# TITLE: Inductive reasoning with large language models: a simulated randomized controlled trial for epilepsy

Article type: Original investigation
Study type: Diagnostic study

Authors

| | |
|---|---|
| Daniel M. Goldenholz, MD, PhD[1,2] | daniel.goldenholz@bidmc.harvard.edu ORCID 0000-0002-8370-2758 |
| Shira R. Goldenholz, MD, MPH[2] | shira.r.g@gmail.com |
| Sara Habib, MD[1,2] | shabib1@bidmc.harvard.edu |
| M. Brandon Westover, MD PhD[1,2] | bwestove@bidmc.harvard.edu |

Affiliation
1 – Department of Neurology, Harvard Medical School, Boston USA
2 – Department of Neurology, Beth Israel Deaconess Medical Center, Boston USA

Corresponding author: Daniel Goldenholz, daniel.goldenholz@bidmc.harvard.edu
330 Brookline Ave, Baker 5
Boston MA 02215
617 632 8930

Keywords: artificial intelligence, large language models, epilepsy, randomized clinical trials
Abstract (max 350): 339
Word counts (max 3000): 1909
Figures/tables: (max 5): 4

Funding - NIH K23NS124656

Potential conflict of interest:
None

## Key Points

**Question**: Can large language models (LLMs) effectively simulate and analyze a randomized clinical trial, accurately summarizing and synthesizing clinical data to evaluate drug efficacy and identify relevant reported symptoms?

**Findings**: In a simulated study using LLMs to generate and analyze clinical notes for a trial comparing a drug to a placebo in epilepsy treatment, AI-driven analyses were found to closely match human expert evaluations. The process demonstrated the ability of LLMs to accurately capture treatment effects and identify reported symptoms, with minimal differences in outcomes between the human and LLM analyses.

**Meaning**: The use of LLMs in simulating and analyzing clinical trials offers a promising approach to developing inductive reasoning systems based on electronic medical records. This could revolutionize the way clinical trials are conducted and analyzed, enabling rapid, accurate assessments of therapeutic efficacy and safety without the need for specialized medical language training.

## Abstract

**Importance**: The analysis of electronic medical records at scale to learn from clinical experience is currently very challenging. The integration of artificial intelligence (AI), specifically foundational large language models (LLMs), into an analysis pipeline may overcome some of the current limitations of modest input sizes, inaccuracies, biases, and incomplete knowledge bases.

**Objective**: To explore the effectiveness of using an LLM for generating realistic clinical data and other LLMs for summarizing and synthesizing information in a model system, simulating a randomized clinical trial (RCT) in epilepsy to demonstrate the potential of inductive reasoning via medical chart review.

**Design**: An LLM-generated simulated RCT based on a RCT for treatment with an anti-seizure medication, cenobamate, including a placebo arm and a full-strength drug arm, evaluated by an LLM-based pipeline versus a human reader.

**Setting**: Simulation based on realistic seizure diaries, treatment effects, reported symptoms and clinical notes generated by LLMs with multiple different neurologist writing styles.

**Participants**: Simulated cohort of 240 patients, divided 1:1 into placebo and drug arms.

**Intervention**: Utilization of LLMs for the generation of clinical notes and for the synthesis of data from these notes, aiming to evaluate the efficacy and safety of cenobamate in seizure control either with a human evaluator or AI-pipeline.

**Measures**: The AI and human analysis focused on identifying the number of seizures, symptom reports, and treatment efficacy, with statistical analysis comparing the 50%-responder rate and median percentage change between the placebo and drug arms, as well as side effect rates in each arm.

**Results**: AI closely mirrored human analysis, demonstrating the drug's efficacy with marginal differences (<3%) in identifying both drug efficacy and reported symptoms.

**Conclusions and Relevance:** This study showcases the potential of LLMs accurately simulate and analyze clinical trials. Significantly, it highlights the ability of LLMs to reconstruct essential trial elements, identify treatment effects, and recognize reported symptoms, within a realistic clinical framework. The findings underscore the relevance

of LLMs in future clinical research, offering a scalable, efficient alternative to traditional data mining methods without the need for specialized medical language training.

## Introduction

It is very challenging to extract knowledge from the electronic medical system[1]. Various approaches, including the use of structured data[2], natural language processing toolboxes[3–5], and others have been shown to hold some promise. Nevertheless, the dream of an AI ingesting hundreds of millions of patient charts to develop "clinical judgement" is currently still not practical. With the advent of highly capable foundational large language models (LLMs)[6–8], this dream may be closer to reality than ever before. Current generation systems are plagued with a variety of constraints, including very modest input size limits confabulations (a.k.a. "hallucinations"), inaccuracies, biases, and incomplete knowledge bases[6]. Despite these limitations, modern LLMs have made important strides both in the realm of generative AI for producing artificial documents, as well as in information extraction and summarization.

In this study, we set out to explore a model system of using an LLM (Figure 1, LLM A) to generate clinical data and other LLMs (Figure 2, LLMs B and C) to summarize and synthesize information. The hypothesis was that a simulated randomized clinical trial could be generated, summarized, and accurately evaluated with the help of LLMs. *Inductive reasoning*, defined here as generalizing knowledge based on a set of observations, is submitted as one of the ways clinicians learn. The purpose of this task was to demonstrate the power of AI-enhanced inductive reasoning applied to medical chart review.

## Methods

### *Building the dataset*

A simulated randomized clinical trial was constructed (Figure 1) based on an actual clinical trial in epilepsy for cenobamate[9]. In that trial, there were 2 months of baseline, and 3 months of maintenance at steady state for the drug. In our simulation, there was a placebo arm and a full-strength drug arm (corresponding to 400mg/day cenobamate). Similar to the cenobamate trial, we included 120 patients per arm. To generate a realistic cohort of simulated patients, a previously validated simulator (CHOCOLATES) was used[10]. CHOCOLATES was designed to account for heterogeneity in seizure

frequencies across patients[11], the "L-relationship" power law within dairies[12], seizure clustering[13,14], seizure susceptibility cycles[15,16], and maximum allowable seizure rates[17]. Based on multiple lines of evidence[10,18–24], we assumed that placebo did not have any intrinsic effect and any measured effect would be due to natural variability and regression to the mean[25]. Similar to the RCT[9], simulated patients needed to have an average rate of 4 seizures per month to be included in the simulated study. Like cenobamate, the simulated drug was 39% more effective than placebo[9]. The precise symptom reporting rates in the placebo and drug arms of the cenobamate trial were simulated as well[9]. A well-characterized, open-source LLM[26] called Llama2:7b was used to generate clinical notes with the temperature parameter set to 1.0 (values >0 increase creativity). The creativity, as well as LLM hallucination, were intentionally part of this study to properly simulate "noise" caused by inaccurate patient reporting and inaccurate documentation. One of four neurologist writing styles was randomly selected at the time of each clinical note generation: 1) a terse minimalist style using bullet points, 2) a complete but brief style, 3) a narrative style in 2-4 paragraphs, and 4) an erudite academic professor with many extraneous details. Each simulated patient had two notes generated, one after the baseline period and one after the blinded maintenance period (480 notes total). Additional random details about the patients' past medical history were added randomly but kept consistent within each patient. In addition to a complete note, each encounter also generated a "ground truth" entry in a data table that indicates what information was used in the prompt to the LLM to generate the clinical note.

*AI analysis of the notes*

An AI pipeline for analysis of the RCT was constructed as follows (Figure 2). Each note was fed individually (due to input size constraints) to a second open-source LLM[27] (Mistral 7B v0.1) set to a temperature of 0.0 to increase precision and decrease extraneous detail. This LLM was selected because it was produced independently of Llama2, and thus would not have the luxury of expecting certain styles or methods of writing. The LLM was asked to summarize the note, specifically indicating the number of

seizures during the observation period and what symptoms were reported by the patient. Due to inaccuracies and incomplete responses from typical open-source LLMs, it was not feasible for the LLM to build the final data table required for statistical analysis. Thus, a set of somewhat poorly formatted but mostly complete summaries was obtained from the second LLM.

A third LLM (Claude 2), was also used. This LLM has an extended data input limit and is able to ingest large numbers of summaries at once, resulting in the ability to produce a well formatted data table, and correctly make synthesis inferences correctly. Claude 2 is freely available via web interface, but the application programming interface (API) requires a paid account. In addition to improving the formatting, the third LLM was asked to indicate, on each row, the number of seizures during each period of the study; it was also asked to indicate if there were symptoms reported in the second encounter that differed from the first encounter (representing new symptoms that started along with the experimental treatment).

*Human analysis of the notes*

The set of 480 generated clinical notes were assessed by one of the authors (SH), a trained neurologist. The relevant features, namely, the number of seizures during the observation period and any symptoms reported, were manually extracted and organized into a data table.

*Statistical analysis of data tables*

Three data tables (the ground truth, the AI, and the human) were analyzed in the same way. The percentage change between average monthly seizure rate during baseline to average monthly seizure rate during the maintenance period was computed for each patient[28]. These percentage change values were used to tally the fraction of 50%-responders in each arm, and then to compute a Fisher Exact test to compare arms (RR50). The same percentage change values were also used to compute a median percentage change (MPC) and the Mann-Whitney U test was used to non-parametrically compare the two arms. Uniquely reported symptoms were tallied up in each arm, and these were summarized.

The TRIPOD reporting checklist[29] is provided (Appendix). Code was prepared in python using langchain and ollama. Open-source code is available at https://github.com/GoldenholzLab/LLM-rct.git.

## Results

Computational time for generation and summarization of notes combined took roughly 20 hours on a single computer; this time would of course be reduced with increased computational resources. The complete set of notes are available for review (Appendix). The human review of the 480 notes required roughly 5 hours. In the placebo group, 9 patients were identified as not having any value reported for seizures in either the baseline or maintenance periods. In the drug group there were an additional 8 such patients. These failures can be attributed to the generative LLM A (Figure 1) that produced the notes. These were not corrected, as these represented examples of undesirable “noise” that prevented perfect reconstruction of the ground truth. When computing the statistics for efficacy, patients with missing numbers were excluded. All patients were included when computing symptom report summaries.

The treatment effect sizes reported for the 50%-responder rate (RR50) and median percentage change (MPC) are shown in Figure 3. The marginal efficacy between drug and placebo are shown in Table 1. All comparisons were statistically significant. The AI and human marginal efficacies differed by 1% in both RR50 and MPC.

The reported symptoms identified from each of the data tables are shown in Figure 4. The maximum differences in symptom rates between tables were: AI vs. truth – 2%, human vs truth – 2%, AI vs. human – 3%.

## Discussion

This study simulated a realistic trial modelled after a recently published randomized drug trial[9], and using AI, was able to reconstruct the important elements that were

reported quantitatively and qualitatively in the clinical notes. The AI pipeline was able to correctly show the marginal drug efficacy (drug vs. placebo) differing from human review by no more than 1%. Similarly, the pipeline was able to identify the relevant symptoms reported in drug and placebo arms, differing with the human by no more than 3%. The use of generative AI allowed us to intentionally inject “noise” (distracting and/or incorrect elements) into our experiment. This deliberate addition was made to help determine if we could teach AI system to learn medical information by induction in the presence of noise. In typical clinical situations, there is virtually always some “noise” generated, whether due to inaccurate reporting by patients or caregivers, or inaccurate recording by clinicians. Our system was able to correctly show a strong effect of the simulated drug and found the appropriate common side effects without being taught to look for something specific. These achievements are all the more remarkable when considering an important point: this entire project did not make use of *any* LLMs specially trained in medical language[30]. Moreover, advanced APIs, necessitating expensive and computationally prohibitive setups, were not required.

Future versions of the present pipeline might employ only a single LLM if it was computationally efficient, reliable, inexpensive and had a very large token size. The advantage of the current approach is that it is not necessary to wait for such advances to be made available.

Like any simulation, this study is only as good as the assumptions made. We assumed we have an adequate model for seizure diaries and trial simulation based on prior work[10,21,24,31,32]. We also assumed that generative LLM clinical notes can represent a first approximation for true clinical notes, and that the conditions presented here are relevant to other inductive learning tasks of interest in clinical settings. Another limitation of this study was a linguistic one: our study was conducted entirely in English. Multilingual open-source models[33] are available to extend the present work to many other languages.

It must also be noted that extremely rare side effects in a randomized controlled trial might be missed by the type of system developed here – for example, if an investigational drug causes a systemic inflammatory reaction in only 1 patient for the whole study, this fact must not be missed by trialists. Whereas the system proposed

here may miss such rare side effects, our goal is to look for larger trends and not "outlier" rare results. Indeed, if such rare reactions were only noted in postmarketing studies, it could take a long time for regulators (and therefore clinicians) to become aware of them, yet an inductively learning AI system could flag situations like this if they happen in low fractions of patients beyond expected levels.

The longer-term purpose of building clinical inductive learning tools is to develop real-time systems that can learn from very large populations and apply this knowledge to uncertain situations. For instance, if a new drug is approved, physicians develop a certain personal clinical "experience" with that drug, and after this they base their prescribing habits on that experience. That personal experience sometimes matches the clinical trials, while sometimes there is a mismatch. This "clinical experience" is one of the ingredients that makes seasoned clinicians more effective at choosing from an uncertain set of choices. If an AI-enhanced system can develop such clinical experience across populations, it will be able to rapidly assist countless clinicians with the most updated experience base possible – vastly larger than any one clinician can accrue in their personal practice.

In conclusion, we demonstrated that known (but hidden) knowledge could be learned by induction with a moderate sample of patient charts. Further studies are needed to expand this capability to broader medical knowledge acquisition and applications.

**Data Sharing Statement**

Open-source code and data is available on: https://github.com/GoldenholzLab/LLM-rct.git.

DMG – project design and oversight, data interpretation, manuscript writing and editing
SRG – manuscript editing
SH – data analysis, manuscript editing
MBW – project oversight, manuscript editing

**Acknowledgements**

Funding for this work came in part from NIH K23NS124656. The authors wish to thank the open-source community for sharing and distributing large language models such as llama2 and mistral, as these tools can help advance biomedical science.

**Conflicts of interest**

None of the authors have any conflicts of interest to declare.

## REFERENCES

1. Yang S, Varghese P, Stephenson E, Tu K, Gronsbell J. Machine learning approaches for electronic health records phenotyping: a methodical review. *J Am Med Inform Assoc*. 2023;30(2):367-381. doi:10.1093/jamia/ocac216
2. Ostropolets A, Hripcsak G, Husain SA, et al. Scalable and interpretable alternative to chart review for phenotype evaluation using standardized structured data from electronic health records. *J Am Med Inform Assoc*. 2023;31(1):119-129. doi:10.1093/jamia/ocad202
3. Fu S, Wen A, Liu H. Clinical Natural Language Processing in Secondary Use of EHR for Research. Published online 2023:433-451. doi:10.1007/978-3-031-27173-1_21
4. Wu S, Roberts K, Datta S, et al. Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc*. 2020;27(3):457-470. doi:10.1093/jamia/ocz200
5. Murphy RM, Klopotowska JE, de Keizer NF, et al. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. *PLoS One*. 2023;18(1):e0279842. doi:10.1371/journal.pone.0279842
6. Lee P, Goldberg C, Kohane I, Bubeck S. The AI revolution in medicine□: GPT-4 and beyond. :200.
7. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. Published online March 22, 2023. Accessed August 8, 2023. https://arxiv.org/abs/2303.12712v5
8. OpenAI. GPT-4 Technical Report. Published online 2023.
9. Krauss GL, Klein P, Brandt C, et al. Safety and efficacy of adjunctive cenobamate (YKP3089) in patients with uncontrolled focal seizures: a multicentre, double-blind, randomised, placebo-controlled, dose-response trial. *Lancet Neurol*. 2020;19(1):38-48. doi:10.1016/S1474-4422(19)30399-0
10. Goldenholz DM, Westover MB. Flexible realistic simulation of seizure occurrence recapitulating statistical properties of seizure diaries. *Epilepsia*. 2023;64(2):396-405. doi:10.1111/epi.17471
11. Ferastraoaru V, Goldenholz DM, Chiang S, Moss R, Theodore WH, Haut SR. Characteristics of large patient-reported outcomes: Where can one million seizures get us? *Epilepsia Open*. 2018;3(3):364-373. doi:10.1002/epi4.12237

12. Goldenholz DM, Goldenholz SR, Moss R, et al. Is seizure frequency variance a predictable quantity? *Ann Clin Transl Neurol*. 2018;5(2). doi:10.1002/acn3.519
13. Chiang S, Haut SR, Ferastraoaru V, et al. Individualizing the definition of seizure clusters based on temporal clustering analysis. *Epilepsy Res*. 2020;163. doi:10.1016/j.eplepsyres.2020.106330
14. Haut SR. Seizure clusters: characteristics and treatment. *Curr Opin Neurol*. 2015;28(2):143-150. doi:10.1097/WCO.0000000000000177
15. Baud MO, Kleen JK, Mirro EA, et al. Multi-day rhythms modulate seizure risk in epilepsy. *Nat Commun*. 2018;9(1):1-10. doi:10.1038/s41467-017-02577-y
16. Karoly PJ, Goldenholz DM, Freestone DR, et al. Circadian and circaseptan rhythms in human epilepsy: a retrospective cohort study. *Lancet Neurol*. 2018;17(11):977-985. doi:10.1016/S1474-4422(18)30274-6
17. Trinka E, Cock H, Hesdorffer D, et al. A definition and classification of status epilepticus - Report of the ILAE Task Force on Classification of Status Epilepticus. *Epilepsia*. 2015;56(10):1515-1523. doi:10.1111/epi.13121
18. Goldenholz DM, Moss R, Scott J, Auh S, Theodore WH. Confusing placebo effect with natural history in epilepsy: A big data approach. *Ann Neurol*. 2015;78(3). doi:10.1002/ana.24470
19. Goldenholz DM, Goldenholz SR. Response to placebo in clinical epilepsy trials-Old ideas and new insights. *Epilepsy Res*. 2016;122. doi:10.1016/j.eplepsyres.2016.02.002
20. Goldenholz DM, Strashny A, Cook M, Moss R, Theodore WH. A multi-dataset time-reversal approach to clinical trial placebo response and the relationship to natural variability in epilepsy. *Seizure*. 2017;53. doi:10.1016/j.seizure.2017.10.016
21. Goldenholz DM, Tharayil J, Moss R, Myers E, Theodore WH. Monte Carlo simulations of randomized clinical trials in epilepsy. *Ann Clin Transl Neurol*. 2017;4(8):544-552. doi:10.1002/acn3.426
22. Karoly PJ, Romero J, Cook MJ, Freestone DR, Goldenholz DM. When can we trust responders? Serious concerns when using 50% response rate to assess clinical trials. *Epilepsia*. 2019;60(9). doi:10.1111/epi.16321
23. Goldenholz DM, Goldenholz SR. Placebo in epilepsy. In: *International Review of Neurobiology*. Academic Press Inc.; 2020. doi:10.1016/bs.irn.2020.03.033
24. Romero J, Larimer P, Chang B, Goldenholz SR, Goldenholz DM. Natural variability in seizure frequency: Implications for trials and placebo. *Epilepsy Res*. 2020;162:106306. doi:10.1016/j.eplepsyres.2020.106306
25. Goldenholz DM, Goldenholz EB, Kaptchuk TJ. Quantifying and controlling the impact of regression to the mean on randomized controlled trials in epilepsy. *Epilepsia*. 2023;64(10):2635-2643. doi:10.1111/epi.17730
26. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models.
27. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. Published online October 10, 2023. Accessed February 26, 2024. https://arxiv.org/abs/2310.06825v1
28. Siddiqui O, Hershkowitz N. Primary Efficacy Endpoint in Clinical Trials of Antiepileptic Drugs: Change or Percentage Change. *Drug Inf J*. 2010;44(3):343-350. doi:10.1177/009286151004400316

29. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol*. 2015;67(6):1142-1151. doi:10.1016/j.eururo.2014.11.025
30. Singhal K, Tu T, Gottweis J, et al. Towards Expert-Level Medical Question Answering with Large Language Models.
31. Goldenholz DM, Tharayil JJ, Kuzniecky R, Karoly P, Theodore WH, Cook MJ. Simulating clinical trials with and without intracranial EEG data. *Epilepsia Open*. 2017;2(2):156-161. doi:10.1002/epi4.12038
32. Oliveira A, Romero JM, Goldenholz DM. Comparing the efficacy, exposure, and cost of clinical trial analysis methods. *Epilepsia*. 2019;60(12):e128-e132. doi:10.1111/epi.16384
33. Workshop B, Le Scao T, Fan A, et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. Published online November 9, 2022. Accessed February 27, 2024. https://arxiv.org/abs/2211.05100v4

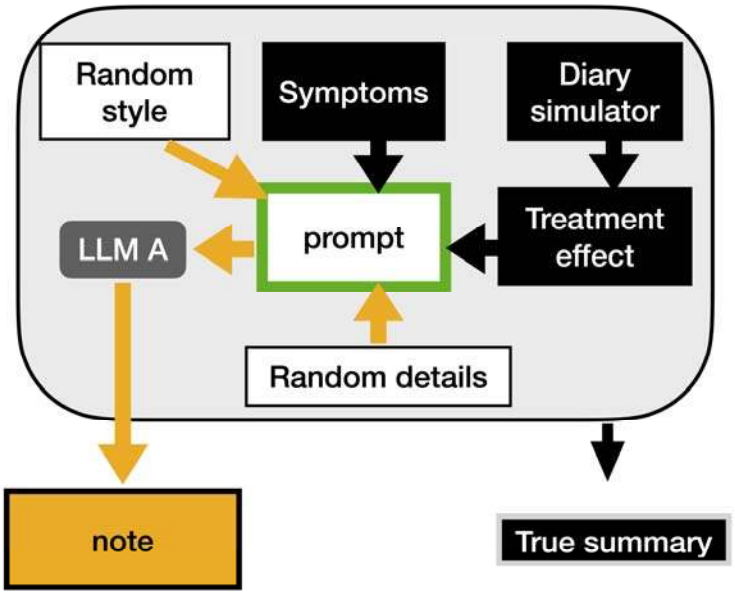



**Figure 1:** Generation of clinical notes. The diary simulator (CHOCOLATES) was used to produce a realistic seizure diary. This was modulated by the treatment effect (0% in placebo arm, and 39% in drug arm) during the experimental maintenance stage. One of four writing styles were chosen, and a random set of reported symptoms were selected (based on previously reported incidence of symptoms for that arm). These items were used to generate the prompt submitted to LLM A (Llama 2:7b). The LLM generated the

clinical note. The true summary was generated based on the original elements used to produce the prompt.

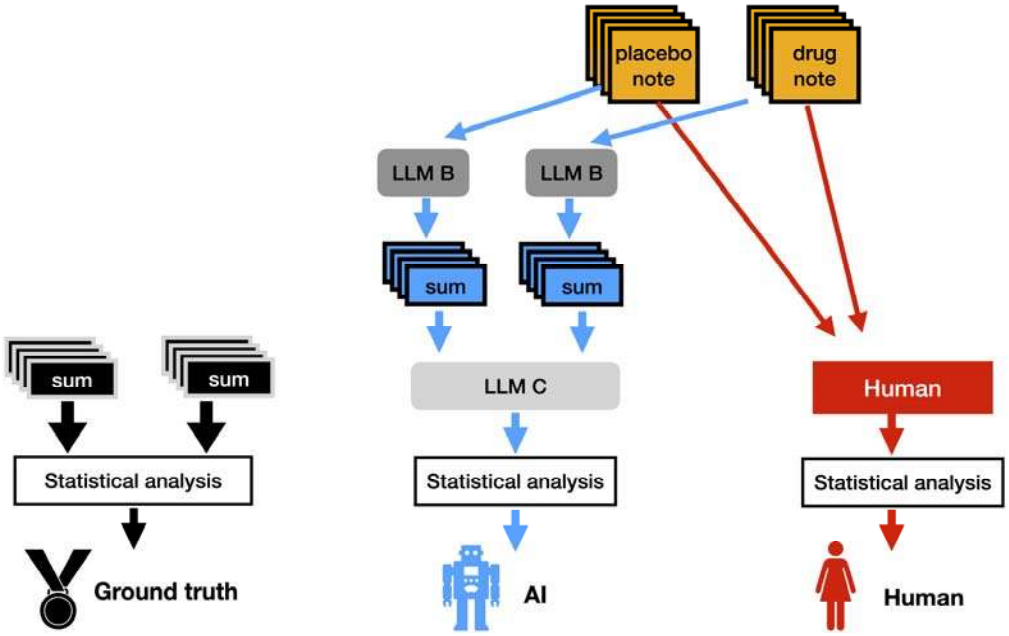



**Figure 2:** Analyzing the trial. The Ground truth summaries (Figure 1) were used directly as a data table. The AI pathway took the clinical notes (Figure 1), and then LLM B (Mistral) produced a summary that indicated the number of seizures and symptoms reported. LLM C (Claude 2) was used to further summarize and synthesize the brief summaries from LLM B into a complete data table. The clinical notes (Figure 1) were manually assessed by the Human to build a data table. The data tables from the Ground truth, the AI and the Human were analyzed in the standard statistical fashion.

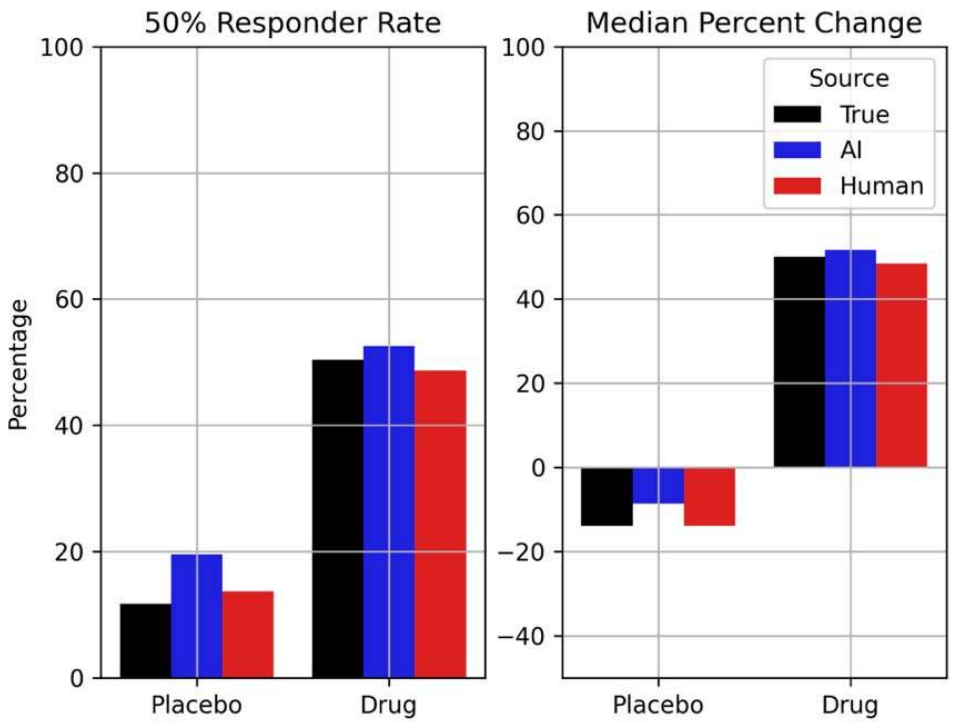



**Figure 3:** Treatment effect. Shown here are the 50% responder rate (RR50) and the median percentage change (MPC) from the placebo and drug arms of the simulated study. Three colors are shown: ground truth (black), AI estimated (blue), and human reviewed (red). All three were similar though not identical. Nevertheless, both the AI and the human would conclude that the drug is dramatically better than placebo.

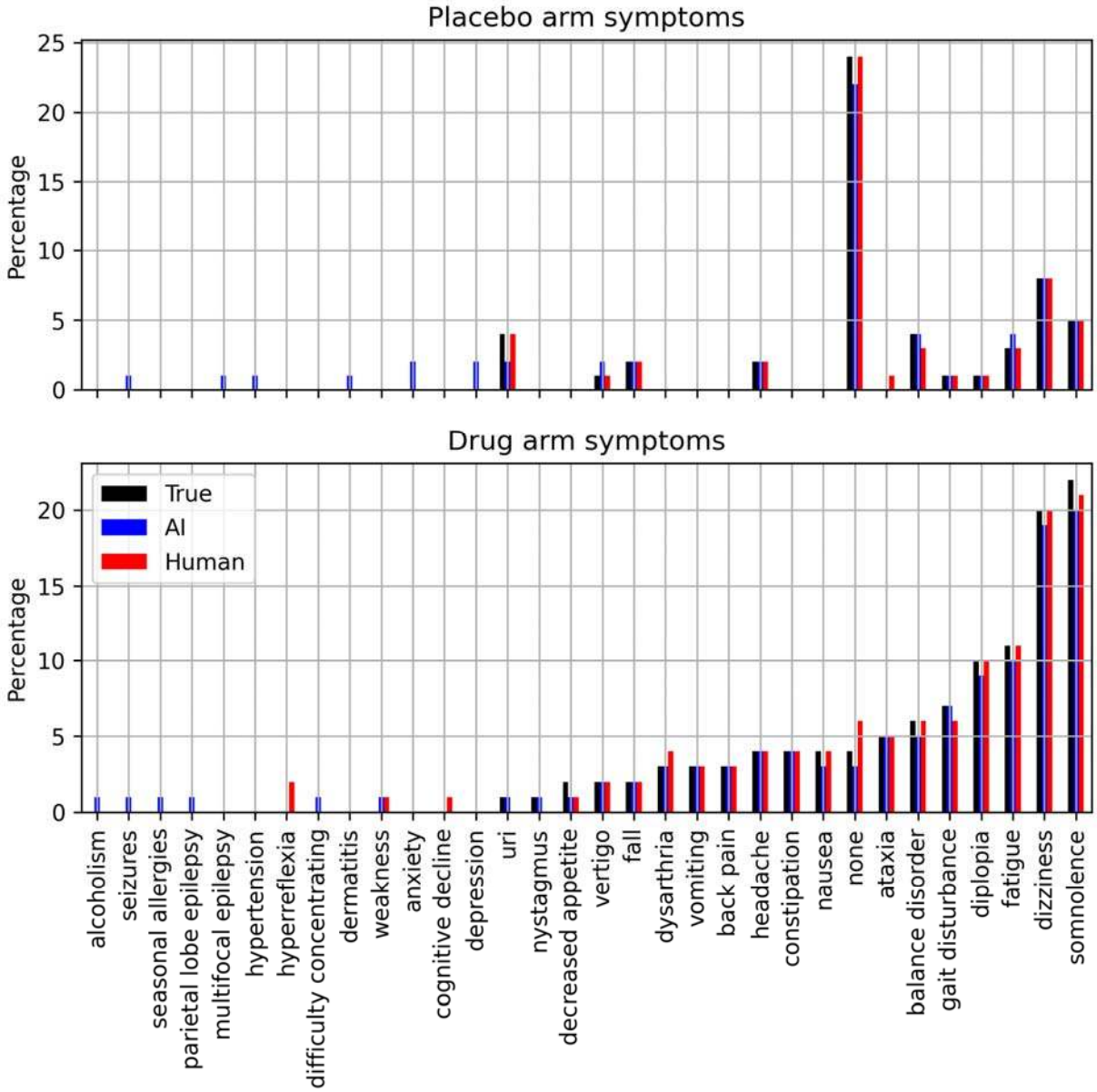



**Figure 4:** Symptom list. Shown here are the symptoms found in either drug or placebo groups. The ground truth (black), AI derived (blue), and human reviewed (red) bars indicate the fraction of each group that reported the specific symptom. Not all bars match, however the general trend is that they are within 3% of each other.

| | Ground truth | AI | Human |
|---|---|---|---|
| RR50[a] | 38%<br>$p=1*10^{-10}$ | 34%<br>$p=3*10^{-7}$ | 35%<br>$p=8*10^{-8}$ |
| MPC[b] | 54%<br>$p=1*10^{-15}$ | 61%<br>$p=8*10^{-11}$ | 62%<br>$p=1*10^{-12}$ |

Table 1: The marginal difference between placebo and drug efficacy using the 50%-responder method (RR50) or the median percentage change (MPC) methods. [a] RR50p values are computed using Fisher Exact Test. [b] MPC p values were computed using Mann-Whitney U test.

Random style

Symptoms

Diary simulator

prompt

LLM A

Treatment effect

Random details

note

True summary

placebo note

drug note

LLM B

LLM B

sum

sum

sum

sum

LLM C

Statistical analysis

Statistical analysis

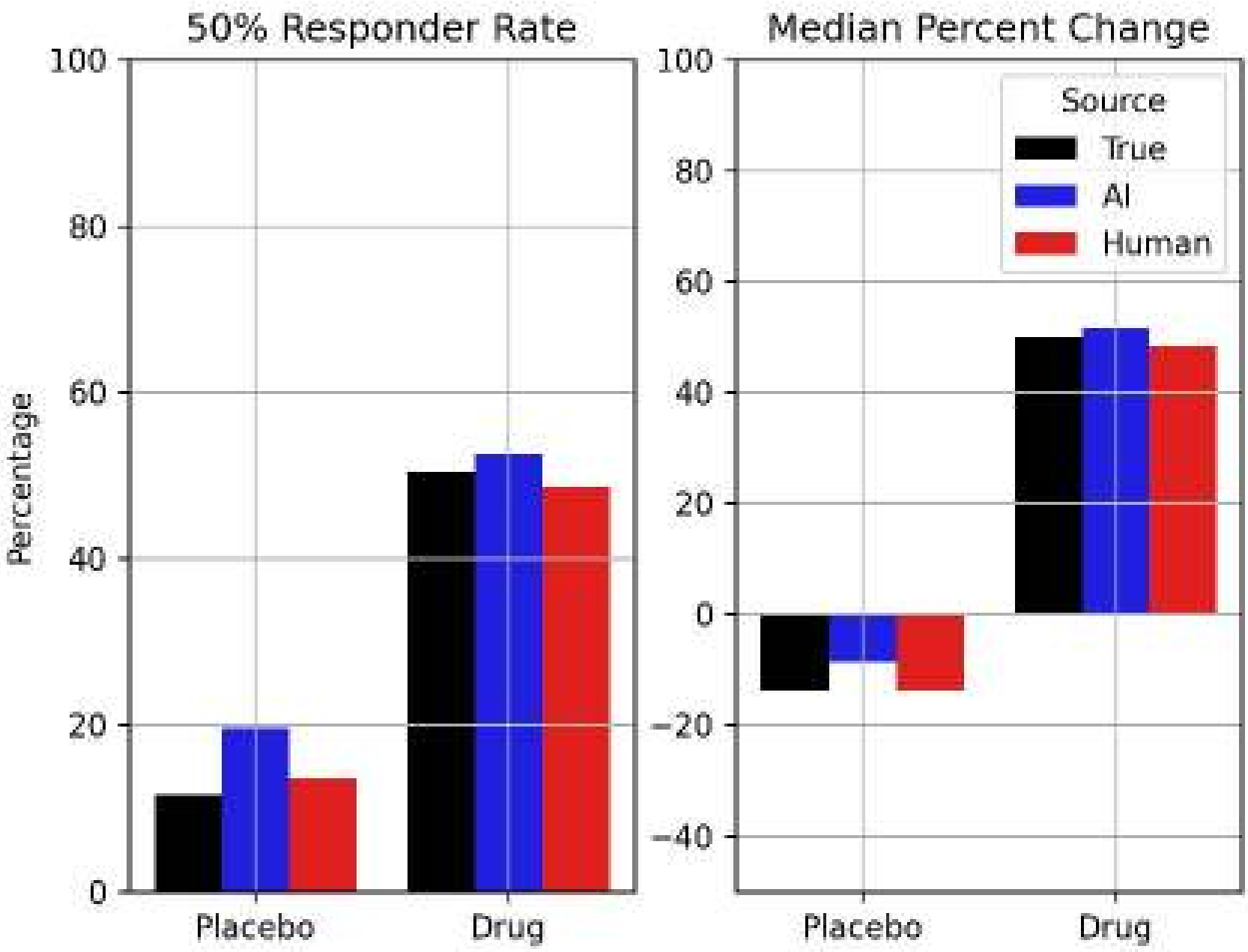
Statistical analysis

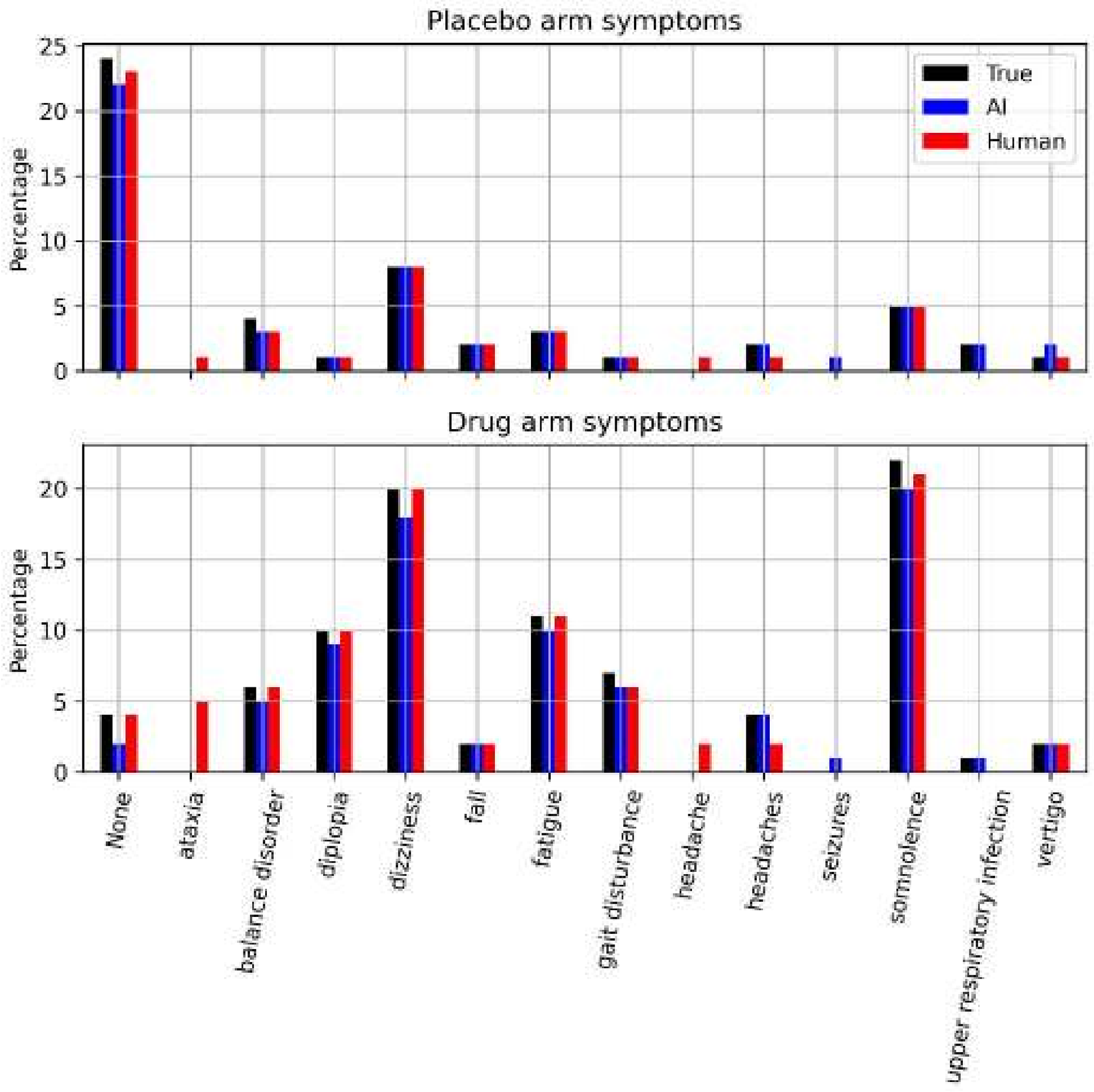
Ground truth

AI

Human

Human


50% Responder Rate
Median Percent Change
Percentage
100
80
60
40
20
0
−20
−40
Placebo
Drug
Source
True
AI
Human


Placebo arm symptoms
Percentage
True
AI
Human
Drug arm symptoms
Percentage
None
ataxia
balance disorder
diplopia
dizziness
fall
fatigue
gait disturbance
headache
headaches
seizures
somnolence
upper respiratory infection
vertigo