The Pulse of Artificial Intelligence in Cardiology: A Comprehensive Evaluation of State-of-the-art Large Language Models for Potential Use in Clinical Cardiology

Andrej Novak[1,2,3], Fran Rode[1], Ante Lisičić[1*], Iskra A. Nola[4], Ivan Zeljković[1], Nikola Pavlović[1], Šime Manola[1]

[1] Dubrava University Hospital, Avenija Gojka Šuška 6, 10 000 Zagreb, Croatia
[2] Department of Physics, Faculty of Science, University of Zagreb, Bijenička c. 32, 10000 Zagreb, Croatia
[3] Luxembourg School of Business, 46 Côte d'Eich, 1450 Luxembourg
[4] School of Medicine, University of Zagreb, Andrija Štampar School of Public Health, 10000 Zagreb, Croatia
[*]A. Lisičić is the corresponding author.

Abstract

Introduction: During the last few years, we have witnessed a surge in the utilization of Large Language Models (LLMs) for diverse applications in clinical medicine. Their utility extends to enhancing ECG interpretation, data analysis, and risk prediction in cardiology. This study aims to evaluate the accuracy of LLMs in answering cardiology-specific questions of various difficulty levels.

Methods: This study undertakes a comparative analysis of three state-of-the-art LLMs: Google Bard, GPT-3.5 Turbo, and GPT-4.0, against four distinct sets of clinical scenarios with increasing complexity. These scenarios cover a range of cardiovascular topics, from prevention to the management of acute illnesses and complex pathologies. The responses generated by the LLMs were assessed for accuracy, understanding of medical terminology, clinical relevance, and appropriateness. The evaluations were conducted by a panel of experienced cardiologists.

Results: All models showed an understanding of medical terminology, but the application of this knowledge varied. GPT-4.0 outperforms Google Bard and GPT-3.5 Turbo across a spectrum of cardiology-related clinical scenarios, demonstrating a strong understanding of medical terminology, contextual understanding, and most proficiently aligning its responses with current guidelines. Limitations were seen in the models' abilities to reference ongoing clinical trials.

Conclusion: LLMs showed promising results in ability to interpret and apply complex clinical guidelines when answering vignette-based clinical queries, with a potential for enhancing patient outcomes through personalized advice. However, they should be utilized with a grain of salt, as supplementary tools in clinical cardiology.

Introduction

Over the last several years, we have witnessed a surge in the utilization of Large Language Models (LLMs) for diverse applications, ranging from basic engineering problems to complex issues in medical research and its applications. LLMs are artificial neural networks typically built with a transformer-based architecture that autonomously learn from data and can produce sophisticated and seemingly intelligent writing after being trained on a massive dataset (1). These models, such as Google Bard, GPT-3.5 Turbo, and GPT-4.0, can be repurposed with minimal effort across numerous domains and diverse tasks (2). Natural language processing powered by pre-trained language models is the key technology behind medical artificial intelligence (AI) systems that utilize clinical narratives (3).

There are numerous potential applications of LLMs in medicine: support in clinical decision-making, knowledge retrieval, summarizing key diagnostic findings, triaging patients' primary care concerns, enhancing patient health literacy, and more (2). They have great potential to modernize academic research by accelerating data analysis, literature reviews, and referencing (4-6).

In a hospital environment, they have the potential to serve as a medium for repetitive tasks, such as assistance with writing discharge letters by summarizing a patient's hospital stay after reading their medical records or providing further recommendations (3,7). LLMs can quickly analyze huge sets of data including electronic health records, or, for example, genomic data (8). This allows them to verify patterns not apparent to humans, identify potential risk factors, and even provide outcome predictions (9).

On the other hand, although these models demonstrate proficiency in generating linguistically precise content, they do not have the capability to comprehend or internalize knowledge of the world in a manner akin to human cognition. Clinical decision-making requires the synthesis of evidence-based medicine, guidelines, and sound clinical judgment (10). Not to be neglected is intuition, a process based on knowledge and care experience (11,12). The rapid evolution of LLMs and the burgeoning variety of available models warrant systematic comparative analyses or benchmarking (13-16). Consensus indicates that LLMs possess the capability to aid in clinical case resolution; however, their application necessitates caution due to inherent imprecision and a propensity for disseminating misinformation (17,18). Overseeing this AI's impending incorporation into daily medical practice is essential to ensure the effective management of knowledge sources and their application.

A notable achievement of GPT has been its performance on the United States Medical Licensing Examination (USMLE), where it attained scores at or near the passing threshold for all three components of the exam (19). This study assessed GPT's clinical reasoning capabilities through its responses to the standardized questions of the USMLE, which emulate aspects of clinical decision-making. In subsequent research (20), GPT was evaluated using the American Heart Association (AHA) Basic Life Support (BLS) and Advanced Cardiovascular Life Support (ACLS) examinations. Although it initially fell short of the passing threshold, GPT's responses were more

pertinent and aligned with established guidelines compared to other AI systems, also offering reasoned explanations for its answers. Another study explored GPT's proficiency in responding to queries from the Ophthalmic Knowledge Assessment Program (OKAP) exam (21). The model demonstrated significant accuracy in two parallel exams, although its performance varied across different ophthalmic subspecialties. It excelled in general ophthalmology while showing modest results in specialized areas such as neuro-ophthalmology and intraocular tumors. These findings, coupled with the heterogeneous results across various topics within the same specialty, have partly motivated the current study.

It's worth noting that a form of AI has been in use for some time in cardiology: electrocardiograms (ECG) are being interpreted by computers daily (22-24) and a study from 2019 (25) has shown that machine learning (ML) algorithms exceeded the diagnostic accuracy of a team of practicing cardiologists in the classification of 12 heart rhythm types from single-lead ECGs. With the increasing complexity of medical data and new knowledge, ML can help with summarizing the key features of clinical data into simpler formats through a process called feature extraction (11,23,26).

ML is already being used to improve upon traditional risk prediction algorithms using available registry data. A study from 2020 demonstrated that ML integrating clinical parameters with coronary artery calcium and automated epicardial adipose tissue quantification significantly improved the prediction of myocardial infarction and cardiac death compared with standard clinical risk assessment (23). A study from February 2023 (27) evaluated the appropriateness of ChatGPT responses to questions regarding the prevention of cardiovascular diseases. The questions were crafted to mirror typical inquiries likely to be raised by individuals without a healthcare background. The majority (84%) of answers provided by ChatGPT were graded as appropriate. The study suggests the potential of education (improved patient health literacy) and counselling regarding the basics of cardiovascular prevention that LLMs could provide to patients. Another recent research highlighted ChatGPT’s potential as an assisted decision support tool for more straightforward clinical questions, but less for questions where general practitioners needed the help of cardiologists in decision-making (27,28).

Finally, let us mention a recent study (29) in which the performance of GPT-3.5 and GPT-4, was evaluated on the Polish Medical Final Examination (MFE) in English and Polish. GPT-4 outperformed GPT-3.5, achieving mean accuracies of 79.7% in both languages and passing all versions of the MFE. In contrast, GPT-3.5 showed lower accuracies and was less consistent in passing the exams. Despite GPT-4's superior performance, its scores were mostly lower than the average scores of medical students, and a significant correlation was observed between the correctness of the answers and the index of difficulty for both models.

The primary aim of this study was to evaluate the clinical utility of three state-of-the-art LLMs - Google Bard, GPT-3.5 Turbo, and GPT-4.0 - by comparing their performance on a series of cardiology-based clinical vignettes. These vignettes were designed to represent a spectrum of

scenarios, from basic cardiovascular care to complex clinical decision-making. Our objective was to assess each model's accuracy, understanding of medical terminology, ability to provide clinically relevant and contextually appropriate responses, and adherence to guidelines. The overarching goal was to provide insights into the readiness of LLMs for application in real-world medical practice. To the best of our knowledge, the applications of LLMs in this context—spanning both vertical (across various levels of difficulty) and horizontal (comparing different LLMs on identical question sets) dimensions, have not been previously explored.

Methods

In this cross-sectional study, we aimed to make a comparative analysis of three state-of-the-art LLMs: Google Bard (based on PaLM 2), GPT-3.5 Turbo, and GPT-4.0 (models 0613) available as text-based chat interfaces provided by OpenAI and Google during July 2023. These models are assessed based on four sets of vignette-based clinical scenarios, differentiated by their levels of complexity, and denoted as test sets A, B, C, and D, respectively. A significant portion of the questions featured five potential responses, among which only one was correct. However, a minor fraction of the queries deviated from this format, presenting three, four, six, or seven potential answers instead, yet still with only one correct answer.

The test set A is composed of 24 queries from the ANA exam (American Nurses Association), encompassing various clinical situations pertinent to fundamental cardiovascular patient care. Test set B includes 25 inquiries from the ACO exam (American College of Osteopathic Internists), encompassing an assortment of clinical circumstances and management strategies in cardiology centering around emergency states and coronary artery disease, along with scoring systems such as the Syntax Score. The test set C is a custom dataset formulated by the study authors, containing 60 questions that mirror the format and objectives of the USMLE Step 2 and 3 exams. Specifically, the set includes questions on general cardiology (18 questions), ischemic heart disease (12 questions), heart failure (9 questions), arrhythmia (12 questions), and a combination of questions dealing with valvular, endocardial, myocardial, and pericardial diseases (9 questions). Lastly, test set D incorporates 90 vignettes, mainly based on current guidelines uniformly sampled from the chapters of established board review reference books (30,31). Test sets B, C, and D feature various patient demographics. Questions exclusively pertaining to images data were excluded, and ECG or cardiac imaging graphical data supporting the vignette were interpreted following a standard format and terminology. Each question sets details together with the results given in Table 1.

We utilized the text-based chat interfaces provided by OpenAI and Google for our inquiries and data collection. For the mobile application 'Dubravka,'[1] developed during this research, we utilized the Application Programming Interface (API) provided by OpenAI to obtain responses from GPT-3.5 Turbo. GPT models were trained on the data until September 2021. The analysis was performed

[1] Dubravka, mobile application source code https://github.com/apptonio/dubravka

with the default temperature parameters set to 1 (on a scale with a minimum of 0 and a maximum of 2 in the OpenAI web application Playground) and 0.7 for Google Bard. The *top p* parameter was always set to 1 (default), as altering both parameters is not recommended[2].

The temperature parameter influences the randomness of the generated text (29). When set to a higher value, LLMs produce more diverse outputs by assigning equal probabilities to a broader range of words or tokens, leading to less predictable and more exploratory outputs. Conversely, a lower temperature setting makes the model more deterministic, favoring higher probability tokens and yielding more coherent, conservative outputs. Variations in the temperature parameter suggest that, within this range (min-to-max), the value primarily affects the overall diversity of the response output rather than the representation of medical knowledge encoded in the model (29). Prompts sent through the models consisted of the exact questions from the test sets, without additional comments or context. Each model's final response was obtained and saved. Final answers from highlighted prompts were stored in Appendix 1, and all questions and answers from test set C were included in Appendix 2.

The accuracy of all three models for each test was calculated by dividing the number of correct answers by the overall number of questions in that test set. Incomplete answers or rare instances where the model yielded two answers as correct were treated as incorrect. The accuracy of the models' responses was analyzed using a chi-square test to assess the likelihood of random occurrences in their performances. An omnibus Chi-squared test was applied to the results obtained from all three LLMs on the test sets, with the significance level set at 0.05. Mobile application 'Dubravka' was programmed in Flutter, an open-source user interface software development kit. Additionally, Python 3, a high-level, general-purpose programming language, was employed for performing calculations and statistical inference through the computing platform Jupyter Notebook 6.4.12[3].

The primary goal of this study was to appraise the clinical relevance of the varying responses generated by the LLMs in response to inquiries seeking medical advice across the clinical continuum, including cardiovascular prevention, general cardiology, heart failure, arrhythmia, acute cardiovascular care management, and finally valvular, myocardial, endocardial and pericardial pathologies with a strong emphasis towards decision-making for complex clinical scenarios. The outputs generated from the LLMs were analyzed by a panel of four cardiologists with over ten years of professional experience in the field.

---

[2] Parameter settings and recommendations could be find on: https://platform.openai.com/docs/api-reference/chat/create.
[3] https://jupyter.org/

Results

Our findings indicate that GPT-4.0 delivered superior performance across all four sets of questions. Moreover, this model exhibited proficiency in integrating complex clinical scenarios, adhering to current cardiology guidelines, computing and interpreting various scores and indices. Conversely, the GPT-3.5 Turbo model yielded comparatively modest results, on several occasions resorting to speculative answers. In contrast, Google Bard underperformed significantly, frequently evading to give a direct answer to the given question. Nonetheless, all three models demonstrated a capacity to provide responses perceived as correct, supplemented by explanatory commentary discounting alternative answers. Detailed performance data for each model across the different sets of questions are presented in Figure 1 and detailed in Table 1.

HERE FIGURE 1.

Table 1: Comparative performance of Google Bard, GPT 3.5 Turbo, and GPT 4.0 on four sets of questions with varied complexity in cardiology.

Test	Level of difficulty	Google Bard	GPT-3.5 Turbo	GPT-4.0
A – ANA	Basic (Cardiac nurse)	19/24	20/24	22/24
B - ACO	Mixed	15/25	16/25	22/25
C - Custom	USMLE-like Step 2/3 Question	30/60	40/60	48/60
D - Difficult	Board reviews (Guidelines)	50/90	51/90	76/90

Note: Each entry denotes the number of correctly answered questions out of the total questions in each respective set. ANA - American Nurses Association; ACO - American College of Osteopathic Internists, USMLE - United States Medical Licensing Examination.

Accuracy. The performance statistics for each model are as follows: Test set A (Basic level): GPT-4.0 achieved a 92% success rate, with Google Bard and GPT-3.5 Turbo scoring 79% and 83% respectively. Test set B (Mixed level): GPT-4.0 had an 88% success rate, higher than the 60% and 64% rates of Google Bard and GPT-3.5 Turbo. Test set C (USMLE-like): GPT-4.0 achieved an 80% success rate, compared to Google Bard's 50% and GPT-3.5 Turbo's 67%. Test set D (Difficult level): GPT-4.0 outperformed others with an 84% success rate, while Google Bard and GPT-3.5 Turbo achieved 55% and 57% success rates.

Understanding of Medical Terminology. Terms related to cardiology pertaining to human anatomy, medical diagnoses, generic drug names, diagnostic tests, and laboratory parameters were correctly identified by all models. GPT-4.0 was the most accurate and proficient, employing medical terminologies appropriately and aligning its answers with current clinical guidelines. GPT-3.5 Turbo demonstrated a good understanding of the terminology, providing a thorough explanation of each treatment option. However, the application of these terms in the context of the

current clinical guidelines was flawed. On the other hand, Google Bard showed gaps in understanding of terms, consequently making several inaccurate statements and recommendations (please refer to Table 2 and Table 4).

Clinical Relevance and Contextual Understanding. GPT-4.0 demonstrated a strong understanding of the clinical context and delivered a relevant response that accurately reflected current AHA/ACC medical guidelines for the management of blood cholesterol (2018 AHA/ACC Guideline on the Management of Blood Cholesterol). In contrast, GPT-3.5 Turbo and Google Bard sometimes strayed from these guidelines (please refer to Table 3, Table 5, and Table 6).

Statistical Analysis. An omnibus Chi-squared test applied to all three LLMs for test sets A ($p = 0.472$) and B ($p = 0.063$) was not statistically significant. In contrast, results on test set C ($p = 0.003$) and set D ($p < 0.001$) were statistically significant. Furthermore, the pairwise testing revealed the statistically significant superiority of GPT-4 over Google Bard on the test set C ($p = 0.001$) while the difference was not significant compared to GPT-3.5 Turbo ($p=0.144$). On the test set D, GPT-4.0 results were statistically significantly better as compared to both GPT-3.5 Turbo ($p < 0.001$) and Google Bard ($p < 0.001$).

Discussion

In the following sections, we will highlight five vignettes that demonstrate specific characteristics of each LLM. We refer the reader to Appendix 1 for further selected vignettes and their interpretations, and to Appendix 2 for a detailed exposition of the models' answers to test set C. Notably, both GPT-4.0, GPT-3.5 Turbo, and Google Bard yield outputs that manifest a high level of comprehensibility, demonstrating the potential utility of these models in relaying complex medical information to diverse audiences.

Model Performance and Proficiency in Medical Terminology. The models' responses are coherent, structured, and detailed, effectively translating the complex nuances of medical jargon into more accessible language. Even though GPT-4 exhibited superior performance across all tests, the limited number of questions in sets A and B could have diminished the statistical power, resulting in a lack of statistical significance. Critically, each model demonstrates an ability to discern important clinical details from the input information, process this information in the context of their underlying medical knowledge, and generate appropriate responses. All models commented on the answers offered with an explanation of why, according to their judgment, a certain answer was correct and others incorrect. For example, as shown in Table 2, GPT-4.0 and GPT-3.5 Turbo presented an exceptional understanding of cardiological terminology. They accurately interpreted the provided ECG and echocardiography findings and deduced the correct diagnosis, consequently suggesting adequate procedures. In contrast to GPT-4.0 and GPT-3.5 Turbo, Google Bard struggled in this field, and while it recognized the reduced ejection fraction

and mitral insufficiency, it neglected the wide QRS complex and the importance of cardiac resynchronization therapy.

HERE TABLE 2.

Variability in Clinical Relevance and Contextual Understanding. The study's examination of clinical relevance and contextual understanding in the responses of GPT-4.0, GPT-3.5 Turbo, and Google Bard revealed distinct variations in their ability to apply medical guidelines accurately. For example, in the vignette dealing with lipid management (Table 3), GTP-4 recommendations to add ezetimibe to the high-intensity statin therapy, owing to the patient's slightly elevated LDL-C level, reflects its proficiency in synthesizing the patient's clinical data and applying it in a clinically relevant manner. Furthermore, this model demonstrated strong contextual understanding, precisely integrating the patient's past medical history, current symptoms, and lipid profile results in its answer. In contrast, while GPT-3.5 Turbo showed a good conceptual understanding by referencing the patient's history, current statin therapy, and the lipid profile in its explanation and explaining each treatment option clearly, it failed to align its final recommendation with current medical guidelines. Google Bard demonstrated some ability to understand the context. In the same example, it accurately recognized that the patient had an acute myocardial infarction and that his LDL-C was above target, but it inaccurately categorized atorvastatin as a moderate-intensity statin and recommended switching to evolocumab, a change in treatment that is not supported by current medical guidelines. Furthermore, in the vignette presented in Table 4, when asked to recommend the necessity for urgent hospital admission according to presented clinical findings, it is interesting to notice that, in addition to giving the incorrect answer, Google Bard provides contradicting arguments that lack clarity and put in question its contextual understanding of a given clinical scenario.

HERE TABLE 3.

HERE TABLE 4.

Finally, all three models occasionally showed a tendency for overtreatment, skipping the steps in guideline-oriented procedures. Each model sometimes turned towards a “safer” option, proposing more advanced diagnostic or therapeutic approaches when they would be unnecessary or anticipatory. As shown in Table 5, GPT-4 managed to avoid the fallacy and suggested intravenous benzodiazepines for cocaine-induced chest pain, suggesting that phentolamine is not the first line of treatment, while GPT-3.5 Turbo and Google Bard chose immediate treatment with phentolamine.

HERE TABLE 5.

Reliability and Response Consistency Across Models. Among the three models, GPT-4.0 shows the highest level of reliability, with consistent application of clinical guidelines and consistency in generating responses. However, GPT-3.5 Turbo and Google Bard have been observed to exhibit fluctuation in their replies when presented with the same query at different times. Intriguingly,

when these models produce an incorrect answer, they frequently alter that answer when queried again later, but the revised response is often also incorrect. This was not observed when the initially generated response was correct, implying a degree of confidence in these decisions. Moreover, there were two questions where GPT-3.5 Turbo began by endorsing one response, only to transition towards an alternate response by the conclusion of the generated reply. Additionally, Google Bard also exhibited instances of internal contradiction within its responses (Table 4).

Validation with Clinical Trials and Guidelines. During this investigation, it was observed that GPT-4.0 rendered responses that adhered to the contemporary AHA/ASA guidelines, a representative example of which is showcased in Table 3. Furthermore, the model's aptness in mirroring the guidelines was found to be accurate within the context of the presented clinical scenario. However, it is important to note that this model's responses were predicated on the guidelines published until September 2021. Yet another example can be seen in Table 6, where both GPT-4 and Google Bard referenced the guidelines, with Google Bard mentioning the exact Expert Consensus Statement. They gave the correct recommendation to start beta-blockers for long QT interval, while GPT-3.5 Turbo suggested an implantable cardioverter defibrillator (ICD) implantation and neglected beta-blockers. Compared to GTP-4.0, Google Bard, despite referencing medical guidelines, exhibited inconsistent application relative to the presented clinical scenario. Finally, let us mention that none of the assessed models demonstrated the capacity to reference or consider ongoing clinical trials within their responses.

HERE TABLE 6.

Speed and Efficiency. Assessing the models' speed and efficiency is challenging due to variability in execution time based on user load. GPT-3.5 Turbo and GPT-4.0 continually stream their responses, providing incremental outputs and potentially reducing the perceived waiting time. In contrast, Google Bard holds back its outputs, delivering the complete text following a slight delay.

Adaptability to Different Sub-Domains. The robust performance of GPT-4.0, GPT-3.5 Turbo, and Google Bard in complex cardiological scenarios suggests their potential applicability to other specialties. Their responses demonstrate a capacity for synthesizing and presenting complex medical information concisely and accurately, equally valuable qualities across all medical sub-domains. However, one must also consider the breadth and depth of the model's training data. As their capabilities are inherently shaped by the corpus they were trained on, their efficacy across different medical domains may vary based on the representation of those domains in the original dataset. Therefore, while we may anticipate a degree of adaptability, evaluating their performance individually within each target sub-domain is advisable to ensure precision and validity (32-35).

HERE FIGURE 2.

Potential for Personalization. Personalization is a crucial attribute for LLMs, focusing on integrating and interpreting user-specific information into their responses (34,35). In this study, we provide ‘Dubravka’, a custom-made mobile application, that interfaces with the GPT-3.5 Turbo

API (Application Programming Interface). This application's primary function is to incorporate user-specific medical data to tailor the LLM's responses to individual health profiles. To achieve this, users are prompted to input basic medical information during their initial interaction with the application. This data is then stored locally on the phone and included as a hidden parameter in the conversation, instructing the LLM to moderate its responses specifically to each patient's medical history, drug therapy, and lifestyle elements. This setup allows the LLM to process individual health data and generate advice based on the user's prompt and the provided medical information, with a primary emphasis on diet, cardiovascular prevention, and lifestyle recommendations. We release this simple prototype in the hope that it will serve as a potential use case for the development of more sophisticated professional-level applications. Additionally, the complete source code is publicly available for those interested in further development or utilization of this framework. In Figure 2, we present the user interface of 'Dubravka' as it appears to a patient with recent deep vein thrombosis, along with her question about the consistent use of compressive stockings. In Appendix 3, we highlight 10 typical patient profiles and their inquiries about their medical conditions. The answers were obtained with the temperature parameter set to zero. Those synthetic patients have been proposed by the authors of this research, reflecting the typical patients from our practice in a high-volume tertiary center.

Limitations

This study, while comprehensive in its assessment of three distinct LLMs, through both horizontal and vertical evaluations, carries several limitations. Firstly, its cross-sectional (observational) design, employing a fixed set of questions with predetermined correct answers, may not fully capture the dynamic and multifaceted nature of clinical decision-making in real-world cardiology. Secondly, it is worth noting that the power of a statistical test, its capacity to detect a genuine effect when it exists, can be influenced by the sample size. This is particularly true for the initial two test sets, where the fewer number of questions might have reduced the sensitivity in detecting significant differences in model performance.

Additionally, the lack of clear and concise guidelines for setting the 'temperature' of LLMs is a notable limitation. The study's methodology, which relied on the knowledge of a clinical cardiologist rather than an IT expert, reflects a real-world scenario where clinicians are typically not versed in adjusting the LLMs' parameter settings, potentially leading to variations in user experience. However, current evidence supports the notion that temperature variations reflect the diversity in phrasing of the answer, rather than the representation of medical knowledge encoded in the model (29).

Another critical aspect concerns the financial accessibility of some LLMs. Our study relied on GPT-4, a paid model, suggesting that routine clinical use might be limited to medical staff who can afford it. This raises questions about equitable access to such advanced AI tools in diverse clinical settings.

The findings of this study are also contingent on the models' knowledge as of their last training update (as of September 2021). Considering the rapid evolution in AI capabilities and medical knowledge, the relevance and accuracy of these findings are subject to change over time.

Lastly, the study did not address the ethical implications and practical challenges of integrating AI in clinical settings, including concerns about patient privacy, data security, and the impact on the physician-patient relationship. While the models demonstrated adaptability, their performance in other medical fields may vary depending on the representation of those domains in their training datasets.

Conclusion

This contribution outlines the benchmarking of three different LLMs, GPT-4.0, GPT-3.5 Turbo, and Google Bard, in responding to complex clinical scenarios in cardiology. The models were evaluated on their understanding and use of medical terminology, clinical relevance, contextual understanding, accuracy, and others.

While all three models displayed considerable capacities in handling complex cardiological scenarios, GPT-4.0 emerged as the most reliable, accurate, and guideline-compliant among them. This capacity for contextual understanding and a clear evidence-based approach underscores its high applicability in a real-world clinical setting. Furthermore, the study also highlights the importance of validating these models across different medical domains and continually updating them to reflect the most recent medical guidelines and clinical trials.

As we find ourselves amidst an accelerating AI revolution, with its influence extending to all aspects of our lives, the need for active engagement, evaluation, and refinement of these technologies becomes ever more apparent. By embracing these advancements *cum grano salis*, we can optimally harness AI's potential to assist medical professionals, rather than replace them.

Acknowledgments

We express our gratitude to the anonymous reviewers whose insightful comments and suggestions have greatly enhanced the quality of this paper. We would also like to extend our gratitude to Antonio Butigan for his contribution to this study. His technical expertise and commitment to excellence were pivotal in the programming of the ‘Dubravka’ mobile application. We gratefully acknowledge the Luxembourg School of Business for providing the technological support for this research. This work was partially supported by the Croatian Science Foundation under project IP-2022-10-7261 (ADESO). Additionally, the authors would like to acknowledge the contribution of the COST Action CA21169 (DYNALIFE), supported by COST (European Cooperation in Science and Technology).

References

1. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature. 2023;614(7947):224–6.

2. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large Language Models Encode Clinical Knowledge. 2022;1–44.

3. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. npj Digit Med. 2022;5(1):1–9.

4. Marchandot B, Matsushita K, Carmona A, Trimaille A, Morel O. ChatGPT: The next frontier in academic writing for cardiologists or a pandora's box of ethical dilemmas. Eur Hear J Open. 2023;3(2):1–3.

5. Biswas S. ChatGPT and the Future of Medical Writing. Radiology. 2023 Apr; 307(2):e223312.

6. Akhter HM, Cooper JS. Acute Pulmonary Edema After Hyperbaric Oxygen Treatment: A Case Report Written With ChatGPT Assistance. Cureus. 2023; 15(2):1–5.

7. Arora A, Arora A. The promise of large language models in health care. Lancet (London, England). 2023;401(10377):641.

8. Javan R, Kim T, Mostaghni N, Sarin S. ChatGPT's Potential Role in Interventional Radiology. Cardiovasc Intervent Radiol. 2023; 821–822.

9. Mann DL. Artificial Intelligence Discusses the Role of Artificial Intelligence in Translational Medicine: A JACC: Basic to Translational Science Interview With ChatGPT. JACC Basic to Transl Sci. 2023; 8(2):221–3.

10. Sara JDS, Toya T, Rihal CS, Lerman LO, Lerman A. Clinical decision-making: Challenging traditional assumptions. International Journal of Cardiology. 2021;326:6–11.

11. Adlung L, Cohen Y, Mor U, Elinav E. Review Machine learning in clinical decision making. Med. 2021;2(6):642–65.

12. Melin-Johansson C, Palmqvist R, Rönnberg L. Clinical intuition in the nursing processand decision-making – A mixed studies review. J Clin Nurs. 2017; 26(23-24):3936-49.

13. Kortemeyer G. Could an Artificial-Intelligence agent pass an introductory physics course? PhysRevPhysEducRes. 2023; 19.010132.

14. Dunn P, Cianflone D. Artificial intelligence in cardiology: Exciting but handle with caution. International Journal of Cardiology. 2023 Jun 9.

15. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. British Journal of Clinical Pharmacology. 2024 Jan;90(1):232-8.

16. Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y., & Radev, D.R., Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. ArXiv. 2023; 2303.18027.

17. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. Does ChatGPT Provide Appropriate and Equitable Medical Advice?: A Vignette-Based, Clinical Evaluation Across Care Contexts. medRxiv. 2023 Jan 1; 2023.03.25.23285475.

18. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare. 2023 Mar 19;11(6):887.

19. Kung TH, Cheatham M, ChatGPT, Medenilla A, Sillos C, Leon L De, et al. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. medRxiv. 2022; 3786:2022.12.19.22283643.

20. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation. 2023;185.

21. Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, Renaud Duval, Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmology Science. Volume 3, Issue 4, 2023.

22. Commandeur F, Slomka PJ, Goeller M, Chen X, Cadet S, Razipour A, et al. Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: A prospective study. Cardiovasc Res. 2020;116(14):2216–25.

23. Quer G, Arnaout R, Henne M, Arnaout R. Machine Learning and the Future of Cardiovascular Care: JACC State-of-the-Art Review. J Am Coll Cardiol. 2021;77(3):300–13.

24. Siontis KC, Liu K, Bos JM, Attia ZI, Cohen-Shelly M, Arruda-Olson AM, Farahani NZ, Friedman PA, Noseworthy PA, Ackerman MJ. Detection of hypertrophic cardiomyopathy by an artificial intelligence electrocardiogram in children and adolescents. International Journal of Cardiology. 2021 Oct 1;340:42-7.

25. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65–9.

26. Kogan E, Didden EM, Lee E, Nnewihe A, Stamatiadis D, Mataraso S, Quinn D, Rosenberg D, Chehoud C, Bridges C. A machine learning approach to identifying patients with pulmonary hypertension using real-world electronic health records. International Journal of Cardiology. 2023 Mar 1;374:95-9.

27. Harskamp RE, De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of

common cardiac conditions (AMSTELHEART-2). medRxiv. 2023; 2023(February): 2023.03.25.23285475.

28. Benz DC, Benetos G, Rampidis G, von Felten E, Bakula A, Sustar A, et al. Validation of deep-learning image reconstruction for coronary computed tomography angiography: Impact on noise, image quality and diagnostic accuracy. J Cardiovasc Comput Tomogr. 2020;14(5):444–51.

29. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Scientific Reports. 2023 Nov 22;13(1):20512.

30. Gaggin HK, Januzzi JL. MGH Cardiology Board Review. Springer; 2nd ed. (2021)

31. Mukherjee D, Moliterno D, Cho L, Lange R, Chatterjee S. 1133 Questions: An Interventional Cardiology Board Review. Wolters Kluwer Health. (2018).

32. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, Sigler C, Knödler M, Keller U, Beule D, Keilholz U. Leveraging large language models for decision support in personalized oncology. JAMA Network Open. 2023 Nov 1;6(11).

33. Cheng SW, Chang CW, Chang WJ, Wang HW, Liang CS, Kishimoto T, Chang JP, Kuo JS, Su KP. The now and future of ChatGPT and GPT in psychiatry. Psychiatry and clinical neurosciences. 2023 Nov;77(11):592-6.

34. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Löffler CM, Schwarzkopf SC, Unger M, Veldhuizen GP, Wagner SJ. The future landscape of large language models in medicine. Communications Medicine. 2023 Oct 10;3(1):141.

35. Arslan S. Exploring the Potential of Chat GPT in Personalized Obesity Treatment. Annals of Biomedical Engineering. 2023 May 5:1-2.

Figure 1. The provided bar graph visually compares the performance of three different Large Language Models: Google Bard, GPT-3.5 Turbo, and GPT-4.0 across four distinct test sets, labeled A, B, C, and D. Each test set represents a different level of difficulty ranging from basic (Cardiac nurse, Test A) to difficult (Board reviews, Test D). The vertical axis of the graph denotes the accuracy percentage, quantifying the proportion of correctly answered questions. For each test set, the graph depicts three bars, each corresponding to one of the three LLMs.

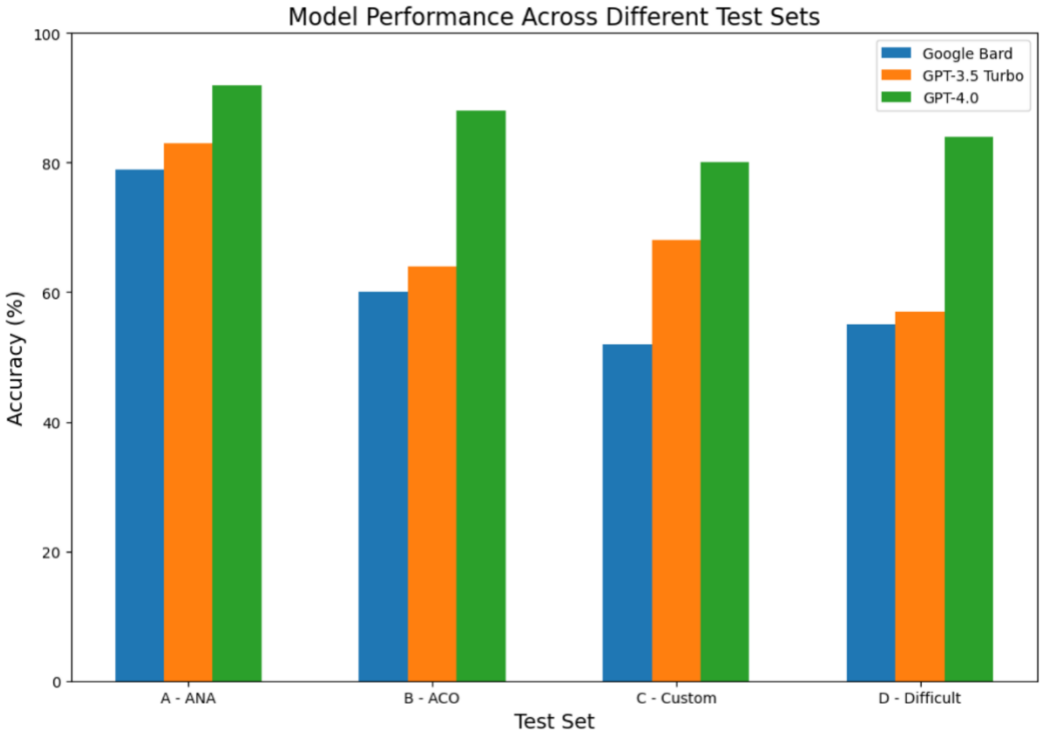
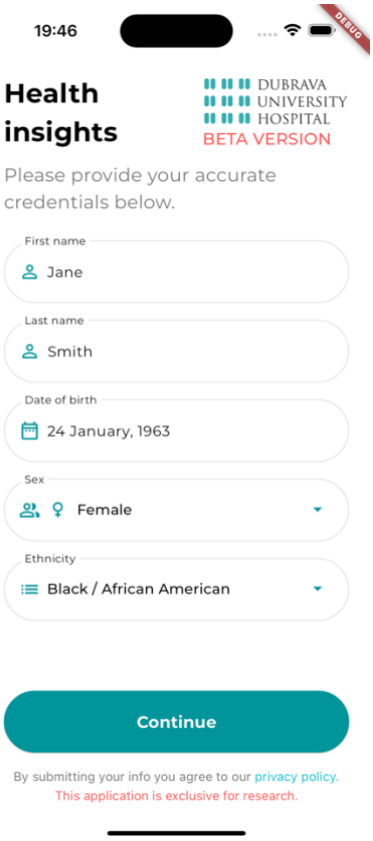
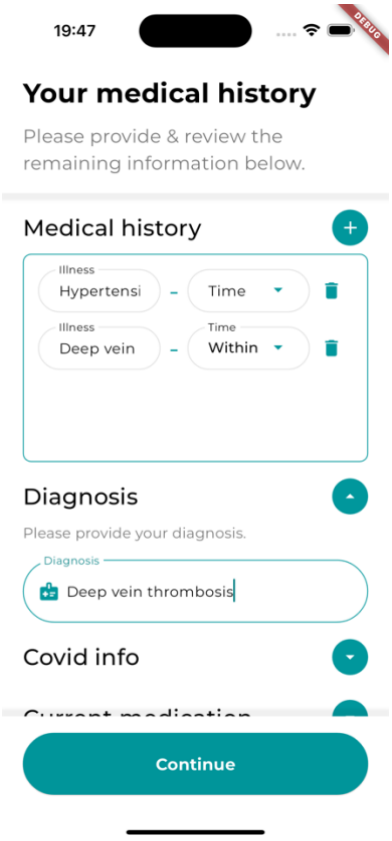
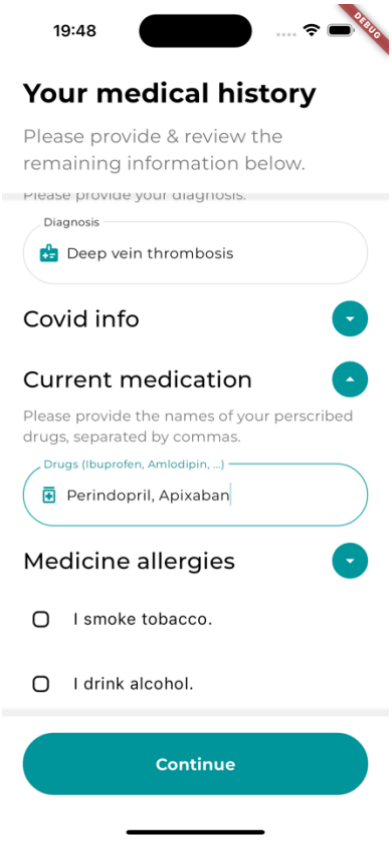
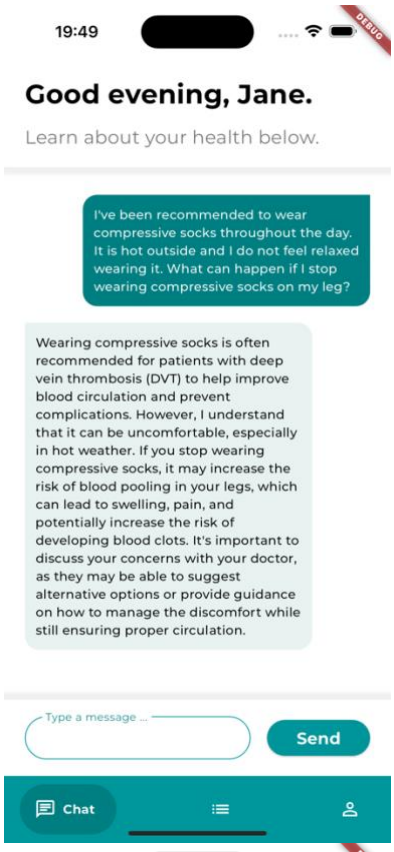


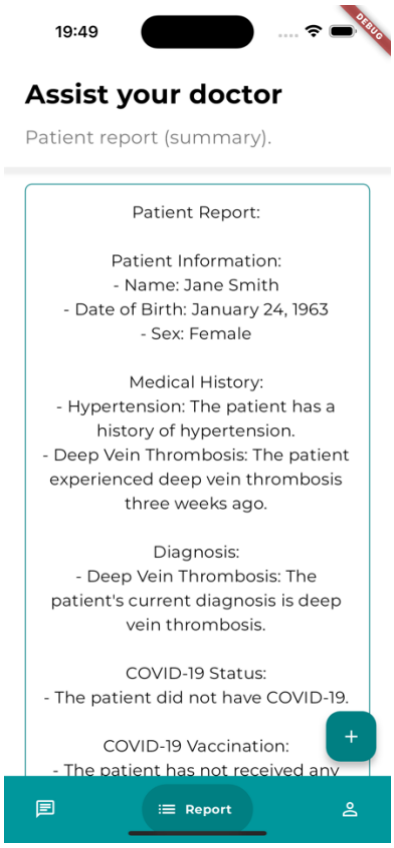
Figure 2. Images represent the user interface of the 'Dubravka' mobile application, specifically designed within this research. This digital health tool incorporates the GPT-3.5 Turbo Application Programming Interface (API), thus enabling the LLM to interact with unique health data from individual users. Consequently, the application has potential to generate tailored advice, with emphasis on areas such as dietary habits, lifestyle modifications, and cardiovascular prevention strategies. In addition, the application includes a feature that facilitates the summarization of a patient's medical history, potentially aiding in the communication between patients and healthcare professionals.


19:46
Health insights
DUBRAVA UNIVERSITY HOSPITAL
BETA VERSION
Please provide your accurate credentials below.
First name
Jane
Last name
Smith
Date of birth
24 January, 1963
Sex
Female
Ethnicity
Black / African American
Continue
By submitting your info you agree to our privacy policy.
This application is exclusive for research.


19:47
Your medical history
Please provide & review the remaining information below.
Medical history
Illness
Hypertensi
Time
Illness
Deep vein
Time
Within
Diagnosis
Please provide your diagnosis.
Diagnosis
Deep vein thrombosis
Covid info
Continue


19:48
Your medical history
Please provide & review the remaining information below.
Diagnosis
Deep vein thrombosis
Covid info
Current medication
Please provide the names of your perscribed drugs, separated by commas.
Drugs (Ibuprofen, Amlodipin, ...)
Perindopril, Apixaban
Medicine allergies
I smoke tobacco.
I drink alcohol.
Continue


19:49
DEBUG
Good evening, Jane.
Learn about your health below.
I've been recommended to wear compressive socks throughout the day. It is hot outside and I do not feel relaxed wearing it. What can happen if I stop wearing compressive socks on my leg?
Wearing compressive socks is often recommended for patients with deep vein thrombosis (DVT) to help improve blood circulation and prevent complications. However, I understand that it can be uncomfortable, especially in hot weather. If you stop wearing compressive socks, it may increase the risk of blood pooling in your legs, which can lead to swelling, pain, and potentially increase the risk of developing blood clots. It's important to discuss your concerns with your doctor, as they may be able to suggest alternative options or provide guidance on how to manage the discomfort while still ensuring proper circulation.
Type a message ...
Send
Chat


19:49
DEBUG
Assist your doctor
Patient report (summary).
Patient Report:
Patient Information:
- Name: Jane Smith
- Date of Birth: January 24, 1963
- Sex: Female
Medical History:
- Hypertension: The patient has a history of hypertension.
- Deep Vein Thrombosis: The patient experienced deep vein thrombosis three weeks ago.
Diagnosis:
- Deep Vein Thrombosis: The patient's current diagnosis is deep vein thrombosis.
COVID-19 Status:
- The patient did not have COVID-19.
COVID-19 Vaccination:
Report

19:49
DEBUG
Assist your doctor
Patient report (summary).
Medication:
- The patient is currently taking the following medications: Perindopril and Apixaban.
Allergies:
- The patient does not have any known allergies.
Lifestyle Factors:
- Smoking: The patient is not a smoker.
- Alcohol Consumption: The patient does not drink alcohol.
Conclusion:
Jane Smith, a 58-year-old female, has a medical history of hypertension and recently experienced deep vein thrombosis. She is currently taking Perindopril and Apixaban as prescribed medications. The patient does n
have any known allergies and ha
Report

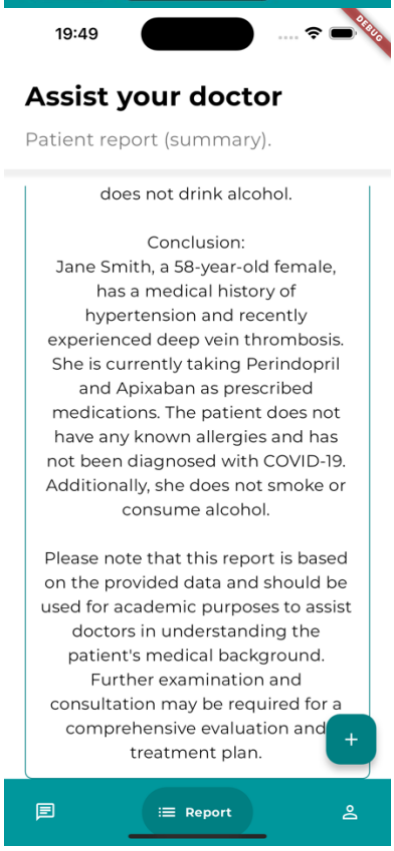

19:49
DEBUG
Assist your doctor
Patient report (summary).
does not drink alcohol.
Conclusion:
Jane Smith, a 58-year-old female, has a medical history of hypertension and recently experienced deep vein thrombosis. She is currently taking Perindopril and Apixaban as prescribed medications. The patient does not have any known allergies and has not been diagnosed with COVID-19. Additionally, she does not smoke or consume alcohol.
Please note that this report is based on the provided data and should be used for academic purposes to assist doctors in understanding the patient's medical background. Further examination and consultation may be required for a comprehensive evaluation and treatment plan.
Report

Table 2. This table presents a case study of a 64-year-old woman with idiopathic dilated cardiomyopathy, her clinical details, and a multiple-choice question on the most suitable treatment option.

Question: A 64 year old woman with a history of idiopathic dilated cardiomyopathy presents for a second opinion. Current medications include carvedilol, ramipril, furosemide, magnesium, potassium and spironolactone which she has been taking for two years. ECG shows sinus rhythm with left bundle branch block. Physical examination reveals a non-elevated jugular venous pressure, a displaced apical impulse and a pansystolic murmur. Blood pressure is 100/78 mmHg, heart rate is 68 bpm. Height is 5 feet, 6 inches. Weight is 130 lbs. Echocardiography reveals severe MR. LVEF is 30%. LV end-systolic dimension is 49 mm. The mitral annulus is dilated. Which of the following statements is most correct? A. Cardiac resynchronization is indicated B. The patient should receive mitral valve replacement C. The patient should receive annuloplasty repair D. Transcatheter mitral valve intervention should be performed
GPT-4: A. Cardiac resynchronization is indicated Cardiac resynchronization therapy (CRT) is beneficial in patients with heart failure with reduced ejection fraction (HFrEF), especially those with left bundle branch block, as it improves cardiac efficiency and can relieve symptoms of heart failure. CRT is indicated in patients with LVEF ≤ 35%, symptomatic heart failure (NYHA class II, III, or ambulatory IV), sinus rhythm, and left bundle branch block with a QRS duration ≥130 ms, despite at least 3 months of guideline-directed medical therapy, like in this patient. The other options may not be the best choices based on the information given: B. Mitral valve replacement might be indicated if mitral valve repair is not possible or if the patient has severe symptomatic mitral regurgitation despite optimal medical therapy, but it's typically not the first choice due to associated surgical risks. C. Mitral valve annuloplasty repair might be indicated if there was severe mitral regurgitation primarily due to annular dilatation, but this needs further evaluation. D. Transcatheter mitral valve intervention might be considered in high surgical risk patients with severe symptomatic mitral regurgitation despite optimal medical therapy, but it's also typically not the first choice, especially in patients with dilated cardiomyopathy where CRT could be beneficial.
GPT-3.5 Turbo: A. Cardiac resynchronization is indicated In this case, the patient has severe mitral regurgitation (MR) in the setting of idiopathic dilated cardiomyopathy. The presence of left bundle branch block and a low ejection fraction (LVEF) of 30% suggests a dyssynchronous ventricular contraction, which can be addressed with cardiac resynchronization therapy (CRT). CRT involves the placement of a biventricular pacemaker to synchronize the contraction of the ventricles and improve overall cardiac function. Option B, mitral valve replacement, is not the first-line treatment for severe MR in the absence of other indications such as significant symptoms or structural valve damage. Option C, annuloplasty repair, may be considered as a surgical treatment option in some cases, but the most appropriate initial step in management is to address the underlying cardiac dysfunction and dyssynchrony with CRT. Option D, transcatheter mitral valve intervention, may be considered in select cases of severe MR, but the primary focus should be on addressing the underlying cardiomyopathy and dyssynchrony with CRT.

Therefore, the most correct statement in this case is that cardiac resynchronization therapy (CRT) is indicated.
Google Bard: The answer is **(B). The patient should receive mitral valve replacement. ** The patient in this question has severe mitral regurgitation, which is a condition in which blood flows backward from the left ventricle into the left atrium during systole. This can cause the left ventricle to enlarge and eventually fail. The patient's echocardiogram also reveals a LVEF of 30%, which is considered to be severely reduced. This means that the left ventricle is not pumping blood as effectively as it should be. The presence of severe mitral regurgitation and severely reduced LVEF is an indication for mitral valve replacement. Mitral valve replacement is a surgical procedure that involves replacing the diseased mitral valve with a mechanical or biological valve. This will improve the patient's symptoms and reduce their risk of complications, such as heart failure and stroke. The other answer choices are not as appropriate in this case. * **Cardiac resynchronization:** Cardiac resynchronization is a procedure that can be used to improve the coordination of the heart's electrical activity. It is not indicated for patients with severe mitral regurgitation. * **Annuloplasty repair:** Annuloplasty repair is a procedure that can be used to tighten the mitral annulus. It is not indicated for patients with severe mitral regurgitation. * **Transcatheter mitral valve intervention:** Transcatheter mitral valve intervention is a minimally invasive procedure that can be used to repair a leaking mitral valve. It is not indicated for patients with severe mitral regurgitation.

Note: GPT-4.0 and GPT-3.5 Turbo agree on option A. They both justify this based on the patient's presentation of left bundle branch block and reduced ejection fraction, which suggests a dyssynchronous ventricular contraction. Cardiac resynchronization therapy (CRT) is beneficial in these patients. On the other hand, Google Bard suggests option B, where the patient should receive mitral valve replacement due to the severe mitral regurgitation and the severely reduced LVEF, which is not the first-line treatment for severe mitral regurgitation in the absence of other indications such as significant symptoms or structural valve damage. Based on the available information, the response of GPT-4.0 and GPT-3.5 Turbo seems more appropriate. Nevertheless, it's crucial to remember that this is a complex clinical situation, and the final decision should always be based on a detailed evaluation by a healthcare provider.

Table 3. The table presents the clinical scenario of a 50-year-old man experiencing chest pain, detailing his medical history, symptoms, treatment, and follow-up. It concludes with a multiple-choice question regarding optimal lipid management strategy.

Question: A 50 year-old man presents to the emergency department with chest pain. While having an emotional argument with his wife, he suddenly developed left-sided substernal chest pain that radiated to his left jaw with associated diaphoresis and shortness of breath. He took aspirin 325 mg at home and called emergency medical services. Upon arrival of emergency services, the patient’s pain improved with two tablets of sublingual nitroglycerin but mild discomfort and dyspnea persisted. His past medical history is notable for hyperlipidemia, for which he takes atorvastatin 40 mg. Upon arrival to the emergency department, his temperature is 37.6 °C, heart rate 92 bpm, blood pressure 174/98 mmHg (similar in both arms), respiratory rate 20/min, and oxygen saturation 98% on room air. He appears uncomfortable. Jugular venous pressure is estimated at 6 cm H2O. Lung fields are clear to auscultation bilaterally. Cardiac exam reveals regular rate and rhythm, normal S1 and S2, and no extra heart sounds. Extremities reveal good distal pulses and are warm without peripheral edema. The ECG is indicative of Wellens’ syndrome. He was promptly taken to the cardiac catheterization lab where he was found to have a 99% proximal LAD stenosis with TIMI-2 flow without other significant coronary artery disease.

Prasugrel 60 mg was administered and an everolimus-eluting stent was successfully placed. He was discharged on atorvastatin 80 mg, in addition to aspirin 81 mg, prasugrel 10 mg, metoprolol XL 50 mg, and lisinopril 10 mg. A repeat lipid panel in outpatient cardiology clinic 6 weeks post-MI revealed and LDL-C of 80 mg/dL, High density lipoprotein cholesterol of 48 mg/dL, and Triglycerides of 189 mg/dL. Which of the following is the best regimen for lipid management?
A. Continue atorvastatin 80
B. Continue atorvastatin 80 and start ezetimibe 10
C. Continue atorvastatin 80 and start fenofibrate 130
D. Stop atorvastatin and start evolocumab

GPT-4.0: This patient presented with an acute myocardial infarction and was found to have significant single-vessel coronary artery disease (CAD). His high-intensity statin therapy (atorvastatin 80 mg daily) was appropriately initiated during hospitalization. The goal of high-intensity statin therapy in patients with clinical CAD is to lower low-density lipoprotein cholesterol (LDL-C) by ≥50% and ideally to a goal of <70 mg/dL.

In this patient, his LDL-C is 80 mg/dL, which is slightly above the ideal target of <70 mg/dL for secondary prevention in patients with atherosclerotic cardiovascular disease (ASCVD). According to the 2018 AHA/ACC cholesterol guidelines, in very high-risk ASCVD patients, if LDL-C remains ≥70 mg/dL despite maximally tolerated statin therapy, ezetimibe is recommended as the next drug to add. Only if the LDL-C level remains ≥70 mg/dL after addition of ezetimibe, or if ezetimibe is not tolerated, is a PCSK9 inhibitor (such as evolocumab) recommended.
Hence, the most appropriate next step would be to continue the current high-intensity statin and add ezetimibe.
The correct answer is:
B) Continue atorvastatin 80 and start ezetimibe 10 mg.

GPT-3.5 Turbo: In this case, the patient presented with an acute myocardial infarction (MI) and received percutaneous coronary intervention (PCI) with stent placement. The subsequent lipid panel results show an LDL-C of 80 mg/dL, HDL-C of 48 mg/dL, and triglycerides of 189 mg/dL. Based on this information, let's analyze the options to determine the best regimen for lipid management:
A. Continue atorvastatin 80.
This option suggests continuing the patient's current regimen of atorvastatin 80 mg. This dose of atorvastatin is considered a high-intensity statin therapy, which is the preferred treatment for patients with established coronary artery disease (CAD). Continuing atorvastatin 80 mg would be appropriate for LDL-C management.
B. Continue atorvastatin 80 and start ezetimibe 10.
Adding ezetimibe to the patient's current atorvastatin regimen is a reasonable option for further LDL-C reduction. Ezetimibe inhibits cholesterol absorption in the intestines and can be used as an adjunct to statin therapy for additional LDL-C lowering.
C. Continue atorvastatin 80 and start fenofibrate 130.
Fenofibrate is primarily used to target elevated triglyceride levels and increase HDL-C levels. However, in this case, the patient's primary concern is LDL-C management rather than elevated triglycerides. While fenofibrate may have some benefits in certain patients, it is not the best choice for lipid management in this scenario.
D. Stop atorvastatin and start evolocumab.
Evolocumab is a PCSK9 inhibitor, a newer class of medications used to further lower LDL-C. However, it is generally reserved for patients with very high LDL-C or familial hypercholesterolemia who have not achieved adequate LDL-C reduction with maximally tolerated statin therapy and other lipid-lowering medications. Stopping atorvastatin, a high-intensity statin, would not be appropriate in this case.

Considering the patient's history of acute MI, the current regimen of atorvastatin 80 mg should be continued, as it is a high-intensity statin and the preferred treatment for patients with established CAD. Option A is the most appropriate choice for lipid management in this scenario. **Therefore, the best regimen for lipid management in this case is:** **A. Continue atorvastatin 80.**
Google Bard: The answer is **D. Stop atorvastatin and start evolocumab.** The patient in the question had an acute myocardial infarction (MI) due to a 99% mid-LAD lesion. He was treated with a stent and discharged on atorvastatin 80 mg, in addition to aspirin 81 mg, prasugrel 10 mg, metoprolol XL 50 mg, and lisinopril 10 mg. A repeat lipid panel in outpatient cardiology clinic 6 weeks post-MI revealed an LDL-C of 80 mg/dL, HDL-C of 48 mg/dL, and triglycerides of 189 mg/dL. These lipid levels are not optimal for a patient who has had an MI. The target LDL-C level for patients who have had an MI is <70 mg/dL. Atorvastatin 80 mg is a moderate-intensity statin. In order to achieve an LDL-C level of <70 mg/dL, the patient may need to be started on a high-intensity statin, such as rosuvastatin or atorvastatin 10 mg. However, the patient is also taking prasugrel, which is a P2Y12 receptor antagonist. Prasugrel can increase the risk of muscle injury, especially when used in combination with high-intensity statins. Evolocumab is a monoclonal antibody that targets PCSK9, a protein that helps to regulate cholesterol levels. Evolocumab is a very effective medication for lowering LDL-C levels. It can be used in combination with statins or as a single agent. In the patient in the question, evolocumab would be a good option for lowering LDL-C levels without increasing the risk of muscle injury. The other answer choices are not as appropriate. * **Option A** is not appropriate because atorvastatin 80 mg is not enough to achieve an LDL-C level of <70 mg/dL. * **Option B** is not appropriate because ezetimibe is a bile acid sequestrant. Bile acid sequestrants can interact with prasugrel, increasing the risk of muscle injury. * **Option C** is not appropriate because fenofibrate is a fibrate. Fibrates can increase the risk of muscle injury, especially when used in combination with statins.

Note: GPT-4.0 provides an accurate and detailed analysis. The patient is post-MI and has an LDL-C level of 80 mg/dL, slightly above the ideal target of <70 mg/dL for secondary prevention in patients with atherosclerotic cardiovascular disease (ASCVD). GPT-4.0 correctly suggests the addition of ezetimibe according to the 2018 AHA/ACC cholesterol guidelines for very high-risk ASCVD patients, if LDL-C remains ≥70 mg/dL despite maximally tolerated statin therapy. On the other hand, GPT-3.5 Turbo begins by accurately detailing the patient's condition and reviewing each of the provided options. However, it chooses option A) as the most appropriate choice, failing to recommend the addition of ezetimibe to reach the target LDL-C of <70 mg/dL. While it correctly identifies the purpose and usage of the other medications, it does not correctly apply the AHA/ACC cholesterol guidelines for very high-risk ASCVD patients in its conclusion. Google Bard makes several mistakes. Firstly, it incorrectly states that atorvastatin 80 mg is a moderate-intensity statin when in fact it is a high-intensity statin. It also suggests that the patient is at risk of muscle injury due to his prasugrel medication, which is inaccurate - prasugrel is an antiplatelet drug and does not increase the risk of muscle injury. Furthermore, Google Bard incorrectly recommends evolocumab (option D) as the appropriate medication to lower LDL-C levels. While evolocumab, a PCSK9 inhibitor, can effectively lower LDL-C, it is usually only recommended when patients fail to achieve LDL-C targets with maximally tolerated statins and ezetimibe. These answers highlight some limitations of these models related to staying current with treatment guidelines and understanding the interactions and indications of various medications.

Table 4. This table presents classical clinical findings (e.g. in the Emergency department) from initial history, physical examination, or electrocardiogram, and asks about the necessity for urgent hospital admission.

Question: All of the following findings on initial history, physical examination, or electrocardiogram should prompt hospital admission for evaluation except: A. ECG demonstrating sinus pause of 2 s B. Physical examination demonstrating elevated jugular venous pressure, rales on lung examination, and lower extremity edema C. Family history of sudden cardiac death D. Physical examination demonstrating parvus et tardus, systolic murmur heard best at upper sternal border and absent S2 E. ECG with Right bundle branch block and ST elevations in V1–V3
GPT-4.0: A. ECG demonstrating sinus pause of 2 s While sinus pauses may be associated with symptoms such as syncope or dizziness, a brief pause of 2 seconds in isolation and in a patient without symptoms (such as the medical student in question) does not typically necessitate hospital admission. This finding might warrant further outpatient evaluation, such as a Holter monitor to assess for more significant bradycardia or other arrhythmias. The other choices represent significant findings that typically would warrant hospital admission for further evaluation: B. Physical exam findings of elevated jugular venous pressure, rales, and lower extremity edema suggest congestive heart failure. C. A family history of sudden cardiac death may suggest an inherited arrhythmia or structural heart disease and warrants further urgent investigation. D. Physical examination findings of parvus et tardus (diminished and delayed carotid pulse), a systolic murmur heard best at the upper sternal border, and an absent S2 suggest aortic stenosis, a serious condition that may require intervention. E. An ECG demonstrating right bundle branch block and ST elevations in V1-V3 is suggestive of a Brugada pattern, which is associated with an increased risk of sudden cardiac death.
GPT-3.5 Turbo: The finding that would not typically require hospital admission for evaluation is: D. Physical examination demonstrating parvus et tardus, systolic murmur heard best at upper sternal border and absent S2 This physical examination finding suggests aortic stenosis, which is a chronic condition that may not necessitate immediate hospital admission unless there are acute complications or hemodynamic instability. Evaluation and management of aortic stenosis can be performed in an outpatient setting. On the other hand, the other options are indicative of more concerning conditions that may require immediate evaluation and treatment: A. ECG demonstrating sinus pause of 2 s: Sinus pause can indicate sinus node dysfunction or other arrhythmias, which may require further evaluation and monitoring in a hospital setting. B. Physical examination demonstrating elevated jugular venous pressure, rales on lung examination, and lower extremity edema: These findings suggest congestive heart failure, which may require admission for further evaluation and management. C. Family history of sudden cardiac death: This raises concern for inherited cardiac conditions and may necessitate evaluation and risk stratification in a hospital setting. E. ECG with Right bundle branch block and ST elevations in V1-V3: This ECG pattern can be indicative of acute myocardial infarction, particularly involving the right ventricle. Immediate evaluation and treatment in a hospital setting is necessary.

Google Bard: The answer is **(B). Physical examination demonstrating elevated jugular venous pressure, rales on lung examination, and lower extremity edema.**

All of the other findings listed should prompt hospital admission for evaluation.

* **ECG demonstrating sinus pause of 2 s:** A sinus pause is a pause in the heart's rhythm that is caused by a temporary interruption in the electrical signal that controls the heart rate. A sinus pause of 2 s is considered to be prolonged and can be a sign of a serious underlying heart condition.

* **Family history of sudden cardiac death:** Sudden cardiac death is the sudden and unexpected death of a person who appears to be in good health. It is often caused by a heart arrhythmia. A family history of sudden cardiac death is a risk factor for developing a heart arrhythmia yourself.

* **Physical examination demonstrating parvus et tardus, systolic murmur heard best at upper sternal border and absent S2:** Parvus et tardus is a type of heart murmur that is caused by a narrowing of the aortic valve. An absent S2 is a sign that the aortic valve is not closing properly. These findings can be a sign of aortic stenosis, a serious heart valve condition.

* **ECG with Right bundle branch block and ST elevations in V1–V3:** Right bundle branch block is a condition that affects the electrical conduction system of the heart. It can cause a heart arrhythmia. ST elevations in V1–V3 are a sign of a heart attack.

Therefore, the only finding that should not prompt hospital admission for evaluation is **physical examination demonstrating elevated jugular venous pressure, rales on lung examination, and lower extremity edema.** These findings are more likely to be caused by a condition other than a heart problem, such as heart failure.

Note: The key to this question is to consider whether each finding could potentially indicate a life-threatening condition or a need for immediate hospital-based treatment. According to GPT-4.0, option A wouldn't necessarily prompt hospital admission for evaluation. While this finding may require further evaluation, such as a Holter monitor to assess for more significant arrhythmias, it doesn't necessarily indicate a need for hospital admission unless it is associated with symptoms like syncope or dizziness. However, GPT-3.5 Turbo believes that option D may not necessitate immediate hospital admission. This finding could suggest aortic stenosis, a chronic condition that may be managed in an outpatient setting unless there are acute complications or hemodynamic instability. On the other hand, Google Bard suggests that option B should not prompt hospital admission. This answer seems less likely to be correct because these symptoms together suggest congestive heart failure, which often does require hospital treatment, especially if the patient is symptomatic or there is potential for acute decompensation.

Table 5. The table presents the case of a 25-year-old woman with chest pain, nausea, and a headache after cocaine use. It offers a detailed account of her medical condition, including vitals and examination results, with a multiple-choice question about the most suitable next step in her treatment.

Question: A 25-year-old woman reports an intense chest pain that is localized in the middle of the chest and the upper sternal area, which lasts for 3 hours. She also describes nausea and a mild occipital headache. She reveales to have inhaled cocaine shortly prior to her episode of chest pain. Her family history is clear of coronary artery disease. Blood pressure is 200/105 mmHg on both arms. Pulse is 105/min and regular. Oxygen saturation is 99% on room air. Heart sounds are normal and no murmurs are heard. Lower extremity pulses are lull and symmetric, there is no peripheral edema. ECG shows sinus tachycardia, otherwise unremarkable. Which of the following is the most appropriate next step in management of this patient?

A. Aspirin and clopidogrel at loading dose
B. Intravenous lorazepam
C. Intravenous metoprolol
D. Intravenous phentolamine
E. Low-molecular-weight heparin

GPT-4.0: The correct answer is: B. Intravenous lorazepam

The presentation is suggestive of cocaine-induced chest pain. Cocaine induces coronary artery vasoconstriction and a hyperadrenergic state which may lead to myocardial ischemia or infarction. It also increases myocardial oxygen demand due to increased heart rate, blood pressure, and myocardial contractility. The management of chest pain following cocaine use is supportive and includes:

1. Oxygen to treat any hypoxemia.
2. Benzodiazepines (like lorazepam) to relieve anxiety and decrease sympathetic output, which can help reduce heart rate and blood pressure.
3. Nitroglycerin to relieve coronary artery vasospasm.
4. Aspirin to inhibit platelet aggregation.

Beta blockers are contraindicated in the acute setting of cocaine-induced chest pain because they can lead to unopposed alpha-adrenergic stimulation and worsen coronary vasoconstriction. So option C should be avoided.

Phentolamine is a nonselective alpha-adrenergic antagonist. While it can theoretically be used to treat cocaine-induced coronary vasoconstriction, it is not first line treatment because of its potential to cause a reflex tachycardia.

Anticoagulation with low molecular weight heparin is typically not necessary in the setting of cocaine-induced chest pain, unless acute coronary syndrome is confirmed. Therefore, option E is also not appropriate as the first next step.

In conclusion, the best immediate management for this patient includes benzodiazepines (option B) for their anxiolytic and sympathetic outflow reduction properties, as well as aspirin and nitroglycerin if ischemia is suspected.

GPT-3.5 Turbo: Based on the given clinical scenario, the most appropriate next step in the management of this patient is: D. Intravenous phentolamine.

The patient's intense chest pain, occipital headache, and history of recent cocaine use suggest the possibility of cocaine-induced myocardial ischemia or infarction. Cocaine can cause vasoconstriction, increased heart rate, and increased blood pressure, leading to increased myocardial oxygen demand and potentially triggering ischemic events.

Intravenous phentolamine, an alpha-adrenergic antagonist, is the treatment of choice for cocaine-induced hypertension and coronary artery vasoconstriction. It helps reverse the vasoconstrictive effects of cocaine by blocking alpha-receptors and reducing peripheral vascular resistance. Prompt administration of phentolamine can help alleviate symptoms and prevent further myocardial damage.

Option A (aspirin and clopidogrel at loading dose) is not the most appropriate choice in this case. While antiplatelet therapy is indicated in acute coronary syndromes, the primary issue in this patient is the acute cocaine-induced vasoconstriction rather than atherosclerotic coronary artery disease.

Option B (intravenous lorazepam) is a benzodiazepine that may help with anxiety or agitation, but it does not directly address the underlying cocaine-induced vasoconstriction and hypertension.

Option C (intravenous metoprolol) is a beta-blocker that is contraindicated in the setting of acute cocaine intoxication. Beta-blockers can exacerbate the vasoconstrictive effects of cocaine and may lead to unopposed alpha-adrenergic effects.

Option E (low-molecular-weight heparin) is not the most appropriate choice in this acute setting. While anticoagulation may be warranted in certain cases of myocardial ischemia or infarction, the primary issue in this patient is the acute cocaine-induced vasoconstriction rather than a thrombotic event.

To summarize, in a patient with intense chest pain, hypertension, and a history of recent cocaine use, the most appropriate next step in management is D. Intravenous phentolamine to reverse cocaine-induced vasoconstriction and hypertension. It is important to address the acute effects of cocaine before considering additional treatments or interventions.

Google Bard: The answer is D. Intravenous phentolamine. The patient in this case is presenting with symptoms of cocaine-induced coronary vasospasm. Cocaine can cause the coronary arteries to constrict, which can lead to chest pain, shortness of breath, and even death. The patient's ECG shows sinus tachycardia, which is a sign of increased heart rate. However, the ECG is otherwise unremarkable, which suggests that there is no significant damage to the heart muscle. The best next step in management of this patient is to administer intravenous phentolamine. Phentolamine is a medication that can be used to relax the coronary arteries and relieve vasospasm. The other answer choices are not as appropriate in this case. Aspirin and clopidogrel are medications that are used to prevent blood clots, but they are not effective in treating cocaine-induced coronary vasospasm. Intravenous lorazepam is a medication that can be used to treat anxiety, but it is not effective in treating cocaine-induced coronary vasospasm. Intravenous metoprolol is a medication that can be used to treat high blood pressure, but it is not effective in treating cocaine-induced coronary vasospasm. Low-molecular-weight heparin is a medication that can be used to prevent blood clots, but it is not effective in treating cocaine-induced coronary vasospasm.

Note: GPT-4.0 correctly identifies that the patient is likely experiencing cocaine-induced chest pain and accurately states that the management of this condition is mostly supportive. It correctly highlights the use of benzodiazepine to decrease sympathetic output and points out that beta-blockers are contraindicated in the acute setting of cocaine-induced chest pain. On the other hand, GPT-3.5 Turbo and Google Bard incorrectly suggest that intravenous phentolamine is the most appropriate next step in management. Phentolamine is a nonselective alpha-adrenergic antagonist that can be used to treat hypertensive emergencies due to catecholamine excess, but it is not the first line of treatment in cocaine-induced chest pain.

Table 6. Provides a case of a 34-year-old asymptomatic woman with an ECG finding of a long QT interval with a multiple-choice question about the most suitable next step in her treatment.

Question: Miss Johnson is an asymptomatic 34 yo woman who presents for breast biopsy and ECG finds long QT interval >470 ms.Which describes the best treatment option? A. Patient needs to be started on beta blocker and followed for symptoms B. Patient needs no further treatment since asymptomatic C. Patient is a candidate for biventricular pacemaker set at heart rate faster than her intrinsic rate D. Patient is a candidate for ICD E. Patient needs ETT first to make further decisions
GPT-4.0: Long QT syndrome can be a serious and potentially life-threatening heart condition. It can cause sudden, uncontrollable, dangerous arrhythmias in response to exercise or stress. **From the options given, the best course of action would be:** **A. "Patient needs to be started on beta blocker and followed for symptoms"** Beta blockers are the first-line treatment for long QT syndrome, even in asymptomatic individuals. They block the effects of adrenaline on your heart, which can prevent the onset of dangerous arrhythmias. Options B and E are not appropriate, as asymptomatic status does not negate the need for treatment or further testing in long QT syndrome. Option C is not typically used in managing long QT syndrome. Biventricular pacing is generally reserved for conditions that cause ventricular dyssynchrony like certain types of heart failure, not primary arrhythmia syndromes like long QT.

Option D, an ICD, might be considered in patients with long QT syndrome who have had a cardiac arrest, or in cases that are refractory to medical management, but it's not generally the first line of treatment.
Finally, all individuals with long QT syndrome should be educated about avoiding QT prolonging medications, maintaining good hydration and electrolyte balance, and promptly addressing any symptoms suggestive of arrhythmias.
Please note, while I aim to provide accurate information, patients with long QT syndrome should be evaluated and managed by a healthcare professional with expertise in this area. This information is based on general guidelines and may not be applicable to all patients.

GPT-3.5 Turbo: In the case of Miss Johnson, an asymptomatic 34-year-old woman with a long QT interval (>470 ms) found on the EKG, the best treatment option would be:
D. Patient is a candidate for an implantable cardioverter-defibrillator (ICD).
A prolonged QT interval on an EKG can increase the risk of life-threatening arrhythmias, such as torsades de pointes, which can lead to sudden cardiac arrest. Therefore, the primary concern is preventing sudden cardiac death in individuals with a prolonged QT interval.
In this scenario, the best treatment option would be an ICD. An ICD is a device that can detect and treat life-threatening arrhythmias by delivering an electric shock to restore normal heart rhythm. As Miss Johnson is asymptomatic but has a prolonged QT interval, she is at risk for potentially dangerous arrhythmias, making her a candidate for an ICD.
The other options mentioned are not appropriate in this context:
A. Patient needs to be started on a beta-blocker and followed for symptoms: While beta-blockers can be part of the treatment for some forms of long QT syndrome, in this scenario, where the patient is asymptomatic but has a prolonged QT interval, an ICD is the more appropriate choice.
B. Patient needs no further treatment since asymptomatic: Asymptomatic individuals with a prolonged QT interval are still at risk of developing life-threatening arrhythmias, and therefore, treatment is necessary.
C. Patient is a candidate for a biventricular pacemaker set at a heart rate faster than her intrinsic rate: Biventricular pacemakers are typically used for patients with heart failure and conduction system abnormalities, and they are not the primary treatment for long QT syndrome.
E. Patient needs an exercise treadmill test (ETT) first to make further decisions: An ETT is typically used to assess exercise capacity and evaluate symptoms related to exercise. In the case of long QT syndrome, the primary concern is the risk of arrhythmias rather than exercise-induced symptoms.
Therefore, the correct answer is:
D. Patient is a candidate for an ICD.

Google Bard: The answer is **A. Patient needs to be started on beta blocker and followed for symptoms.**
Long QT syndrome is a genetic disorder that affects the heart's electrical system. It can cause a prolonged QT interval on an ECG, which can lead to arrhythmias such as torsades de pointes.
Patients with long QT syndrome are at increased risk of sudden cardiac death (SCD). The 2020 ACC/AHA/HRS Expert Consensus Statement on Long QT Syndrome recommends that patients with long QT syndrome be started on beta blockers. Beta blockers can help to shorten the QT interval and reduce the risk of SCD.
Patients with long QT syndrome should also be followed for symptoms. Symptoms of long QT syndrome can include syncope (fainting), palpitations, and seizures. If a patient with long QT syndrome experiences symptoms, they should be seen by a doctor immediately.
The other answer choices are incorrect. Option B is incorrect because patients with long QT syndrome are at increased risk of SCD, even if they are asymptomatic. Option C is incorrect because biventricular pacemakers are not used to treat long QT syndrome. Option D is incorrect because ICDs are not the first-line treatment for long QT syndrome. Option E is incorrect because an ETT (exercise treadmill test) is not necessary to make a decision about treatment for long QT syndrome.
Therefore, the patient in this question should be started on beta blockers and followed for symptoms.

Note: GPT-4.0, and Google Bard, agree that the patient should be started on a beta-blocker and closely monitored for symptoms. They agree on this approach primarily because beta blockers, as first-line treatment, can mitigate the effects of adrenaline on the heart, thus preventing potentially dangerous arrhythmias in patients with Long QT syndrome. Both models emphasize the importance of treating asymptomatic individuals and dismiss the use of a biventricular pacemaker or an implantable cardioverter-defibrillator (ICD) as initial treatment. However, GPT-4.0 acknowledges that an ICD could be considered for patients with refractory conditions. Conversely, GPT-3.5 Turbo suggests that Miss Johnson should receive an ICD due to her increased risk of potentially lethal arrhythmias. This model prioritizes the prevention of sudden cardiac death and views ICD as the more suitable option despite the patient's asymptomatic status. This contrast in responses between GPT-3.5 Turbo and the other two models underscores the critical importance of context-specific and patient-centered approaches in healthcare decisions.