

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*}, Ivan Zeljković¹, Fran Rode¹, Ante Lisičić¹, Iskra A. Nola⁴, Nikola Pavlović¹, Šime Manola¹

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

Abstract

Introduction: Over the past two years, the use of Large Language Models (LLMs) in clinical medicine has expanded significantly, particularly in cardiology, where they are applied to ECG interpretation, data analysis, and risk prediction. This study evaluates the performance of five advanced LLMs—Google Bard, GPT-3.5 Turbo, GPT-4.0, GPT-4o, and GPT-o1-mini—in responding to cardiology-specific questions of varying complexity.

Methods: A comparative analysis was conducted using four test sets of increasing difficulty, encompassing a range of cardiovascular topics, from prevention strategies to acute management and diverse pathologies. The models' responses were assessed for accuracy, understanding of medical terminology, clinical relevance, and adherence to guidelines by a panel of experienced cardiologists.

Results: All models demonstrated a foundational understanding of medical terminology but varied in clinical application and accuracy. GPT-4.0 exhibited superior performance, with accuracy rates of 92% (Set A), 88% (Set B), 80% (Set C), and 84% (Set D). GPT-4o and GPT-o1-mini closely followed, surpassing GPT-3.5 Turbo, which scored 83%, 64%, 67%, and 57%, and Google Bard, which achieved 79%, 60%, 50%, and 55%, respectively. Statistical analyses confirmed significant differences in performance across the models, particularly in the more complex test sets. While all models demonstrated potential for clinical application, their inability to reference ongoing clinical trials and some inconsistencies in guideline adherence highlight areas for improvement.

Conclusion: LLMs demonstrate considerable potential in interpreting and applying clinical guidelines to vignette-based cardiology queries, with GPT-4.0 leading in accuracy and guideline alignment. These tools offer promising avenues for augmenting clinical decision-making but should be used as complementary aids under professional supervision.

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 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 exemplify the adaptability of LLMs, requiring minimal reconfiguration to excel across numerous domains and 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, including 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). LLMs can assist with repetitive hospital tasks, such as writing discharge letters and analyzing large datasets to identify patterns, risk factors, and outcome predictions (3,7-9). In pursuit of deploying these advanced models into real-world clinical settings, we developed a prototype mobile application, 'Dubravka,' that integrates user-specific medical data and interfaces with a Large Language Model. This application is intended to serve as a stepping stone towards personalized, guideline-oriented patient care within the cardiology domain (10-18).

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.

AI has been used in cardiology for some time, with computers interpreting electrocardiograms (ECGs) daily (22-24). A 2019 study showed that machine learning (ML) algorithms exceeded the diagnostic accuracy of a team of practicing cardiologists in classifying 12 heart rhythm types

from single-lead ECGs, and ML can even be used to predict pathologies from ECG recordings (25, 26). However, the use of LLMs for ECG interpretation remains scarce. A study from early 2023 (27) evaluated the appropriateness of ChatGPT's 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 for LLMs to enhance education, improve patient health literacy, and provide counseling on the basics of cardiovascular prevention. Another recent study highlighted ChatGPT's potential as an assisted decision support tool for straightforward clinical questions, though it was less effective 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 both 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. 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 five State-of-the-Art Large Language Models (LLMs)—Google Bard, GPT-3.5 Turbo, GPT-4.0, GPT-4o, and GPT-o1-mini—by comparing their performance on a series of cardiology-focused clinical vignettes. These vignettes were designed to represent a spectrum of scenarios, from basic cardiovascular care to 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 conduct a comparative analysis of five state-of-the-art LLMs: Google Bard (based on PaLM 2), GPT-3.5 Turbo, GPT-4.0 (model 0613), GPT-4o-mini (model gpt-4o-mini-2024-07-18), and GPT-4o (model gpt-4o-2024-08-06), all available as text-based chat interfaces provided by OpenAI and Google. These models were 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, 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 queries 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, including 15 questions from reference (30) and others created to follow the structure and outcomes of references (30, 31) while being aligned with current AHA guidelines. The questions featured different background scenarios, altered demographics, and reported anamnestic data, maintaining the core concepts of the board-type questions. Questions in both Test set C and Test set D were reviewed and cross-validated by a panel of four senior cardiologists to ensure clinical accuracy, relevance, and appropriate complexity.

Test sets B, C, and D feature various patient demographics. Questions exclusively pertaining to image data were excluded, and ECG or cardiac imaging graphical data supporting the vignette were incorporated into the questions textually, describing these findings following a standard format and terminology. Details of each question set, and the corresponding results are provided 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,’ which was developed as part of this research at Dubrava University Hospital, we employed the OpenAI GPT-3.5 Turbo API to generate patient-tailored cardiovascular prevention advice. The application, programmed in Flutter, stores basic patient information locally and provides them as hidden parameters to the LLM prompt, enabling personalization of outputs. The analysis was performed 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 recommended¹. 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

¹ Parameter settings and recommendations could be found on: <https://platform.openai.com/docs/api-reference/chat/create>

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.

The accuracy of all five models—GPT-4.0, GPT-3.5 Turbo, Google Bard, GPT-4o, and GPT-o1-mini—was calculated for each test by dividing the number of correct answers by the total number of questions in the respective test set. Performance differences among the models were analyzed using a Friedman test, a non-parametric alternative to repeated measures ANOVA, to account for the ranking nature of the data. Post-hoc pairwise comparisons were performed with Bonferroni correction to identify statistically significant differences between the models when the omnibus test indicated significance. The significance level for all tests was 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 6.4.12².

The outputs generated from the LLMs were evaluated based on correct and incorrect responses. Incomplete answers or rare instances where the model yielded two answers as correct were treated as incorrect. The previously mentioned panel of four senior cardiologists (three interventional cardiologists and one imaging subspecialist) with over ten years of professional experience in the field, evaluated the answers, i.e., the full explanations given by the LLMs, following these criteria: accuracy, understanding of medical terminology, and clinical relevance. They also considered the depth of the perceived understanding and explanation provided by the LLMs for each response. These were not scored numerically. The panel discussed their observations in detail.

The primary goal of this study was to appraise the clinical relevance of the 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.

Results

Our findings demonstrate that GPT-4.0 consistently delivered superior performance across all four sets of questions, achieving the highest scores in integrating clinical scenarios, adhering to current cardiology guidelines, and computing and interpreting various scores and indices. The

² <https://jupyter.org/>

GPT-4o and GPT-o1-mini models also performed well, following GPT-4.0 by a narrow margin, particularly excelling in USMLE-like Step 2/3 questions and ABIM board reviews. These models demonstrated strong reasoning abilities and reliable adherence to clinical guidelines, making them effective tools for complex medical queries. GPT-3.5 Turbo, while performing reasonably well, lagged behind the GPT-4 family of models, often resorting to speculative or less evidence-based answers, particularly in more difficult scenarios. On the other hand, Google Bard underperformed significantly across all test levels, frequently failing to provide direct answers or offering overly simplistic explanations that lacked depth and precision.

Despite their varying levels of accuracy, all models demonstrated an ability to provide responses that were perceived as correct and supplemented by explanatory commentary, often discounting alternative answers. Detailed performance data for each model across the different sets of questions are presented in Figure 1 and Table 1.

HERE FIGURE 1.

Table 1: Comparative performance of Google Bard, GPT 3.5 Turbo, and GPT 4.0, GPT o1-mini and GPT-4o on four sets of questions with varied complexity in cardiology.

Test	Level of difficulty	Google Bard	GPT-3.5 Turbo	GPT-4.0	GPT-o1-mini	GPT-4o
A – ANA	Basic (Cardiac nurse)	19/24	20/24	22/24	21/24	21/24
B - ACO	Mixed	15/25	16/25	22/25	18/25	19/25
C - Custom	USMLE-like Step 2/3 Question	30/60	40/60	48/60	47/60	46/60
D - Difficult	ABIM board reviews	50/90	51/90	76/90	50/90	71/90

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, ABIM - American Board of Internal Medicine cardiology board.

Accuracy. The performance statistics for each model are as follows: Test Set A (Basic Level): GPT-4.0 achieved a 92% success rate, with GPT-4o and GPT-o1-mini scoring 88% each. GPT-3.5 Turbo followed with 83%, while Google Bard scored 79%. Test Set B (Mixed Level): GPT-4.0 had an 88% success rate, higher than the 76% and 72% rates of GPT-4o and GPT-o1-mini, respectively. GPT-3.5 Turbo and Google Bard scored 64% and 60%, respectively. Test Set C (USMLE-like): GPT-4.0 achieved an 80% success rate, compared to 78% and 77% for GPT-4o and GPT-o1-mini. GPT-3.5 Turbo scored 67%, and Google Bard with 50%. Test Set D (Difficult Level): GPT-4.0 outperformed others with an 84% success rate, followed by GPT-4o at 79% and GPT-o1-mini at 76%. GPT-3.5 Turbo and Google Bard scored 57% and 55%, respectively.

Understanding of Medical Terminology. Terms related to cardiology, including those pertaining to human anatomy, medical diagnoses, generic drug names, diagnostic tests, and laboratory parameters, were correctly identified by all models. GPT-4.0, GPT-4o, and GPT-o1-mini were the most accurate and proficient, consistently employing medical terminologies appropriately and aligning their answers with current clinical guidelines. GPT-3.5 Turbo demonstrated a solid understanding of the terminology and provided detailed explanations of each treatment option. However, its application of these terms within the framework of current clinical guidelines was occasionally flawed. Google Bard, by contrast, showed significant gaps in understanding medical terms, leading to multiple inaccurate statements and recommendations, as evidenced in Table 2 and Table 4.

Clinical Relevance and Contextual Understanding. GPT-4.0, GPT-4o, and GPT-o1-mini demonstrated a strong understanding of clinical context and consistently adhered to established guidelines, including the 2018 AHA/ACC Guideline on the Management of Blood Cholesterol. These models occasionally referenced other notable guidelines, such as the ACC/AHA/HFSA guidelines for heart failure management and ESC guidelines for cardiac resynchronization therapy, aligning their responses with evidence-based practices. However, none of the models cited specific expert papers or primary literature directly. GPT-3.5 Turbo, while generally detailed in its explanations, occasionally deviated from guideline recommendations, leading to less accurate clinical advice. Google Bard showed significant gaps in guideline adherence, providing inconsistent recommendations that often conflicted with established clinical standards (see Table 3, Table 5, and Table 6).

Statistical Analysis. A Friedman test was conducted to assess performance differences among five models—GPT-4.0, GPT-3.5 Turbo, Google Bard, GPT-4o, and GPT-o1-mini—across four test sets. For Test set A, the analysis did not reveal any statistically significant differences across the models ($p = 0.112$). Similarly, for Test set B, no significant differences were observed ($p = 0.068$). However, significant differences were found for Test set C ($p = 0.004$) and Test set D ($p < 0.001$), indicating variation in performance across models on more complex test sets.

Post-hoc pairwise comparisons provided additional insights. For Test set A, no significant differences were identified between any of the models after Bonferroni correction. On Test set B, while differences were not statistically significant, GPT-4.0 displayed a trend of higher performance compared to Google Bard and GPT-3.5 Turbo. On Test set C, GPT-4.0 significantly outperformed Google Bard (adjusted $p = 0.002$), but the difference between GPT-4.0 and GPT-3.5 Turbo was not statistically significant (adjusted $p = 0.062$). Additionally, GPT-4.0 showed no significant performance difference compared to GPT-4o or GPT-o1-mini for this test set. For Test set D, GPT-4.0 demonstrated statistically significant superiority over both Google Bard and GPT-3.5 Turbo (adjusted $p < 0.001$ for both). GPT-4.0 also significantly outperformed GPT-4o (adjusted $p = 0.038$), although no significant difference was detected between GPT-4.0 and GPT-o1-mini (adjusted $p = 0.177$).

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. Notably, GPT-4.0, GPT-4o, GPT-o1-mini, 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 medical information to diverse audiences.

Model Performance and Proficiency in Medical Terminology.

The models' responses were coherent, structured, and detailed, effectively translating complex medical jargon into accessible language. GPT-4.0 consistently outperformed other models across all test sets, demonstrating a superior ability to integrate clinical information and adhere to current medical guidelines. However, the relatively smaller number of questions in test sets A and B may have reduced the statistical power, leading to a lack of significant differences for those sets despite observable performance trends. Critically, each model showed the capacity to identify and process important clinical details from the input, apply medical knowledge, and generate contextually appropriate responses.

All models included explanations for their selected answers, rationalizing why certain options were correct and others were not. For instance, as shown in Table 2, GPT-4.0 and GPT-3.5 Turbo displayed a strong grasp of cardiological terminology, accurately interpreting ECG and echocardiographic findings, and recommending appropriate procedures. Conversely, while Google Bard failed to recognize key details such as the wide QRS complex and the role of cardiac resynchronization therapy. GPT-4o and GPT-o1-mini also showed strong performance in clinical reasoning, often closely aligning with GPT-4.0, though subtle differences were evident in certain nuanced scenarios. In more challenging scenarios (Appendix 1, Set D, Question 3), GPT-4.0's alignment with guideline-based care outshone the less consistent recommendations of Google Bard, reinforcing that complexity levels reveal pronounced model disparities.

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-4o, GPT-o1-mini, GPT-3.5 Turbo, and Google Bard revealed distinct variations in their ability to apply medical guidelines accurately. For instance, in the vignette addressing lipid management (Table 3), GPT-4.0, GPT-4o, and GPT-o1-mini consistently recommended adding ezetimibe to high-intensity statin therapy based on the patient's slightly elevated LDL-C level. These models demonstrated proficiency in synthesizing the patient's clinical data and aligning their recommendations with current medical guidelines. GPT-o1-mini, while closely matching the performance of GPT-4.0, occasionally provided slightly less detailed explanations but maintained adherence to clinical standards.

GPT-3.5 Turbo displayed a solid conceptual understanding, referencing the patient's history, current statin therapy, and lipid profile in its explanation. However, it failed to align its final recommendation with current medical guidelines, reflecting a gap in contextual application despite clear reasoning. Conversely, Google Bard exhibited only partial contextual understanding. In the lipid management vignette, it correctly identified the patient's acute myocardial infarction and noted that their LDL-C was above target, but it incorrectly categorized atorvastatin as a moderate-intensity statin and recommended switching to evolocumab. This suggestion is inconsistent with guideline-directed management and highlights limitations in its application of evidence-based practices.

In the vignette presented in Table 4, which involved recommending the necessity for urgent hospital admission based on clinical findings, GPT-4.0, GPT-4o, and GPT-o1-mini provided accurate and contextually relevant responses. In contrast, Google Bard not only offered an incorrect answer but also provided contradictory arguments that lacked clarity, further underscoring deficiencies in its contextual understanding of clinical scenarios. Furthermore, for cocaine-induced chest pain (Appendix 1, Table 5), GPT-4.0 and GPT-o variants selected benzodiazepines as the appropriate initial step, reflecting a prudent adherence to first-line treatments, unlike models that prematurely escalated care.

HERE TABLE 3.

HERE TABLE 4.

Finally, all models occasionally demonstrated a tendency toward overtreatment, at times bypassing steps in guideline-oriented procedures. This inclination often involved proposing more advanced diagnostic or therapeutic approaches that were unnecessary or anticipatory. As shown in Table 5, GPT-4.0, GPT-4o, and GPT-o1-mini avoided this pitfall by correctly recommending intravenous benzodiazepines as the first-line treatment for cocaine-induced chest pain. This approach aligns with clinical guidelines that emphasize the role of benzodiazepines in reducing sympathetic stimulation, which is the underlying mechanism in such cases. In contrast, GPT-3.5 Turbo and Google Bard opted for immediate treatment with phentolamine, which, while effective for addressing hypertensive emergencies in cocaine toxicity, is not the first-line recommendation for chest pain management in this context.

HERE TABLE 5.

Reliability and Response Consistency Across Models. Among the five models analyzed, GPT-4.0, GPT-4o, and GPT-o1-mini demonstrated the highest levels of reliability, consistently applying clinical guidelines and generating responses aligned with evidence-based practices. GPT-3.5 Turbo and Google Bard, on the other hand, exhibited notable fluctuations in their answers when presented with the same query at different times. Notably, when GPT-3.5 Turbo and Google Bard produced an incorrect response initially, their revised answers upon subsequent queries often remained incorrect. Additionally, GPT-3.5 Turbo displayed a unique tendency to endorse one response initially and shift to an alternate response by the conclusion of its output,

creating confusion about its position. Google Bard also exhibited instances of internal contradiction, where it simultaneously justified and dismissed certain options within the same response (as observed in Table 4). Conversely, the GPT-4o series models showed a marked improvement in consistency, with minimal fluctuations or contradictions in their reasoning, further highlighting their reliability compared to GPT-3.5 Turbo and Google Bard.

Validation with Clinical Trials and Guidelines. During our evaluation, GPT-4.0, GPT-4o, and GPT-o1-mini consistently applied established guidelines—such as AHA/ACC recommendations for lipid management (e.g., adding ezetimibe for elevated LDL-C in Table 3) and expert consensus on arrhythmia care (e.g., beta-blockers for long QT in Table 6)—with a high degree of accuracy. However, it should be noted that the responses from GPT-4.0 and related models are predicated on medical literature and guidelines available until their respective knowledge cutoffs in 2023, potentially limiting their applicability to very recent advances. Google Bard (e.g. Table 6), while occasionally referencing an Expert Consensus Statement, applied these guidelines less consistently. In contrast, GPT-3.5 Turbo deviated from the guidelines by recommending an implantable cardioverter-defibrillator (ICD) without considering beta-blockers as the first-line treatment, showcasing a gap in guideline alignment.

Lastly, none of the models demonstrated the ability to incorporate or reference ongoing clinical trials within their responses. This limitation highlights an opportunity for future models to integrate real-time clinical trial data for enhanced decision-making and relevance.

HERE TABLE 6.

Speed and Efficiency. Assessing the speed and efficiency of the models is inherently challenging due to variability in execution time influenced by user load and server conditions. GPT-4.0, GPT-4o, GPT-o1-mini, and GPT-3.5 Turbo exhibited steady response generation, with incremental streaming of outputs, which helped reduce perceived waiting time and allowed for real-time assessment of answers. 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-4o, and GPT-o1-mini, alongside the solid contributions of GPT-3.5 Turbo and Google Bard in various cardiological scenarios, highlights their potential applicability to other medical specialties. These models demonstrate a capacity to synthesize and present complex medical information concisely and accurately, qualities that are invaluable across diverse medical sub-domains. LLMs can quickly analyze large sets of data, including electronic health records and genomic data (8). This allows them to identify patterns not apparent to humans, recognize potential risk factors, and even provide outcome predictions (9). 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. In a hospital environment, they have the potential to serve as a medium for repetitive tasks, such as assisting with writing discharge letters by

summarizing a patient's hospital stay after reading their medical records or providing further recommendations (3,7). 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). 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).

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). Intuition, based on knowledge and care experience, also plays a crucial role (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 the impending incorporation of this AI into daily medical practice is essential to ensure the effective management of knowledge sources and their application.

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. Notably, the open-source architecture allows users to modify the application to interface with more advanced models, such as GPT-4, enabling customization based on individual preferences and budgetary constraints. 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 2, 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. Exploring the Dubravka prototype demonstrates that LLMs have the potential to be integrated into patient-facing tools, streamlining patient education, medical history summaries, and prevention guidance. Although currently focused on lifestyle and prevention, this approach highlights LLMs' adaptability for personalized clinical support.

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. Moreover, while the questions contained information derived from investigative methods such as ECGs and sonography, LLMs were not asked to interpret images directly. 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, GPT-4o and GPT-o1-mini a paid models, 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 study benchmarks the performance of five large language models (LLMs)—GPT-4.0, GPT-4o, GPT-o1-mini, GPT-3.5 Turbo, and Google Bard—on their ability to respond to diverse clinical scenarios in cardiology. The evaluation focused on understanding medical terminology, clinical relevance, contextual accuracy, and guideline adherence.

Among these models, GPT-4.0 consistently demonstrated superior performance, showing high reliability, accuracy, and compliance with contemporary medical guidelines, making it the most suitable for real-world clinical applications. GPT-4o and GPT-o1-mini also performed well, closely following GPT-4.0, while GPT-3.5 Turbo displayed a solid grasp of medical concepts but occasionally diverged from guideline-based recommendations. Google Bard, despite some strengths, showed inconsistencies in contextual understanding and a tendency to deviate from established guidelines.

This study underscores the importance of continual validation of these models across medical specialties, ensuring they remain updated with the latest clinical guidelines and evidence. Moreover, the incorporation of open-source frameworks offers a pathway for cost-effective solutions, enabling broader accessibility while allowing customization for specific clinical settings.

As AI continues to transform healthcare, this research reinforces the critical role of LLMs as tools to augment, not replace, medical professionals. By approaching these technologies with critical engagement and a commitment to ongoing refinement, we can harness their potential to enhance patient care and support clinical decision-making.

Acknowledgments

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 numbers HRZZ-MOBODL-2023-08-7617 and IP-2022-10-7261 (ADESO), and by the University of Zagreb Support, Grant number 10106-24-1505. Additionally, the author(s) would like to acknowledge the contribution of the COST Action CA21169, supported by COST (European Cooperation in Science and Technology).

References

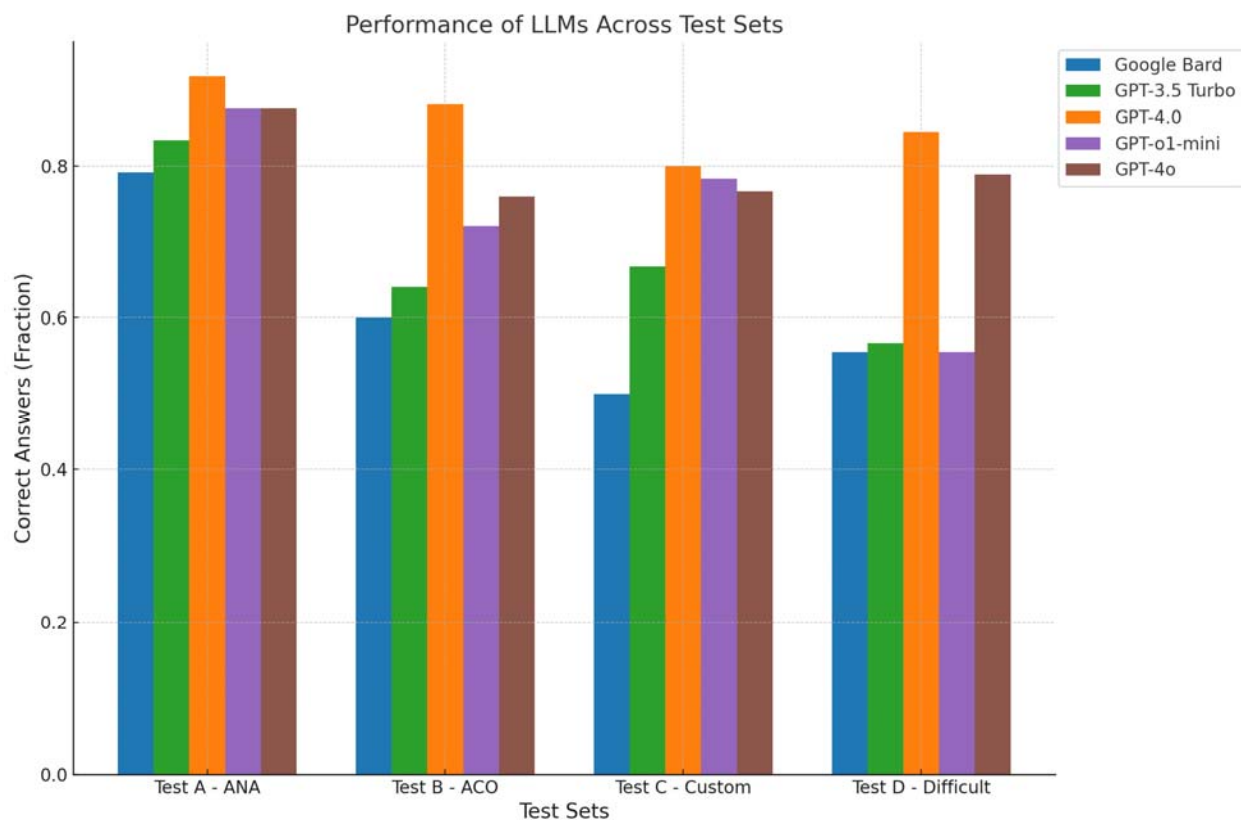
1. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224-6.
2. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large Language Models Encode Clinical Knowledge. 2022;1-44.

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):194.
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; 307(2):e223312.
6. Li Z, Yin S, Wang S, Wang Y, Qiang W, Jiang J. Transformative applications of oculomics-based AI approaches in the management of systemic diseases: A systematic review. *Journal of Advanced Research*. 2024 Nov 13.
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; 46(6):821-2.
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, Rönnerberg L. Clinical intuition in the nursing process and 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. *Int J Cardiol*. 2023;388:131117.
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;90(1):232-8.
16. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. 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; 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;11(6):887.
19. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198.
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;185:109732.
21. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmol Sci*. 2023;3(4):100324.
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, et al. Detection of hypertrophic cardiomyopathy by an artificial intelligence electrocardiogram in children and adolescents. *International Journal of Cardiology*. 2021;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, et al. A machine learning approach to identifying patients with pulmonary hypertension using real-world electronic health records. *International Journal of Cardiology*. 2023;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). *Acta Cardiol*. 2024;79(3):358-66.
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, Gašior 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;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, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*. 2023;6(11):e2343689
33. Cheng SW, Chang CW, Chang WJ, Wang HW, Liang CS, Kishimoto T, et al. The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and clinical neurosciences*. 2023;77(11):592-6.
34. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. *Communications Medicine*. 2023;3(1):141.
35. Arslan S. Exploring the Potential of Chat GPT in Personalized Obesity Treatment. *Ann Biomed Eng*. 2023;51(9):1887-8.

Figure 1. Model Performance Across Different Test Sets. 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.



ANA - American Nurses Association exam queries; ACO - American College of Osteopathic Internists exam queries; Custom - custom dataset queries formulated by the study authors, that mirror the format and objectives of the United States Medical Licensing Examination (USMLE) Step 2 and 3 exams; Difficult - incorporates 90 board-like queries created to follow the structure and outcomes of established American Board of Internal Medicine (ABIM) cardiology board review reference books (30, 31) while being aligned with current AHA guidelines.

Figure 2. The interface of the 'Dubravka' mobile application. 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 the potential to generate tailored advice, with an 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. The images in the first row depict the information that the patient enters, while the second row shows communication with the LLM, and the essential medical history summarized by the LLM based on the given data (first row). The images have been cropped to contain only the relevant information.

The image displays two screenshots of the 'Dubravka' mobile application. The top screenshot shows a registration form titled 'Health insights' for 'DUBRAVA UNIVERSITY HOSPITAL BETA VERSION'. The form includes fields for 'First name' (Jane), 'Last name' (Smith), 'Date of birth' (24 January, 1963), 'Sex' (Female), and 'Ethnicity' (Black / African American). A 'Continue' button is at the bottom, with a note: 'By submitting your info you agree to our [privacy policy](#). This application is exclusive for research.' The bottom screenshot shows the 'Your medical history' section. It contains a 'Medical history' list with two entries: 'Hypertensi' (Time) and 'Deep vein' (Within). Below this is a 'Diagnosis' section with the text 'Deep vein thrombosis'. The 'Covid info' and 'Current medication' sections are collapsed. The 'Current medication' section shows 'Perindopril, Apixaban'. The 'Medicine allergies' section has two radio button options: 'I smoke tobacco.' and 'I drink alcohol.' A 'Continue' button is at the bottom.

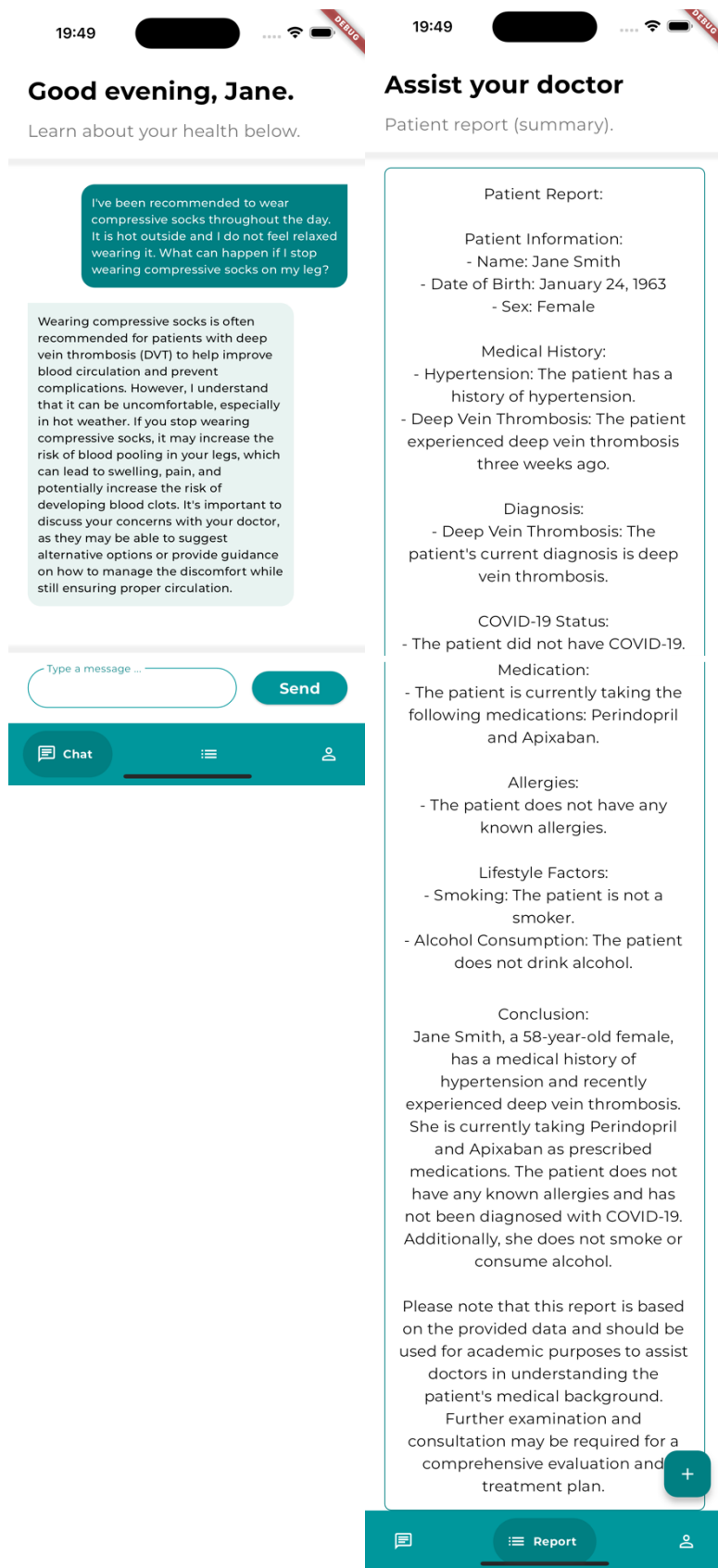


Table 2. (Test set D, Question 3, Chapter 21 in (30)) This table presents a case of a 64-year-old woman with idiopathic dilated cardiomyopathy, her clinical details, and a multiple-choice question on the most suitable treatment option.

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 4o: Correct Answer: A. Cardiac resynchronization is indicated

Reasoning:

- This patient has idiopathic dilated cardiomyopathy with:
- LVEF of 30% (HFrEF).
- Severe functional mitral regurgitation (MR) due to left ventricular (LV) dilation.
- Left bundle branch block (LBBB) on ECG.
- Symptomatic heart failure (NYHA class likely II-III based on medication and findings).

Why Cardiac Resynchronization Therapy (CRT) is Indicated:

- CRT is strongly recommended in patients with HFrEF, LBBB, and QRS ≥ 150 ms to improve LV synchrony and reduce MR severity by addressing the underlying ventricular dyssynchrony (Class I recommendation, ACC/AHA/HFSA and ESC guidelines).
- CRT has been shown to improve symptoms, reduce MR, and enhance reverse remodeling in such patients.

Why Other Options Are Incorrect:

B. The patient should receive mitral valve replacement:

- Incorrect because:
- The MR is functional (secondary) to LV dilation and systolic dysfunction, not a primary valvular issue.
- Addressing the underlying LV dysfunction with CRT is the first-line approach, and valve replacement is not indicated unless severe MR persists after optimal medical/CRT therapy.

C. The patient should receive annuloplasty repair:

- Incorrect because:
- Annuloplasty is a surgical intervention typically reserved for refractory functional MR after failing guideline-directed medical therapy and CRT.
- In this patient, addressing dyssynchrony with CRT is likely to reduce MR severity without the need for surgery.

D. Transcatheter mitral valve intervention should be performed:

- Incorrect because:
- Transcatheter options (e.g., MitraClip) are primarily used in patients with prohibitive surgical risk and persistent severe MR despite optimized medical therapy and CRT.

- This patient has no signs of prohibitive surgical risk and has not yet undergone CRT.

Summary: The most appropriate first-line intervention is CRT to address dyssynchrony and potentially improve both LV function and secondary MR. Options B, C, and D are either invasive or unnecessary at this stage, given the likelihood of improvement with CRT.

GPT-01-mini: The most correct answer is A. Cardiac resynchronization is indicated.

- A. Cardiac resynchronization is indicated: Correct. The patient has LBBB and HFrEF (LVEF <35%), both of which are indications for CRT. CRT improves symptoms, quality of life, and survival in patients with LBBB and reduced ejection fraction, and can reduce mitral regurgitation.
- B. The patient should receive mitral valve replacement: Incorrect. MVR is typically considered if medical therapy fails or in cases of severe symptomatic MR. This patient is stable with medical therapy, so MVR is not the first-line treatment.
- C. The patient should receive annuloplasty repair: Incorrect. Annuloplasty is considered for functional MR, but in patients with severe MR and a dilated mitral annulus, it may not be sufficient. The first-line treatment is usually CRT.
- D. Transcatheter mitral valve intervention should be performed: Incorrect. TMVr is an option for patients who are not surgical candidates. However, this patient is a candidate for CRT and possible surgical intervention, so TMVr is not the primary option.

Conclusion: the most correct answer is A. Cardiac resynchronization is indicated. The patient has LBBB with severe heart failure and a low ejection fraction (<35%), which meets the indications for CRT according to current guidelines (e.g., AHA/ACC/HRS).

B (mitral valve replacement) and C (annuloplasty repair) are possible future options, but not necessarily first-line treatments in the context of a dilated mitral annulus and low ejection fraction. D (transcatheter mitral valve intervention) would be considered only if surgical options are contraindicated or ineffective.

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 \leq 35%, symptomatic heart failure (NYHA class II, III, or ambulatory IV), sinus rhythm, and left bundle branch block with a QRS duration \geq 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, GPT-4o, GPT-o1-mini, and GPT-3.5 Turbo all agree on option A. They justify this choice based on the patient's presentation of left bundle branch block (LBBB) and reduced ejection fraction (LVEF), indicating dyssynchronous ventricular contraction. Cardiac resynchronization therapy (CRT) is recommended in such cases as it can improve left ventricular synchrony, reduce mitral regurgitation (MR), and enhance overall cardiac function. On the other hand, Google Bard suggests option B, advocating for mitral valve replacement due to the severe MR and significantly reduced LVEF. However, this approach overlooks that the MR is secondary (functional) to left ventricular dilation and not a primary valvular issue. Current guidelines prioritize CRT for addressing the underlying dyssynchrony and improving symptoms before considering invasive surgical interventions.

Table 3. (Test set D, Question 4, Chapter 4 in (30)) 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 325 mg of aspirin 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 4o: Correct Answer: B. Continue atorvastatin 80 and start ezetimibe 10

Explanation: This patient has experienced an acute coronary syndrome (ACS) with proximal LAD stenosis, successfully treated with PCI and stent placement. Management of dyslipidemia post-ACS follows established guidelines, emphasizing intensive LDL-C lowering to reduce future cardiovascular risk.

Option Analysis:

A. Continue atorvastatin 80:

- Incorrect: Current LDL-C is 80 mg/dL, which does not meet the recommended target of <55 mg/dL for very high-risk patients (e.g., ACS, prior MI).
- Although atorvastatin 80 mg is part of high-intensity statin therapy, adding another agent (e.g., ezetimibe) is required to achieve further LDL-C lowering.

B. Continue atorvastatin 80 and start ezetimibe 10:

- Correct: Per ACC/AHA and ESC guidelines:
- In very high-risk patients who do not achieve LDL-C <55 mg/dL with high-intensity statin alone, ezetimibe is the preferred second-line therapy.
- Ezetimibe inhibits intestinal cholesterol absorption, works synergistically with statins, and lowers LDL-C by an additional ~20%.
- This regimen is the first step before considering PCSK9 inhibitors like evolocumab if LDL-C remains above target.

C. Continue atorvastatin 80 and start fenofibrate 130:

- Incorrect: Fenofibrate targets hypertriglyceridemia and is not used to reduce LDL-C. While the patient's triglycerides (189 mg/dL) are slightly elevated, they do not meet the threshold for fibrate initiation (>500 mg/dL, to reduce pancreatitis risk).
- Addressing LDL-C remains the priority in post-ACS management.

D. Stop atorvastatin and start evolocumab:

- Incorrect: PCSK9 inhibitors (like evolocumab) are reserved for patients who fail to achieve LDL-C targets with maximal statin plus ezetimibe therapy or are intolerant to statins.
- Stopping atorvastatin would remove the cornerstone of therapy and is not indicated.

Conclusion: The best approach is to continue atorvastatin 80 mg and add ezetimibe 10 mg, optimizing LDL-C reduction to achieve the target <55 mg/dL while maintaining established guideline-recommended therapy.

GPT-o1-mini: The best option is B. Continue atorvastatin 80 and start ezetimibe 10.

- A. Continue atorvastatin 80: Partially correct. Atorvastatin 80 mg is appropriate for secondary prevention in a patient with acute coronary syndrome (ACS). However, the patient's LDL-C level of 80 mg/dL is not low enough to optimally reduce cardiovascular risk, especially given that current guidelines suggest an LDL target of <70 mg/dL after ACS.
- B. Continue atorvastatin 80 and start ezetimibe 10: Correct. This combination therapy is appropriate. Ezetimibe can further lower LDL-C by inhibiting cholesterol absorption in the intestine, providing additional benefit beyond statin therapy. This has been shown to reduce major adverse cardiovascular events (MACE) in patients after an ACS, especially in those who do not reach optimal LDL-C levels with statin therapy alone.
- C. Continue atorvastatin 80 and start fenofibrate 130: Incorrect. Fenofibrate is primarily used to lower triglycerides and increase HDL-C, but elevated triglycerides (189 mg/dL) alone are not an indication for fibrates unless there is a concern for pancreatitis or severe hypertriglyceridemia. It is also less effective than ezetimibe in reducing LDL-C, which is the primary target for secondary prevention in ACS.
- D. Stop atorvastatin and start evolocumab: Incorrect. Evolocumab (a PCSK9 inhibitor) is appropriate for patients with severe hyperlipidemia or those who cannot tolerate statins. However, this patient is on atorvastatin 80 mg, which is already a high-intensity statin. Evolocumab should be considered if the patient cannot achieve LDL-C goals with statins alone or has statin intolerance, neither of which applies in this case.

Conclusion: The best regimen is B. Continue atorvastatin 80 and start ezetimibe to further lower LDL-C and achieve better lipid control post-MI, based on current guidelines for secondary prevention in ACS.

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 $\geq 50\%$ 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, GPT-4o, GPT-o1-mini, and GPT-3.5 Turbo all provide accurate analyses, recognizing the patient is post-myocardial infarction (MI) with an LDL-C level of 80 mg/dL, which is above the target of <70 mg/dL for secondary prevention in patients with atherosclerotic cardiovascular disease (ASCVD). GPT-4.0, GPT-4o, and GPT-o1-mini correctly suggest adding ezetimibe to atorvastatin 80 mg, aligning with the 2018 AHA/ACC cholesterol guidelines for very high-risk ASCVD patients when LDL-C remains ≥ 70 mg/dL despite maximally tolerated statin therapy. In contrast, GPT-3.5 Turbo reviews the options thoroughly but incorrectly selects option A (continue atorvastatin 80 mg alone) without recommending the addition of ezetimibe, failing to meet guideline-directed targets. Google Bard makes several errors. First, it inaccurately classifies atorvastatin 80 mg as a moderate-intensity statin when it is actually high-intensity. Additionally, it incorrectly asserts that prasugrel increases the risk of muscle injury, which is not supported as prasugrel is an antiplatelet drug without direct myotoxic effects. Google Bard also suggests evolocumab (option D) as the preferred therapy, overlooking the stepwise approach recommended by guidelines, which prioritize ezetimibe before PCSK9 inhibitors in cases where LDL-C remains elevated despite statin and ezetimibe therapy.

Table 4. (Test set D, Question 8, Chapter 17 in (30)) This table presents 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 4o: Correct Answer: C. Family history of sudden cardiac death

Explanation: Hospital admission for cardiac evaluation is indicated when clinical findings suggest acute or life-threatening conditions or significant pathology requiring urgent investigation or intervention. A family history of sudden cardiac death is a significant risk factor but does not, by itself, necessitate immediate hospital admission without other concerning symptoms, signs, or ECG findings.

Option Analysis:

A. ECG demonstrating sinus pause of 2 s:

- Requires Admission:
- A sinus pause ≥ 2 seconds is abnormal and could indicate sick sinus syndrome or conduction system disease, warranting inpatient telemetry monitoring and further evaluation to assess the risk of syncope or sudden cardiac arrest.

B. Physical examination demonstrating elevated jugular venous pressure, rales on lung examination, and lower extremity edema:

- Requires Admission:
- These findings suggest acute decompensated heart failure, which necessitates hospital admission for diuresis, hemodynamic stabilization, and further workup for underlying causes.

C. Family history of sudden cardiac death:

- Does Not Require Immediate Admission:
- While significant, this finding alone does not indicate an acute issue. It warrants outpatient evaluation, including risk stratification with ECG, echocardiography, and potentially advanced imaging or genetic testing if there are concerning findings.

D. Physical examination demonstrating parvus et tardus, systolic murmur heard best at upper sternal border, and absent S2:

- Requires Admission:
- These findings are consistent with severe aortic stenosis, which can lead to syncope, heart failure, or sudden death. Hospital evaluation is necessary if symptomatic or if severe stenosis is suspected.

E. ECG with right bundle branch block (RBBB) and ST elevations in V1–V3:

- Requires Admission:
- This pattern is characteristic of Brugada syndrome, a condition associated with ventricular arrhythmias and sudden cardiac death. Patients require inpatient evaluation with telemetry

monitoring and potentially an electrophysiology study.

Conclusion: While a family history of sudden cardiac death is a significant risk factor, it does not independently require immediate hospitalization without accompanying acute findings. Other options involve clinical features indicative of potentially life-threatening cardiac conditions necessitating urgent evaluation.

GPT-o1-mini: The correct answer is C. Family history of sudden cardiac death.

Explanation:

- A. ECG demonstrating sinus pause of 2 s: Correct. A sinus pause of 2 seconds can be indicative of sinus node dysfunction, which may predispose to more prolonged pauses and lead to syncope or arrhythmias. This can be a serious condition, and the patient should be admitted for further evaluation, including monitoring and possibly pacemaker placement.
- B. Physical examination demonstrating elevated jugular venous pressure, rales on lung examination, and lower extremity edema: Correct. These findings suggest heart failure, which requires urgent evaluation and management. Admission is warranted for stabilization, monitoring, and treatment (e.g., diuretics, inotropes, etc.).
- C. Family history of sudden cardiac death: Incorrect. A family history of sudden cardiac death is concerning for genetic arrhythmias (e.g., arrhythmogenic right ventricular cardiomyopathy, long QT syndrome, hypertrophic cardiomyopathy), but it is not an immediate reason for hospital admission unless the patient demonstrates signs or symptoms of a current arrhythmic event, such as syncope or palpitations. Admission would be indicated if the patient had symptoms or electrocardiographic findings of arrhythmias.
- D. Physical examination demonstrating parvus et tardus, systolic murmur heard best at upper sternal border and absent S2: Correct. These findings suggest severe aortic stenosis. This can lead to significant hemodynamic compromise and may require urgent intervention (e.g., balloon valvuloplasty or valve replacement), making hospital admission necessary for further management.
- E. ECG with Right bundle branch block and ST elevations in V1–V3: Correct. ST elevations in V1-V3 with right bundle branch block suggest the possibility of Brugada syndrome or acute myocardial injury and can be associated with ventricular arrhythmias. Admission is necessary for monitoring and further evaluation of the underlying cause (e.g., arrhythmias, myocardial infarction, or Brugada syndrome).

Conclusion: C. Family history of sudden cardiac death is the correct answer as it does not immediately require hospital admission unless accompanied by symptoms or significant arrhythmias on examination. The other findings indicate potentially life-threatening conditions that necessitate hospitalization for further evaluation and management.

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 and GPT-o1-mini, option C (family history of sudden cardiac death) does not require immediate hospital admission. While significant as a risk factor for inherited arrhythmias or structural heart disease, it does not indicate an acute or life-threatening condition requiring urgent inpatient evaluation. GPT-4o supports this reasoning, emphasizing that admission is warranted only when acute symptoms or concerning findings accompany the family history. GPT-3.5 Turbo, however, identifies option D (parvus et tardus and related findings) as not requiring immediate hospital admission, suggesting that this could represent chronic aortic stenosis, which might be managed on an outpatient basis unless acute complications arise. In contrast, GPT-4.0, GPT-4o, and GPT-o1-mini correctly point out that severe aortic stenosis can lead to life-threatening complications and warrants urgent evaluation if symptomatic or if severe stenosis is suspected. Google

Bard suggests option B (elevated jugular venous pressure, rales, and lower extremity edema) as not requiring hospital admission, which is less likely to be correct. These findings strongly suggest acute decompensated heart failure, a condition that typically requires immediate hospital admission for stabilization and further management.

Table 5. (Test set C) 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 reveals 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 full 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 4o: Correct Answer: D. Intravenous phentolamine

Explanation: This patient presents with chest pain, severe hypertension, and a history of recent cocaine use, which strongly suggests cocaine-induced vasospasm (acute coronary syndrome secondary to vasoconstriction). Management focuses on reversing vasospasm, controlling blood pressure, and providing symptomatic relief. Let's analyze each option:

Option Analysis:

A. Aspirin and clopidogrel at loading dose:

- Incorrect: While antiplatelet agents are crucial in managing typical acute coronary syndromes, they do not address the underlying cocaine-induced vasospasm. This patient has no ECG or clinical evidence of acute thrombosis or ischemia requiring these medications initially.

B. Intravenous lorazepam:

- Partially Correct: Benzodiazepines like lorazepam are effective in reducing sympathetic stimulation, alleviating anxiety, and lowering blood pressure in cocaine toxicity.
- However, they are not sufficient alone to reverse vasospasm or manage the hypertensive crisis in this case.

C. Intravenous metoprolol:

- Incorrect: Beta-blockers (even cardioselective ones like metoprolol) are contraindicated in cocaine-related chest pain due to the risk of unopposed alpha-adrenergic vasoconstriction, which can worsen vasospasm and lead to ischemia.

D. Intravenous phentolamine:

- Correct: Phentolamine, an alpha-adrenergic blocker, is the treatment of choice for severe hypertension and vasospasm associated with cocaine use.
- It directly counteracts cocaine-induced vasoconstriction and reduces blood pressure effectively.

E. Low-molecular-weight heparin:

- Incorrect: Anticoagulation is not indicated unless there is evidence of a thrombotic

event such as acute myocardial infarction, which is not supported by this patient's clinical presentation or ECG findings.

Conclusion: The most appropriate next step is intravenous phentolamine to counteract cocaine-induced vasospasm and manage hypertensive urgency. Lorazepam may be used adjunctively for anxiety, but beta-blockers and anticoagulants should be avoided unless there is clear evidence of ischemia or thrombosis.

GPT o1-mini: The correct answer is D. Intravenous phentolamine.

Explanation: This patient presents with chest pain, hypertension, and cocaine use. Cocaine is known to increase sympathetic activity, which leads to vasoconstriction, increased heart rate, and elevated blood pressure. It can also lead to coronary artery vasospasm, which can cause myocardial ischemia and chest pain. The key clinical finding in this case is severe hypertension (200/105 mmHg), sinus tachycardia, and symptoms consistent with cocaine-induced chest pain.

Now, let's break down the options:

- A. Aspirin and clopidogrel at loading dose:
 - This would typically be appropriate in cases of acute coronary syndrome (ACS) (e.g., myocardial infarction), but the patient's ECG does not show any significant ischemic changes (e.g., ST-segment elevations or depressions). Moreover, cocaine-induced chest pain is more commonly related to coronary vasospasm, not atherothrombotic occlusion. Thus, this option is not the most appropriate at this moment.

- B. Intravenous lorazepam:
 - This is used for the management of severe anxiety or benzodiazepine withdrawal, but it is not the first line in treating cocaine-induced chest pain. While cocaine use may contribute to anxiety, the hypertensive crisis and coronary vasospasm need to be addressed more urgently.

- C. Intravenous metoprolol:
 - Beta-blockers like metoprolol are typically used to control heart rate and reduce myocardial oxygen demand in acute coronary syndrome. However, beta-blockers can worsen vasospasm in the setting of cocaine-induced chest pain, because cocaine also blocks the reuptake of norepinephrine, which further increases sympathetic tone. Therefore, beta-blockers should be avoided in this scenario.

- D. Intravenous phentolamine:
 - Phentolamine is an alpha-adrenergic blocker used to counteract the effects of sympathomimetic drugs like cocaine. It relieves vasoconstriction by blocking alpha-1 adrenergic receptors, thus helping to reduce hypertension and prevent coronary vasospasm. This is the most appropriate treatment for cocaine-induced chest pain with severe hypertension and suspected coronary vasospasm.

- E. Low-molecular-weight heparin:
 - This would be used if there was a concern for acute coronary syndrome (ACS) with myocardial infarction or thrombosis, but there is no strong evidence in this case of an atherosclerotic event or thrombosis. The primary issue here is cocaine-induced vasospasm, not a thrombotic event.

Conclusion: D. Intravenous phentolamine is the most appropriate next step in this patient's management because it directly addresses the cocaine-induced vasospasm and hypertension, making it the correct option.

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-4o, GPT-o1-mini, 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. (Test set B) 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 4o: Correct Answer: A. Patient needs to be started on beta blocker and followed for symptoms

Explanation:

Miss Johnson has a prolonged QT interval (>470 ms), which raises concerns about Long QT Syndrome (LQTS), a condition associated with a risk of torsades de pointes and sudden cardiac death. Although she is currently asymptomatic, appropriate management is crucial to reduce the risk of life-threatening arrhythmias.

Option Analysis:

A. Patient needs to be started on beta blocker and followed for symptoms:

- Correct: Beta blockers (e.g., propranolol or nadolol) are first-line therapy for LQTS to reduce the risk of arrhythmias by blunting adrenergic stimulation.
- Even asymptomatic patients with a prolonged QT interval benefit from beta blockers as they significantly reduce mortality and arrhythmic events.
- Close follow-up is essential to monitor symptoms and QT interval changes.

B. Patient needs no further treatment since asymptomatic:

- Incorrect: Asymptomatic status does not eliminate the risk of sudden cardiac death in LQTS.
- Treatment is indicated for individuals with a prolonged QT interval >470 ms, even in the absence of symptoms.

C. Patient is a candidate for biventricular pacemaker set at heart rate faster than her intrinsic rate:

- Incorrect: Biventricular pacemakers are not the standard treatment for LQTS.
- Pacemakers may be considered in specific cases where there is bradycardia-induced QT prolongation, but this is rare and not the primary management strategy.

D. Patient is a candidate for ICD:

- Incorrect: ICDs are reserved for high-risk patients with LQTS who have:
- A history of cardiac arrest or ventricular arrhythmias.
- Syncope despite optimal beta blocker therapy.
- Asymptomatic patients without high-risk features do not require ICD placement.

E. Patient needs ETT first to make further decisions:

- Incorrect: Exercise treadmill testing (ETT) is sometimes used to assess QT interval dynamics under stress, but it is not a prerequisite for initiating treatment in patients with a clearly prolonged QT interval on baseline ECG.

Conclusion:

The best treatment option for this patient is starting beta blockers and close follow-up for symptoms or further QT interval prolongation. This proactive approach significantly reduces the risk of fatal arrhythmias in LQTS. Other interventions, such as ICDs, are reserved for symptomatic or high-risk individuals.

GPT o1-mini: The correct answer is A. Patient needs to be started on beta blocker and followed for symptoms.

Explanation: Miss Johnson is a 34-year-old woman with long QT interval >470 ms on her ECG, which suggests long QT syndrome (LQTS), a condition that increases the risk of torsades de pointes, a potentially life-threatening arrhythmia. The management of LQTS typically involves:

1. Beta-blockers: The primary treatment for asymptomatic long QT syndrome is to start beta-blockers, particularly in patients who are at risk for arrhythmias. Beta-blockers are used to prevent arrhythmic events by reducing sympathetic stimulation, which is a trigger for QT interval prolongation.
2. Follow-up for symptoms: Even if the patient is currently asymptomatic, beta-blockers are indicated because they can help reduce the risk of sudden cardiac death and arrhythmias. The patient should be monitored for any future symptoms, and adjustments to therapy can be made if necessary.

Let's break down the options:

- A. Patient needs to be started on beta blocker and followed for symptoms:
 - This is correct because beta-blockers are first-line treatment for patients with long QT syndrome, even if asymptomatic. These drugs help to prevent arrhythmias by reducing sympathetic tone and preventing excessive prolongation of the QT interval during physical or emotional stress.
 - B. Patient needs no further treatment since asymptomatic:
 - This is incorrect because even asymptomatic patients with prolonged QT intervals are at risk for arrhythmias, including torsades de pointes. Treatment with beta-blockers is recommended to reduce the risk, even in the absence of symptoms.
 - C. Patient is a candidate for biventricular pacemaker set at heart rate faster than her intrinsic rate:
 - This is incorrect because biventricular pacemakers are generally used for heart failure with reduced ejection fraction or conduction abnormalities like heart block. Long QT syndrome does not typically require a pacemaker unless there is an associated conduction defect or significant symptomatic bradycardia, which is not indicated here.
 - D. Patient is a candidate for ICD:
 - This is incorrect because the implantable cardioverter-defibrillator (ICD) is typically indicated for patients who have had sustained arrhythmias, a history of syncope, or high-risk features for arrhythmias (such as a history of cardiac arrest or family history of sudden cardiac death). This patient is asymptomatic, and an ICD is not recommended without further symptomatic evidence or a history of arrhythmic events.
 - E. Patient needs ETT first to make further decisions:
 - This is incorrect. While an exercise tolerance test (ETT) might be useful in some clinical situations to evaluate the heart's response to stress, it is not necessary for the management of long QT syndrome. The primary treatment decision is based on the QT interval and the presence of symptoms or arrhythmic risk.

Conclusion: The most appropriate treatment for this patient is starting beta-blockers and monitoring for any future symptoms, making A the correct choice.

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, GPT-4o, GPT-o1-mini, 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.