Leveraging GPT-4o for Structured Data Extraction in Clinical Trials: Accuracy, Efficiency, and Research Implications

Brandon Rodriguez[1]

[1]brandonrodriguez@berkeley.edu

ABSTRACT

The advent of accessible and user-friendly pretrained large language models (LLMs), such as ChatGPT, Meta AI, or Google Gemini has opened the possibility for use in research, particularly in labor-intensive tasks requiring multiple researchers to manually parse data. One such example of this is in use in systematic reviews and meta-analysis (SR/MA), which both summarize large amounts of relevant studies on a particular topic in order to answer a research question. This project aims to evaluate the use of GPT-4o (ChatGPT-4o) to parse large amounts of clinical trial papers and extract relevant characteristics for each paper such as interventions, sample size, and attrition rate for each intervention. This is done using the data from a recent meta-analysis on medications for generalized anxiety disorder (GAD), using the authors manually parsed data to judge accuracy. Using several packages for text extraction, table extraction, and optical character recognition as well OpenAI's API for batch querying ChatGPT-4o prompt engineering was conducted on training data before a final prompt was tested on test data. The output from the test data, organized into a JSON was then compared with the manually parsed data to judge accuracy. Using this workflow, ChatGPT-4o was shown the ability to extract data with an overall accuracy of 94%. This is an increase in overall accuracy from prior studies using ChatGPT-3.5 on structured data extraction, and support the use of pretrained LLMs in data extraction tasks without prior model training.

Introduction

As of 2024, Large Language Models (LLMs)[1] have become increasingly accessible by researchers across various sectors through the use of conversational applications. This front-facing application allows researchers to prompt complex tasks to a LLM as if talking to another person, allowing for feedback, corrections, and memory recollection. Notable examples of these conversational applications include ChatGPT[2,3], Google Gemini[4], and Meta AI[5]. Several studies and reviews have been made for the use of LLMs in helping automate parts healthcare research[6–8], emphasizing the need for accuracy and ethics while acknowledging the superiority over traditional natural language processing methods.

One study in 2024[6] assessed the use of ChatGPT for extracting structured data from clinical notes, in particular pathological

classifications from pathology reports. These pathology reports were first transformed from a PDF file to a text format, and then ChatGPT was queried to extract structured data using the API from it's parent company, OpenAI. Data was split into two sets for training and testing, and different kinds of queries were tested on the training data before a final query was used for testing. This editing of queries used is known as prompt engineering and was found to significantly impact accuracy. The final query was found to extract pathological classifications with an overall accuracy of 89%, which outperformed the performance of two traditional NLP methods (WordPiece and Named Entity Recognition (NER)). The results from this study showcase the feasibility of using ChatGPT and other conversational applications to infer and extract complex data from healthcare text.

This project aims to assess the abilities of one of the latest OpenAI GPT models (GPT-4o) to extract structured data features from clinical trial reports. The clinical trials come from a recent meta-analysis on clinical trials on medications for Generalized Anxiety Disorder (GAD)[9]. The data was split in two groups in order to develop the final prompt and general training on the first group and test the final prompt on the second group. The accuracy was evaluated by comparing the data extracted by ChatGPT-4o to manually parsed characteristics. The ChatGPT-4o output and errors were then analyzed to understand how the LLM interprets data from clinical trials, and which characteristics it struggles most to extract accurately.

Results

Data Context

This project aims to assess the use of ChatGPT-4o in another form of healthcare text, clinical trial reports, to assess its accuracy in extracting important features such as sample size and attrition rate. For this project I selected the clinical trial data used in Kong 2020's[9] "Comparative Remission Rates and Tolerability of Drugs for Generalized Anxiety Disorder: A Systematic Review and Network Meta-analysis of Double-Blind Randomized Controlled Trials," a meta-analysis on treatments for Generalized Anxiety Disorder (GAD), using the authors own manually data extractions as the basis for the comparison data set. The data is comprised of 31 studies that each test one or more interventions for GAD in Double-Blind Randomized Controlled Trials, using the Hamilton Anxiety Rating Scale (HAM-A). Interventions that occurred only once or twice in studies, such as Silexan or Pexacerfont, were originally omitted from the meta-analysis, but were re-added as a part of the manually extracted data set. Using random assignment, 10 of the studies were assigned for prompt engineering and general training, and 21 of the studies were assigned to be used to test the final prompt. Of the training set of studies, two were omitted due to data issues, bringing the training set to eight studies. Of the testing set of studies, two were omitted due to data issues, bringing the testing set to 19 studies.

The specific task given to ChatGPT-4o is to identify the main author and year of publication, interventions (and dosing), sample size, main race, percentage of female participants, mean HAMA, attrition rate, sponsor and diagnosis criteria from a clinical trial report. For each characteristic other than publication, intervention, sample size, and diagnosis criteria, the characteristics extracted should represent the intervention-specific metric. For example, for Lenox-Smith 2003[10], while each intervention row has '244' as the sample size (being the entire study sample size), the Venlafaxine 75-150 mg/day intervention

row has ’61.5’ as the percentage of female participants, while the Placebo intervention row has ’56.6’ for the same characteristic. ChatGPT-4o was instructed to put an ’NA’ whenever there were only study-wide characteristics when an intervention-specific characteristic was requested. ChatGPT-4o was also instructed to put an ‘NA’ wherever data was unavailable or if it was unsure of an answer.

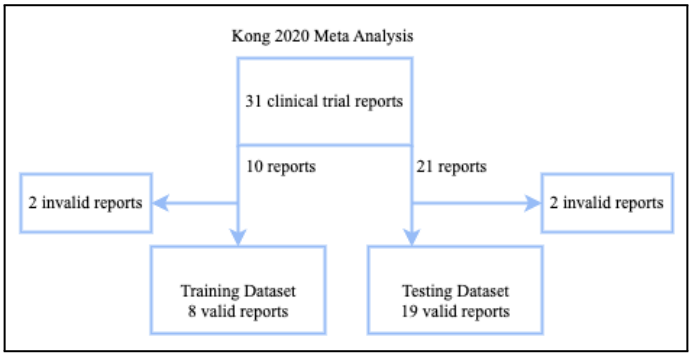
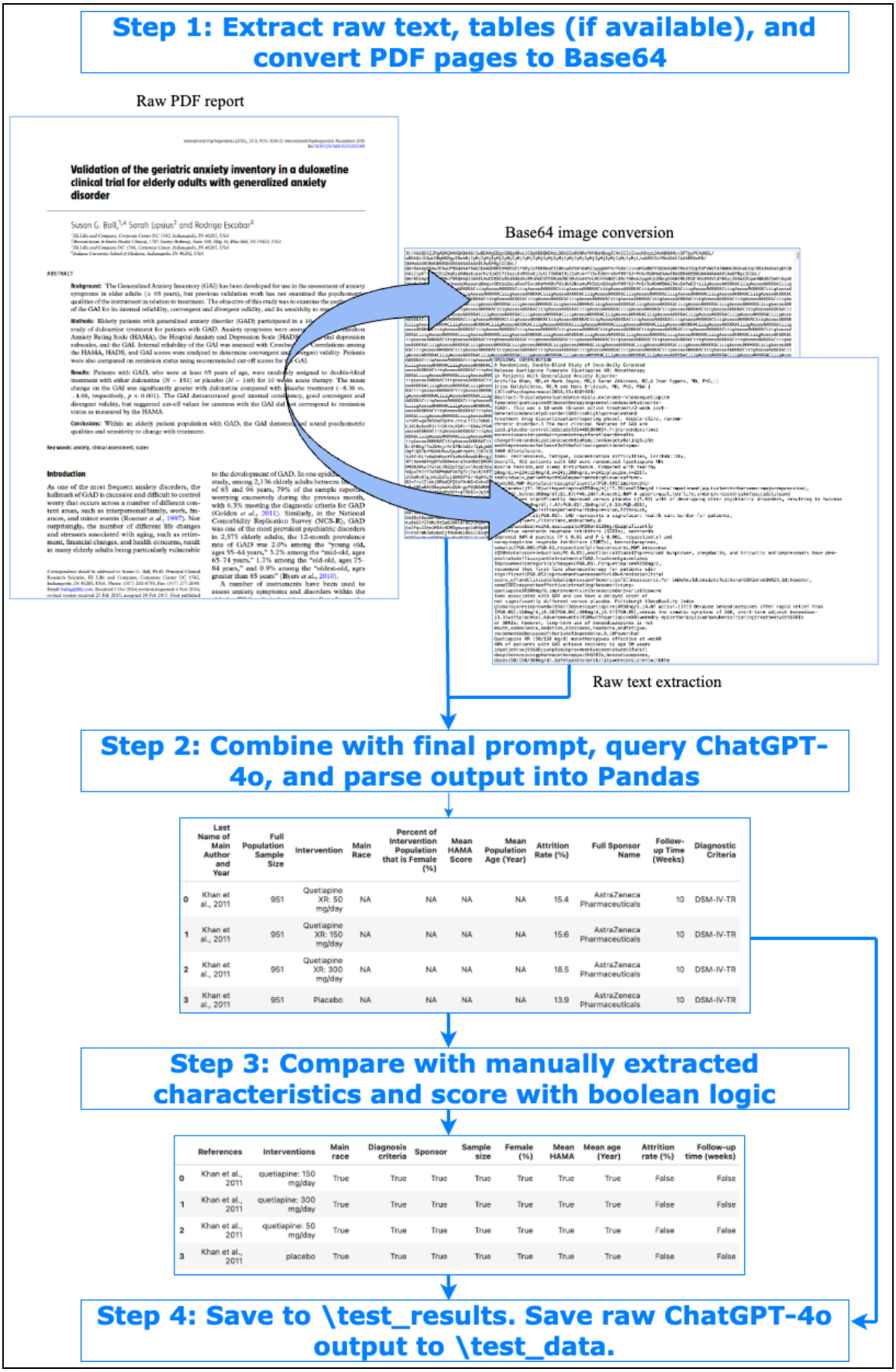


Figure 1. The flow of studies after random assignment. Studies were omitted due to data issues or discrepancies.

Overall Performance

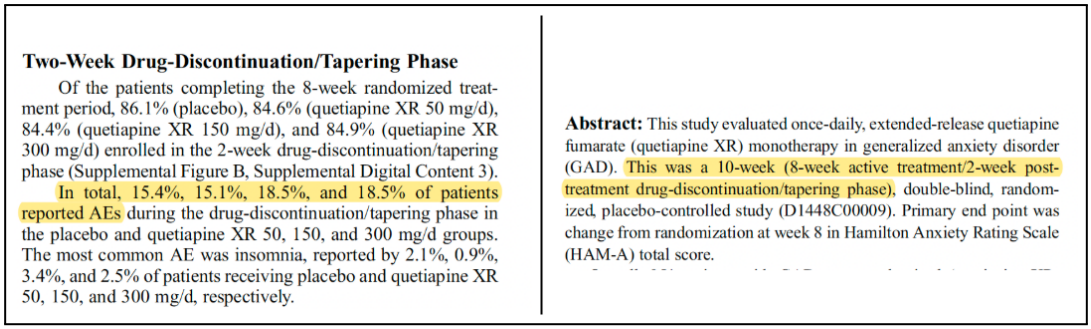
Using the training dataset comprised of 8 randomly assigned studies, accuracy testing was done with iterative prompt engineering to create a final prompt. This final prompt was used to evaluate the GPT-4o model using the testing dataset comprised of 18 randomly assigned studies, and can be found in Figure 1. The accuracy scores of interventions, sample size, main race, percentage of the intervention population that is female, mean HAMA score, mean population age, attrition rate, full sponsor name, follow-up time, and diagnosis criteria were 1.00, 1.00, 1.00, 0.92, 0.96, 0.92, 0.87, 0.96, 0.83 and 0.92, respectively. The overall accuracy for all characteristics was 0.94, and all accuracy metrics are summarized in Table 1.



	Last Name of Main Author and Year	Full Population Sample Size	Intervention	Main Race	Percent of Intervention Population that is Female (%)	Mean HAMA Score	Mean Population Age (Year)	Attrition Rate (%)	Full Sponsor Name	Follow-up Time (Weeks)	Diagnostic Criteria
0	Khan et al., 2011	951	Quetiapine XR: 50 mg/day	NA	NA	NA	NA	15.4	AstraZeneca Pharmaceuticals	10	DSM-IV-TR
1	Khan et al., 2011	951	Quetiapine XR: 150 mg/day	NA	NA	NA	NA	15.6	AstraZeneca Pharmaceuticals	10	DSM-IV-TR
2	Khan et al., 2011	951	Quetiapine XR: 300 mg/day	NA	NA	NA	NA	18.5	AstraZeneca Pharmaceuticals	10	DSM-IV-TR
3	Khan et al., 2011	951	Placebo	NA	NA	NA	NA	13.9	AstraZeneca Pharmaceuticals	10	DSM-IV-TR

	References	Interventions	Main race	Diagnosis criteria	Sponsor	Sample size	Female (%)	Mean HAMA	Mean age (Year)	Attrition rate (%)	Follow-up time (weeks)
0	Khan et al., 2011	quetiapine: 150 mg/day	True	True	True	True	True	True	True	False	False
1	Khan et al., 2011	quetiapine: 300 mg/day	True	True	True	True	True	True	True	False	False
2	Khan et al., 2011	quetiapine: 50 mg/day	True	True	True	True	True	True	True	False	False
3	Khan et al., 2011	placebo	True	True	True	True	True	True	True	False	False

Figure 2. Diagram of the workflow used on the Testing dataset, including the raw inputs and outputs, as well as the location of outputs. The data in this diagram comes from Khan 2011[11]



Two-Week Drug-Discontinuation/Tapering Phase

Of the patients completing the 8-week randomized treatment period, 86.1% (placebo), 84.6% (quetiapine XR 50 mg/d), 84.4% (quetiapine XR 150 mg/d), and 84.9% (quetiapine XR 300 mg/d) enrolled in the 2-week drug-discontinuation/tapering phase (Supplemental Figure B, Supplemental Digital Content 3).

In total, 15.4%, 15.1%, 18.5%, and 18.5% of patients reported AEs during the drug-discontinuation/tapering phase in the placebo and quetiapine XR 50, 150, and 300 mg/d groups. The most common AE was insomnia, reported by 2.1%, 0.9%, 3.4%, and 2.5% of patients receiving placebo and quetiapine XR 50, 150, and 300 mg/d, respectively.

Abstract: This study evaluated once-daily, extended-release quetiapine fumarate (quetiapine XR) monotherapy in generalized anxiety disorder (GAD). This was a 10-week (8-week active treatment/2-week post-treatment drug-discontinuation/tapering phase), double-blind, randomized, placebo-controlled study (D1448C00009). Primary end point was change from randomization at week 8 in Hamilton Anxiety Rating Scale (HAM-A) total score.

Figure 3. Specific places in Khan 2011[11] where ChatGPT-4o may have pulled from erroneously.

Inference and Interpretation (Case Report)

To better understand how ChatGPT-4o interprets the extracted data, this section will go over the results for Khan 2011[11], of which the workflow can be seen in Figure 2.

Khan 2011[11] is different from other studies in this paper in that it contains a supplementary patient demographics file embedded as a link in the paper. Since ChatGPT-4o cannot access links by default, then it cannot extract the data, which is why the manually extracted data set is largely NA's. As seen in the results towards the bottom, ChatGPT-4o extracted all charecteristics correctly except for Attrition rate (%) and Follow-up time (weeks).

Attrition rate is all NA in the manually extracted dataset, and yet ChatGPT-4o returned unique values for this characteristic. Looking closer at the text of Khan 2011[11], there appears to be a section related to adverse effects discontinuations which has the exact values which ChatGPT-4o pulled as attrition rate, highlighted in Figure 3. Adverse effects discontinuations are apart of the total attrition rate, but are not the total attrition rate (as patients may have discontinued for other reasons). ChatGPT-4o did not consider this in it's reasoning, and pulled the values incorrectly.

For Follow-up time (weeks), it is 8 in the manually extracted dataset, and yet ChatGPT-40 returned 10. Looking at the abstract section, this is where it is stated to be a 10 week study, however only 8 weeks were the treatment, and 2 weeks were a drug discontinuation period, highlighted in Figure 3. In the final prompt, ChatGPT-4o was instructed to ignore any discontinuation periods, however this was phrased as "washout periods." It's possible this is the reason why the entire 10 week study period was returned by ChatGPT-4o.

Characteristic Level Error Analysis

Characteristic	Accuracy
Interventions	1.00
Main race	1.00
Sample size	1.00
Female (%)	0.92
Mean HAMA	0.96
Mean age (Year)	0.92
Attrition rate (%)	0.87
Sponsor	0.96
Follow-up time (weeks)	0.83
Diagnosis criteria	0.92
Overall Accuracy	0.94
NA Accuracy	0.70

Table 1. Characteristics and corresponding accuracy of extraction of the Testing set.

Interventions

"Interventions" refers to the intervention(s) being tested in each clinical trial, for example "Duloxetine 75 mg/day", and ChatGPT-4o was tasked with returning not only the name of the intervention but the exact dosing in a measurement of mg/day. In both the training and testing set, ChatGPT-4o scored 1.00 in accuracy, making this an easy characteristic for ChatGPT-4o to extract. While this may seem like an obvious characteristic, certain studies, such as Ball 2015[12], involved complicated dosing schedules where patients could increase their dose depending on their reaction the intervention, resulting in a range of doses for a single intervention that could appear as separate interventions to a casual reader. What aided in this robustness was an additional instruction in the final prompt that specifically asked for either a single value for an intervention, or a range of doses for an intervention.

Main Race

"Main race" refers to the main race of the study-wide population, often reported in a patient demographics section, although occasionally omitted completely. ChatGPT-4o scored 0.88 in accuracy on the training set, and 1.00 in accuracy on the testing set. While the accuracy for this characteristic remains high in both training and testing sets, the increase in accuracy is likely related the additional instruction on the final prompt related to inputting 'NA' whenever uncertain of an answer. One issue noticed during a review of the manually extracted data set was that certain studies, such as Nicolini 2009[13], have linked in their report a supplemental file that contained patient demographics. ChatGPT-4o–without specialized software–cannot open links, meaning it would be impossible for ChatGPT-4o to extract this data, while in manual extraction by a researcher, they could open these links and extract this data. ChatGPT-4o can access links by using specialized software known as a "plugin" named "Link Reader", however since this feature is still in beta and requires prior knowledge of ChatGPT-4o to implement, it was not used in this project.

Sample size

"Sample size" refers to the size of the study-wide population, in particular the number of individuals randomized to an intervention. ChatGPT-4o scored 1.00 on both the training and testing set, making this an easy characteristic for ChatGPT-4o to extract. While prompt engineering, the most important factor to the ensure ChatGPT-4o extracted the correct metric was the additional instruction that explained what was expected to be pulled, as well as asking ChatGPT-4o to pull the "Full Population Sample Size" instead of simply the "Sample Size." The additional instruction was especially important for certain studies that reported the sample size at each step of development, from interviewing potential participants, to randomization, to the actual size of patients that began each intervention. In early prompt engineering without additional instruction, ChatGPT-4o would occasionally pull a sample size other than the randomized sample size.

Female (%)

"Female (%)" refers to the percentage of female participants for a specific intervention in the clinical trial, often reported in the demographics section. ChatGPT-4o was tasked with returning the intervention-specific characteristic for each intervention, and returning NA if the information was unavailable, or only available as a study-wide characteristic. ChatGPT-4o scored 0.71 the training set, and 0.92 on the testing set. This increase in accuracy was likely due to the final additional instruction, that instructed ChatGPT-4o to return an NA if it does not find intervention-specific characteristics. Despite this instruction, on the 4% of observations it failed to extract correctly, on half it instead put the study-wide characteristic when it should have returned an NA. This may be related to the phenomenon of "artificial hallucination," in which a deep learning-based generation "hallucinates" an incorrect answer and reports it confidently. This is especially pertinent as in the final additional instruction, it is instructed to report 'NA' if it is unsure of any extracted characteristic.

Mean HAMA

"Mean HAMA" refers to the mean HAMA score of participants for a specific intervention in the clinical trial, prior to beginning the intervention. ChatGPT scored 0.96 on both the training and testing set. While HAMA reduction is often used as a metric for intervention efficacy, it's not the only one used, which led to large amounts of NA values that ChatGPT-4o had to predict. ChatGPT-4o still maintained a high accuracy on both sets of data, likely due to the additional instruction given to ensure that ChatGPT-4o returned the intervention specific intervention. However, despite this additional instruction, on the 4% of observations it failed to extract correctly, it was due to returning the study-wide characteristic, instead of returning NA.

Mean age (Year)

"Mean age (Year)" refers to the mean age in years of participants for a specific intervention in the clinical trial. ChatGPT-4o scored 0.88 on the training set, and 0.92 on the testing set. Similar to intervention-specific characteristics such as Female (%) and Mean HAMA, the 8% of observations it failed to extract correctly, it was due to returning a study-wide characteristic instead of returning an NA. The increase in accuracy from 0.88 to 0.92 was likely due to the final additional instruction that reinforced NA handling, as well as the additional image input, as many characteristics related to demographics (Female (%),

Mean HAMA) are stored in graphs or tables that are difficult to parse from text input alone.

Attrition rate (%)

"Attrition rate (%)" refers to the percentage of patients who failed to complete a specific intervention after random assignment, for any reason. ChatGPT-4o scored 0.58 on the training set, and 0.87 on the testing set. Out of all the characteristics, this one has the most complicated logic to understand, and most studies contain deceptive sections that would cause even a manual researcher to fail to extract correctly. This characteristic was manually extracted from one of two methods: either the authors of the study explicitly mention it, or it must be manually calculated using one or more flowcharts/visuals. Depending on the study, this characteristic is available from one of these two methods, both, and sometimes neither, which would be where a 'NA' would be inputted. Almost all studies would, however, contain a section dedicated to adverse events (AEs), and the attrition rates specific to these adverse events. ChatGPT-4o would have to ignore this section and use one of the two methods to return the correct intervention-specific attrition rate or return a 'NA' if unavailable or unsure. In the training set, which had a much simpler prompt and lack of image input, this is significantly more difficult for ChatGPT-4o to return correctly, and this is reflected in the low accuracy score. In the testing set with a more descriptive prompt and image input, ChatGPT-4o performed much better, but still struggled compared to other characteristics. The change in accuracy from training to testing supports the use of multimodal inputs to increase accuracy and interpretation, further discussed in the 'Prompt engineering and training' section.

Sponsor

"Sponsor" refers to the full name of the sponsor of the clinical trial, reported by the author and occasionally omitted. ChatGPT-4o scored 0.92 on the training set and 0.96 on the testing set. While this is a fairly simple extraction, the 4% of observations that ChatGPT-4o missed were due to the author not reporting any sponsor, and ChatGPT-4o assuming the authors institution is the sponsor, and returning this. For example, in Nimatoudis 2004[14], there is no sponsor reported, and ChatGPT-4o instead returned the institution of the author, Wyeth Hellas. The increase in accuracy from the training set is likely due to the additional instruction regarding 'NA' handling, but may have been due to chance due to the infrequency of there being no sponsor reported by the author. Additionally, this may add evidence to the phenomenon of "artificial hallucination", with ChatGPT-4o "hallucinating" a sponsor despite there being none reported.

Follow-up time (weeks)

"Follow-up time (weeks)" refers to the total length of the treatment part of the study, including follow-ups but omitting washout/discontinuation periods. ChatGPT-4o scored 0.83 on the training set and the testing set. Similar to the "Attrition rate (%)" characteristic, this is a complex characteristic to extract even through manual extraction, as each study may have unique phrasings and protocol that impacts interpretability. For example, in Khan 2011[11], the abstract clearly states that it is a 10 week study, however in parenthesis notes that it is actually a 8 week treatment, and 2 week discontinuation period. Since the wording "discontinuation" was not included in the final prompt (washout was used instead), this likely was the reason ChatGPT-4o failed to return the correct follow-up time (8) and instead returned the total length of the study (10). The same accuracy in both the

training and testing set means that additional specification needs to be added on top of the additional instructions section of the final prompt, likely using multiple phrasings.

Diagnosis criteria

"Diagnosis criteria" refers to the diagnostic criteria used to diagnose patients with GAD. In all studies, this was a variation of the Diagnostic and Statistical Manual of Mental Disorders, which had a new publication in 1993, 2013, and 2022, being the 4th, 5th, and 5th version Text Revision (TR), respectively. Thus, the majority of cases were DSM-IV, with some being DSM-V and a very small amount being DSM-V-TR. ChatGPT-4o scored 1.00 on the training set and 0.92 on the testing set. This was a very easy characteristic for ChatGPT-4o to extract, and the loss in accuracy is due to the random organization of studies, and not a failure in prompt engineering. This is due to the manually extracted data set from Kong 2020[9] using DSM-IV for each study, despite some not being so. Due to this, this characteristic was omitted from the final accuracy calculations.

NA Accuracy

For the testing set, I decided to do an additional set of scoring known as "NA Accuracy," which is the accuracy scores for each characteristic as well as overall accuracy on observations where the true value was 'NA'. Several characteristics did not have any 'NA' values, and thus were omitted from this scoring. The NA accuracy for main race, sponsor, female (%), mean HAMA, mean age (year), and attrition rate (%) were 0.70, 0.00, 0.83, 0.80, 0.83, and 0.40, respectively. The overall NA accuracy was 0.70. While the characteristic-specific NA accuracy values have too small of a sample size to interpret, the overall NA accuracy of 0.70 comes from the 66 NA values present in the testing dataset. This NA accuracy is significantly lower than the overall accuracy of 0.94, and could be related to the phenomenon of "artificial hallucination" discussed in certain characteristics. An NA value in the dataset represents a complete lack of information for a certain characteristic in the dataset as defined by the prompt, and yet with 30% of these observations, ChatGPT-4o returned an answer with confidence, suggesting a "hallucination rate" of 30%. This suggests that "artificial hallucination" is still an observable issue in the GPT-4o model. However, in a recent meta-analysis on this issue in LLMs[15], GPT-3.5, a model prior to GPT-4o made by OpenAI, had a hallucination rate of 39.6%, suggesting an improvement from the GPT-3.5 model to the GPT-4o model.

Prompt engineering and training

The 8 studies randomly assigned for training were experimented with to improve the prompt and improve the parsing into Python after generation. The training data was also tested using a semi-final prompt, and received an overall accuracy of 0.88. The two metrics that GPT-4o struggled with the most during training were Female % and Attrition Rate, where it scored an accuracy of 0.71 and 0.58, respectively. The full training data extraction can be found in Train_Data_Extraction.ipynb[16], and the full training accuracy scoring can be found in Train_Data_Analysis.ipynb[16], both available on the GitHub repository for this project. Below are some of the largest improvements from the first prompt and final prompt, and how that impacted the output generated by GPT-4o.

Clinical Trial Report Analysis:

Extracted Text:
<EXTRACTED TEXT HERE>

Extracted Tables:
<EXTRACTED TABLES HERE>

This is a clinical trial report. For EACH intervention in the trial (including placebo), please extract the following characteristics and format the response as valid JSON using this exact example structure:

Example format:

```
{
  "Last Name of Main Author and Year": "Doe et al., 2021",
  "Full Population Sample Size": "451",
  "Intervention": "Duloxetine: 50 mg/day",
  "Main Race": "White",
  "Percent of Intervention Population that is Female (%)": "61.5",
  "Mean HAMA Score": "24.5",
  "Mean Population Age (Year)": "43.2",
  "Attrition Rate (%)": "30.2",
  "Full Sponsor Name": "ABC Pharmaceuticals",
  "Follow-up Time (Weeks)": "10",
  "Diagnostic Criteria": "DSM-IV"
}
```

'Full Population Sample Size' should refer to the TOTAL population enrolled in the study and randomized, across all interventions and groups, not just the population size for the specific intervention.

'Intervention' should be in mg/day, not any other unit of measurement. If a single value is provided, extract it as a single value (e.g., "duloxetine: 75 mg/day"). If a range is specified, extract the full range in the format "lower value-upper value mg/day" (e.g., "duloxetine: 70-150 mg/day"). Do not combine separate interventions into a range.

'Follow-up Time' should refer to total length of the treatment period and any follow-ups, omitting washout periods.

'Mean HAMA' should be the mean HAMA score at the beginning of the study for the specific intervention.

'Attrition Rate' should be % of patients who failed to complete the treatment after assignment.

Make sure each JSON object follows this format exactly. If a characteristic is only reported at the study-wide level and not linked to a specific intervention, input 'NA' for that intervention. Only include characteristics under specific interventions if the information explicitly ties them to that intervention. If you are unsure about an answer, input 'NA'.

Figure 4. Final prompt used in the Testing data set. In an actual run, the text, table, and image extracts are included.

Output format

The initial prompt given to ChatGPT simply provided extracted text a certain study, and asked it to extract a list of certain characteristics from the text "in a structured format". This simple prompt led to ChatGPT returning output in extremely varied formats and interpretations, such as bringing together treatments into one summarized output, or choosing to omit the Placebo results. While these outputs were technically meeting the instructions provided, it made it impossible to compare results across studies and judge it for accuracy by comparing it with the manually extracted characteristic. Note while this sort of varied formats is acceptable for manual review in small quantities, it's difficult to analyze further without carefully providing the structure required. To address this issue, an example structure was specified as well as the data-interchange format (JSON) for parsing into Python for future analysis. An example of this is This can be seen in the final prompt beginning at "Example Format:", with an example provided being "Last Name of Main Author and Year: 'Doe et al., 2021'". This led to fairly standardized outputs from ChatGPT, making it much easier to parse and compare to manually extracted data.

Manually Extracted Data Review

The original manually extracted data set used to compare the generated data set was pulled directly from a figure in the meta-analysis by Kong 2020[9], as this is where the selected studies are from. The comparison with the generated output in this study was under the assumption that there were no omissions (as there are none mentioned in the paper), and all the data in the manually extracted data could be found in the PDF for each study. This turned out to be false upon testing, as for example, for Coric 2010[17], GPT-4o correctly identified Pexacerfont 100 mg/day as one of the interventions in this study, but it was absent in the manually extracted data set. This omission is not noted anywhere in the Kong 2020[9] study, and had to be added into the manually extracted data set. There were also cases were data for a certain study in from the manually extracted data set could not be found upon manual review of the study, and thus could not be found by GPT-4o, such as in Mezhebovsky 2012[18], leading to a omission. These changes were documented in the Train_Data_Extraction file provided in the GitHub repository. To ensure this would not occur in the testing set, the manually extracted data set and all the studies listed were manually reviewed and edited if necessary. The manually extracted data set lightly edited for training, and the manually extracted data set reviewed completely before testing are available as train_truth_excel.xlsx[16] and test_truth_excel.xlsx[16] in the GitHub repository, respectively.

Additional Instructions

Certain characteristics extracted from a study, such as intervention or sample size, benefit from additional instruction to ensure a format or interpretation. For example, certain interventions are divided into three doses a day, and are thus written in the report as "Drug 10 mg/tid" with tid being an abbreviation for "three times a day". Since in the manually extracted data set, all doses were converted to once a day, or "mg/day", this caused the GPT-4o to fail at identifying the correct interventions. To fix this, adding an additional instruction for this characteristic after the example format helped fix this issue. For the case of intervention dosing, the additional instruction was "'Intervention' should be in mg/day, not any other unit of measurement". Additional instructions were included for ensuring the placebo was included and the sample size referred to the sample size for the entire study, and can be seen in the final prompt as the lines after the example format.

Table and Image Extraction

One of characteristics GPT-4o struggled the most with was attrition rate. This is defined as the % of participants in every intervention that dropped out before the study was over. In most cases, GPT-4o instead reported the percentage of participants that dropped out due to adverse effects, as often a section in most clinical trials. While this was partially improved by additional instruction, there was a clear need for additional extracted input, as attrition rate was often put in tables and visuals. For these reasons, the final prompt includes not just extracted text from the clinical trial, but extracted tables and full pdf images of the clinical trials, allowing GPT-4o to consider this in its reasoning. This is seen in the final prompt as the input before the example format, and is discussed more fully in the Inference and Interpretation (Case Study) section.

Discussion

LLMs with conversational applications, such as ChatGPT, Google Gemini, and Meta AI have been tested in the last couple years for use in labor intensive research tasks, such as parsing and organizing large amounts of data[8,19]. These LLMs may find a new use as a tool in meta-analysis studies that require researchers to manually read through thousands of clinical trials to assess relevance and extract key data. This project aimed to assess the use of ChatGPT-4o in parsing clinical trial reports in order to understand the overall accuracy and effect of prompt engineering on accuracy.

In clinical research, a meta-analysis uses statistical techniques to combine the results of multiple clinical trials to generate an average result that may better estimate the effects of a particular intervention[20]. Meta-analysis is an especially important tool for understanding the efficacy of medications for major depressive disorder (MDD) and generalized anxiety disorder (GAD), as the specific pharmacodynamics behind the first line treatment, selective serotonin reuptake inhibitors (SSRIs), is still not fully understood[21–23].Additionally, in one controversial 2008 study, it was estimated that 82% of the effect of an anti-depressant medication is due to placebo[24]. These two factors make it difficult to prove the efficacy of novel medications for psychiatric conditions. The issue of the placebo effect is exacerbated in clinical trials with a small sample size and uncertainty regarding effective medication dosage[25,26]. Meta-analysis allows researchers to pool the results from multiple clinical studies in order to have a better understanding of the true effect of a medication for MDD and GAD, as well as compare the efficacy of a potential SSRI compared to other drugs in its medication class using individual studies.

One of the largest issues stopping meta-analysis is the labor required in manually reading though hundreds or thousands of clinical trials reports in order to extract data. A meta-analysis usually starts by querying large databases, such as PubMed or Google Scholar, with general queries that aim to capture as many relevant studies as possible. A reference system, such as EndNote library, is then used to remove duplicates, and then reports are screened manually by researchers to exclude reports based on criteria[20]. A meta-analysis on meta-analysis studies found that the average time to complete and publish a meta-analysis is approximately 67.3 weeks, with a mean yield rate of studies reviewed to be 2.94%[27]. ChatGPT, the conversational application behind the GPT models, may be able to automate these manual processes, without requiring the researcher using it prior knowledge of LLMs[19]. However, there is the large issue of accuracy and bias, especially if the meta-analysis is used for advising clinical practice. In a prior study, ChatGPT was found to be able to automate the first manual process of selecting clinical trials using the titles and abstracts, with a recall greater than 0.9 when using the prompt that performed the best during training, reaching equivalent recall rates to manual selection[28]. This project aimed to evaluate the use of ChatGPT in the second manual process, the extraction of key features from the selected papers, as well as to further understand the effect of prompt engineering on task accuracy.

Our data supports the use of ChatGPT to accurately process large volumes of clinical trial data and organize it into a structured format without requiring model training or human annotation. These results support the use of LLMs by researchers to parse large amounts of clinical trials reports, lessening the need for manual review and data extraction. Further work is required to understand how accuracy may differ across different types of data and subjects.

In conclusion, this project and several other recent studies show that ChatGPT and other conversational application LLMs are powerful and accessible tools for researchers to analyze large amounts of healthcare documents. With further studies confirming the accuracy and efficacy of these models, they may become a regular tool of researchers to transform and analyze data.

Methods

Data processing

Each clinical trial report was downloaded as a PDF document and then placed into a folder labeled "testing" or "training" after random assignment. These folders were located on a local computer, upon which text, tables, and images were extracted from using Python packages before given to GPT-4o along with a prompt using the OpenAI API.

Python package openai[29] was used to connect to the OpenAI API, select the model (GPT-4o), and prompt ChatGPT to analyze all the input data sent and extract certain characteristics in a specific format. Common Python packages, such as pandas, pdfplumber, and io were used to extract text, tables and images from studies. The manually extracted dataset used to compare the generated output to was pulled in the Python environment using a package named openpyxl, and the functions used to parse and score the generated output were built fully in the Python environment. The prompt was developed and tested on a training set of randomly assigned studies, before a final prompt was reached.

The final prompt used was selected by iteratively adding and removing subtasks and phrases based on the accuracy on the "training" studies, as well as literature review. This iterative process is known as "prompt engineering," and the initial prompt was heavily inspired by the works of Huang[6] and Cai[28], who each created papers querying ChatGPT for research purposes. The prompt was loosely structured in three parts. The first part is explaining the objective of the task, and all the extracted text, tables, and images from the clinical trial. The second part is a detailed explanation of the output structure wanted with examples. The third part is additional instructions to ensure a standard format and interpretation, as well as greater define the characteristics being extracted. This study used the latest LLM model from OpenAI, "GPT-4o", which was queried using the OpenAI API. GPT-4o scored higher on benchmark tests compared to prior models GPT-4 and GPT-3.5[3]. GPT-4 was estimated to have around 1.8 trillion parameters, and GPT-4o is estimated to have a similar or possibly larger amount. All GPT models are pre-trained on various public and authored documents. The exact code used to select the "GPT-4o" model and query it is available at the GitHub repository[16] for this project.

Model evaluation

The performance of ChatGPT-4o was evaluated by comparing its generated output with the manually extracted data provided in the meta-analysis and edited for completeness and accuracy. Accuracy was recorded and returned for every characteristic besides "References", which is the name of the author and year of publication that was used for grouping purposes. Overall accuracy for all characteristics was 0.94, and included all characteristics besides "References" and "Diagnostic criteria" and was discussed in full in the Overall Performance section.

Data availability

In total 31 studies were used as input data for this project, each of which testing one or more interventions for GAD in Double-Blind Randomized Controlled Trials using the Hamilton Anxiety Rating Scale (HAM-A). The 10 studies that comprised the training data set were Stein 2017[30], Bidzan 2012[31], Coric 2010[17], Stein 2014[32], Feltner 2003[33], Koponen 2007[34], Montgomery 2008[35], Pollock 2008b[36], Bandelow 2010[37], and Mezhebovsky 2013[18]. Pollock 2008b[36] and Mezhebovsky 2013[18] were omitted from the training data set upon manual review due to data issues. The 21 studies that comprised the testing data set were Rynn 2008[38], Lennox 2003[10], Kasper 2014[39], Hartford 2007[40], Boyer 2004[41], Merideth 2012[42], Mahablesh 2013[43], Davidson 2004[44], Pollock 2001[45], Nicolini 2009[13], Allgulander 2004[46], Pollock 2008a[46], Wu 2011[47], Bose 2008[48], Rickels 2003[49], Rothschild 2012[50], Stein 2008[51], Khan 2011[11], Alaka 2014[52], Ball 2015[12], Nimatoudis 2004[14]. Pollock 2008a[53] and Nicolini 2009[13] were omitted from the testing data set due to data issues.

Code availability

The code used in every method tested in this project are available in the GitHub repository for this project[16]. Any additional information or data required is available from the corresponding author upon request.

References

1. Radford, A. & et al. Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (2018). OpenAI.

2. OpenAI. Gpt-4o overview. https://openai.com/index/hello-gpt-4o/. Accessed date not specified.

3. OpenAI. Gpt-4 technical report. https://arxiv.org/pdf/2303.08774.pdf (2023). ArXiv.

4. Team, G. *et al.* Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2024).

5. Tan, C. W. Meta ai's open pretrained transformer (opt): The future of text generation. https://www.researchgate.net/profile/Chi-Wee-Tan/publication/370582580_Meta_AI's_Open_Pretrained_Transformer_OPT_The_Future_of_Text_Generation/links/6456fbe75762c95ac378d8be/Meta-AIs-Open-Pretrained-Transformer-OPT-The-Future-of-Text-Generation.pdf (2023). ResearchGate preprint.

6. Huang, J., Yang, D. M., Rong, R. & et al. A critical assessment of using chatgpt for extracting structured data from clinical notes. *npj Digit. Medicine* **7**, 106, DOI: 10.1038/s41746-024-01079-8 (2024).

7. Kwong, J. C. C., Wang, S. C. Y., Nickel, G. C. & et al. The long but necessary road to responsible use of large language models in healthcare research. *npj Digit. Medicine* **7**, 177, DOI: 10.1038/s41746-024-01180-y (2024).

8. Zia, A. *et al.* Artificial intelligence-based medical data mining. *J. Pers. Medicine* **12**, 1359, DOI: 10.3390/jpm12091359 (2022).

9. Kong, W. *et al.* Comparative remission rates and tolerability of drugs for generalised anxiety disorder: A systematic review and network meta-analysis of double-blind randomized controlled trials. *Front. Pharmacol.* **11**, 580858, DOI: 10.3389/fphar.2020.580858 (2020).

10. Lenox-Smith, A. J. & Reynolds, A. A double-blind, randomised, placebo controlled study of venlafaxine XL in patients with generalised anxiety disorder in primary care. *Br. J. Gen. Pract.* **53**, 772–777 (2003).

11. Khan, A. *et al.* A randomized, double-blind study of once-daily extended release quetiapine fumarate (quetiapine xr) monotherapy in patients with generalized anxiety disorder. *J. Clin. Psychopharmacol.* **31**, 418–428, DOI: 10.1097/jcp.0b013e318224864d (2011).

12. Ball, S. G., Lipsius, S. & Escobar, R. Validation of the geriatric anxiety inventory in a duloxetine clinical trial for elderly adults with generalized anxiety disorder. *Int. Psychogeriatrics* **27**, 1533–1539, DOI: 10.1017/s1041610215000381 (2015).

13. Nicolini, H. *et al.* Improvement of psychic and somatic symptoms in adult patients with generalized anxiety disorder: examination from a duloxetine, venlafaxine extended-release and placebo-controlled trial. *Psychol. Medicine* **39**, 267–276, DOI: 10.1017/s0033291708003401 (2009).

14. Nimatoudis, I. *et al.* Remission rates with venlafaxine extended release in greek outpatients with generalized anxiety disorder. a double-blind, randomized, placebo controlled study. *Int. Clin. Psychopharmacol.* **19**, 331–336, DOI: 10.1097/00004850-200411000-00003 (2004).

15. Ji, Z. *et al.* Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, DOI: 10.1145/3571730 (2023).

16. Rodriguez, B. Leveraging gpt-4o for structured data extraction in clinical trials. https://github.com/brandonorodriguez/Leveraging-GPT-4o-for-Structured-Data-Extraction-in-Clinical-Trials (2024). GitHub repository.

17. Coric, V. *et al.* Multicenter, randomized, double-blind, active comparator and placebo-controlled trial of a corticotropin-releasing factor receptor-1 antagonist in generalized anxiety disorder. *Depress. Anxiety* **27**, 417–425, DOI: 10.1002/da.20695 (2010).

18. Mezhebovsky, I., Mägi, K., She, F., Datto, C. & Eriksson, H. Double-blind, randomized study of extended release quetiapine fumarate (quetiapine xr) monotherapy in older patients with generalized anxiety disorder. *Int. J. Geriatr. Psychiatry* **28**, 615–625, DOI: 10.1002/gps.3867 (2013).

19. Michelson, M. & Reuter, K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemp. Clin. Trials Commun.* **16**, 100443, DOI: 10.1016/j.conctc.2019.100443 (2019).

20. Tawfik, G. M. *et al.* A step by step guide for conducting a systematic review and meta-analysis with simulation data. *Trop. Medicine Heal.* **47**, 46, DOI: 10.1186/s41182-019-0165-6 (2019).

21. Stahl, S. M. Mechanism of action of serotonin selective reuptake inhibitors. serotonin receptors and pathways mediate therapeutic effects and side effects. *J. Affect. Disord.* **51**, 215–235, DOI: 10.1016/s0165-0327(98)00221-3 (1998).

22. Turner, E. H. & Rosenthal, R. Efficacy of antidepressants. *BMJ (Clinical Res. Ed.* **336**, 516–517, DOI: 10.1136/bmj.39510.531597.80 (2008).

23. Moncrieff, J., Cooper, R. E., Stockmann, T. & et al. The serotonin theory of depression: A systematic umbrella review of the evidence. *Mol. Psychiatry* **28**, 3243–3256, DOI: 10.1038/s41380-022-01661-0 (2023).

24. Kirsch, I. *et al.* Initial severity and antidepressant benefits: a meta-analysis of data submitted to the food and drug administration. *PLoS Medicine* **5**, e45, DOI: 10.1371/journal.pmed.0050045 (2008).

25. Raman, R. Statistical methods in handling placebo effect. In Witek, N. P., Goetz, C. G. & Stebbins, G. T. (eds.) *International Review of Neurobiology*, vol. 153, 103–120, DOI: 10.1016/bs.irn.2020.04.004 (Academic Press, 2020).

26. Khan, A., Redding, N. & Brown, W. A. The persistence of the placebo response in antidepressant clinical trials. *J. Psychiatr. Res.* **42**, 791–796, DOI: 10.1016/j.jpsychires.2007.10.004 (2008).

27. Borah, R., Brown, A. W., Capers, P. L. & Kaiser, K. A. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ Open* **7**, e012545, DOI: 10.1136/bmjopen-2016-012545 (2017).

28. Cai, X. *et al.* Utilizing chatgpt to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. *medRxiv* DOI: 10.1101/2023.09.06.23295072 (2023).

29. PyPI. Openai package. https://pypi.org/project/openai/. Accessed December 7, 2024.

30. Stein, D. J. *et al.* Efficacy and safety of agomelatine (10 or 25 mg/day) in non-depressed out-patients with generalized anxiety disorder: a 12-week, double-blind, placebo-controlled study. *Eur. Neuropsychopharmacol.* **27**, 526–537, DOI: 10.1016/j.euroneuro.2017.02.007 (2017).

31. Bidzan, L., Mahableshwarkar, A. R., Jacobsen, P., Yan, M. & Sheehan, D. V. Vortioxetine (lu aa21004) in generalized anxiety disorder: results of an 8-week, multinational, randomized, double-blind, placebo-controlled clinical trial. *Eur. Neuropsychopharmacol.* **22**, 847–857, DOI: 10.1016/j.euroneuro.2012.07.012 (2012).

32. Stein, D. J. *et al.* Agomelatine in generalized anxiety disorder: an active comparator and placebo-controlled study. *J. Clin. Psychiatry* **75**, 362–368, DOI: 10.4088/jcp.13m08433 (2014).

33. Feltner, D. *et al.* A randomized, double-blind, placebo-controlled, fixed-dose, multicenter study of pregabalin in patients with generalized anxiety disorder. *J. Clin. Psychopharmacol.* **23**, 240–249, DOI: 10.1097/01.jcp.0000084032.22282.ff (2003).

34. Koponen, H. *et al.* Efficacy of duloxetine for the treatment of generalized anxiety disorder: implications for primary care physicians. *Prim. Care Companion to J. Clin. Psychiatry* **9**, 100–107, DOI: 10.4088/pcc.v09n0203 (2007).

35. Montgomery, S., Chatamra, K., Pauer, L., Whalen, E. & Baldinetti, F. Efficacy and safety of pregabalin in elderly people with generalised anxiety disorder. *Br. J. Psychiatry* **193**, 389–394, DOI: 10.1192/bjp.bp.107.037788 (2008).

36. Pollack, M. H., Tiller, J., Xie, F. & Trivedi, M. H. Tiagabine in adult patients with generalized anxiety disorder: results from 3 randomized, double-blind, placebo-controlled, parallel-group studies. *J. Clin. Psychopharmacol.* **28**, 308–316, DOI: 10.1097/jcp.0b013e318172b45f (2008).

37. Bandelow, B. *et al.* Extended-release quetiapine fumarate (quetiapine xr): a once-daily monotherapy effective in generalized anxiety disorder. data from a randomized, double-blind, placebo- and active-controlled study. *Int. J. Neuropsychopharmacol.* **13**, 305–320, DOI: 10.1017/s1461145709990423 (2010).

38. Rynn, M. *et al.* Efficacy and safety of duloxetine in the treatment of generalized anxiety disorder: a flexible-dose, progressive-titration, placebo-controlled trial. *Depress. Anxiety* **25**, 182–189, DOI: 10.1002/da.20271 (2008).

39. Kasper, S. *et al.* Lavender oil preparation Silexan is effective in generalized anxiety disorder–a randomized, double-blind comparison to placebo and paroxetine. *Int. J. Neuropsychopharmacol.* **17**, 859–869, DOI: 10.1017/s1461145714000017 (2014).

40. Hartford, J. *et al.* Duloxetine as an SNRI treatment for generalized anxiety disorder: results from a placebo and active-controlled trial. *Int. Clin. Psychopharmacol.* **22**, 167–174, DOI: 10.1097/yic.0b013e32807fb1b2 (2007).

41. Boyer, P., Mahé, V. & Hackett, D. Social adjustment in generalised anxiety disorder: a long-term placebo-controlled study of venlafaxine extended release. *Eur. Psychiatry* **19**, 272–279, DOI: 10.1016/j.eurpsy.2004.05.010 (2004).

42. Merideth, C., Cutler, A. J., She, F. & Eriksson, H. Efficacy and tolerability of extended release quetiapine fumarate monotherapy in the acute treatment of generalized anxiety disorder: a randomized, placebo controlled and active-controlled study. *Int. Clin. Psychopharmacol.* **27**, 40–54, DOI: 10.1097/yic.0b013e32834d9f49 (2012).

43. Mahableshwarkar, A. R., Jacobsen, P. L., Chen, Y. & Simon, J. S. A randomised, double-blind, placebo-controlled, duloxetine-referenced study of the efficacy and tolerability of vortioxetine in the acute treatment of adults with generalised anxiety disorder. *Int. J. Clin. Pract.* **68**, 49–59, DOI: 10.1111/ijcp.12328 (2014).

44. Davidson, J. R. T., Bose, A., Korotzer, A. & Zheng, H. Escitalopram in the treatment of generalized anxiety disorder: double-blind, placebo controlled, flexible-dose study. *Depress. Anxiety* **19**, 234–240, DOI: 10.1002/da.10146 (2004).

45. Pollack, M. H. *et al.* Paroxetine in the treatment of generalized anxiety disorder: Results of a placebo-controlled, flexible-dosage trial. *J. Clin. Psychiatry* **62**, 350–357, DOI: 10.4088/jcp.v62n0508 (2001).

46. Allgulander, C. *et al.* Efficacy of sertraline in a 12-week trial for generalized anxiety disorder. *Am. J. Psychiatry* **161**, 1642–1649, DOI: 10.1176/appi.ajp.161.9.1642 (2004).

47. Wu, W. Y., Wang, G., Ball, S. G., Desaiah, D. & Ang, Q. Q. Duloxetine versus placebo in the treatment of patients with generalized anxiety disorder in china. *Chin. Med. J.* **124**, 3260–3268, DOI: 10.3760/cma.j.issn.0366-6999.2011.20.010 (2011).

48. Bose, A., Korotzer, A., Gommoll, C. & Li, D. Randomized placebo-controlled trial of escitalopram and venlafaxine XR in the treatment of generalized anxiety disorder. *Depress. Anxiety* **25**, 854–861, DOI: 10.1002/da.20355 (2008).

49. Rickels, K. *et al.* Paroxetine treatment of generalized anxiety disorder: a double-blind, placebo-controlled study. *Am. J. Psychiatry* **160**, 749–756, DOI: 10.1176/appi.ajp.160.4.749 (2003).

50. Rothschild, A. J., Mahableshwarkar, A. R., Jacobsen, P., Yan, M. & Sheehan, D. V. Vortioxetine (lu AA21004) 5 mg in generalized anxiety disorder: results of an 8-week randomized, double-blind, placebo-controlled clinical trial in the united states. *Eur. Neuropsychopharmacol.* **22**, 858–866, DOI: 10.1016/j.euroneuro.2012.07.011 (2012).

51. Stein, D. J., Ahokas, A. A. & de Bodinat, C. Efficacy of agomelatine in generalized anxiety disorder: a randomized, double-blind, placebo-controlled study. *J. Clin. Psychopharmacol.* **28**, 561–566, DOI: 10.1097/jcp.0b013e318184ff5b (2008).

52. Alaka, K. J. *et al.* Efficacy and safety of duloxetine in the treatment of older adult patients with generalized anxiety disorder: a randomized, double-blind, placebo-controlled trial. *Int. J. Geriatr. Psychiatry* **29**, 978–986, DOI: 10.1002/gps.4088 (2014).

53. Pollack, M. H. *et al.* Early improvement during duloxetine treatment of generalized anxiety disorder predicts response and remission at endpoint. *J. Psychiatr. Res.* **42**, 1176–1184, DOI: 10.1016/j.jpsychires.2008.02.002 (2008).