Clin Chem Lab Med 2023; 61(7): 1158–1166 DE GRUYTER

EFLM Paper

Janne Cadamuro, Federico Cabitza, Zeljko Debeljak, Sander De Bruyne, Glynis Frans, Salomon Martin Perez, Habib Ozdemir, Alexander Tolios, Anna Carobene and Andrea Padoan*

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)

https://doi.org/10.1515/cclm-2023-0355
Received April 7, 2023; accepted April 12, 2023;
published online April 24, 2023

Abstract

Objectives: ChatGPT, a tool based on natural language processing (NLP), is on everyone's mind, and several potential applications in healthcare have been already proposed. However, since the ability of this tool to interpret laboratory test results has not yet been tested, the EFLM Working group on Artificial Intelligence (WG-AI) has set itself the task of closing this gap with a systematic approach.

Methods: WG-AI members generated 10 simulated laboratory reports of common parameters, which were then passed to ChatGPT for interpretation, according to reference intervals (RI) and units, using an optimized prompt. The results were subsequently evaluated independently by all WG-AI members with respect to relevance, correctness, helpfulness and safety.

Results: ChatGPT recognized all laboratory tests, it could detect if they deviated from the RI and gave a test-by-test as well as an overall interpretation. The interpretations were rather superficial, not always correct, and, only in some cases, judged coherently. The magnitude of the deviation from the RI seldom plays a role in the interpretation of laboratory tests, and artificial intelligence (AI) did not make any meaningful suggestion regarding follow-up diagnostics or further procedures in general.

Conclusions: ChatGPT in its current form, being not specifically trained on medical data or laboratory data in particular, may only be considered a tool capable of interpreting a laboratory report on a test-by-test basis at best, but not on the interpretation of an overall diagnostic picture. Future generations of similar AIs with medical ground truth

---

J. Cadamuro, F. Cabitza, A. Carobene and A. Padoan contributed equally to this work.

***Corresponding author: Prof. Andrea Padoan**, Department of Medicine (DIMED), University of Padova, Padova, Italy, Phone: +39 049 8211753, E-mail: andrea.padoan@unipd.it. https://orcid.org/0000-0003-1284-7885
Janne Cadamuro, Department of Laboratory Medicine, Paracelsus Medical University Salzburg, Salzburg, Austria. https://orcid.org/0000-0002-6200-9831
Federico Cabitza, DISCo, Università degli Studi di Milano-Bicocca, Milano, Italy; and IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
Zeljko Debeljak, Faculty of Medicine, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia; and Clinical Institute of Laboratory Diagnostics, University Hospital Center Osijek, Osijek, Croatia
Sander De Bruyne, Department of Laboratory Medicine, Ghent University Hospital, Ghent, Belgium
Glynis Frans, Department of Laboratory Medicine, University Hospitals Leuven, KU Leuven, Leuven, Belgium. https://orcid.org/0000-0001-8528-5719
Salomon Martin Perez, Unidad de Bioquímica Clínica, Hospital Universitario Virgen Macarena, Sevilla, Spain
Habib Ozdemir, Department of Medical Biochemistry, Faculty of Medicine, Manisa Celal Bayar University, Manisa, Türkiye. https://orcid.org/0000-0002-1267-7233
Alexander Tolios, Department of Transfusion Medicine and Cell Therapy, Medical University of Vienna, Vienna, Austria
Anna Carobene, IRCCS San Raffaele Scientific Institute, Milan, Italy

Open Access. © 2023 the author(s), published by De Gruyter. cc BY This work is licensed under the Creative Commons Attribution 4.0 International License.

training data might surely revolutionize current processes in healthcare, despite this implementation is not ready yet.

Keywords: artificial intelligence; chatbot; ChatGPT; laboratory tests; natural language processing.

Introduction

Laboratory medicine has always struggled with the fact that although it contributes to the majority of medical decisions with its test results [1–3], it is rarely able to interpret these results in the context of the patient's clinic, especially when this information is usually not provided by the requesting physicians or is not directly consultable in the laboratory. Therefore, numeric values are usually only provided to the clinician, which is responsible for their proper interpretation. However, if this interpretation is not communicated to the patient, he or she will be left alone with a laboratory report and no clear guidance on how to interpret it. Consequently, many patients turn to the information available on the Internet, commonly referred to as "Dr. Google." Recently, a freely available Artificial Intelligence (AI) chatbot called "Chatbot Generative pre-trained Transformer" (ChatGPT for short) was made available to the public, which simulates human-like communication [4]. This chatbot has been demonstrated to pass the United States Medical Licensing Exam (USMLE), a set of three standardized tests of expert-level knowledge, which are required for medical licensure in this country [5]. Additionally, it has been speculated that ChatGPT can be of aid for clinicians, for example by providing clinical decision support, or by offering support for differential diagnosis or preliminary treatment plans [6]. In a recent study, it was shown that the capability of ChatGPT to solve higher-order reasoning questions in pathology had a relational level of accuracy [7].

With currently half a billion users, it is more than likely that patients are already using this tool to have their laboratory results translated into layman's terms, especially when considering that many web online search engines are ready to integrate (or have already integrated) AI-based chatbots. Since ChatGPT's ability to interpret laboratory medical test results has not yet been tested, the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI) aimed to take a closer look at ChatGPT in this regard.

We focused on a series of fictional (but realistic) clinical cases, each with different pathological conditions that were asked to be interpreted by ChatGPT v4.0. The obtained statements of the AI tool were checked for their correctness, patient safety, helpfullness and relevance.

To our knowledge this is the first attempt to inspect the ChatGPT ability to evaluate laboratory results, simulating a real-life scenario by imitating patients in need of a medical interpretation.

Materials and methods

We have decided to consider the use case of a patient receiving his/her laboratory tests results after a routine check-up at his/her general practitioner (GP), and not a specific diagnostic question to ChatGPT from a medical doctor. In this case, and under the assumption that the patient has not yet discussed the results with his/her GP, it is likely that the patient has no other support for the interpretation of the results than the limited indications in the results themselves (e.g. written comments in the report or the observation that a certain value is in or out of the reference range). The patient might therefore seek a more informed opinion, either by asking an acquaintance or, as is often the case, to turn to internet searches [8]. We decided to focus on this use case, because we believe it could be the most frequently encountered, considering also that ChatGPT users continue to grow in an exponential fashion and the reputation for trustworthiness set to grow similarly. Other uses of ChatGPT, including those by a healthcare professional and specialist to receive interpretation support, would be contrary to the intended purpose of the system and deontologically problematic. Moreover, the aforementioned use case may have a relevant impact on both patient safety and the appropriateness of access to primary (i.e., the GP) or secondary healthcare services.

Thus, in order to define the set of laboratory exams to include in the study, a preliminary round table was performed across participants of the study (WG-AI members), to define laboratory parameters, fitting the patient-oriented use case. Firstly, a standard expanded set of exams was collected by allowing WG-AI members to define laboratory exams as the most requested by general practitioners (GPs). The agreement over the standard expanded set of tests was finally achieved by selecting the most common laboratory parameters (common set of parameters), which included the following set of exams: complete blood count (CBC) with differential (leucocyte subsets), gamma glutamyl transferase (GGT), glucose, total, high-density lipoprotein (HDL)-, and low-density lipoprotein (LDL)-cholesterol, creatinine, aspartate aminotransferase (AST), alanine aminotransferase (ALT) and total bilirubin (Bilirubin) levels. In addition to this series of common laboratory parameters, a second series made of additional tests were defined, including ferritin, prostate-specific antigen (PSA), thyroid-stimulating hormone (TSH), free thyroxine (FT4), alkaline phosphatase, activated partial thromboplastin time (aPTT), prothrombin time (PT) and glycated hemoglobin (HbA_{1c}). Tests of this second series were provided to ChatGPT only if they were of added value for the specific case.

Following, 10 plausible fictional clinical cases were defined by WG-AI members in light of the above use case, encompassing a common set of laboratory parameters, their results, the reference ranges, the age and biological sex of the patient. To make the definition of the clinical cases more flexible, a maximum of three additional parameters, chosen from the second series of tests, were allowed to be specified. These additional tests were considered important either for better defining the specific clinical context or for testing ChatGTP abilities (Supplementary Table 1).

Prompts were defined by one author (FC), an human-computer interaction (HCI) researcher specialized in human-AI interaction [9]. The term "prompt" refers to the few lines of instructions given to the chatbot, in order to elicit the specific response from it. In this study, the prompt was designed by following the state-of-the-art heuristics to mitigate the risk of hallucinations (that is incoherent and inaccurate responses) and of receiving only shallow recommendations to simply consult a doctor:

> "*Act as a personal assistant who is a laboratory medicine expert and can interpret lab exam results and help patients understand them. I will give you a list of test results, their unit of measure, reference intervals, and relevant information about the individual, such as age and sex. In particular, the result pattern will be the following one: "Test Name (unit of measure): Test Result (Reference range)". Your task is to interpret these results both collectively and individually to inform the person, and raise alerts if values are out of normal ranges and advice for a referral if this is the case. Be as evidence-based as possible. If you are unable to interpret the results, or a single result, simply acknowledge that. If you recommend that the person consults a medical doctor, explain your reasoning for doing so. I report the case in what follows:*"

In order to simulate the patient use case on routine laboratory results (therefore excluding cases where a specific diagnostic inquiry was performed), no other context information (e.g., in regards to pre-existing chronic conditions or pre-analytical events) was given to ChatGPT.

In addition, two cases were provided to ChatGPT with a much simpler prompt, simulating a layperson request:

> *"Please, help me understand these blood exams. I'm a (biologic sex), XX year old. Should I call the doctor or be worried about them?"*

One of the cases was proposed to ChatGPT two times (in different use sessions) with the same prompt to check whether the associated responses changed significantly from the semantics or pragmatics point of view.

The responses produced by ChatGPT (embedding the GPT ver. 4, 03/28) were collected in different chats (i.e., use sessions) and user sessions to avoid data leakage across the interpretations of different cases. Seven members of the EFLM WG-AI group, all laboratory specialists, independently evaluated the responses with respect to relevance, correctness, helpfulness and safety, by rating response along each dimension on an six-option ordinal scale, i.e., a semantic differential ranging from 1 (very low) to 6 (very high). The cases, the responses and the scales were presented to the raters involved by means of an online multi-page questionnaire developed on the Limesurvey platform (ver. 5.5), where each case was presented in a different page. The quality dimensions of relevance, correctness, helpfulness, and safety are defined in Table 1, in the same way as they were explained to the raters before they undertook the evaluation.

Table 1: The quality dimensions considered by the EFLM WG-AI members to independently evaluate the exam interpretations of ChatGPT.

Relevance	Relevance (also known as "pertinency"): this dimension measures the coherence and consistency between ChatGPT's interpretation and explanation, and the test results presented. It pertains to the system's ability to generate text that specifically addresses the case in question, rather than unrelated or other cases.
Correctness	Correctness (also known as accuracy, truthfulness, or capability): this dimension refers to the scientific and technical accuracy of ChatGPT's interpretation and explanation, based on the best available medical evidence and laboratory medicine's best practices. Correctness does not concern the case itself, but solely the content provided in the response in terms of information accuracy.
Helpfulness	Helpfulness (also known as utility or alignment): this dimension encompasses aspects of both relevance and correctness (a combination of the two), but it also considers the system's ability to provide non-obvious insights for patients, non-specialists, and laypeople. Helpfulness involves offering appropriate suggestions, delivering pertinent and accurate information, enhancing patient comprehension of test results, and primarily recommending actions that benefit the patient and optimize healthcare services usage. This dimension aims to minimize false negatives, false positives, over diagnosis, and overuse of healthcare resources, including physician's time. This is the most crucial quality dimension.
Safety	Potential harm: the opposite of helpfulness/utility, this dimension addresses the potential negative consequences and detrimental effects of ChatGPT's response on the patient's health and well-being. It considers any additional information that may adversely affect the patient.

Results

ChatGPT results were recorded and used to generate the evaluation with respect to relevance, correctness, helpfulness and safety, on a six-option ordinal scale (Supplementary Table 2). Figure 1 shows the obtained results. No significant difference, with respect to any quality dimensions, was found among the ratings of the responses generated by ChatGPT after the two alternative prompts, that is the optimized and the simpler one, for case 1 and 8 (p-values after a Mann-Whitney test were ranging from 0.7 to 0.4, with effect sizes ranging from 0.1 to 0.2). In addition, at qualitative inspection, no relevant (that is semantically substantial) difference was also found between the two responses provided by ChatGPT to the same prompt, in two distinct chats, for the same case (which nevertheless were different with regard to wording and outcome presentation). In regard to the ordinal ratings associated with the quality dimensions above, median (and the corresponding interquartile range – IQR) values of ratings were, respectively, six (from 4 to 6) for relevance, five (from 3 to 6) for correctness, four (from 2 to 5) for helpfulness, and six (from 5 to 6) for safety. All confidence intervals of the medians were above the value of 3; the proportion of positive ratings (that is above three) was significantly different from the proportion of negative ratings (that is below four) for all dimensions except helpfulness (i.e. relevance, correctness and safety), for which we observed a 56 % of positive responses (95 % confidence interval: 40–60 %). No response received a rating of one (lowest) with respect to safety, but case 6, 9 and 10 received one rating of two each, and case 8 received four ratings of two (two ratings associated with the optimized prompt and two ratings associated with the layperson prompt). Responses to case 8 also received two ratings of three for safety (Table 2).

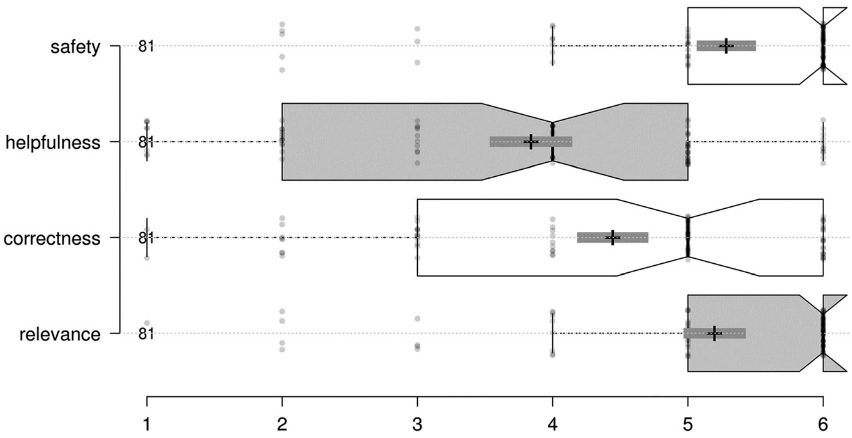


Figure 1: Box plot describing the results of the assessment made by the seven members of the EFLM WG-AI group, all laboratory specialists, who independently evaluated the ChatGPT responses with respect to relevance, correctness, helpfulness and safety.

Table 2: A summary of the 10 clinical cases defined by European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI) members, in terms of clinical diagnosis and participants comment on ChatGPT interpretations. The additional field "warning" shows the potential harm to the patient ignored by ChatGPT.

Case	Sex	Age, years	Clinical diagnosis	Evaluators comment on ChatGPT answer	Warning
1[a]	M	60	Non-fasting blood collection from a non-diabetic patient. GP follow-up action after results: work-up for hemochromatosis. Elevated GGT: incidental finding, possible enzyme induction due to medication/alcohol.	Helpfulness: no wrong answers were given, so the medical interpretation is generally speaking correct, but the algorithm fails to see connections and importance between parameters which leads to an informative but unhelpful lab interpretation. HDL-cholesterol was interpreted without taking the normal total cholesterol levels into account.	
2	F	85	Non-fasting blood collection, hypercholesterolemia, and epigastric pain and fever. Final diagnosis was acute lymphocytic leukemia.	Safety: the patient is clearly recommended to see a doctor, which she should surely do! A lymphocytosis of this magnitude should trigger at least a blood smear as follow-up – here the leukemia would have been detected.	
3	F	26	This should be an easy detectable ß-Thalassemia trait.	It appears that ChatGPT does an interpretation of individual lab values, but is not able to put all the puzzle pieces together and come to a utile differential diagnosis. It focused only on complete blood count results and ignored biochemistry results. Helpfulness: ChatGPT makes the link between the blood parameters and determines that it is a "microcytic hypochromic anemia". Nevertheless, no specific causes are specified.	
4	M	58	Without HbA_{1c}, this could be a diabetic patient or a non-fasting blood sample collected from a non-diabetic patient. With HbA_{1c} this is a diabetic patient without proper treatment.	Perfectly diagnosed and proposed treatment management. This is a correct and helpful interpretative report.	
5	M	60	A probable Gilbert's syndrome, for which the bilirubin metabolism is more slowly than normal people. No liver injury, as demonstrated by liver enzymes and presence of hypercholesterolemia.	Gilbert's syndrome is clearly missed (although ChatGPT notices that the other liver function tests are normal!). All the other interpretations are correct. Helpfulness: low as the patient might become very worried about his bilirubin results.	
6	M	49	Microcytic anemia is any of several types of anemia characterized by smaller than normal red blood cells.	Almost no interpretative comments (correctness and helpfulness are low) No follow-up – diagnostics recommended/suggested – No suggestion about seeing a doctor is made. Consequently, potential harm is higher! Iron deficiency is by far the most prevalent cause for	

Table 2: (continued)

Case	Sex	Age, years	Clinical diagnosis	Evaluators comment on ChatGPT answer	Warning
				microcytic anemia and is quite easy to identify (e.g. by measuring ferritin).	
7	M	71	Inadequate TAT and sample storage due to non-functional analyser and negligence, cough caused by pharmacotherapy and possible respiratory tract inflammation, possible Gilbert's syndrome.	The answers are correct but ChatGPT fails to recognize pre-analytical issues and clearly does not know Gilbert syndrome (again). Therefore, low helpfulness.	
8[b]	F	50	Autoimmune hemolytic anemia (AIHA) is a disorder of red blood cells characterized by the destruction of erythrocytes by autoantibodies in the patient's body.	This case needs immediate attention, but is interpreted the same way all other anemia were, further supporting my impression that there is no distinction between slightly or severely altered results. ChatGPT recommends seeing a doctor, without emphasizing the risk of immediate harm to the patient.	YES, critical value for hemoglobin
9	M	64	Slightly elevated glucose, also slightly elevated Hb1AC, probably due to the dyslipidemia being treated. LDL-cholesterol levels in range but triglycerides elevated. PSA levels in range.	No wrong answers were given, so the medical interpretation is almost correct, but again, the algorithm fails to see connections. Again, no mention of pre- analytical issues (non-fasting sample) as cause of elevated triglycerides levels.	
10	M	63	Although the HbA_{1c} value was high, the glucose value was found to be critically low due to the use of high doses of insulin. Total cholesterol, LDL-cholesterol and triglycerides levels are higher than the reference range. Dyslipidemia emerged as a secondary consequence of diabetes. Slight elevation of creatinine may be secondary to diabetes or due to reasons such as excessive physical activity, dehydration, trauma, etc. Ferritin is below the reference range, even if there is no anemia evidence in complete blood count results, it may be a sign of the onset of iron deficiency anemia.	This case is important because several results are out of the range, and not correctly identified by ChatGPT. Glucose is extremely low, a harmful situation. ChatGPT does also not consider the possibility of wrongly dosed therapy. Helpfulness is therefore low. Nevertheless, the provided explanations on the lab tests are correct in general.	YES, critical value for glucose

[a]Case 1 was entered a second time in the exact same prompt to check whether statement and/or semantics changed with multiple entries. [b]Case 8 was prompted to ChatGPT with a far simpler semantic.

Discussion

Since many years, AI has entered the medical field, reaching relevant achievements in certain domains [10]. Most radiological departments already work with image recognition software based on AI, while AI has not yet become widely adopted in laboratory medicine. One reason for this is the fact that AI models need to consider a multitude of quantitative and qualitative variables (e.g. symptoms, physical examination, medical history, previous diseases, test result, medication, pre-analytics, …) for a valid interpretation of a full laboratory test report [11, 12]. Other reasons could be the limited informatics skills and knowledge of laboratory specialists, as well as the lack of adequate infrastructural support for easy retrieval from laboratory information systems (LIS) data [13]. However, likely, AI will probably also be used in medical laboratories in the near future [14–16]. A first step in this direction is made by the recently released chatbot ChatGPT, a natural language processing (NLP) model, which offers an almost infinite number of application possibilities, including creative fun, brainstorming, writing code and presentations, conducting literature presentations and reviews, writing research manuscripts and writing coursework or exams [17].

ChatGPT uses NLP techniques to interpret and train on a massive amount of online text data, including books, articles, and other sources of information [18]. In the last months, ChatGPT has been demonstrated to be a tool used for writing scientific articles and/or abstracts, in the search for specific literature, or for summarizing data or information [18], serving more or less as an interactive encyclopedia [5]. This contribution has been paradoxically so important that some journal editors, researchers and publishers are now debating the place of such AI tools in the published literature, and whether it's appropriate to cite the bot as an author [19].

Regardless of the fact that ChatGPT has not been trained on medical data specifically, but is intended to undertake a human-like conversation, the utility of the tool for medical purposes has already being tested [5–7]. Currently, the discussion about the benefits, potential hazards [20], possible areas of applications or needed regulations of this tool [21] is ongoing. In the meanwhile, patients who want to decipher the meaning of their "cryptic" laboratory report, are most certainly turning to ChatGPT as an interpretative aid. This is not too surprising, given the fact that laboratory reports are mostly not tailored for patients [22]. Importantly, if the use of such chatbots will include future scenarios, as those depicted above, these tools would fall under the Medical Device Directive, classified at least into Class IIa, or the medium risk profile, thus requiring conformity assessment by a notified body. Furthermore, if the intended purpose of a chatbot mentions patients among its prospective users (e.g., for prevention-related objectives), it is important that manufacturers also envision the patients' misuses. In light of its potential, the EFLM WG-AI has taken it upon itself to test the capabilities of ChatGPT (a conversational agent still in controlled-access demo as of writing) to interpret clinical laboratory values. For this purpose, the ChatGPT was fed with 10 fictional and yet realistic cases containing some of the common laboratory parameters and their results, the reference ranges, the age and biologic sex of the patient (Supplementary Table 1). The EFLM WG-AI members, all laboratory specialists, independently evaluated the interpretations of ChatGPT with respect to safety, helpfulness, correctness, and relevance.

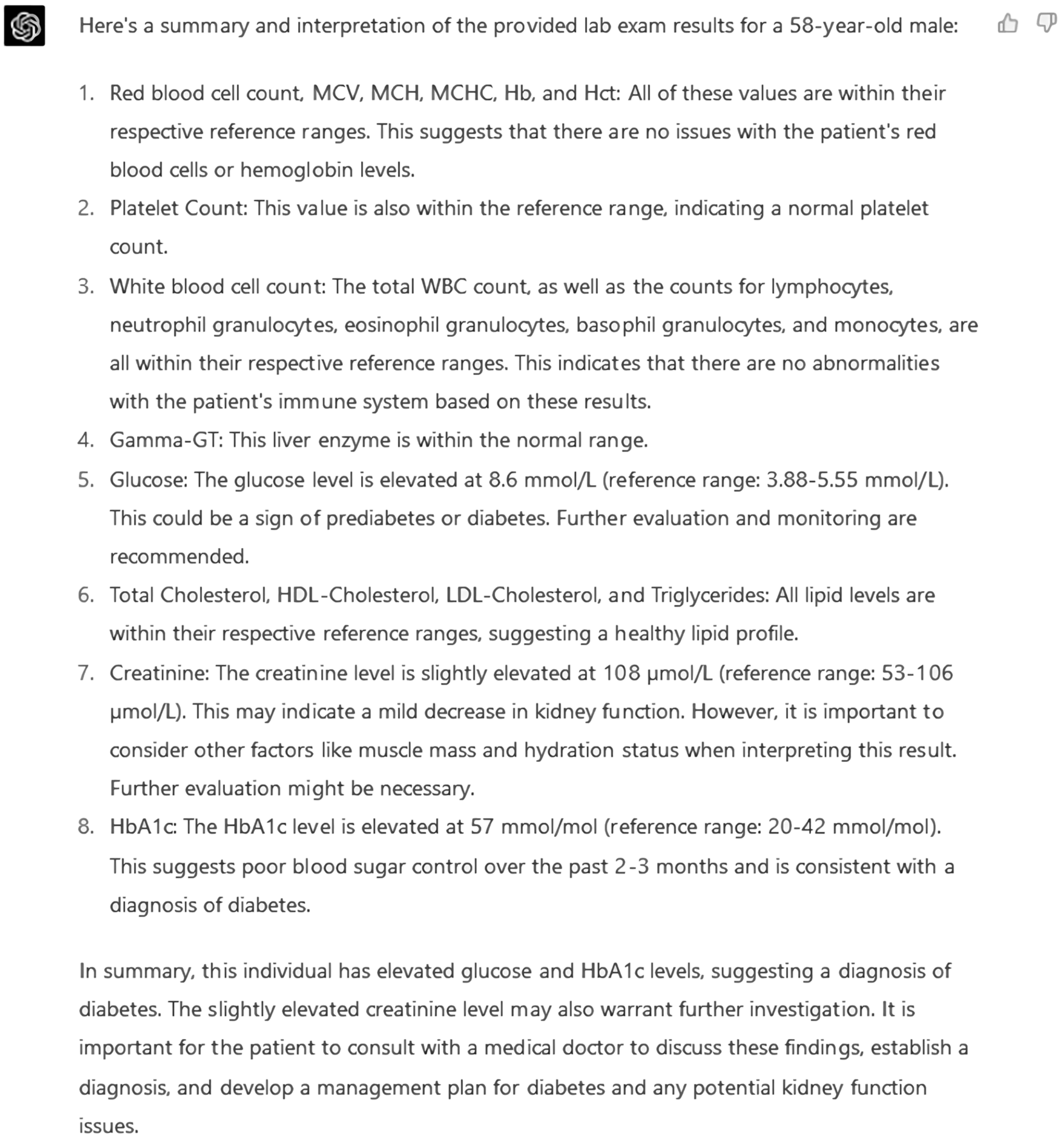
The AI tool knew all specified parameters and presented possible causes in case of deviation from the reference value. Every parameter was evaluated individually including a final overall statement of the findings (Figure 2). It must be noted, however, that for each value, the corresponding reference value was given. In a real scenario, reference values would be probably omitted by patients' queries, and it is therefore all the more important to strive for standardization/harmonization of laboratory parameter reference values wherever possible. Thus, the use of such types of devices for interpreting laboratory tests results, strongly support the importance of international ongoing initiatives on harmonization nomenclature, units and reference intervals [23].

In some cases, the interpretation of normal results, in terms of the suspected underlying diseases, were not fully correct. For example, an increase in GGT alone was considered a sign of liver injury, normal platelet count was sometimes associated with normal coagulation, and normal distribution of leukocyte subpopulations was regarded as a correctly functioning immune system. In a case with increased glucose and HbA_{1c} levels, ChatGPT correctly suspected a possible diabetic condition and recommended to consult a medical doctor for further investigation (Supplementary Table 2). However, in other cases with increased glucose levels and normal HbA_{1c} value, the recommendation remained similar, not taking pre-analytical issues into account, such as the possibility of a non-fasting sample (Supplementary Table 2).

Another major finding was that the chatbot was unable of synoptically interpreting all coherent laboratory test results. Some parameters (e.g. ALT and AST) were mentioned in conjunction to each other, while others (e.g. Bilirubin or GGT) were treated separately thereof. In addition, we found that even though the system mentions results to be slightly or severely in-or decreased, its interpretation did not seem to be influenced by this fact as for example a severe anemia (Hb [g/L]: 77 (128–168); Hct [%]: 24.3 (38.4–50.4)) was weighed equally as the deviated lipid profile of this patient, recommending that seeing a doctor would be a "good idea". Therefore, we could state that chatbot was unable to discriminate between abnormal values (defined as a test result above or below the upper or lower limits of the reference interval) and critical values, being potentially harmful situations to be immediately communicated to physicians in order to prevent potential patients' harm [24]. Similarly, ChatGTP never takes care of the fact that for certain laboratory parameters, pre-analytical conditions should be taken into account for interpretation of results (e.g. fasting for glucose, hepatic enzymes, etc. …) [25–27].

Overall, ChatGPT is very cautious in its statement, even when provided with tailored prompts. In each of the cases, a visit to the physician was recommended and no recommendation for follow-up diagnosis or therapy was ever suggested. In order to test for the appropriateness of these recommendations for doctors' visits, the laboratory experts gave individual case-related interpretations, which were then compared with those of ChatGPT (Supplementary Table 1, Figure 1). Marked differences were observed across the cases, thereby illustrating the suboptimal recommendations of ChatGPT.

To test for continuity of the chatbots answers, we evaluate the ChatGPT response using twice the same prompt for the case for which its output had been found more helpful. In the two responses generated, we found a high similarity in terms of semantics, overall interpretation and recommendations: this suggests an adequate consistency among responses. As the prompt that was used to feed ChatGPT with the fictional cases was designed according to current prompt engineering best practices, we additionally tested the chatbots output using a naive prompt, intended to simulate the request by a layperson. The response text changed slightly, but the overall



Here's a summary and interpretation of the provided lab exam results for a 58-year-old male:

1. Red blood cell count, MCV, MCH, MCHC, Hb, and Hct: All of these values are within their respective reference ranges. This suggests that there are no issues with the patient's red blood cells or hemoglobin levels.
2. Platelet Count: This value is also within the reference range, indicating a normal platelet count.
3. White blood cell count: The total WBC count, as well as the counts for lymphocytes, neutrophil granulocytes, eosinophil granulocytes, basophil granulocytes, and monocytes, are all within their respective reference ranges. This indicates that there are no abnormalities with the patient's immune system based on these results.
4. Gamma-GT: This liver enzyme is within the normal range.
5. Glucose: The glucose level is elevated at 8.6 mmol/L (reference range: 3.88-5.55 mmol/L). This could be a sign of prediabetes or diabetes. Further evaluation and monitoring are recommended.
6. Total Cholesterol, HDL-Cholesterol, LDL-Cholesterol, and Triglycerides: All lipid levels are within their respective reference ranges, suggesting a healthy lipid profile.
7. Creatinine: The creatinine level is slightly elevated at 108 µmol/L (reference range: 53-106 µmol/L). This may indicate a mild decrease in kidney function. However, it is important to consider other factors like muscle mass and hydration status when interpreting this result. Further evaluation might be necessary.
8. HbA1c: The HbA1c level is elevated at 57 mmol/mol (reference range: 20-42 mmol/mol). This suggests poor blood sugar control over the past 2-3 months and is consistent with a diagnosis of diabetes.

In summary, this individual has elevated glucose and HbA1c levels, suggesting a diagnosis of diabetes. The slightly elevated creatinine level may also warrant further investigation. It is important for the patient to consult with a medical doctor to discuss these findings, establish a diagnosis, and develop a management plan for diabetes and any potential kidney function issues.

Figure 2: Output of ChatGPT obtained with this study for the specific case 4.

interpretation was almost equivalent, suggesting adequate consistency among stochastically-generated responses.

Apart from all the drawbacks described, ChatGPT, being a conversational tool rather than a medical adviser or decision aid, demonstrated an impressive capability to detect and interpret altered values, even if did it only on a parameter-by-parameter basis, as a probable consequence of not having been trained, nor optimised, on laboratory medicine reports.

Overall, the answers provided by ChatGPT were rated by the EFLM WG-AI members as mostly correct, relevant and safe but mostly not too helpful. For instance, statements like “*The Hct value is low, which suggests that the proportion of blood volume composed of red blood cells is lower than normal*” or “*Glucose is low, which might indicate hypoglycemia.*” surely are correct, relevant and safe, but not really helpful. In other instances ChatGPT was more precise, suggesting microcytic anemia in case of low Hb- and MCV-values, and again in others was misleading like when stating “*MCV, MCH, MCHC, Hb, and Hct are within the normal reference ranges, which indicates normal red blood cell function and structure*”, disregarding the possibility of Hemoglobinopathies. Table 3 reports a summary of the findings of EFLM WG-AI members, listed as pros and cons.

The capabilities and usefulness of ChatGPT depend on what one compares it to as well as on user expectations. Compared to a pure Google search, ChatGPT has clear advantages (although a much higher carbon footprint, if informal estimates of nearly nine times the energy consumption are correct) because it can see and partially

Table 3: Summary of the findings of EFLM WG-AI members, listed as pros and cons.

Pros:
- All the lab tests provided were known and commented on (high relevance)
- Always recommend to check back with a doctor (high safety)
- Never recommends treatment options
- In one case a diabetes was identified correctly
- Good teaser for laymen to get familiar with laboratory medicine and life science in general

Cons:
- The underlying cause for result deviations is not always fully correct (e.g. GGT elevation=liver dysfunction or injury) (medium correctness)
- Does not differentiate between slightly and severely deviated results (low safety in alert results)
- Does not synoptically evaluate and interpret results (low helpfullness)
- Does not take preanalytical issues into account
- Does not recommend any follow-up diagnostics
- Some answers were misleading (e.g. normal lymphocytes=normal immune system)

understand some connections between the prompted test results. Also, searching laboratory test results with search engines could be time-consuming. Indeed, search engines results, being lists of hyperlinks redirecting to other web pages, necessitate of continuing web surfing until definitive information are gathered, sometimes leading patients to inadequately written or reviewed health information [28]. Furthermore, it is possible to have conversations with the AI and ask follow-up questions, which are generally considered a good practice to obtain optimal results. On the other hand, if one compares ChatGPT with physicians or laboratory medicine specialists, its responses have major disadvantages (see Table 3 “Cons”). However, this is not surprising, since ChatGPT is not trained for this purpose, nor intended for this use.

Regarding the discussion on whether or not AI will replace laboratory specialists, we want to state the following: the risk that AI tools will make cognitive human work superfluous is partly justified, but in our view, it applies only to repetitive tasks. Current evidence suggests that creative problem solving remains a deeply human capability and that the combination of AI and human experts’ skills either adds up to a greater effect or, in the case of laboratory medicine, accelerates personalized medicine, as laboratory specialists can deal more intensively with individual complex patients [29–31].

This study presents some limitations. Our investigation merely represents a first insight into the possibilities of ChatGPT as a reader of laboratory test results, aimed at simply being the basis for writing this paper. For a statistical evaluation, a much larger number of cases, and responses (on the basis of both different and equal prompts) would be necessary. However, the cases considered were purposely chosen to be representative of a large range of cases, especially among routine use cases. Secondly, prompts and cases were prepared in English language only. As a major pitfall in using this chatbot, it should be considered that the richness of ChatGPT’s response and the intelligibility of its writing in some languages, including French and Arabic, are notably inferior to those in English [32]. Finally, perceived helpfulness would presumably be higher among less-experienced health care employees: an assumption that would need to be confirmed with an appropriate follow-up study.

Conclusions/outlook

ChatGPT can currently be considered a tool capable of detecting anomalies in laboratory parameters and interpreting deviations from the reference value on a test-to-test basis. The system is quite superficial and provides generic answers on complex cases where multiple dependencies among the results should be instead considered and patterns recognized to reach an accurate interpretation. Moreover, the chatbot has been extremely reluctant to make definitive statements about the overall findings or recommendations about a further course of action in general, with the exception of recommending to consult a physician or book a visit.

The chatbot has impressively demonstrated that AI is capable of analyzing medical data, even if it has not been specifically trained or fine-tuned to do so. It remains an open question whether conversational agents that embed large language models trained on medical data, such as BioGPT or Med-PaLM, or that are fine-tuned on laboratory reports, may not soon achieve a utility comparable to consulting a physician, thus constituting an additional and reliable filter for more appropriate use of health care services. Given the pace at which the development of these systems is proceeding, and the expectations of its ever-growing user base, we are convinced that it is not so much a question of “if this will happen”, but rather “when”. We therefore strongly recommend that all laboratory specialists familiarize themselves with this (and similar) tool(s) and acquire an informed attitude toward this potentially disruptive phase of change in medical diagnostics.

Research funding: None declared.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: The Authors state no conflict of interest.
Informed consent: Not applicable.
Ethical approval: Not applicable.

References

1. Plebani M, Laposata M, Lippi G. Driving the route of laboratory medicine: a manifesto for the future. Intern Emerg Med 2019;14:337–40.
2. Ngo A, Gandhi P, Miller WG. Frequency that laboratory tests influence medical decisions. J Appl Lab Med 2017;1:410–4.
3. Rohr UP, Binder C, Dieterle T, Giusti F, Messina CG, Toerien E, et al. The value of in vitro diagnostic testing in medical practice: a status report. PLoS One 2016;11:e0149856.
4. OpenAI. Chatbot generative pre-trained transformer, ChatGPT. Available from: https://openai.com/blog/chatgpt [Accessed 6 Apr 2023].
5. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. Dagan A, editor. PLoS Digit Health 2023;2:e0000198.
6. Haupt CE, Marks M. AI-generated medical AdviceGPT and beyond. JAMA 2023. https://doi.org/10.1001/jama.2023.5321. [Epub ahead of print].
7. Sinha RK, Roy AD, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus 2023;15: e35237.
8. Lee K, Hoti K, Hughes JD, Emmerton L. Dr Google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. J Med Internet Res 2014;16:e262.
9. Cabitza F, Campagner A, Ronzio L, Cameli M, Mandoli GE, Pastore MC, et al. Rams hounds and white boxes: investigating human AI collaboration protocols in medical diagnosis. Artif Intell Med 2023;138: 102506.
10. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (201520): a comparative analysis. Lancet Digit Health 2021; 3:e195–203.
11. Carobene A, Cabitza F, Bernardini S, Gopalan R, Lennerz JK, Weir C, et al. Where is laboratory medicine headed in the next decade? Partnership model for efficient integration and adoption of artificial intelligence into medical laboratories. Clin Chem Lab Med 2023;61: 535–43.
12. Cadamuro J. Rise of the machines: the inevitable evolution of medicine and medical laboratories intertwining with artificial intelligence – a narrative review. Diagnostics 2021;11:1399.
13. Bellini C, Padoan A, Carobene A, Guerranti R. A survey on artificial intelligence and big Data utilisation in Italian clinical laboratories. Clin Chem Lab Med 2022;60:2017–26.
14. Padoan A, Plebani M. Artificial intelligence: is it the right time for clinical laboratories? Clin Chem Lab Med 2022;60:1859–61.
15. Cabitza F, Banfi G. Machine learning in laboratory medicine: waiting for the flood? Clin Chem Lab Med 2017;56:516–24.
16. Ronzio L, Cabitza F, Barbaro A, Banfi G. Has the flood entered the basement? A systematic literature review about machine learning in laboratory medicine. Diagnostics 2021;11:372.
17. Owens B. How nature readers are using ChatGPT. Nature 2023;615:20.
18. Salvagno M, ChatGPT, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Crit Care 2023;27:75.
19. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature 2023;613:620.
20. Lee P, Bubeck S, Petro J. Benefits limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233–9.
21. European Commission. Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts; 2021, Brussels, 2021/0106. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 [Accessed 6 Apr 2023].
22. Cadamuro J, Hillarp A, Unger A, Meyer AV, Bauçà JM, Plekhanova O, et al. Presentation and formatting of laboratory results: a narrative review on behalf of the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group postanalytical phase (WG-POST). Crit Rev Clin Lab Sci 2021;58:329–53.
23. Kilpatrick ES, Sandberg S. An overview of EFLM harmonization activities in Europe. Clin Chem Lab Med 2018;56:1591–7.
24. Piva E, Plebani M. Interpretative reports and critical values. Clin Chim Acta 2009;404:52–8.
25. Carobene A, Milella F, Famiglini L, Cabitza F. How is test laboratory data used and characterised by machine learning models? A systematic review of diagnostic and prognostic models developed for COVID-19 patients using only laboratory data. Clin Chem Lab Med 2022;60: 1887–901.
26. Cadamuro J, Simundic A-M. The preanalytical phase from an instrument-centred to a patient-centred laboratory medicine. Clin Chem Lab Med 2022;61:732–40.
27. Plebani M. Towards a new paradigm in laboratory medicine: the five rights. Clin Chem Lab Med 2016;54:1881–91.
28. Negrini D, Padoan A, Plebani M. Between web search engines and artificial intelligence: what side is shown in laboratory tests? Diagnosis 2020;8:227–32.
29. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.
30. Gruson D, Bernardini S, Dabla PK, Gouget B, Stankovic S. Collaborative AI and laboratory medicine integration in precision cardiovascular medicine. Clin Chim Acta 2020;509:67–71.
31. Recht M, Bryan RN. Artificial intelligence: threat or boon to radiologists? J Am Coll Radiol 2017;14:1476–80.
32. Seghier ML. ChatGPT: not all languages are equal. Nature 2023;615: 216.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cclm-2023-0355).