Drug Safety (2023) 46:711–713
https://doi.org/10.1007/s40264-023-01315-2

EDITORIAL

Future of ChatGPT in Pharmacovigilance

Hanyin Wang[1] · **Yanyi Jenny Ding**[1] · **Yuan Luo**[1]

Accepted: 27 April 2023 / Published online: 12 June 2023

Developed by OpenAI, ChatGPT is a sophisticated large language model (LLM) capable of generating responses that resemble human language when presented with written prompts, making it suitable for various applications. Since its inception, ChatGPT has shown the ability to transform how humans and machines interact, inspiring various applications across multiple domains, including pharmacovigilance.

Pharmacovigilance's primary objective is to ensure the safe and efficient utilization of medications while safeguarding public health and patient safety. An integral component of pharmacovigilance is collecting and analyzing safety information related to drugs [1]. ChatGPT's proficiency in handling copious amounts of textual data and its capability to engage in instantaneous conversations with users presents an opportunity to enhance the reporting of adverse drug reactions (ADRs) and improve the accuracy and promptness of pharmacovigilance operations.

With its training on a vast 570 GB corpus of diverse online resources, ChatGPT can function as a database for pharmacovigilance lexicon. The significance of ChatGPT lies in its capability to identify ADRs using real-world evidence sourced from nontraditional platforms, such as social media. The US FDA has approximated that a mere 1–10% of all ADRs are reported to the FDA Adverse Event Reporting System (FAERS) [2], whereas conversations about ADRs happen more frequently on social media. Previous research has showcased the capacity of natural language processing (NLP) models in mining ADRs from social media platforms through textual analysis [3]. ChatGPT's ability to utilize its vast general language knowledge gained from training to quickly adapt to new domains with minimal fine-tuning suggests that it would perform better in identifying ADR keywords on platforms such as social media, where informal language is frequently used to describe ADRs [4]. As a preliminary assessment (2 March 2023), we examined ChatGPT's ability to detect drug abuse risk in tweets by comparing its performance against the examples in a published study [4]. When supplying the same set of examples in Table 1 of the study, ChatGPT demonstrated evaluations of drug abuse risk that conform with the results in the table.

ChatGPT also demonstrated its ability to provide concise summaries in response to ADR-related inquiries about frequently used medications, with most of the content corroborated by published information. As demonstrated in a test case (2 March 2023), ChatGPT provided a list of adverse effects aligned with a published study for Lasix, a diuretic medication commonly prescribed since 1966 [5]. However, the effectiveness of ChatGPT is significantly influenced by the phrasing of the inquiry. When referring to the drug's brand name, Lasix, ChatGPT can produce a precise ADR inventory. However, when the International Union of Pure and Applied Chemistry (IUPAC) name of the ingredient (4-chloro-[(2-furan-2-ylmethyl)amine]-5-sulfamoylbenzoic acid) was utilized, ChatGPT erroneously associated it with an antibiotic called furazolidone. Nevertheless, a basic internet search by humans using the IUPAC name was able to obtain the intended outcome. Additionally, despite being marketed as a multilingual tool, ChatGPT lacks adequate training data for pharmaceuticals in languages other than English. This deficiency was evident when attempting to retrieve ADR information for the widely used medication Motrin (ibuprofen) in Chinese. In two separate attempts (2 March 2023), ChatGPT was unable to identify Motrin and erroneously linked it with aspirin.

The efficacy of ChatGPT in various pharmacovigilance tasks pertaining to less prevalent or newly approved medications that require urgent postmarket monitoring depends on the scientific rigor necessary for each task. Alpelisib, approved for metastatic breast cancer treatment just before ChatGPT's training data cut-off in 2021 [6, 7], was chosen

Hanyin Wang and Yanyi Jenny Ding contributed equally to this work.

✉ Yuan Luo
Yuan.Luo@Northwestern.edu

[1] Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, 750 N. Lake Shore Drive, 11-189, Chicago, IL 60611, USA

as a test case. In our assessment dated 16 March 2023, ChatGPT identified high blood sugar, gastrointestinal issues, and skin rash as the top three ADRs associated with alpelisib. These corresponded precisely with the most frequent grade 3/4 ADRs observed in the phase III study [8]. The FDA label [6] and European Medicine Agency (EMA) assessment report [7] confirmed lung problems and urinary tract infections, which are next on ChatGPT's list. The capacity of ChatGPT to aggregate and prioritize data could thus prove helpful in pharmacovigilance. Notably, although ChatGPT mentioned liver abnormalities, we did not find any evidence of them in the publications we reviewed. However, the EMA report did investigate liver failure laboratory tests. It is difficult to determine if ChatGPT discovered or fabricated the postmarket signal through online sources, but it could serve as a helpful lead for future investigations.

While ChatGPT was able to compile a comprehensive list of relevant ADRs, it fell short in determining the root causes. In these situations, human pharmacovigilance experts possess superior knowledge and expertise. Regarding determining the causes tied to the ADRs, ChatGPT only identified one specific cause related to alpelisib, namely the 'inhibition of PI3K pathway'. However, other causes, such as 'individual variability', are broad and do not provide much useful information. The results were inconclusive when requesting scientific evidence to support the only specific cause identified by ChatGPT. A phase I trial with incorrect citation was returned [9], with fabricated content about insulin that did not appear in the original publication. Despite the issues with one of the citations, the other source provided by ChatGPT did accurately discuss the mechanism of PI3K inhibition, which blocks insulin signaling and glucose uptake [10]. Therefore, while extra precaution should be exercised when searching for scientific evidence through ChatGPT, it still provides a starting point for research.

Regarding another crucial aspect of pharmacovigilance, the efficacy of ChatGPT as a drug–drug interaction (DDI) database remains uncertain. While supporting evidence can be found for all the DDIs for alpelisib returned by ChatGPT [11, 12], some were noted as low clinical concerns [7]. Moreover, the list was not comprehensive as one of the DDIs listed by the FDA [6], BCRP inhibitors, was not included by ChatGPT. The selective coverage of ChatGPT has limited its functionality as a comprehensive database for DDI.

After all, humans, although well-educated, are not walking encyclopedias that can learn everything and retrieve anything with a simple query. It is thus challenging for pharmacovigilance teams to identify every potential drug that may lead to a particular ADR, even after a holistic literature review. ChatGPT, as a generative model, is expected to be skilled at summarization. Yet, from a brief query example (2 March 2023), the amount of information ChatGPT could return was still limited, despite being trained with a large amount of data. When asked for an exhaustive list of drugs that cause dizziness, ChatGPT only returned 15 drugs. This list is far from complete compared with a peer-reviewed study performed by humans [13]. Only three of the 15 drugs from ChatGPT were included in the peer-reviewed study, and only four of 12 categories from the peer-reviewed study were also included by ChatGPT. While we could not rule out the possibility that ChatGPT brought in insights humans had missed, we also wanted to acknowledge its insufficiency to cover the full range of what humans could do.

Instead of using the tool as an exhaustive ADR or DDI database, a more realistic case is summarization based on given information, which mimics an adverse drug event case reviewer working for pharmaceutical companies or regulatory agencies to review clinical trial data [14, 15]. This task has two components, named entity recognition and relation extraction. The former task aims to detect the mentioning of drug names and phrases describing ADRs, and the latter task aims to abstract the relationship among the entities. As an experiment (2 March 2023), we supplied ChatGPT with one instance from the n2c2 NLP research data sets from 2018 (Track 2)—Adverse Drug Events and Medication Extraction [16]. This dataset consists of clinical notes with relations between drugs and adverse effects annotated by domain experts. ChatGPT was asked to identify adverse drug events from one piece of de-identified clinical notes. ChatGPT identified all the ADR–drug pairs in the medical history portion annotated by experts. Furthermore, additional ADRs scattered throughout the other parts of the text were also highlighted by ChatGPT in its response; therefore, the reply was comprehensive and rather exhaustive. However, we want to refrain from making a definite conclusion based on this single test case and encourage researchers in the field to conduct systematic studies to evaluate ChatGPT's ability to perform medical text summarization.

Admittedly, embracing the advancement of LLMs could introduce possibilities for more efficient pharmacovigilance, and extra precaution should be taken when it comes to real-world applications. Although ChatGPT can leverage information from social media and FAERS, regulations prohibiting the uploading of protected health information on platforms such as ChatGPT can constrain its utilization in the drug safety industry. Furthermore, at the time this article was written, the source code for ChatGPT had not been released, which hinders domain-specific pretraining or fine tuning. Although the pretraining corpus should be large enough to cover publicly accessible pharmacovigilance information, it did not systematically incorporate domain knowledge from protected databases, which are significant contributors to drug safety knowledge bases. Its performance on less commonly used drugs is less adequate due to insufficient training data. However, while ChatGPT lacks the ability to continuously update with new information

or fine tuning, other alternatives such as GPT-4 offer such customization, expanding the scope of utilizing LLMs in pharmacovigilance.

ChatGPT was trained to be generative, reflected by the different answers it returned when the same prompt was given. The tool cannot provide highly reliable scientific evidence or consistently summarize fact-based questions. Despite these limitations, ChatGPT's responses can serve as valuable starting points for downstream validation, which is often more feasible than searching for entirely new evidence. ChatGPT may also identify patterns and unravel novel signals from ADRs reported on various online platforms not easily captured by traditional methods. However, relying solely on ChatGPT's responses for decision making would compromise the scientific rigor of drug safety research and jeopardize patients' health. ChatGPT's outputs should always be considered with human expertise and existing knowledge. Therefore, while ChatGPT suffers from accuracy and coverage issues, the authors remain hopeful about its future applications in the field of pharmacovigilance.

Acknowledgments All the experiments were carried out utilizing the 13 February 2023 Version of ChatGPT API developed by OpenAI without additional engineering to the model.

Declarations

Funding No funding was provided for the authoring of this paper.

Conflicts of interest Hanyin Wang, Yanyi Jenny Ding, and Yuan Luo have no conflicts of interest to declare.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material/Code availability Not applicable.

Authors' contributions All authors contributed to conceptualization of the study and writing the manuscript, and read and approved the final version.

References

1. Meyboom RH, Egberts AC, Gribnau FW, Hekster YA. Pharmacovigilance in perspective. Drug Saf. 1999;21:429–47.
2. Heinrich J. Adverse events surveillance systems for adverse events and medical errors. United States General Accounting Office, Washington, D.C., US. 2000. https://www.gao.gov/assets/t-hehs-00-61.pdf. Accessed 27 May 2023.
3. Carpenter KA, Altman RB. Using GPT-3 to build a lexicon of drugs of abuse synonyms for social media pharmacovigilance. Biomolecules. 2023;13(2):387.
4. Hu H, Phan N, Chun SA, Geller J, Vo H, Ye X, et al. An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning. Comput Soc Netw. 2019;6(1):1–19.
5. Greger R, Wangemann P. Loop diuretics. Kidney Blood Press Res. 1987;10(3–4):174–83.
6. PIQRAY® (alpelisib) tablets, for oral use, Novartis. 2019.
7. European Medicines Agency. Assessment report - Piqray. Amsterdam: European Medicines Agency; 2020.
8. Rugo H, André F, Yamashita T, Cerda H, Toledano I, Stemmer S, et al. Time course and management of key adverse events during the randomized phase III SOLAR-1 study of PI3K inhibitor alpelisib plus fulvestrant in patients with HR-positive advanced breast cancer. Ann Oncol. 2020;31(8):1001–10.
9. Jain S, Shah AN, Santa-Maria CA, Siziopikou K, Rademaker A, Helenowski I, et al. Phase I study of alpelisib (BYL-719) and trastuzumab emtansine (T-DM1) in HER2-positive metastatic breast cancer (MBC) after trastuzumab and taxane therapy. Breast Cancer Res Treat. 2018;171:371–81.
10. Hopkins BD, Pauli C, Du X, Wang DG, Li X, Wu D, et al. Suppression of insulin feedback enhances the efficacy of PI3K inhibitors. Nature. 2018;560(7719):499–503.
11. Rocca A, Maltoni R, Bravaccini S, Donati C, Andreis D. Clinical utility of fulvestrant in the treatment of breast cancer: a report on the emerging clinical evidence. Cancer management and research. 2018;10:3083.
12. Zhou S-F. Drugs behave as substrates, inhibitors and inducers of human cytochrome P450 3A4. Curr Drug Metab. 2008;9(4):310–22.
13. Chimirri S, Aiello R, Mazzitello C, Mumoli L, Palleria C, Altomonte M, et al. Vertigo/dizziness as a Drugs' adverse reaction. J Pharmacol Pharmacother. 2013;4(1 Suppl):S104–9.
14. Schmider J, Kumar K, LaForest C, Swankoski B, Naim K, Caubel PM. Innovation in pharmacovigilance: use of artificial intelligence in adverse event case processing. Clin Pharmacol Ther. 2019;105(4):954–61.
15. Coloma PM, Trifirò G, Patadia V, Sturkenboom M. Postmarketing safety surveillance: where does signal detection using electronic healthcare records fit into the big picture? Drug Saf. 2013;36:183–97.
16. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. 2020;27(1):3–12.