UDK 007.52:616.71-007.15-053.3 **Original Scientific Paper**
Doi: 10.7251/SANUS2604045S

APPLICATION OF ARTIFICIAL INTELLIGENCE IN THE INTERPRETATION OF LABORATORY FINDINGS

Đorđe Stojisavljević*[1], Darija Knežević[2], Duška Jović[2], Igor Paravac[2], Ljubiša Preradović[3]

[1]University Computer Center, Bulevar vojvode Petra Bojovića 1A, 78000, Banja Luka, the Republic of Srpska, Bosnia and Herzegovina;
[2]Faculty of Medicine, University of Banja Luka, Banja Luka, Save Mrkalja 19 Banja Luka, the Republic of Srpska, Bosnia and Herzegovina,
[3]Faculty of Architecture, Civil Engineering and Geodesy, Bulevar vojvode Petra Bojovića 1A, 78000, the Republic of Srpska, Bosnia and Herzegovina

Abstract. *Interpretation of laboratory results represents one of the key components of clinical decision-making and is an integral part of everyday medical practice. Laboratory medicine provides objective, quantitative, and qualitative data on a patient's physiological and pathological state; however, these data alone do not constitute a diagnosis and do not have full clinical significance without appropriate medical interpretation. This paper analyzes the level of agreement in the interpretation of laboratory test results between the artificial intelligence system MedGPT and physicians under conditions with and without reported clinical symptoms. The study was designed as an observational and comparative analysis. Laboratory parameters were selected to include common biochemical and hematological markers routinely used in clinical practice with established diagnostic value. The analysis was conducted on a controlled set of simulated laboratory results, both with and without additional clinical symptoms. The results demonstrate a high level of concordance between MedGPT and physicians when interpreting isolated laboratory data, with no statistically significant differences in assessments. Agreement between artificial intelligence assessments of laboratory findings with and without symptoms was moderate, with a statistically significant difference between the two conditions ($p = 0.001$). Agreement between physician assessments of laboratory findings with and without symptoms was also moderate and statistically significant ($p = 0.001$). A high and statistically significant level of agreement was observed between artificial intelligence and physicians for findings without clinical symptoms ($p < 0.001$). The inclusion of clinical symptoms leads to reduced agreement and greater variability in interpretations, but without significant differences in the final conclusions. These findings indicate that artificial intelligence systems can serve as a useful support tool in the post-analytical phase of laboratory medicine, while maintaining the central role of physicians' clinical judgment.*

Key words: *artificial intelligence, laboratory medicine, interpretation of laboratory results*

Introduction

Laboratory findings should be viewed as information that supports clinical reasoning rather than as isolated indicators of disease, with their significance contingent upon the clinical context, the individual characteristics of the patient, and the dynamics of

parameter changes over time [1]. Traditional interpretation of laboratory findings in clinical practice is based primarily on the individual clinical reasoning of the physician, who interprets laboratory test results in the context of the patient's medical history, physical examination, and personal experience. The challenges physicians face are multifaceted and extend beyond medical knowledge to encompass organizational and informational factors, including the complexity of test selection, varying nomenclature for identical analyses, insurance limitations, patient costs, and unclear or inconsistent laboratory report formats [2,3]. An experimental study conducted by ALQahtani et al. [4] provides clear empirical evidence that time pressure exerts a negative impact on physicians' diagnostic accuracy. Under controlled conditions, participants exposed to perceived time constraints arrived at diagnoses considerably more quickly, but with significantly lower accuracy compared to colleagues working without such pressure. The authors demonstrated that physicians under time pressure made 37% more diagnostic errors on average, emphasizing a clear trade-off between speed and quality of clinical reasoning.

The application of artificial intelligence (AI) in laboratory medicine represents one of the most significant technological advances in contemporary diagnostics. Unlike conventional statistical and expert systems, modern machine learning and deep learning models enable the analysis of large, heterogeneous, and dynamic datasets, directly contributing to improvements in the accuracy, efficiency, and clinical relevance of laboratory findings [5]. Current research indicates that AI can be applied across all phases of the laboratory process: pre-analytical, analytical, and post-analytical. In the pre-analytical phase, machine learning algorithms are used to optimize sample processing, reduce patient identification errors, and predict unacceptable samples. In the analytical phase, AI supports quality control, automated result verification, and the interpretation of complex analytical patterns. In the post-analytical phase, AI plays a particularly important role in the clinical interpretation of laboratory findings and the integration of results with clinical data [6]. Lippi and Plebani [7] demonstrate that AI models can achieve diagnostic performance comparable to, and in certain cases superior to, that of human experts, particularly in tasks requiring the recognition of subtle patterns within large and complex datasets. In laboratory medicine, this advantage is most evident in areas such as automated morphological analysis of blood cells through digital morphology, urinary sediment interpretation, recognition of microbiological patterns, and screening for malignant, metabolic, and genetic disorders.

Although the application of AI in laboratory medicine offers numerous advantages, its integration into routine clinical practice is accompanied by significant technical, organizational, ethical, and legal challenges. Inadequate consideration of these aspects may lead to misinterpretation of results, erosion of trust, and potential risks to patient safety [8]. Besides the increasingly prevalent use of chatbot systems by patients, we have observed growing interest among physicians in the use of AI assistants in everyday clinical practice in the recent years. Unlike patients, who most commonly use chatbot systems to obtain a basic understanding of laboratory findings, physicians employ these tools as auxiliary aids to support clinical reasoning,

education, information synthesis, and workflow optimization, but not as a substitute for their own professional judgment [9]. The literature clearly emphasizes that chatbot systems, including advanced models such as ChatGPT, have significant limitations when used in a clinical context. Their responses may be convincing but are not necessarily accurate, and the absence of genuine clinical reasoning and causal understanding poses a risk in complex or atypical cases. For this reason, the prevailing view in the literature is that chatbot systems should be used exclusively as a support tool, with the final decision remaining the responsibility of the physician, who must consider the complete clinical context, laboratory data, and the individual characteristics of the patient [10–12].

Emphasis is placed on the importance of developing specialized, domain-restricted chatbot systems designed exclusively for medical use, given that the clinical context demands high reliability, interpretability, and alignment with medical knowledge. Unlike general-purpose language models, systems such as MedGPT show potential to be better suited to clinical needs, having been trained on large volumes of medical data and designed to process temporally organized health information from electronic health records [13]. Although MedGPT is not a direct clinical chatbot for everyday use, its concept represents an important step toward the development of more reliable AI assistants that could, in the future, be integrated into laboratory and clinical information systems [14].

The aim of this study is to determine the degree of agreement in the assessment of laboratory findings between an artificial intelligence system (MedGPT) and physicians, both in the absence of stated clinical symptoms and in the presence of stated clinical symptoms.

Material and Methods

The study was designed as a controlled, simulation-based comparative study, aimed at evaluating the degree of agreement and differences in the interpretation of laboratory findings between an AI system and physicians. The analysis was performed on a controlled set of simulated laboratory findings, both with and without added clinical symptoms. Laboratory parameters were selected to encompass commonly used biochemical and hematological indicators routinely employed in clinical practice and carrying diagnostic significance.

Each of the 30 laboratory sets, in both variants (with and without symptoms), was subjected to independent assessment by the artificial intelligence system MedGPT (customized version of ChatGPT configured for the interpretation of medical and laboratory information) and by a physician with clinical experience and who is employed at the University Clinical Center of the Republic of Srpska.

Both the AI system and the physician assigned each set one of five predefined categorical interpretation ratings:

1. Normal findings: all parameters within reference range values, with no signs of disease.
2. Mild deviation: minimal deviations without apparent clinical significance.
3. Suspected disease: abnormalities present that require additional analyses.
4. Confirmed disease (pathological condition): the finding indicates a specific disorder or disease.
5. Multiple abnormalities (comorbidity or syndrome): several interrelated disorders.

Statistical analysis was performed using IBM SPSS Statistics version 26. The level of statistical significance was set at $\alpha = 0.05$. The degree of agreement between individual assessments was analyzed using Cohen's κ coefficient. This coefficient was used to evaluate the consistency of AI system ratings under conditions without symptoms and with symptoms, as well as to assess the consistency of physician ratings under the same conditions. Additionally, Cohen's κ was applied to examine the agreement between ratings provided by the AI system and those provided by the physician, separately for laboratory findings without clinical symptoms and for findings with stated clinical symptoms. Cohen's κ values were interpreted in accordance with standard criteria, distinguishing levels of poor, moderate, good, and very good agreement. The Wilcoxon signed-rank test was used to examine statistically significant differences in ratings depending on the presence of clinical symptoms.

Results and Discussion

The agreement between AI ratings for laboratory findings without symptoms and with symptoms was moderate, with a statistically significant difference observed between the two assessments ($p = 0.001$). These results indicate that the presence of clinical symptoms significantly affected the manner in which the AI system interpreted the same laboratory findings. The Wilcoxon signed-rank test demonstrated a statistically significant difference between AI ratings without symptoms and with symptoms ($Z = 2.652$; $p = 0.008$), further confirming that clinical symptoms led to a systematic shift in the rating of laboratory findings (Table 1).

Table 1. Agreement of artificial intelligence assessments with and without symptoms

Measure of agreement	Value	Asymptotic standard error	Approximate T-value	p-value
Cohen's kappa coefficient	0.327	0.122	3.337	0.001

Table 2 presents the results of the agreement analysis for physician ratings of laboratory findings with and without clinical symptoms. The agreement between physician ratings for laboratory findings without symptoms and with symptoms was moderate and statistically significant ($p = 0.001$). The Wilcoxon signed-rank test likewise demonstrated a statistically significant difference between physician ratings under conditions without symptoms and with symptoms ($Z = 2.496$; $p = 0.013$). These

results indicate that the presence of clinical symptoms significantly influenced the physician's final interpretation of laboratory findings (Table 2).

Table 2. Agreement of physicians' assessments with and without symptoms

Measure of agreement	Value	Asymptotic standard error	Approximate T-value	p-value
Cohen's kappa coefficient	0.365	0.142	3.262	0.001

Table 3 presents the results of the agreement analysis between AI system ratings and physician ratings for laboratory findings without clinical symptoms. The analysis of agreement between AI system and physician ratings for laboratory findings without stated clinical symptoms demonstrated a high degree of concordance, with this agreement reaching statistical significance ($p < 0.001$). The Wilcoxon signed-rank test did not reveal a statistically significant difference between AI system and physician ratings under conditions without symptoms ($Z = 0.000$; $p = 1.000$), indicating that ratings were identical or very similar in the majority of cases (Table 3).

Table 3. Agreement of artificial intelligence and physicians' assessments without symptoms

Measure of agreement	Value	Asymptotic standard error	Approximate T-value	p-value
Cohen's kappa coefficient	0.745	0.113	6.541	< 0.001

The agreement between AI system and physician ratings for laboratory findings with stated clinical symptoms was moderate, with a statistically significant association between ratings ($p < 0.001$). The Wilcoxon signed-rank test did not reveal a statistically significant difference between AI system and physician ratings under conditions with symptoms ($Z = 1.134$; $p = 0.257$), although a somewhat greater degree of divergence was observed compared to the condition without symptoms (Table 4).

Table 4. Agreement of artificial intelligence and physicians' assessments with symptoms

Measure of agreement	Value	Asymptotic standard error	Approximate T-value	p-value
Cohen's kappa coefficient	0.536	0.132	4.716	< 0.001

The results of this study indicate a high degree of concordance between the interpretation of laboratory findings provided by the MedGPT AI system and that of the physician, particularly in situations where laboratory findings were not accompanied by clinical symptoms. Under such conditions, good to very good agreement was recorded between ratings, with no statistically significant differences between the two approaches. This result suggests that, in the domain of interpreting isolated laboratory data, MedGPT is capable of generating conclusions comparable to physician clinical reasoning, and this discovery is of particular relevance in the context of the post-analytical phase of laboratory medicine.

These results are consistent with the findings of previous studies examining the capabilities of large language models in controlled clinical scenarios. D'Urso et al. [15] demonstrated that AI systems achieve a high level of accuracy in the interpretation of laboratory findings when input data are clearly defined and limited to quantitative parameters, without complex clinical information. Similarly, Girton et al. [14] observed that AI-generated responses frequently exhibit greater structure and consistency compared to human interpretations, a pattern particularly pronounced in laboratory reports containing a larger number of parameters.

It is especially significant that no statistically significant difference was observed between MedGPT and physician ratings for laboratory findings without symptoms. This further supports the thesis that AI systems may play an important role in the standardized interpretation of laboratory data, especially in situations where clinical context is unavailable or minimal. In practice, such scenarios are common in laboratory medicine, where laboratory specialists and clinicians frequently receive limited patient information, a circumstance that underscores the potential utility of AI assistants under such conditions.

On the other hand, the results of this study clearly demonstrate that the presence of clinical symptoms significantly influences the process of interpreting laboratory findings, both for physicians and for the AI system. Both MedGPT and the physician were observed to modify their ratings when symptoms were included, with the degree of agreement between ratings without and with symptoms being moderate. This indicates that clinical context constitutes a key component of the interpretation process.

Although this study demonstrated MedGPT's sensitivity to the presence of symptoms, the agreement between MedGPT and the physician was lower under conditions where symptoms were included compared to conditions without symptoms. Nevertheless, no statistically significant differences in ratings were recorded in this case either, suggesting that, despite greater variability, MedGPT remains relatively aligned with physician judgment. This result is consistent with the discoveries of Meyer et al. [16], who demonstrated that chatbot systems, in more complex clinical scenarios, more frequently produce generalized or conservative assessments, in an apparent effort to avoid potentially hazardous conclusions.

Such behavior on the part of AI systems may be interpreted as a consequence of the manner in which large language models are trained, given that they do not possess genuine clinical reasoning but instead rely on statistical patterns derived from data. In this context, clinical symptoms, which frequently require nuanced interpretation and assessment of their relative significance, represent a considerable challenge for AI systems. This aspect was further highlighted by Gvajaia et al. [17], who demonstrated that the inclusion of broader clinical context improves the diagnostic accuracy of GPT-4, while differences persist in the way AI systems and physicians weight individual pieces of information.

The discoveries of Yan et al. [18] provide additional explanation for these results, indicating that large language models are sensitive to changes in key medical data, but that this sensitivity is not always consistent and may vary depending on the formulation of the input information. This may explain why, in the present study, agreement between MedGPT and the physician decreases in the presence of symptoms, even though the final ratings do not show statistically significant differences.

The systematic review and meta-analysis by Shan et al. [19] further confirm that, although large language models (LLMs) demonstrate significant diagnostic potential, their overall accuracy and clinical reliability have, on average, yet to reach the level of clinical professionals, particularly in complex and atypical clinical situations. The results of the present study are consistent with these conclusions, indicating that MedGPT can serve as a highly useful tool in certain domains, but cannot fully replace physician clinical reasoning.

The results of this study indicate that AI systems such as MedGPT hold significant potential as support tools for the interpretation of laboratory findings, particularly in situations requiring rapid, consistent, and standardized data interpretation. Their application may contribute to reducing the workload of physicians and laboratory specialists, as well as improving the efficiency of clinical workflows. However, in line with the recommendations of numerous authors, including D'Urso et al. [15] and the EFLM Working Group [10], these systems should not be used as a substitute for clinical reasoning, but exclusively as a complement to physician decision-making. The greatest value of MedGPT lies in the domain of interpreting isolated laboratory findings and supporting the clinical process, while the integration of symptoms and broader clinical context remains an area in which physician experience, clinical intuition, and professional accountability continue to play an indispensable role.

Conclusion

A high degree of agreement was established in the interpretation of laboratory findings provided by the MedGPT AI system under conditions without stated clinical symptoms, while under conditions with symptoms this agreement decreased to a moderate level. Clinical symptoms play a significant role in the process of interpreting laboratory results, both for physicians and for the AI system, with their influence being more pronounced in the case of physicians. The research results confirm that MedGPT can serve as a useful support tool for the interpretation of laboratory findings, but not as a substitute for physician clinical reasoning.

Literature

[1] Paranjape K, Schinkel M, Hammer RD, Schouten B, Nannan Panday RS, Elbers PWG, et al. The Value of Artificial Intelligence in Laboratory Medicine. Am J Clin Pathol. 2021;155(6):823-31. doi: 10.1093/ajcp/aqaa170.

[2] Zhu X, Bratanow NC. Physician Perspective and Challenges Ordering/Interpreting Laboratory Tests Related to the Opioid Crisis in America. Ther Drug Monit. 2021;43(1):6-13. doi: 10.1097/FTD.0000000000000840.

[3] Hickner J, Thompson PJ, Wilkinson T, Epner P, Sheehan M, Pollock AM, et al. Primary care physicians' challenges in ordering clinical laboratory tests and interpreting results. J Am Board Fam Med. 2014;27(2):268-74. doi: 10.3122/jabfm.2014.02.130104.

[4] ALQahtani DA, Rotgans JI, Mamede S, ALAlwan I, Magzoub ME, Altayeb FM, et al. Does Time Pressure Have a Negative Effect on Diagnostic Accuracy? Acad Med. 2016;91(5):710-6. doi: 10.1097/ACM.0000000000001098.

[5] Xie H, Jia Y, Liu S. Integration of artificial intelligence in clinical laboratory medicine: Advancements and challenges. Interdisciplinary Medicine. 2024;2(3):1-13. doi:10.1002/INMD.20230056.

[6] Dodig S, Čepelak I, Dodig M. Are we ready to integrate advanced artificial intelligence models in clinical laboratory? Biochem Med (Zagreb). 2025;35(1):010501. doi: 10.11613/BM.2025.010501.

[7] Lippi G, Plebani M. Lights and shadows of artificial intelligence in laboratory medicine. Adv Lab Med. 2025;6(1):1-3. doi: 10.1515/almed-2025-0024.

[8] Pennestrì F, Banfi G. Artificial intelligence in laboratory medicine: fundamental ethical issues and normative key-points. Clin Chem Lab Med. 2022;60(12):1867-74. doi: 10.1515/cclm-2022-0096.

[9] Maity S, Saikia MJ. Large Language Models in Healthcare and Medical Applications: A Review. Bioengineering (Basel). 2025;12(6):631. doi: 10.3390/bioengineering12060631.

[10] Cadamuro J, Carobene A, Cabitza F, Debeljak Z, De Bruyne S, van Doorn W, et al.; European Federation of Clinical Chemistry and Laboratory Medicine Working Group on Artificial Intelligence. A comprehensive survey of artificial intelligence adoption in European laboratory medicine: current utilization and prospects. Clin Chem Lab Med. 2024;63(4):692-703. doi: 10.1515/cclm-2024-1016.

[11] Prégent J, Chung VH, El Adib I, Désilets M, Hudon A. Applications of Artificial Intelligence in Psychiatry and Psychology Education: Scoping Review. JMIR Med Educ. 2025;11:e75238. doi: 10.2196/75238.

[12] Cabral BP, Braga LAM, Conte Filho CG, Penteado B, Freire de Castro Silva SL, Castro L, et al. Future Use of AI in Diagnostic Medicine: 2-Wave Cross-Sectional Survey Study. J Med Internet Res. 2025;27:e53892. doi: 10.2196/53892.

[13] Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artif Intell Med. 2021;117:102083. doi: 10.1016/j.artmed.2021.102083.

[14] Girton MR, Greene DN, Messerlian G, Keren DF, Yu M. ChatGPT vs Medical Professional: Analyzing Responses to Laboratory Medicine Questions on Social Media. Clin Chem. 2024;70(9):1122-39. doi: 10.1093/clinchem/hvae093.

[15] D'Urso F, Paladini F, Pollini M, Broccolo F. Use of Artificial Intelligence Chatbots in Interpretation of Clinical Chemistry and Laboratory Medicine Reports: A Standardized Approach. Applied Sciences. 2025; 15(8):4232. doi: 10.3390/app15084232

[16] Meyer A, Soleman A, Riese J, Streichert T. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. Clin Chem Lab Med. 2024;62(12):2425-2434. doi: 10.1515/cclm-2024-0246.

[17] Gvajaia N, Kutchava L, Alavidze L, Yandamuri SP, Pestvenidze E, Kupradze V. Artificial intelligence in the medical field: diagnostic capabilities of GPT-4 in comparison with physicians. Journal of Medical Artificial Intelligence.2025;8:1-15. doi: 10.21037/jma
[18] Yan C, Fu X, Xiong Y, Wang T, Hui SC, Wu J, et al. LLM sensitivity evaluation framework for clinical diagnosis. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Di Eugenio B, Schockaert S, editors. Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi (UAE): Association for Computational Linguistics; 2025;3083-94.
[19] Shan G, Chen X, Wang C, Liu L, Gu Y, Jiang H, Shi T. Comparing Diagnostic Accuracy of Clinical Professionals and Large Language Models: Systematic Review and Meta-Analysis. JMIR Med Inform. 2025;13:e64963. doi: 10.2196/64963.

PRIMJENA VJEŠTAČKE INTELIGENCIJE U INTERPRETACIJI LABORATORIJSKIH NALAZA

Đorđe Stojisavljević*[1], Darija Knežević[2], Duška Jović[2], Igor Paravac[2], Ljubisa Preradović[3]

[1]Univerzitetski računarski centar, Bulevar vojvode Petra Bojovića 1A, 78000, Banja Luka, Republika Srpska;

[2]Medicinski fakultet Univerziteta u Banjoj Luci, Save Mrkalja 19 Banja Luka, Republika Srpska,

[3]Arhitektonsko-građevinsko-geodetski fakultet, Bulevar vojvode Petra Bojovića 1A, 78000, Banja Luka, Republika Srpska

Sažetak. *Interpretacija laboratorijskih rezultata predstavlja jedan od ključnih segmenata kliničkog odlučivanja i predstavlja dio svakodnevne medicinske prakse. Laboratorijska medicina obezbjeđuje objektivne, kvantitativne i kvalitativne podatke o fiziološkom i patološkom stanju pacijenta, ali ti podaci sami po sebi ne predstavljaju dijagnozu, niti imaju pun klinički značaj bez adekvatne medicinske interpretacije. Ovaj rad analizira stepen saglasnosti u interpretaciji laboratorijskih nalaza između sistema vještačke inteligencije MedGPT i ljekara u uslovima sa i bez navedenih kliničkih simptoma. Istraživanje je osmišljeno kao opservaciono i komparativno. Laboratorijski parametri su izabrani tako da obuhvate uobičajene biohemijske i hematološke pokazatelje koji se rutinski koriste u kliničkoj praksi i imaju dijagnostički značaj. Analiza je izvedena na kontrolisanom setu simuliranih laboratorijskih nalaza, sa i bez dodatih kliničkih simptoma. Rezultati pokazuju visok nivo podudarnosti između MedGPT i ljekara pri interpretaciji izolovanih laboratorijskih podataka, bez statistički značajnih razlika u ocjenama. Saglasnost između ocjena vještačke inteligencije za laboratorijske nalaze bez simptoma i sa simptomima bila je umjerena, pri čemu je uočena statistički značajna razlika između ove dvije procjene ($p = 0,001$). Saglasnost između ocjena koje je dao ljekar za laboratorijske nalaze bez simptoma i sa simptomima bila je umjerena i statistički značajna ($p = 0,001$). Između ocjena vještačke inteligencije i ljekara za nalaze bez kliničkih simptoma utvrđena je visoka i statistički značajna saglasnost ($p < 0,001$). Ovi rezultati ukazuju da je prisustvo kliničkih simptoma značajno uticalo na način na koji je sistem vještačke inteligencije interpretirao iste laboratorijske nalaze. Nalazi ukazuju da sistemi vještačke inteligencije mogu predstavljati koristan alat za podršku u postanalitičkoj fazi*

laboratorijske medicine, uz neophodno zadržavanje centralne uloge ljekarskog kliničkog rasuđivanja.

***Ključne riječi**: vještačka inteligencija, laboratorijska medicina, interpretacija laboratorijskih nalaza.*