Artificial intelligence generates proficient Spanish obstetrics and gynecology counseling templates

Rachel L. Solmonovich, MD; Insaf Kouba, MD; Oscar Quezada, MD; Gianni Rodriguez-Ayala, MD; Veronica Rojas, MD; Kevin Bonilla, MD; Kevin Espino, MD; Luis A. Bracero, MD

BACKGROUND: Effective patient counseling in Obstetrics and gynecology is vital. Existing language barriers between Spanish-speaking patients and English-speaking providers may negatively impact patient understanding and adherence to medical recommendations, as language discordance between provider and patient has been associated with medication noncompliance, adverse drug events, and underuse of preventative care. Artificial intelligence large language models may be a helpful adjunct to patient care by generating counseling templates in Spanish.

OBJECTIVES: The primary objective was to determine if large language models can generate proficient counseling templates in Spanish on obstetric and gynecology topics. Secondary objectives were to (1) compare the content, quality, and comprehensiveness of generated templates between different large language models, (2) compare the proficiency ratings among the large language model generated templates, and (3) assess which generated templates had potential for integration into clinical practice.

STUDY DESIGN: Cross-sectional study using free open-access large language models to generate counseling templates in Spanish on select obstetrics and gynecology topics. Native Spanish-speaking practicing obstetricians and gynecologists, who were blinded to the source large language model for each template, reviewed and subjectively scored each template on its content, quality, and comprehensiveness and considered it for integration into clinical practice. Proficiency ratings were calculated as a composite score of content, quality, and comprehensiveness. A score of >4 was considered proficient. Basic inferential statistics were performed.

RESULTS: All artificial intelligence large language models generated proficient obstetrics and gynecology counseling templates in Spanish, with Google Bard generating the most proficient template ($p<0.0001$) and outperforming the others in comprehensiveness (P=.03), quality (P=.04), and content (P=.01). Microsoft Bing received the lowest scores in these domains. Physicians were likely to be willing to incorporate the templates into clinical practice, with no significant discrepancy in the likelihood of integration based on the source large language model (P=.45).

CONCLUSIONS: Large language models have potential to generate proficient obstetrics and gynecology counseling templates in Spanish, which physicians would integrate into their clinical practice. Google Bard scored the highest across all attributes. There is an opportunity to use large language models to try to mitigate the language barriers in health care. Future studies should assess patient satisfaction, understanding, and adherence to clinical plans following receipt of these counseling templates.

Key words: anthropic Claude, artificial intelligence, ChatGPT, counseling templates, Google Bard, language barrier, large language models, Microsoft Bing, obstetrics and gynecology, patient counseling, Spanish-speaking patients

From the Northwell, New Hyde Park, NY (Solmonovich, Kouba, Quezada, Rodriguez-Ayala, Rojas, Bonilla, Espino, and Bracero); Department of Obstetrics and Gynecology, South Shore University Hospital, Bay Shore, NY (Solmonovich, Kouba, Rojas, Bonilla, Espino, and Bracero); Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY (Solmonovich, Rodriguez-Ayala, Rojas, and Bracero); Department of Obstetrics and Gynecology, Peconic Bay Medical Center, Riverhead, NY (Quezada); Department of Obstetrics and Gynecology, Huntington Hospital, Huntington, NY (Rodriguez-Ayala)

This project did not receive any funding.

The authors report no conflict of interest.

Corresponding author: Rachel Solmonovich, MD RSolmonovich@northwell.edu

2666-5778/$36.00

http://dx.doi.org/10.1016/j.xagr.2024.100400

Introduction

There are over 40 million Spanish speakers in the United States,[1] and they make up a significant portion of our patient population. Based on the 2019 US Census data, 39% of those who speak Spanish at home do not speak English well.[2] Language discordance between provider and patient has been shown to be associated with medication noncompliance, adverse drug events, and underuse of preventative care,[3–5] while language concordance has been shown to improve health outcomes,[5–7] However, only 39.7% of physicians report being multilingual, with 35.5% of them speaking Spanish.[8]

The gap between the number of Spanish-speaking physicians versus patients necessitates additional tools to provide comprehensive care that patients understand and retain. Patient information brochures have been shown to enhance patients' medical knowledge,[9] increase patient intention to speak with physicians about medical problems,[10] and improve their understanding of hospital admissions and newly prescribed medications.[11] In pregnancy, educational pamphlets significantly increased maternal perception of the safety and benefit of the influenza vaccine, as well as the overall uptake.[12] The benefits are apparent, but the cost and time spent developing pamphlets can be substantial.[13]

Artificial intelligence (AI) large language models (LMM) have become popular, and multiple publications have shown their potential in medicine. AI involves computer science and linguistics to create machines capable of

AJOG Global Reports at a Glance

Why was this study conducted?

To determine if large language models (LLMs) can generate proficient Spanish obstetrics and gynecology (Ob-Gyn) counseling templates. To compare the content, quality, and comprehensiveness of generated templates between different LLMs, and to assess which have potential for integration into clinical practice.

Key findings

LLMs generate proficient Spanish counseling templates on Ob-Gyn topics. Google Bard outperformed the other LLMs for proficiency, comprehensiveness, quality, and content, and Microsoft Bing scored the lowest for these attributes. Practicing physicians would integrate the counseling templates into their clinical practices. Spanish-speaking experts on the topics should review the length, formatting, medical jargon, and word choice before integration into practice.

What does this study add to what is already known?

LLMs generate counseling templates with accurate Ob-Gyn information for Spanish-speaking patients.

performing tasks that normally require human intelligence.[14] LLMs are a type of AI model that are trained on massive text datasets and can interact in a dialogue format through human-like responses.[15] Benefits include improved scientific writing, data analysis, and language review, personalized learning,[16–20] documentation,[21–23] and generating responses quickly, therefore saving time.[24–27] ChatGPT, one such LLM, has been found to pass various medical licensing exams[22,28–30] and appropriately respond to medical prompts across multiple specialties,[31–35] suggesting it may be applicable for clinical care.

This technology may be a useful tool for generating patient information materials. ChatGPT showed promise for developing consent forms in simple words that patients can easily understand.[36] It also demonstrated capability for answering specialty medical questions in Spanish,[37] scoring above passing on a national access exam to specialized medical training in Spain,[38] and achieving statistically comparable results regardless of Spanish versus English prompt language.[39,40]

Obstetrics and gynecology (Ob-Gyn) encompasses a range of complex medical conditions and procedures, and comprehensive, understandable counseling is essential for informed decision-making and patient-physician collaboration. Language barriers negatively impact patient quality of care and safety[41,42] and LLMs may be able to improve the communication between provider and patient by providing fast and accurate translations.[16] Therefore, the primary objective of this study was to determine if LLMs can generate proficient counseling templates in Spanish on obstetric and gynecology topics. Secondary objectives were to (1) compare the content, quality, and comprehensiveness of generated templates between the different LLMs, (2) compare the proficiency ratings among the large language model generated templates, and (3) assess which generated templates physicians would integrate into their practice.

Materials and methods

Free open-access LLMs (ChatGPT-3.5, Microsoft Bing, Claude, and Google Bard) generated counseling templates on December 7, 2023 using the prompt in English “Please provide brief counseling in Spanish on the topic of ‘x’ from the perspective of a physician counseling a patient on an 8th grade level” provided by the primary author. An 8th grade level was chosen to ensure it would be suitable even for patients with limited health literacy. The 4 selected topics were (1) Group B Strep in pregnancy, (2) gestational diabetes (Supplement 1), (3) pap smear and human papilloma virus, and (4) Tetanus-Diphtheria-Pertussis vaccination in pregnancy. The 6 native Spanish-speaking Ob-Gyn study authors, who were blinded to the source LLM for each template, then reviewed and scored each template on its content, quality, and comprehensiveness on a 5-point Likert scale (very poor, poor, fair, good, excellent), and considered it for integration into clinical practice, a binary variable of yes/no. Each source LLM thus received 24 scores for each evaluated domain. A composite score of the average sum of all ratings for content, quality, and comprehensiveness was used to generate a proficiency rating.

This cross-sectional study followed the STROBE reporting guidelines. It did not require institutional review board approval because no human participants were recruited.

Outcomes

The primary outcome was the LLM ability to generate proficient Ob-Gyn counseling templates in Spanish. The secondary outcomes were the levels of proficiency, and its individual components, of each LLM, as well as the reviewer’s willingness to integrate the generated templates into their clinical practices. A score of >4, reflecting good or excellent on the 5-point Likert scale, was considered proficient, solid content, high quality, and comprehensive.

Statistical analysis

The 5-point Likert scale was converted to numerical values 1−5, with 1 = very poor and 5 = excellent, and means with standard deviations were tabulated via Microsoft Excel for analysis. The Chi-square and ANOVA tests were used for categorical and continuous variables, respectively, using OpenEpi, Version 3, open-source calculator. A *P* value <.05 was considered statistically significant.

Results

The 6 Ob-Gyn authors scoring the LLM-generated templates spoke a variety of Spanish dialects including Colombian, Salvadorian, Mexican, Puerto Rican, and Peruvian at a native level. Their ages ranged from 32

−71 years old (median 40.5 years), and they were between 1−41 years out of residency. Two completed fellowships, one in minimally invasive gynecologic surgery and one in maternal fetal medicine.

Primary outcomes

All LLMs generated proficient counseling templates in Spanish on the 4 selected Ob-Gyn topics.

Secondary outcomes

Bard generated templates were the most proficient, with a score of 4.6, compared to Claude, Chat, and Bing, which scored 4.2, 4.2, and 4.1, respectively (*P*<.0001), consistently outperforming the other LLMs in the individual domains, averaging 4.6 (*P*=.03) for comprehensiveness, 4.6 for quality (*P*=.04), and 4.7 for content (*P*=.01). Table lists the average scores for each LLM.

All reviewers would integrate LLM-generated counseling templates into their practice, but rates varied among the different LLMs. ChatGPT-generated templates showed the highest integration potential, with 79.2% of the templates scoring "yes." Authors stated that the templates were clear, concise, and brief but may lack certain important details. Authors were least likely to want to integrate Bing-generated templates, with only 58.3% of the generated templates receiving a "yes" score, commenting that they had good content, but at times included too much medical jargon, physiology, and Spanish words that were unfamiliar to the evaluators. Those with ≤8 years since residency appreciated the thoroughness and question and answer format generated by Bard, while the two most experienced reviewers (23 and 41 years since residency) would not integrate Bard-generated templates into their practice, stating that they had excessive detail and were too long. The authors agreed that Claude provided concise and simple templates that used basic language and would be easy for patients to understand.

Comment

Principal findings

The 6 authors reviewing the LLM-generated templates were practicing Ob-Gyn physicians who spoke a variety of Spanish dialects at a native level. They all agreed that the LLMs generated proficient counseling templates on the prompted Ob-Gyn topics. Bard received the highest scores across all domains, and Bing scored the lowest across all attributes.

Results in the context of what is known

ChatGPT has gained rapid popularity,[43] and several recent publications have demonstrated its capabilities across medical licensing exams[22,28−30] and specialties,[31−35] even in Spanish.[37−40] In the field of Ob-Gyn, it produced appropriate responses to fertility prompts.[44] Our study, too, found the LLMs generated accurate Ob-Gyn information for counseling templates in Spanish.

Clinical implications

Effective patient counseling can be challenging, given time constraints and language limitations. Our study found that LLMs generated templates that were high quality, with solid content, and comprehensive, which Ob-Gyn physicians would be willing to integrate into clinical care. Using this tool to provide informative handouts for patients to review in their native language and in their own time may improve their understanding of Ob-Gyn topics and adherence to physician recommendations.

Although all LLMs included in the study were found to generate useful templates, none were perfect. Length, format, medical jargon, and Spanish word choice were brought up as issues, revealing that integration into clinical care should involve oversight and modification by experts in the field. Still, a significant amount of time may be saved, and a large step towards overcoming the language barrier may be taken, by using LLMs for this purpose.

Research implications

The reviewers were all practicing Ob-Gyn physicians. Future research could evaluate whether the LLM-generated Ob-Gyn templates in Spanish are viewed as positively from the perspectives of those without expert clinical knowledge in the field. Furthermore, studies should assess patient satisfaction, understanding, and adherence with the clinical plan following receipt of these LLM-generated templates. In addition, LLMs could be prompted with more Ob-Gyn topics, to expound on their abilities and confirm their usefulness as an adjunct to patient counseling.

Strengths and limitations

Although we had a small number of reviewers assessing the LLM-generated templates, all 6 were practicing obstetricians and gynecologists, with the expert level knowledge necessary to evaluate the templates for integration into clinical care based on multiple attributes.

TABLE
Large language model scores across domains

Domains	ChatGPT-3.5	Microsoft Bing	Anthropic Claude	Google Bard	*P* value
Proficiency	4.17±0.63	4.1±0.65	4.18±0.72	4.65±0.73	<.0001
Comprehensiveness	4.13±0.45	4.08±0.72	4.17±0.76	4.63±0.77	.03
Quality	4.21±0.72	4.08±0.65	4.13±0.68	4.63±0.77	.04
Content	4.17±0.7	4.13±0.61	4.25±0.74	4.71±0.69	.01
Integrate					0.45
Yes	19 (79.17)	14 (58.33)	15 (62.5)	16 (66.67)	
No	5 (20.83)	10 (41.67)	9 (37.5)	8 (33.33)	

Data are presented as mean±standard deviation or number (percentage).

They were also native Spanish speakers, who spoke a variety of dialects, ensuring the templates' suitability for a diverse Spanish-speaking population and increasing the generalizability of our results. Our study was limited in that we only requested templates on a few well-known, general Ob-Gyn topics. LLM prompts involving other specialties, more specialized topics, or less familiar conditions may not produce the same level of template proficiency appropriate for Spanish-speaking patients.

Conclusions

Our study demonstrates the potential of LLMs to generate proficient Ob-Gyn counseling templates in Spanish, which practicing physicians would integrate into their clinical practice. There is some discrepancy in the comprehensiveness, quality, and content between the LLMs studied, with Google Bard scoring the highest, but all LLMs had an average rating of good or excellent in each domain. There is opportunity to take advantage of LLMs in this manner, to improve English-speaking physicians and Spanish-speaking patients communication and reduce the negative effects of the language barrier. Future studies could evaluate whether patients truly benefit from integration of such templates into clinical practice. ■

CRediT authorship contribution statement

Rachel L. Solmonovich: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Insaf Kouba:** Writing – review & editing, Methodology, Formal analysis, Data curation, Conceptualization. **Oscar Quezada:** Investigation, Formal analysis, Data curation. **Gianni Rodriguez-Ayala:** Investigation, Formal analysis, Data curation. **Veronica Rojas:** Investigation, Formal analysis, Data curation. **Kevin Bonilla:** Investigation, Formal analysis, Data curation. **Kevin Espino:** Investigation, Formal analysis, Data curation. **Luis A. Bracero:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization.

Tweetable statement

LLMs may be used to mitigate language barriers in obstetrics and gynecology by generating counseling templates in Spanish about conditions and procedures with comprehensive, quality content for patients.

Patient consent statement

Consents were not applicable as no human participants were recruited. The manuscript authors reviewed the AI-generated data themselves.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.xagr.2024.100400.

REFERENCES

1. Balch B. The United States needs more Spanish-speaking physicians. AAMCNEWS. webpage: Association of American Medical Colleges; 2023 https://www.aamc.org/news/united-states-needs-more-spanish-speaking-physicians.
2. Dietrich, S. and E. Hernandez. Nearly 68 million people spoke a language other than English at home in 2019. 2022 4/18/24].
3. Gandhi TK, Burstin HR, Cook EF, et al. Drug complications in outpatients. Journal of General Internal Medicine 2000;15(3):149–54.
4. Jih J, Vittinghoff E, Fernandez A. Patient-physician language concordance and use of preventive care services among limited English Proficient Latinos and Asians. Public Health Reports 2015;130(2):134–42.
5. Cano-Ibáñez N, Zolfaghari Y, Amezcua-Prieto C, Khan KS. Physician–patient language discordance and poor health outcomes: a systematic scoping review. Front Public Health 2021;9:629041.
6. Lopez Vera A, Thomas K, Trinh C, Nausheen F. A case study of the impact of language concordance on patient care, satisfaction, and comfort with sharing sensitive information during medical care. J Immigrant Minority Health 2023;25(6):1261–9.
7. Hsueh L, Hirsh AT, Maupomé G, Stewart JC. Patient–provider language concordance and health outcomes: a systematic review, evidence map, and research agenda. Med Care Res Rev 2021;78(1):3–23.
8. Ortega P, Felida N, Avila S, Conrad S, Dill M. Language profile of the US physician workforce: a descriptive study from a National Physician Survey. J Gen Int Med 2023;38(4):1098–101.
9. Adirim T, Chafranskaia A, Nyhof-Young J. Investigating the impact of socioeconomic status on the effectiveness of a pamphlet on achieving and maintaining bone health in breast cancer survivors: a patient education resource development primer. J Cancer Edu 2012;27(1):54–8.
10. Donald RA, Arays R, Elliott JO, Jordan K. The effect of an educational pamphlet on patient knowledge of and intention to discuss screening for obstructive sleep apnea in the acute ischemic stroke population. J Neurosci Nurs 2018;50(3):177–81.
11. Thomas NE, Edwards L, Mcardle P. Knowledge is Power. A quality improvement project to increase patient understanding of their hospital stay. BMJ Qual Improv Rep 2017;6(1):u207103-w3042.
12. Meharry MP, Cusson RM, Stiller R, Vázquez M. Maternal influenza vaccination: evaluation of a patient-centered pamphlet designed to increase uptake in pregnancy. Maternal Child Health J 2014;18(5):1205–14.
13. Papadakos J, Samoil D, Giannopoulos E, et al. The cost of patient education materials development: opportunities to identify value and priorities. J Cancer Educ 2020;37:834–42.
14. Sarker IH. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Comput Sci 2022;3(2):158.
15. OpenAI, Introducing ChatGPT. OpenAI.com.
16. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Comm Med 2023;3(1):141.
17. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11(6):887.
18. Hutson M. Could AI help you to write your next paper? Nature 2022;611:3.
19. Khera R, Butte AJ, Berkwits M, et al. AI in medicine-JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. JAMA 2023;330(9):818–20.
20. Doshi RH, Bajaj S. Promises—and pitfalls—of ChatGPT-assisted medicine. Stat News; 2023 https://www.statnews.com/2023/02/01/promises-pitfalls-chatgpt-assisted-medicine/#:%7E:text=Clinical%20applications%20The%20use%20of%20ChatGPT%20in%20clinical,generating%20medical%20charts%2C%20progress%20notes%2C%20and%20discharge%20instructions.
21. Decker H, Trang K, Ramirez J, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023;6(10):e2336997.
22. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot

for medicine. N Engl J Med 2023;388(13):1233–9.

23. Bala S, Keniston A, Burden M. Patient perception of plain-language medical notes generated using artificial intelligence software: pilot mixed-methods study. JMIR Form Res 2020;4(6):e16670.

24. Deng JL, Yijia. The benefits and challenges of ChatGPT: an overview. Front Comput Intell Syst 2023;2:81–3.

25. Salvagno M, Taccone SF, Gerli GA. Can artificial intelligence help for scientific writing? Critical Care 2023;27(1):75.

26. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023;307(2):e230163.

27. Clynch N, Kellett J. Medical documentation: part of the solution, or part of the problem? A narrative review of the literature on the time spent on and value of medical documentation. Int J Med Inform 2015;84(4):221–8.

28. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023;2(2):e0000198.

29. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine (Baltimore) 2023;102(32):e34673.

30. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312.

31. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Network Open 2023;6(8):e2330320.

32. Ayers JW, Zhu Z, Poliak A, et al. Evaluating artificial intelligence responses to public health questions. JAMA Network Open 2023;6(6):e2317517.

33. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open 2023;6(10):e2336483.

34. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectrum 2023;7(2):pkad015.

35. Schubert CM, Wick W, Venkataramani V. Performance of large language models on a neurology board−style examination. JAMA Network Open 2023;6(12):e2346721.

36. Decker H, Trang K, Ramirez J, et al. Large language model−based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Network Open 2023;6(10):e2336997.

37. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep 2023;13(1):22129.

38. Carrasco JP, García E, Sánchez DA, et al. ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España. Revista Española de Educación Médica 2023;4(1):55–69.

39. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency Entrance Examination (MIR): promising horizons for AI in clinical medicine. Clin Pract 2023;13(6):1460–87.

40. Alfertshofer M, Hoch CC, Funk PF, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Annals Biomed Eng 2023;52:1542–5.

41. Moissac DD, Bowen S. Impact of language barriers on quality of care and patient safety for official language minority francophones in Canada. J Patient Exp 2019;6(1):24–32.

42. Smid CM, Dorman FK, Boggess AK. Lost in translation? English- and Spanish-speaking women's perceptions of gestational weight gain safety, health risks and counseling. J Perinatol 2015;35(8):585–9.

43. Teubner T, Flath CM, Weinhardt C, van der Aalst W, Hinz O. Welcome to the era of ChatGPT et al. Bus Inform Syst Eng 2023;65(2):95–101.

44. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. Fertil Steril 2023;120(3):575–83.