Original Article

Evaluating Accuracy and Readability of Responses to Midlife Health Questions: A Comparative Analysis of Six Large Language Model Chatbots

Himel Mondal, Devendra Nath Tiu[1], Shaikat Mondal[2], Rajib Dutta[3], Avijit Naskar[4], Indrashis Podder[5]

Department of Physiology, All India Institute of Medical Sciences, Deoghar, [1]Department of Physiology, Sheikh Bhikhari Medical College, Hazaribagh, Jharkhand, [2]Department of Physiology, Raiganj Government Medical College and Hospital, Raiganj, [3]Department of Gynecology and Obstetrics, Diamond Harbour Government Medical College and Hospital, Diamond Harbour, [4]Department of General Medicine, Baruipur Sub-Divisional Hospital, Baruipur, [5]Department of Dermatology, College of Medicine and Sagore Dutta Hospital, Kolkata, West Bengal, India

Submitted: 06-Oct-2024
Revised: 22-Nov-2024
Accepted: 02-Dec-2024
Published: 05-Apr-2025

ABSTRACT

Background: The use of large language model (LLM) chatbots in health-related queries is growing due to their convenience and accessibility. However, concerns about the accuracy and readability of their information persist. Many individuals, including patients and healthy adults, may rely on chatbots for midlife health queries instead of consulting a doctor. In this context, we evaluated the accuracy and readability of responses from six LLM chatbots to midlife health questions for men and women. **Methods:** Twenty questions on midlife health were asked to six different LLM chatbots – ChatGPT, Claude, Copilot, Gemini, Meta artificial intelligence (AI), and Perplexity. Each chatbot's responses were collected and evaluated for accuracy, relevancy, fluency, and coherence by three independent expert physicians. An overall score was also calculated by taking the average of four criteria. In addition, readability was analyzed using the Flesch-Kincaid Grade Level, to determine how easily the information could be understood by the general population. **Results:** In terms of fluency, Perplexity scored the highest (4.3 ± 1.78), coherence was highest for Meta AI (4.26 ± 0.16), accuracy of responses was highest for Meta AI, and relevancy score was highest for Meta AI (4.35 ± 0.24). Overall, Meta AI scored the highest (4.28 ± 0.16), followed by ChatGPT (4.22 ± 0.21), whereas Copilot had the lowest score (3.72 ± 0.19) ($P < 0.0001$). Perplexity showed the highest score of 41.24 ± 10.57 in readability and lowest in grade level (11.11 ± 1.93), meaning its text is the easiest to read and requires a lower level of education. **Conclusion:** LLM chatbots can answer midlife-related health questions with variable capabilities. Meta AI was found to be highest scoring chatbot for addressing men's and women's midlife health questions, whereas Perplexity offers high readability for accessible information. Hence, LLM chatbots can be used as educational tools for midlife health by selecting appropriate chatbots according to its capability.

KEYWORDS: *Artificial intelligence, chatbots, health education, large language models, midlife health, patient education, patient queries*

INTRODUCTION

Large language models (LLMs) are increasingly being integrated into various sectors, including health care, due to their ability to process and generate human-like text.[1] With advancements in artificial intelligence (AI), chatbots powered by LLMs have become popular tools for providing quick, accessible responses to a wide range of medical queries.[2-4] For many individuals, these chatbots represent a convenient alternative to medical consultations, offering instant advice on health-related concerns without the need for appointments or in-person visits.[5]

Access this article online

Quick Response Code:

Website: https://journals.lww.com/jomh

DOI: 10.4103/jmh.jmh_182_24

Address for correspondence: Dr. Himel Mondal, Department of Physiology, All India Institute of Medical Sciences, Deoghar - 814 152, Jharkhand, India.
E-mail: himelmkcg@gmail.com

How to cite this article: Mondal H, Tiu DN, Mondal S, Dutta R, Naskar A, Podder I. Evaluating accuracy and readability of responses to midlife health questions: A comparative analysis of six large language model chatbots. J Mid-life Health 2025;16:45-50.

As men and women transition through midlife, they may encounter new or evolving health challenges, such as hormonal changes, cardiovascular issues, and metabolic shifts.[6] While many may seek information from healthcare professionals, others turn to digital sources, including AI-driven chatbots, for quick answers.[7] However, the quality of these chatbot responses, particularly their medical accuracy and readability, remains uncertain. Inaccurate or poorly presented information can lead to confusion, misinformation, or even harmful decisions.[8]

Despite the growing reliance on chatbots, few studies have rigorously assessed the performance of LLM chatbots in providing reliable health advice.[9-13] Other studies still think that the models need further improvement to act as a health guide. Given the critical nature of accurate health information, it is essential to evaluate how well these tools perform in delivering medically sound and easy-to-understand responses.[14,15]

With this background, this study aimed to address the gap by evaluating the accuracy and readability of responses from six LLM chatbots that provide free access to the population.

Methods

Type and setting

This study is a cross-sectional study and was conducted in a virtual setting where questions were posted to the chatbots, and responses were systematically analyzed for accuracy and readability. As this study does not involve any human or animal participants, the study does not require any ethical clearance. The study was conducted in September 2024.

Framing questions

A set of 20 questions was developed, focusing on key aspects of midlife health for both men and women. These questions covered common concerns in midlife, including hormonal changes, cardiovascular risks, bone health, metabolic changes, mental health, and lifestyle recommendations. The questions were designed to reflect real-world queries that individuals in midlife might ask a healthcare provider. The questions were set by three physicians and they came to a consensus to finalize the questions [Annexure 1].

Choosing chatbots

We only included the chatbot that provides free or limited free access to any users. Six LLM chatbots were selected based on their popularity, accessibility, and relevance in the health sector. These chatbots have previously been explored for healthcare-related studies. The chatbots are arranged according to the first alphabet. They were ChatGPT, Claude, Copilot, Gemini, Meta AI, and Perplexity.[16-18]

Generating response

Each of the 20 midlife health questions was posted to six LLM chatbots in an identical format. The responses were collected in real time and stored for further analysis. To maintain consistency, the same questions were asked on the same day (September 10, 2024) across all chatbots to avoid potential updates or changes in LLM algorithms over time. No follow-up questions or clarifications were asked; only the initial response to each question was evaluated.

Rating by physicians

A panel of three expert physicians evaluated the accuracy, relevancy, fluency, and coherence of the responses. Each physician independently rated the responses based on a predefined Likert-like scale for each of four attributes as shown in Figure 1. The physicians were blinded to the identity of the chatbots during the evaluation process to prevent bias.

Readability analysis

The readability of the responses was calculated by the "Flesch-Kincaid Readability Calculator" available free at https://goodcalculators.com/flesch-kincaid-calculator. This calculator provides Flesch Reading Ease score (measures how readable the text is) and Flesch-Kincaid Grade Level (measures the educational level a person needs to understand the

Parameter	Characteristic	Score
Fluency	Exceptionally fluent with precise and clear language	5
	Very fluent but with minor language issues	4
	Generally fluent, but occasional language concerns	3
	Noticeable language issues affecting comprehension	2
	Poor fluency, making the content hard to understand	1
Coherence	Highly logical and well-structured	5
	Mostly logical, with minor coherence issues	4
	Generally logical, but some areas lack coherence	3
	Noticeable coherence issues impacting understanding	2
	Poor logical flow, making the content difficult to follow	1
Accuracy	Highly accurate, supported by robust scientific evidence	5
	Mostly accurate, with minor factual errors	4
	Generally accurate, but with notable inaccuracies	3
	Several factual errors impacting reliability	2
	Inaccurate information, undermining credibility	1
Relevancy	Highly relevant and applicable to the target audience	5
	Mostly relevant, with some areas needing improvement	4
	Generally relevant, but some content may not be useful	3
	Limited relevance; substantial improvements needed	2
	Content lacks relevance and usefulness	1
Overall score: (Score of Fluency + Coherence + Accuracy + Relevancy)/4		

Figure 1: Rating criteria and score used by the human raters to rate the text generated by chatbots

text) along with other characteristics such as number of sentences, words, average words per sentence, and average syllables per word.

Data analysis

The ratings from the three expert physicians were averaged for each response to create an "average" score for each chatbot. For example, to calculate the average fluency score, the final score for each question was the average of three individual scores. The overall fluency level of a chatbot was then determined by averaging the scores of all 20 questions. The "overall" score was calculated by taking the average score of four characteristics as shown in Figure 1. Descriptive statistics were used to summarize the accuracy scores and readability metrics. Comparative analysis was conducted using ANOVA to determine if significant differences existed between the accuracy and readability of the six LLM chatbots. Correlation analysis was also performed to explore any relationship between accuracy and readability. We used Microsoft Excel 2021 and GraphPad Prism 9.5.0 (GraphPad Software, Boston, USA) for statistical analysis.

Results

The average score of each of the four attributes is shown in Figure 2a-d. In terms of fluency, Perplexity scored the highest (4.3 ± 1.78), followed by Meta AI (4.28 ± 0.15), while Gemini scored the lowest (3.81 ± 0.37) ($P < 0.0001$). Coherence was highest for Meta AI (4.26 ± 0.16) and lowest for Gemini (3.74 ± 0.41) ($P < 0.0001$). The accuracy of responses was highest for Meta AI (4.25 ± 0.31), followed by ChatGPT (4.24 ± 0.35), with the lowest score for Copilot (3.59 ± 0.32) ($P < 0.0001$). The relevancy score was highest for Meta AI (4.35 ± 0.24), followed by ChatGPT (4.24 ± 0.24), with the lowest for Copilot (3.49 ± 0.48) ($P < 0.0001$).

The overall score is shown in Figure 3. Meta AI scored the highest (4.28 ± 0.16), followed by ChatGPT (4.22 ± 0.21), whereas Copilot had the lowest score (3.72 ± 0.19) ($P < 0.0001$).

The linguistic characteristics are shown in Table 1. The ChatGPT generates the most sentences, followed by Perplexity. On the other hand, Copilot produces the fewest sentences, indicating a significantly more concise output. In terms of word count, ChatGPT leads again, while Claude follows. Copilot also generates the fewest words. Perplexity showed the highest score of 41.24 in readability, meaning its text is the easiest to read. This is followed by Gemini and Meta AI. ChatGPT, while generating more content, has the lowest Ease score at 21.28, making its responses harder to read compared to the other chatbots. The grade level required to understand the responses further supports this, with ChatGPT's text needing a higher education level (average grade 14.12 ± 1.23), closely followed by Copilot (14.28 ± 1.64). Conversely, Perplexity generates the simplest text, requiring a lower grade level, making it the most accessible chatbot in terms of readability.

Discussion

The evaluation across six chatbots showed that Meta AI consistently performed the best in coherence, accuracy, and relevancy, making it the most suited chatbot for getting answers related to midlife health. This makes it particularly suitable for users seeking comprehensive and well-structured content. In terms of overall performance,

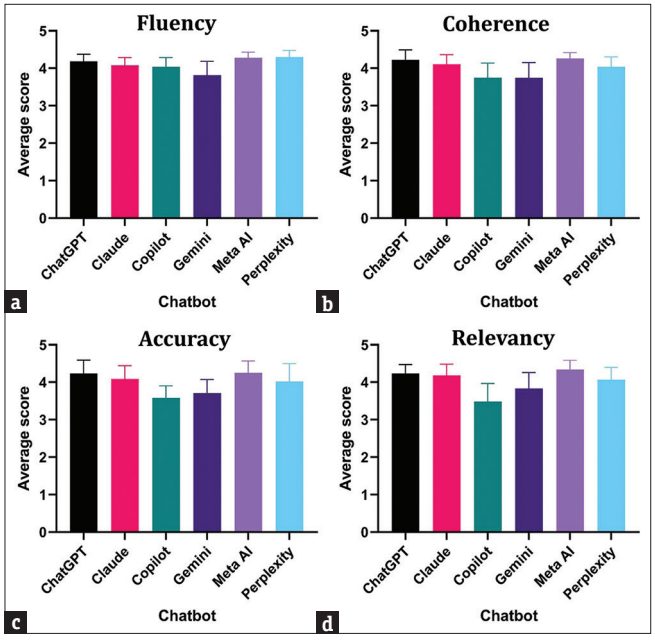

Figure 2: Score of chatbots in fluency (a), coherence (b), accuracy (c), and relevancy (d)

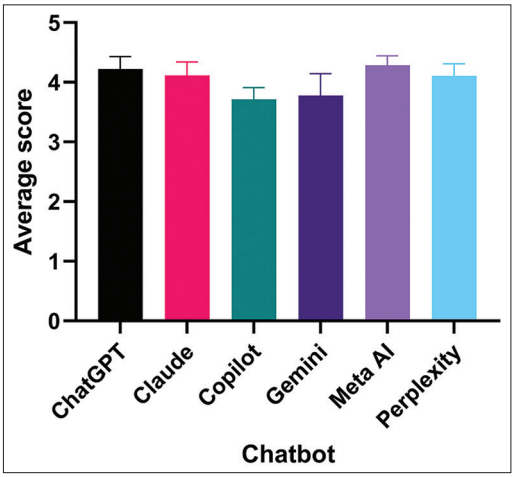

Figure 3: Overall score of chatbots

Table 1: Linguistic characteristic of the responses of chatbots

Chatbot	Sentence	Words	Reading ease score*	Grade†
ChatGPT	20.45±6.04	278.7±68.24	21.28±7.46	14.12±1.23
Claude	17.35±4.72	208.1±19	27.77±11.1	12.52±1.7
Copilot	4.7±1.38	68.9±17.59	24.91±11.61	14.28±1.64
Gemini	13.85±6	168.8±62.41	36.91±12.51	11.44±1.78
Meta AI	15.15±4.48	172.75±34.9	33.99±15.7	11.6±2.95
Perplexity	18.05±11.52	195±116.19	41.24±10.57	11.11±1.93
ANOVA *P* value	<0.0001	<0.0001	<0.0001	<0.0001

*Flesch Reading Ease Score=206.835–1.015 × (total words/total sentences) − 84.6 × (total syllables/total words); score range 0–100; higher the score, easier to read, †Flesch-Kincaid Grade Level=0.39 × (total words/total sentences) + 11.8 × (total syllables/total words) − 15.59; United States grade levels of education; score indicates year of education needed to understand the text. Calculation was carried out from: https://goodcalculators.com/flesch-kincaid-calculator. AI: Artificial intelligence

Meta AI ranked the highest, followed closely by ChatGPT. Copilot, on the other hand, underperformed in accuracy and relevancy, earning the lowest overall score. While ChatGPT produced more sentences and words than the others, it also had the lowest readability score, whereas Perplexity generated the most readable and accessible text. Perplexity's high readability makes it ideal for users who prioritize ease of understanding, whereas ChatGPT, despite generating more content, might be less useful for users seeking concise and easy-to-read responses. Copilot's lower performance suggests that it may not be the best option for generating complex, coherent, or highly accurate content. Other studies have reported variable levels of quality of response.[19-21]

Meta AI's strong performance likely results from advanced language models trained on diverse, high-quality data, enabling it to generate well-structured and accurate responses. Perplexity's focus on readability suggests that it may prioritize clarity over depth, which contributes to its more user-friendly output. ChatGPT's worldliness, while generating more content, compromises readability, which may be due to an overemphasis on providing detailed information.[22] However, this may not be suitable for common people. Copilot's lower scores could be due to limitations in its language processing capabilities, resulting in less coherent and relevant responses. However, none of the above chatbots are dedicated to healthcare information.[23] Despite that, they provide adequate information which makes them suitable for health education tools. The major difference between these chatbots is the capability of answering customized answers and interaction with follow-up questions which was not possible with Internet search engines like Google.[24]

Previous studies also reported the applicability of the LLM chatbot in health education.[12,25-27] Using chatbot responses as a health education tool for midlife health has the potential to empower women to better understand complex health issues, such as menopause, cardiovascular risk, and mental health challenges, through accurate and coherent guidance. Many women may not have access to healthcare facilities that provide detailed guidance about their midlife health issues.[28,29] It is similar for men in that they may face challenges in accessing healthcare and related advice.[30] For them, the chatbots can be helpful at various stages of their healthcare needs.[7,31] However, the varying performance of different chatbots highlights the importance of selecting the right tool to ensure that the information provided is both accurate and clear, making the chatbot an effective health education resource in this context.

This study has several limitations. We evaluated only six popular chatbots with 20 midlife-related questions. The evaluation criteria focused primarily on fluency, coherence, accuracy, relevancy, and readability. Further studies are required for user satisfaction, engagement, and contextual relevance of responses. The cross-sectional design and evolving technology limit the long-term efficacy of chatbot interactions in promoting health education. The study's reliance on quantitative metrics may overlook qualitative aspects of chatbot interactions that could provide deeper insights into user experiences and educational outcomes.

Conclusion

LLM chatbots can answer midlife-related health questions with variable capabilities. Meta AI was found to be the highest-scoring chatbot for addressing midlife health inquiries, whereas Perplexity offers high readability, ensuring that the information is accessible and easy to understand. Thus, LLM chatbots can serve as valuable educational tools for midlife health, provided that users select the appropriate chatbot based on its specific capabilities. It is essential for healthcare providers and users to consider both the accuracy and readability of chatbot responses to enhance health education. By doing so, men and women in midlife

can receive tailored, reliable information that supports their health and well-being during this crucial life stage.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-Year timeline and perspectives. J Med Syst 2024;48:22.
2. Andrew A. Potential applications and implications of large language models in primary care. Fam Med Community Health 2024;12:e002602.
3. Lang S, Vitale J, Fekete TF, Haschtmann D, Reitmeir R, Ropelato M, *et al.* Are large language models valid tools for patient information on lumbar disc herniation? The spine surgeons' perspective. Brain Spine 2024;4:102804.
4. Mondal H, Mondal S, Podder I. Using ChatGPT for writing articles for patients' education for dermatological diseases: A pilot study. Indian Dermatol Online J 2023;14:482-6.
5. Clark M, Bailey S. Chatbots in Health Care: Connecting Patients to Information: Emerging Health Technologies. Ottawa (ON): Canadian Agency for Drugs and Technologies in Health; 2024.
6. Santoro NF, Coons HL, El Khoudary SR, Epperson CN, Holt-Lunstad J, Joffe H, *et al.* NAMS 2021 Utian translational science symposium September 2021, Washington, DC Charting the path to health in midlife and beyond: The biology and practice of wellness. Menopause 2022;29:504-13.
7. Garg R, Munshi A. Revolutionizing menopause management: Harnessing the potential of artificial intelligence. J Midlife Health 2024;15:53-4.
8. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: Rapid review. J Med Internet Res 2024;26:e56930.
9. Goodman RS, Patrinely JR, Stone CA Jr., Zimmerman E, Donald RR, Chang SS, *et al.* Accuracy and reliability of Chatbot responses to physician questions. JAMA Netw Open 2023;6:e2336483.
10. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, *et al.* Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29:721-32.
11. Hassona Y, Alqaisi D, Al-Haddad A, Georgakopoulou EA, Malamos D, Alrashdan MS, *et al.* How good is ChatGPT at answering patients' questions related to early detection of oral (mouth) cancer? Oral Surg Oral Med Oral Pathol Oral Radiol 2024;138:269-78.
12. Mondal H, Dash I, Mondal S, Behera JK. ChatGPT in answering queries related to Lifestyle-related diseases and disorders. Cureus 2023;15:e48296.
13. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. ChatGPT versus web search for patient questions: What does ChatGPT do better? Eur Arch Otorhinolaryngol 2024;281:3219-25.
14. Halaseh FF, Yang JS, Danza CN, Halaseh R, Spiegelman L. ChatGPT's role in improving education among patients seeking emergency medical treatment. West J Emerg Med 2024;25:845-55.
15. Shah YB, Ghosh A, Hochberg AR, Rapoport E, Lallas CD, Shah MS, *et al.* Comparison of ChatGPT and traditional patient education materials for Men's health. Urol Pract 2024;11:87-94.
16. Najafali D, Reiche E, Camacho JM, Morrison SD, Dorafshar AH. Let's chat about Chatbots: Additional thoughts on ChatGPT and its role in plastic surgery along with its ability to perform systematic reviews. Aesthet Surg J 2023;43:P591-2.
17. Mihalache A, Grad J, Patil NS, Huang RS, Popovic MM, Mallipatna A, *et al.* Google Gemini and bard artificial intelligence Chatbot performance in ophthalmology knowledge assessment. Eye (Lond) 2024;38:2530-5.
18. Sarangi PK, Datta S, Swarup MS, Panda S, Nayak DS, Malik A, *et al.* Radiologic decision-making for imaging in pulmonary embolism: Accuracy and reliability of large language models-bing, Claude, ChatGPT, and Perplexity. Indian J Radiol Imaging 2024;34:653-60.
19. Iannantuono GM, Bracken-Clarke D, Karzai F, Choo-Wosoba H, Gulley JL, Floudas CS. Comparison of large language models in answering immuno-oncology questions: A cross-sectional study. Oncologist 2024;29:407-14.
20. Rahsepar AA, Tavakoli N, Kim GH, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT versus Google bard. Radiology 2023;307:e230922.
21. Irmici G, Cozzi A, Della Pepa G, De Berardinis C, D'Ascoli E, Cellina M, *et al.* How do large language models answer breast cancer quiz questions? A comparative study of GPT-3.5, GPT-4 and Google Gemini. Radiol Med 2024;129:1463-7.
22. Mondal H, Mondal S. ChatGPT in academic writing: Maximizing its benefits and minimizing the risks. Indian J Ophthalmol 2023;71:3600-6.
23. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, *et al.* Applications and concerns of ChatGPT and other conversational large language models in health care: Systematic review. J Med Internet Res 2024;26:e22769.
24. Morita PP, Lotto M, Kaur J, Chumachenko D, Oetomo A, Espiritu KD, *et al.* What is the impact of artificial intelligence-based chatbots on infodemic management? Front Public Health 2024;12:1310437.
25. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the possible use of AI Chatbots in public health education: Feasibility study. JMIR Med Educ 2023;9:e51421.
26. Yalla GR, Hyman N, Hock LE, Zhang Q, Shukla AG, Kolomeyer NN. Performance of artificial intelligence Chatbots on glaucoma questions adapted from patient brochures. Cureus 2024;16:e56766.
27. Monje S, Ulene S, Gimovsky AC. Identifying ChatGPT-written patient education materials using text analysis and readability. Am J Perinatol 2024;41:2229-31.
28. Johnson PJ, Jou J, Upchurch DM. Psychological distress and access to care among midlife women. J Aging Health 2020;32:317-27.
29. Leone T. Women's mid-life health in low and middle income countries: A comparative analysis of the timing and speed of health deterioration in six countries. SSM Popul Health 2019;7:100341.
30. Fadaei Dehcheshmeh N, Emamian Fard SM, Roghani T, Mohammadi P, Faraji-Khiavi F. Challenges of middle-aged men in utilizing new health services from primary health care providers' perspective: A qualitative study. BMC Prim Care 2022;23:318.
31. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered Chatbots in medical education: Potential applications and implications. Cureus 2023;15:e43271.

Annexure

Annexure 1:

Questions about women's midlife health

1. What are the common symptoms of menopause?
2. How can I manage weight gain during menopause?
3. What are the risks of osteoporosis, and how can I prevent it?
4. How does menopause affect my sexual health?
5. What are the best treatment options for menopausal symptoms?
6. How can I reduce my risk of heart disease after menopause?
7. What mental health changes can occur during midlife?
8. Is hormone replacement therapy safe for me?
9. How can I maintain my skin and hair health during midlife?
10. What screenings should I be getting at this stage of my life?

Questions about men's midlife health

1. What are the symptoms of andropause, and how does it affect me?
2. How can I maintain muscle mass and strength as I age?
3. How can I prevent or manage weight gain during midlife?
4. What are the signs and symptoms of prostate problems?
5. How can I maintain sexual health and function?
6. What are the best ways to manage stress and mental health during midlife?
7. How can I reduce my risk of heart disease?
8. What should I know about testosterone replacement therapy?
9. What dietary changes should I make as I age?
10. What screenings and checkups should I be getting at this stage of life?