, et al. Effect of Intensive vs Standard Blood Pressure Control on Probable Dementia: A Randomized Clinical Trial. JAMA. 2019;321: 553–561.

54. Pragmatic Evaluation of Events And Benefits of Lipid-lowering in Older Adults - Full Text View - ClinicalTrials.Gov. [cited 27 Oct 2021]. Available: https://clinicaltrials.gov/ct2/show/NCT04262206

55. NIA. NIA-funded active Alzheimer's and related dementias clinical trials and studies. In: NIA [Internet]. 2021 [cited 20 Apr 2021]. Available: https://www.nia.nih.gov/research/ongoing-AD-trials

56. Science. In: AAAS [Internet]. [cited 10 Jul 2023]. Available:

https://www.science.org/content/article/another-alzheimers-drug-flops-pivotal-clinical-trial

57. Drug Approval Package: Aduhelm (aducanumab-avwa). [cited 31 Oct 2021]. Available: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2021/761178Orig1s000TOC.cfm

58. Manly JJ, Glymour MM. What the Aducanumab Approval Reveals About Alzheimer Disease Research. JAMA Neurol. 2021. doi:10.1001/jamaneurol.2021.3404

59. Folstein MF, Folstein SE, McHugh PR. Mini-Mental State Examination. J Psychiatr Res. 1975. doi:10.1037/t07757-000

60. Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology. 1993. pp. 2412–2412. doi:10.1212/wnl.43.11.2412-a

61. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005;53: 695–699.

62. ADNI. 2021 [cited 1 Nov 2021]. Available: http://adni.loni.usc.edu/data-samples/adni-participant-demographic/

63. Azure OpenAI Service content filtering - Azure OpenAI. [cited 10 Jul 2023]. Available: https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/content-filter

64. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76: 378–382.

65. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol. 2012;8: 23.

66. Maxwell AE. Coefficients of Agreement Between Observers and Their Interpretation. Br J Psychiatry. 1977;130: 79–83.

67. Function calling and other API updates. [cited 7 Jul 2023]. Available: https://openai.com/blog/function-calling-and-other-api-updates

68. Beekly DL, Ramos EM, van Belle G, Deitrich W, Clark AD, Jacka ME, et al. The National Alzheimer's Coordinating Center (NACC) Database: an Alzheimer disease database. Alzheimer Dis Assoc Disord. 2004;18: 270–277.

69. Cooper GF, Bahar I, Becich MJ, Benos PV, Berg J, Espino JU, et al. The center for causal discovery of biomedical knowledge from big data. J Am Med Inform Assoc. 2015;22: 1132–1136.

70. Kleinberg S, Hripcsak G. A review of causal inference for biomedical informatics. J Biomed Inform. 2011;44: 1102–1112.

71. Johnson KW, Glicksberg BS, Hodos RA, Shameer K, Dudley JT. Causal inference on electronic health records to assess blood pressure treatment targets: an application of the parametric g formula. Biocomputing 2018. WORLD SCIENTIFIC; 2017. pp. 180–191.

72. Schulam P, Saria S. Reliable decision support using counterfactual models. Adv Neural Inf Process Syst. 2017;30. Available:

https://proceedings.neurips.cc/paper/2017/hash/299a23a2291e2126b91d54f3601ec162-Abstract.html

73. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data. 2015;3: 277–287.

74. Liu J, Zhang Z, Razavian N. Deep EHR: Chronic Disease Prediction Using Medical Notes. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1808.04928

75. Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal laboratory tests. Machine Learning for Healthcare. 2016. Available: http://www.jmlr.org/proceedings/papers/v56/Razavian16.pdf

76. Razavian N, Sontag D. Temporal Convolutional Neural Networks for Diagnosis from Lab Tests. arXiv [cs.LG]. 2015. Available: http://arxiv.org/abs/1511.07938

77. Hammond R, Athanasiadou R, Curado S, Aphinyanaphongs Y, Abrams C, Messito MJ, et al. Predicting childhood obesity using electronic health records and publicly available data. PLoS One. 2019;14: e0215571.

78. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572: 116–119.

79. Bahadori MT, Lipton ZC. Temporal-Clustering Invariance in Irregular Healthcare Time Series. arXiv [cs.LG]. 2019. Available: http://arxiv.org/abs/1904.12206

80. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25: 44–56.

81. Choi E, Bahadori MT, Song L, Stewart WF. GRAM: graph-based attention model for healthcare representation learning. Proceedings of the 23rd. 2017. Available: https://dl.acm.org/doi/abs/10.1145/3097983.3098126?casa_token=INfp-TEjFLEAAAAA:mr_jWB7QVMoRDuT7fydn63JnSmADd1tA8U2cC5-WO6Fm-Og06vOM7X9NBIgxZxRbTqk81a8DG4Qt

82. Albers DJ, Elhadad N, Claassen J, Perotte R, Goldstein A, Hripcsak G. Estimating summary statistics for electronic health record laboratory data for use in high-throughput phenotyping algorithms. J Biomed Inform. 2018;78: 87–101.

83. Wang T, Qiu RG, Yu M. Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks. Sci Rep. 2018;8: 1–12.

84. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010;26: 1205–1210.

85. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12: 417–428.

86. Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. A Genome-Wide Association Study of Red Blood Cell Traits Using the Electronic Medical Record. PLoS One. 2010;5: e13011.

87. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc. 2010;17: 568–574.

88. Polubriaginof FCG, Vanguri R, Quinnies K, Belbin GM, Yahi A, Salmasian H, et al. Disease heritability inferred from familial relationships reported in medical records. Cell. 2018;173: 1692–1704.e11.

89. Ananthakrishnan AN, Cagan A, Cai T, Gainer VS, Shaw SY, Savova G, et al. Identification of Nonresponse to Treatment Using Narrative Data in an Electronic Health Record Inflammatory Bowel Disease Cohort. Inflamm Bowel Dis. 2015;22: 151–158.

90. Schmittdiel JA, Adams SR, Segal J, Griffin MR, Roumie CL, Ohnsorg K, et al. Novel use and utility of integrated electronic health records to assess rates of prediabetes recognition and treatment: brief report from an integrated electronic health records pilot study. Diabetes Care. 2014;37: 565–568.

91. Pathak J, Simon G, Li D, Biernacka JM, Jenkins GJ, Chute CG, et al. Detecting Associations between Major Depressive Disorder Treatment and Essential Hypertension using Electronic Health Records. AMIA Summits Transl Sci Proc. 2014;2014: 91.

92. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, et al. Rapid-Learning System for Cancer Care. J Clin Oncol. 2010;28: 4268.

93. Chen IY, Szolovits P, Ghassemi M. Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA Journal of Ethics. 2019;21: 167–179.

94. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. J Biomed Inform. 2014;51: 24–34.

95. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak. 2014;14: 1–9.

96. Gijsberts CM, Groenewegen KA, Hoefer IE, Eijkemans MJC, Asselbergs FW, Anderson TJ, et al. Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events. PLoS One. 2015;10: e0132321.

97. Horsky J, Drucker EA, Ramelson HZ. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. AMIA Annu Symp Proc. 2017;2017: 912.

98. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. J Public Health . 2011;34: 138–148.

99. Mountcastle SB, Joyce AR, Sasinowski M, Costello N, Doshi S, Zedler BK. Validation of an administrative claims coding algorithm for serious opioid overdose: A medical chart review. Pharmacoepidemiol Drug Saf. 2019;28: 1422–1428.

100. Juniat V, Athwal S, Khandwala M. Clinical coding and data quality in oculoplastic procedures. Eye . 2019;33: 1733–1740.