Sedat Abusoglu*, Muhittin Serdar, Ali Unlu and Gulsum Abusoglu

Comparison of three chatbots as an assistant for problem-solving in clinical laboratory

https://doi.org/10.1515/cclm-2023-1058
Received September 22, 2023; accepted December 5, 2023;
published online December 14, 2023

Abstract

Objectives: Data generation in clinical settings is ongoing and perpetually increasing. Artificial intelligence (AI) software may help detect data-related errors or facilitate process management. The aim of the present study was to test the extent to which the frequently encountered pre-analytical, analytical, and postanalytical errors in clinical laboratories, and likely clinical diagnoses can be detected through the use of a chatbot.

Methods: A total of 20 case scenarios, 20 multiple-choice, and 20 direct questions related to errors observed in pre-analytical, analytical, and postanalytical processes were developed in English. Difficulty assessment was performed for the 60 questions. Responses by 4 chatbots to the questions were scored in a blinded manner by 3 independent laboratory experts for accuracy, usefulness, and completeness.

Results: According to Chi-squared test, accuracy score of ChatGPT-3.5 (54.4 %) was significantly lower than CopyAI (86.7 %) (p=0.0269) and ChatGPT v4.0. (88.9 %) (p=0.0168), respectively in cases. In direct questions, there was no significant difference between ChatGPT-3.5 (67.8 %) and WriteSonic (69.4 %), ChatGPT v4.0. (78.9 %) and CopyAI (73.9 %) (p=0.914, p=0.433 and p=0.675, respectively) accuracy scores. CopyAI (90.6 %) presented significantly better performance compared to ChatGPT-3.5 (62.2 %) (p=0.036) in multiple choice questions.

Conclusions: These applications presented considerable performance to find out the cases and reply to questions. In the future, the use of AI applications is likely to increase in clinical settings if trained and validated by technical and medical experts within a structural framework.

Keywords: artificial intelligence; clinical laboratory; assistant; machine learning

Introduction

The Chat Generative Pre-Trained Transformer (ChatGPT) is an artificial intelligence (AI) model; this was developed by Open AI (OpenAI, L.L.C., San Francisco, CA, USA) as software that can provide human-like text. ChatGPT includes a vast diversity of language owing to the conversational large language model (LLM) [1]. Although the first version published in 2018 had a transformer structure on model use with a 40 GB dataset and 1.5 B parameters, the later and advanced GPT-3 reached to 570 GB dataset and 175 B parameter coverage [2]. ChatGPT v4.0 became remarkably popular, especially in the field of education, and the program was reported to achieve 57–78 % success in board exams, including the United States Medical Licensing Exam (USMLE), Multiple-Choice Question Answering (MedMCQA), and Biomedical Research Question Answering (PubMedQA) and passed the exams [3]. It was quickly adopted by Khan Academy and Duolingo for use in the field of education [4]. ChatGPT v4.0 is one of the leading artificial intelligence programs and can have applications in clinical diagnostic support systems, data management, and patient education. Regarding the content generation category, AI alternative bots have also been developed, including WriteSonic and Copy AI, similar to ChatGPT v4.0 [5].

The aim of the present study was to test the extent to which the frequently encountered pre-analytical, analytical, and postanalytical errors in clinical laboratories, and likely clinical diagnoses can be detected through the use of chatbots such as ChatGPT-3.5, ChatGPT v4.0., WriteSonic, and Copy AI. To the best knowledge of the authors, this was the first study to investigate the performance of different chatbots in clinical settings.

Sedat Abusoglu and Muhittin Serdar contributed equally to this work.

***Corresponding author: Prof. Sedat Abusoglu**, PhD, Department of Biochemistry, Selcuk University Faculty of Medicine, Alaaddin Keykubat Campus, Postal Code: 42075 Selcuklu, Konya, Türkiye, Phone: +905370212647, E-mail: sedatabusoglu@yahoo.com. https://orcid.org/0000-0002-2984-0527

Muhittin Serdar, Department of Biochemistry, Acıbadem Mehmet Ali Aydinlar University Faculty of Medicine, İstanbul, Türkiye, E-mail: maserdar@hotmail.com. https://orcid.org/0000-0002-3014-748X

Ali Unlu, Department of Biochemistry, Selcuk University Faculty of Medicine, Konya, Türkiye, E-mail: aunlu@selcuk.edu.tr. https://orcid.org/0000-0002-9991-3939

Gulsum Abusoglu, Department of Medical Laboratory Techniques, Selcuk University Vocational School of Medicine, Konya, Türkiye, E-mail: tekinglsm@gmail.com. https://orcid.org/0000-0003-1630-1257

Materials and methods

Study design

This cross-sectional, observational study was performed in June 2023 at the Department of Biochemistry, Faculty of Medicine, Selçuk University, Konya, Turkey. Three different formats of questions related to clinical laboratory were developed based on the quality indicators as prescribed by the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) Working Group "Laboratory Errors and Patient Safety" project. According to this project, quality indicators were defined for three different categories such as pre-analytical, intra-analytical and post-analytical phases to cover total testing process. Preanalytical indicators were classified as misidentification errors (mislabeled or unlabelled samples), test transcription errors (requests with erroneous data entry for in- or out-patients), incorrect sample type or fill level, unsuitable samples for transportation and storage problems, contaminated, haemolysed and clotted samples, inappropriate test requests, inappropriate time in sample collection, unintelligible and inappropriate requests. Intra-analytical indicators were presented as test with inappropriate ICQ (Internal quality control) performances, tests not covered by an EQA-PT (External quality assessment-proficiency testing) control, unacceptable performances in EQA-PT schemes and data transcription errors. Post-analytical indicators consist of inappropriate turnaround times, incorrect laboratory reports, notification of critical values, interpretative comments, results notification; outcome measures such as sample recollection and inaccurate results; support processes such as employee competence, client relationships and efficiency of laboratory information system [6].

In this study, eight cases were classified into pre-analytical category described with codes and definitions as Pre-MisS: Misidentified samples, Pre-InTime: Samples collected at inappropriate time of sample collection, Pre-NotSt: Samples not properly stored before analysis, Pre-InTem: Samples transported at inappropriate temperature, Pre-WroCo: Samples collected in wrong container, Pre-Hem: Samples with free Hb >0.5 g/L (clinical chemistry), Pre-MicCon: Contaminated samples, Pre-WroTy: Samples of wrong or inappropriate type. Ten cases were matched (Seven cases with no match for appropriate code) with intra-analytical phase as Intra-Var: Tests with CV% higher than selected target (Two different cases within same code) and Intra-Unac: Unacceptable performances in EQAS-PT schemes. Two cases were defined in post-analytical step with codes as Out-InacR: Inaccurate results released and Post-IncRep: Incorrect reports issued by the laboratory.

In the present study, 20 case scenarios, 20 multiple-choice, and 20 direct questions related to commonly encountered errors in pre-analytical, analytical, and postanalytical processes in the clinical laboratory setting were developed in English language. These case scenarios were not developed in the format of a multiple-choice question. The prepared case scenarios were completely fictional and did not include real patient data. The content thereof was not quoted from existing sources. Laboratory results, clinical findings, and pre-analytical-analytical-postanalytical issues were identified with regard to above-mentioned case scenarios. The current iteration of the chatbots used in the study was until 2021; thus, the reference source for the multiple-choice and direct questions was considered as Tietz 6th Edition [7]. Although the questions were being entered into the program, they were converted into text format, considering that they might have the ability to synthesize information from narrative texts into medical results. Format 1: In the case Format (n=20) (Supplementary Material Case List), Format 2: Non-case, direct spot questions without multiple choices (n=20) (Supplementary Material Question List); Format 3: Multiple choice questions (n=20) (Supplementary Material Multiple Choice Question List). Prompts were developed for each of the three categories (Supplementary Material Prompt List). In above prompts, the chatbots were instructed to answer in the form of "The correct answer to this question is:" with an aim to obtain clear answers for the question. In the study, a chat session was conducted with four online chatbots (ChatGPT-3.5, ChatGPT v4.0., WriteSonic, and Copy.ai) and questions were asked in a random order and the responses were accordingly recorded.

Data collection

To avoid the influence of prior answers on questions, a new session was initiated for each question. For all questions, the chatbots were asked whether their responses were from an evidence-based reference source, and if so, the source was recorded. A statistical evaluation was performed for the responses of each of the four chatbots to the scenarios. The first response by the programs without using the "regenerate response" button was considered the final response. A difficulty assessment (1: Easy; 2: Medium; and 3: Difficult) was performed for the total 60 questions and responses from four chatbots were scored in a blinded manner by three independent laboratory experts with respect to accuracy (1: Totally incorrect; 2: Almost Correct; and 3: Totally Correct), usefulness (1: Totally useless; 2: Almost useful; 3: Totally useful), and completeness (1: Totally incomplete, 2: Almost complete; 3: Totally complete).

Accuracy (also called as correctness or truthfulness): This expression adress to the scientific and technical correctness of chatbots' statements, with regard to valid clinical evidence and clinical laboratory implementations. Correctness does not include only the right answer about the case, also the explanation given in the response from the point of information consistency. In direct or multiple-choice case questions, accuracy refers to accurate representation and truthfulness with evidence-based source (textbook) information.

Usefullness (also named as usability or practicability): This concept defines chatbots' ability to give beneficial insights to patients, laboratory professionals and scientists and comprises all sides about the information provided by chatbots. This term includes giving proper recommendations, presenting relevant and right comments, increasing laboratory professionals' understanding of test results and interpretation. This term helps laboratory experts to minimize time to reach the correct information.

Completeness (also known as entireness or integrity): This dimension explains the unity between the responses of chatbots and the actual evidence-based information about the cases and questions. This phenomenon covers all aspects about correct presentation of pre-, post- and analytical considerations, clinical data and demonstrates chatbots' utility to fully present the entire information.

To assess the accuracy of the cases, the Reflective Testing (European Federation of Clinical Chemistry and Laboratory Medicine [EFLM]) website and other reported case training references were used as sources [8, 9].

The chatbots were evaluated according to hallucinations. Open-domain hallucinations: It is a type of hallucination due to wrong external data source. The chatbot confidentially presents wrong information about the question. Close-domain hallucinations: It is a type of hallucination due to inability of checking out keywords, inappropriate

interpretation of question, wrong analysis, providing non-sense or absent information even though the external data source or reference is reliable [10].

Statistical analysis

All statistical calculations (Inter rater-agreement and Chi-squared test) were performed using the MedCalc Statistical Software version 19.2.6 (MedCalc Software bv, Ostend, Belgium; https://www.medcalc.org; 2020); and Microsoft 365 Excel spreadsheets (v16.0.16501.20074). Two-sided p-values <0.05 were considered as statistically significant. The scores from the blinded assessment of three independent laboratory experts were calculated as a total score for each category.

Results

Total score percentages for accuracy, helpfulness, completeness were presented in Figure 1. There were 60 questions in total, 20 in each of the three groups and the maximum score was 180. In terms of accuracy, the highest scores in the case scenarios in their respective order were as follows: ChatGPT v4.0. (total score=160), CopyAI (total score=156), WriteSonic (total score=129), and ChatGPT-3.5 (total score=98). According to Chi-squared test, accuracy percentage of ChatGPT-3.5 (54.4 %) was significantly lower than CopyAI (86.7 %) (p=0.0269) and ChatGPT v4.0. (88.9 %) (p=0.0168), respectively. There was no statistical difference between ChatGPT-3.5 (54.4 %) and WriteSonic (71.7 %) percentage values (p=0.263).

The highest scores in the direct question group in respective order were as follows: ChatGPT v4.0. (total score=142), CopyAI (total score=133), WriteSonic (total score=125), and ChatGPT-3.5 (total score=122). According to comparison of accuracy percentages, there was no statistical difference between ChatGPT-3.5 (67.8 %) and WriteSonic (69.4 %), ChatGPT v4.0. (78.9 %) and CopyAI (73.9 %) (p=0.914, p=0.433 and p=0.675, respectively).

However, in the multiple-choice question group, the highest scores in respective order were as follows: CopyAI (total score=163), WriteSonic (total score=157), ChatGPT v4.0. (total score=140), and ChatGPT-3.5 (total score=112). While there was no statistically significant change between score points percentages of ChatGPT-3.5 (62.2 %)- ChatGPT v4.0. (77.8 %) (p=0.287), ChatGPT-3.5 (62.2 %)- WriteSonic (77.8 %) (p=0.072), CopyAI (90.6 %) presented significantly better performance compared to ChatGPT-3.5 (62.2 %) (p=0.036).

Multiple choice question (total score=142), direct question (total score=126), and case (total score=117) received the highest scores according to the questions' difficulty levels.

According to inter-rater agreement, Kappa values were ranged between 0.40 and 0.53; 0.21–0.40 and 0.08–0.37 for

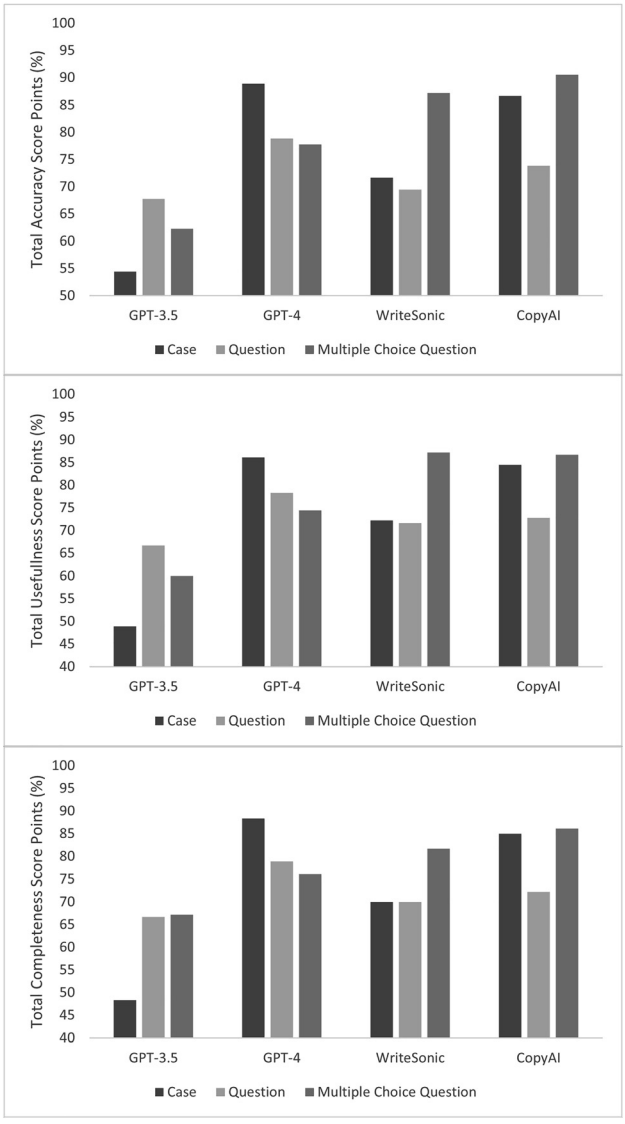

Figure 1: Total accuracy, usefullness, completeness score rates for all chatbots.

case, direct question and multiple-choice questions, respectively.

ChatGPT-3.5 did not provide reference sources for 12 answers (60 %) of the 20 questions related to case scenarios, whereas this rate was 35 and 10 % for WriteSonic and CopyAI, respectively. ChatGPT v4.0. provided reference sources in all case questions.

Regarding the multiple-choice questions, ChatGPT-3.5 provided references for all of the answers, while ChatGPT v4.0., WriteSonic, and CopyAI did not provide references for 25 %, 70 %, and 15 % of the answers, respectively.

Regarding the direct questions, ChatGPT-3.5 and ChatGPT v4.0. provided references for all questions, whereas WriteSonic and CopyAI did not provide references for 25 and 5 % of the questions, respectively.

Open-domain/close-domain hallucination percentages in cases were 55/15; 5/15; 25/25; 5/10 % for ChatGPT-3.5, ChatGPT v4.0., WriteSonic, and CopyAI, respectively. In case group, 2 and 4 of ChatGPT v4.0. and CopyAI references were not found (Supplementary Material Case List).

Discussion

The European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI) assessed the performance of ChatGPT v4.0. over 10 clinical scenarios. Seven laboratory experts evaluated the responses by ChatGPT v4.0. for relevance, accuracy, usefulness, and safety. Although the responses by ChatGPT v4.0. had the lowest scores in terms of helpfulness, relationship building and safety parameters showed the highest scores [11]. The prompt prepared for the purposes of the present study was designed for the program to respond to a specific command. It is likely that posing the direct question without any prompt would prove to not be a limiting factor considering the creativity of the program. The researchers concluded that ChatGPT v4.0. was capable of assessing laboratory data in a test–test context and detect deviations from the reference range and abnormal results. All laboratory professionals are recommended to familiarize themselves with this and other similar AI applications [11]. In the present study, the prompts were directed in combination with the question. Nevertheless, considering that the chatbots provided responses to confirm the diagnosis, the clinicians received different scores during the expert evaluations.

In a Japanese study [12] ChatGPT v4.0. was reported to have higher accuracy scores (79.9 %) than ChatGPT-3.5 (50.8 %) for all questions. Wang et al. [10] reported better accuracy rates for ChatGPT v4.0. (82–86 %) than ChatGPT-3.5 (56–76 %) in national medical examinations.

In another study [13], ChatGPT v4.0. performed better accuracy scores (80.6 %) compared to ChatGPT-3.5 (61.3) and Google Bard (54.8 %) ($p<0.009$). These results were consistent with the results of the present study. These findings represent a remarkable progress in the development of ChatGPT 4 from the older version ChatGPT-3.5.

Munoz-Zuluaga et al. [14] reported the percentage of correct answers to 65 questions by ChatGPT v4.0 version and ChatGPT-3.5 version was 50.7 %, and 26 %, respectively; furthermore, the bots failed to identify the hook effect. Similarly, in the present study, ChatGPT v4.0. version had a higher rate of correct responses than ChatGPT-3.5 version (88.9 vs. 54.4 %). Similarly in the present study, Case 17 was associated with substrate depletion like hook effect. Nevertheless, none of the chatbots provided comments on the serum lipase value, which was randomly set to a low level, in the case with a definitive diagnosis of acute pancreatitis. It was surprising to get the exact true results from the chatbots for those common technical cases in a laboratory setting, including Case 18 (Macroprolactin), Case 19 (Biotin interference), Case 10 (Volume Displacement Effect), Case 8 (Carry-over), and Case 14 (Paraprotein interference). In Question 18, for the serum prolactin/total prolactin ratio after precipitation to confirm the diagnosis of macroprolactin during the assessment, the laboratory experts reported that the ratio provided by the chatbots, but not the source reference, was the ratio they used in practice.

ChatGPT was reportedly not aware that the pediatric blood lead cut-off level was 3.5 µg/dL (falsely reported as 3.5 mg/dL in Ref. [14]) because this information was updated after 2021. In the present study, for the adult blood lead levels indicative of severe exposure in Question 13, different cut-off values were observed in different sources and the study date of the reference given in the question was not up to date. Similarly, ChatGPT was unable to answer this question correctly, whereas Copy AI provided the information, i.e., “Council of State and Territorial Epidemiologists’ (CSTE) blood lead reference value is 3.5 µg/dL,” without providing a reference. This was suggestive of the fact that different chatbots might have access to up-to-date information. World Health Organisation (WHO) reported the same value as 5 µg/dL in 2021 [15]. This might be considered as an advantage to use chatbots due to availability to recent data about some specific concerns.

As an AI language model, ChatGPT v4.0. can only provide information based on the data it is trained on; thus, its answers should only be considered as a source of information and not as an alternative to professional medical advice [16]. Above all, clinical laboratories must ensure that laboratory data is accurate and reliable to avoid the risk of the use of inaccurate data by sophisticated systems, including those used by Machine Learning (ML) and AI, leading to inaccurate and potentially harmful information [17].

Although machine learning-based artificial intelligence applications currently have limited capabilities, they have the potential to synthesize the behavioral patterns of laboratory experts via deep learning in the database and to support clinical decision-making. The bots lack human creativity, perception, developing approaches to complex biological patterns, finding solutions to cases with emotional intelligence, and reflecting professional experiences on decision-making processes. Artificial intelligence-based machine learning applications should be trained and validated by technical and medical experts within a structural framework described as “Human-in-the-loop” [18]. Clinical laboratories must provide precise and correct laboratory data to prevent

wrong interpretations produced from artificial intelligence based machine learning applications that can result with inaccurate information about the patient care [19].

In addition to the infra-structural development of the process, control and healthy maturation of the learning steps from human source may be achieved. Within this loop, human factor can enable the transfer of comprehensions as well as the development of different algorithms and designing the bridges between conditions and contexts. Therefore, it is possible to use these auxiliary programs in a controlled manner in the process of static or adaptive algorithm development and learning and to develop evidence-based approaches with the support of laboratory experts.

It is obvious that AI-based chatbots use different algorithms to approach and evaluate scientific data. For this reason, it can be considered that especially laboratory experts can highly contribute to the development of these applications. Since these tools can develop advanced learning skills over time, their periodical re-evaluation at every stage will contribute to the improvement of a more reliable structure. Limited number of independent evaluators, single and same source for all questions might be considered as a selection bias for this study.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

References

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare 2023;11:1–20.
2. OpenAI. Chatbot generative pre-trained transformer, ChatGPT. https://openai.com/blog/chatgpt [Accessed 6 May 2023].
3. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47:33.
4. Swathi K. GPT 4 used in Khan academy and Duolingo to enhance their AI learning platforms. https://www.analyticsinsight.net/gpt-4-is-being-used-by-khan-academy-and-duolingo/ [Accessed 18 Apr 2023].
5. Chen TJ. ChatGPT and other artificial intelligence applications speed up scientific writing. J Chin Med Assoc 2023;86:351–3.
6. Sciacovelli L, Lippi G, Sumarac Z, West J, Del Pino Castro IG, Vieira KF, et al. Quality indicators in laboratory medicine: the status of the progress of IFCC working group “laboratory errors and patient safety” project. Clin Chem Lab Med 2017;55:348–57.
7. Rifai N, Horvath AR, Wittwer C. Tietz textbook of clinical chemistry and molecular diagnostics, 6th ed. St. Louis, Missouri, USA: Elsevier; 2018.
8. Oosterhuis WP, Verboeket-van de Venne WPHG. Reflective testing in primary care. http://www.reflectivetesting.com/uk/index.htm [Accessed 25 Apr 2023].
9. Allen LC, Dominiczak MH, Pulkki K, Pazzagli M. Clinical case material for teaching clinical chemistry and laboratory medicine. Clin Chem Lab Med 2001;39:875–89.
10. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inf 2023;177:105173.
11. Cadamuro J, Cabitza F, Debeljak Z, Bruyne SD, Frans G, Perez SM, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European federation of clinical chemistry and laboratory medicine (EFLM) working group on artificial intelligence (WG-AI). Clin Chem Lab Med 2023;61:1158–66.
12. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 2023;9:e48002.
13. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, et al. Benchmarking large language models’ performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770.
14. Munoz-Zuluaga C, Zhao Z, Wang F, Greenblatt MB, Yang HS. Assessing the accuracy and clinical utility of ChatGPT in laboratory medicine. Clin Chem 2023;69:939–40.
15. World Health Organization. WHO guideline for the clinical management of exposure to lead. Geneva: CC BY-NC-SA 3.0 IGO; 2021.
16. Temsah O, Khan SA, Chaiah Y, Senjab A, Alhasan K, Jamal A, et al. Overview of early ChatGPT’s presence in medical literature: insights from a hybrid literature review by ChatGPT and human experts. Cureus 2023;15:e37281.
17. Padoan A, Plebani M. Artificial intelligence: is it the right time for clinical laboratories? Clin Chem Lab Med 2022;60:1859–61.
18. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. Artif Intell Rev 2023;56:3005–54.
19. Plebani M. Artificial intelligence in laboratory medicine: lights and shadows. Biochim Clin 2023;47:217–9.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cclm-2023-1058).