

## ORIGINAL RESEARCH ARTICLE

# Artificial intelligence-enabled digital reproductive health education: corpus-based strategies for strengthening Spanish health communication capacity

DOI: 10.29063/ajrh2026/v30i9s.2

Licui Zhu\* and Liang Xu

Foreign Language College Changchun University, China

\*For Correspondence: Email: 233645136@qq.com

## Abstract

This study examines how corpus-grounded artificial intelligence (AI) can strengthen Spanish reproductive health communication capacity within China's digital health ecosystem. A mixed-methods design was employed, combining corpus linguistics, AI-assisted message generation, expert-informed evaluation, and quantitative user assessment. A domain-specific Spanish reproductive health corpus was constructed from 3,052 documents, including clinical guidelines, patient education materials, and FAQs/user queries, yielding 1,065,110 tokens for linguistic analysis and AI grounding. Corpus-derived readability benchmarks, lexical simplification rules, and discourse patterns were integrated into an AI content generation pipeline to produce reproductive health messages, which were then compared with non-corpus-grounded AI outputs. The user evaluation phase was conducted among 240 Spanish-speaking or Spanish-proficient adults in selected Chinese metropolitan cities. Data were collected through a structured questionnaire measuring Corpus-Grounded AI Generation, Spanish Communication Quality, and Health Communication Capacity, and were analyzed using reliability testing, correlation analysis, and structural equation modeling. The findings showed that corpus-grounded AI significantly improved Spanish communication quality, while communication quality had the strongest effect on users' comprehension, confidence, and help-seeking intention. Mediation analysis further demonstrated that Spanish communication quality significantly mediated the relationship between corpus-grounded AI generation and health communication capacity. The study concludes that linguistically informed AI design can enhance the clarity, accessibility, and effectiveness of reproductive health education and offers a practical framework for multilingual digital health communication in sensitive healthcare contexts. (*Afr J Reprod Health* 2026; 30 [9s]: 15-27).

---

**Keywords:** Artificial intelligence, Reproductive health education, Corpus linguistics, Spanish health communication, Digital health platforms

---

## Résumé

Cette étude visait à examiner le rôle de l'IA guidée par la linguistique de corpus dans l'amélioration de la qualité de la communication en espagnol sur la santé reproductive au sein de l'écosystème numérique chinois. En utilisant une approche à méthodes mixtes, la recherche a combiné l'analyse de corpus de textes espagnols sur la santé reproductive, la génération de contenu basée sur l'IA et l'évaluation par les utilisateurs. Un corpus espagnol spécifique au domaine de la santé reproductive a été construit et intégré dans un pipeline de génération d'IA ancré dans le corpus pour optimiser la lisibilité, l'exactitude et la sensibilité culturelle. La qualité de la communication et la capacité de communication en santé ont été mesurées à l'aide d'une analyse de fiabilité, de tests de corrélation et de modélisation par équations structurelles. Les résultats ont révélé que l'IA ancrée dans le corpus améliore significativement la qualité de la communication en espagnol, ce qui améliore à son tour la compréhension des utilisateurs, l'auto-efficacité et les intentions de recherche d'aide. La qualité de la communication est apparue comme un mécanisme médiateur clé entre la conception de l'IA et les résultats en matière de communication santé. Sur le plan théorique, l'étude établit une base pour intégrer la linguistique de corpus dans les systèmes d'éducation à la santé pilotés par l'IA, tandis que sur le plan pratique, elle propose un cadre évolutif pour les plateformes multilingues d'éducation à la santé reproductive. Ces résultats ont des implications importantes pour les politiques mondiales de santé numérique, la gouvernance de l'IA et les stratégies de communication en santé reproductive. (*Afr J Reprod Health* 2026; 30 [9s]: 15-27).

---

**Mots-clés:** Intelligence artificielle, Éducation à la santé reproductive, Linguistique de corpus, Communication en santé espagnole, Plateformes de santé numérique

---

## Introduction

Digital reproductive - health education has become an indispensable component of public - health systems, particularly in contexts where barriers of access to health care, health literacy and communication continue to exist. In China, the quick digitalisation of health services has heightened the use of mobile-health (mHealth) platforms, social-media-based interventions and artificial intelligence (AI)-assisted educational tools for sexual and reproductive health education<sup>1,2</sup>. These developments are part of wider policy efforts to improve reproductive health awareness, maternal outcomes and informed decision-making using scalable digital solutions<sup>3</sup>.

AI has changed the face of health education by making it possible to automate content creation, personalize learning paths, use conversational agents and large-scale analysis of health communication data<sup>4,5</sup>. AI enabled systems have also demonstrated great promise in maternal health reproductive settings such as the fertility education, prenatal care, breastfeeding support and sexual health counselling through mobile applications and social media platforms in China<sup>6</sup>. Such systems also provide efficient health information dissemination activities and customise information to the informational needs and literacy levels of the users.

However, the success of AI-driven reproductive health education is not limited to the technological capability of the AI but also relies on the linguistic and communicative quality of the transmitted messages. Health- communication research consistently shows that complicated medical language, poor readability and cultural mismatch are harmful to understanding and to the effectiveness of digital health interventions<sup>7</sup>. The challenge is greater in multilingual and transnational contexts, where information must be available in cross-linguistic and cross-cultural frames. Spanish, as one of the world's most widely spoken languages, plays an important role in global health communication capacity, especially in the case of reproductive health education in a variety of populations<sup>8</sup>.

Corpus-based approaches form a strong methodological basis to advance AI-enabled health

communication to the next level. By analyzing large corpora of health texts, social - media discourse and patient education materials, corpus linguistics facilitates systematic identification of lexical complexity, discourse patterns and readability gaps<sup>9</sup>. Recent developments in natural - language - processing (NLP) and machine - learning have propelled the development of multilingual corpora and text - simplification systems, including resources that are explicitly geared towards Spanish health communication<sup>10</sup>. These tools help AI systems to create linguistically accessible, culturally sensitive and context appropriate reproductive health content. Telemedicine and AI adoption have also led to policies in areas like South Asia for bridging the gap in early cancer detection especially for low-income groups<sup>11</sup>.

Despite such advances, such scholarship remains fragmented across disciplines. Digital-reproductive- health research in China is mostly focused on the effectiveness of interventions and user engagement, and less on the optimization of language use and multilingual communication strategies. On the contrary, corpus and AI health communication research in Spanish has been very restricted to readability, lexical simplification and evaluation measures, and has often been disallowed from actual reproductive health education contexts<sup>12</sup>. Few studies combine AI enabled digital - health education with corpus based linguistic strategies to enhance communication capacity across languages.

This study addresses that lacuna by proposing an integrated framework for AI enabled digital reproductive-health education that is grounded on corpus-based strategies to build Spanish health-communication capacity with China as primary research and implementation context. By connecting Chinese digital - reproductive - health platforms with corpus - informed artificial intelligence methods for Spanish health communication, the research project fosters a new interdisciplinary approach connecting the fields of public - health, artificial intelligence and applied linguistics. The findings are aimed at informing the design of more inclusive, linguistically accessible and scalable reproductive-health education systems in an increasingly globalized digital health landscape.

## **Literature review**

### **AI-enhanced digital reproductive health education in China**

The rapid expansion of digital health infrastructure in China has significantly transformed the landscape of reproductive health education and service delivery. Mobile health (mHealth) platforms, social media applications and artificial intelligence (AI) - driven decision support systems are being increasingly used to disseminate sexual and reproductive health information, boost health literacy, and facilitate reproductive decision-making<sup>13,14</sup>. These developments coincide with the national public health strategies with a focus on digital innovation as a means of tackling inequalities in access to reproductive and maternal healthcare, especially among women of reproductive age.

Artificial intelligence has enhanced the functionality and scalability of digital reproductive health education in China by offering a personalized content delivery, automated counselling and real-time interaction<sup>15</sup>. AI supported chatbots, recommendation algorithms and intelligent question-answering systems have been implemented in areas such as fertility education, in-vitro fertilization counselling, prenatal guidance, breastfeeding support and postnatal care<sup>9</sup>. Empirical studies show that AI-powered applications, especially those incorporated into ubiquitous applications like WeChat, increase user engagement, continuity of care and self-efficacy in reproductive health decision-making<sup>16</sup>.

Despite these advances, research has been focused mostly on behavioral outcomes, clinical indicators and user satisfaction, whereas the linguistic and communicative aspects of AI-mediated reproductive health education have attracted comparatively little attention. Many digital interventions assume that access to information implies understanding, and, thus, often overlook the role of language complexity, discourse structure, and cultural framing of information in understanding and trust<sup>17,18</sup>. This oversight is particularly relevant in the reproductive health arena, where medical terminology, stigma, and sensitive issues can hinder effective communication.

Moreover, although AI systems in China increasingly use large-scale textual data, such as user queries, health education materials and social media

discourse, very few studies explicitly incorporate corpus-based linguistic analysis into system design or evaluation<sup>19</sup>. As a result, AI enabled reproduction health education poses the risk of reproducing existing communication barriers rather than systematically addressing health literacy gaps. This highlights the need for linguistically informed AI strategies that optimize for message clarity, message accessibility and cross-linguistic transferability.

### **Corpus based AI based strategies for Spanish health communication**

Corpus-based approaches have long been recognized as irreplaceable tools for analysing and enhancing health communication. By analyzing systematically large sets of health-related texts, corpus linguistics allows identification of patterns of lexical complexity, syntactic density, discourse framing and readability that effect patient understanding<sup>20,21</sup>. Recent advances in natural language processing and machine learning have helped to reinforce information about the integration of corpus methods into AI-driven health communication systems.

Spanish has been identified as an important language in health communication research using corpora because of its worldwide dissemination and the reported difficulties with health literacy in digital health communication settings. Investigations of Spanish language patient education materials have consistently found high levels of lexical complexity and readability that fall outside recommended health communication standards<sup>22,23</sup>. These findings have been a motivation for the development of Spanish-specific health corpora, lexical simplification benchmark, and automatic readability assessment tools.

AI enabled corpus-based strategies such as lexical simplification and sentence restructuring and adaptive content generation have proven to increase comprehension of medical information without a loss of informational accuracy<sup>24</sup>. Multilingual and Spanish-centric corpora have been used in the training and testing of AI systems that can create "easy-to-read" health texts, which aids in accessibility for diverse audiences. Importantly, these approaches focus on capacity building in health communication by giving AI systems empirically-grounded linguistic knowledge as opposed to using generalized language models<sup>25</sup>.

Nevertheless, corpus-based Spanish health communication research has developed mostly outside the context of reproductive health education and is often applied retrospectively, when materials are evaluated instead of applied prospectively to the design of AI-driven educational interventions<sup>26</sup>. Furthermore, most of the studies focus on Western healthcare systems and there has been only little investigation of how corpus-based strategies of communication in Spanish could be incorporated into digital health platforms designed in non-Spanish-speaking contexts for example China<sup>27</sup>.

### ***Combining AI, corpus linguistics and reproductive health education***

The literature reveals a notable gap at the nexus of AI-enabled reproductive health education, corpus-based linguistic optimization and Spanish health communication capacity. While China is at the forefront of digital reproductive health innovation, the design and linguistic aspects of communication (especially concerning multi-lingual and transnational communication) are underdeveloped<sup>28</sup>. On the other hand, corpus based Spanish health communication studies provide strong methodological tools, yet do not connect with artificial intelligence (AI) driven reproductive health systems and real-world digital health systems<sup>29</sup>.

Addressing this gap requires an interdisciplinary approach that draws on corpus-based linguistic insights to drive AI-divided reproductive health education design. By incorporating the knowledge of Spanish health communication corpus and text simplification into artificial intelligence systems that have been developed as part of the digital health ecosystem in China, it becomes possible to enhance health communication capacity across the world while maintaining context sensitivity and scalability in technology<sup>30,31</sup>.

In this regard, the present study contributes to the existing scholarship by linking public health, artificial intelligence, and applied linguistics. It places corpus-based Spanish health communication at a step higher than the evaluative tool, but as a strategic tool for boosting the effectiveness, inclusivity, and transnationality of AI-enabled digital reproductive health education.

### ***Conceptual framework***

The conceptual framework illustrates how corpus linguistics-informed artificial intelligence can

enhance digital reproductive health education by strengthening Spanish health communication capacity within China's digital health ecosystem. The framework positions China as the implementation context, where widespread use of mobile health platforms and AI-enabled services enables large-scale delivery of reproductive health education.

The framework illustrates how corpus-based linguistic resources inform corpus-grounded AI generation to improve Spanish communication quality, leading to enhanced comprehension, health literacy, and help-seeking outcomes, with contextual moderators shaping effectiveness. At the input level, corpus-based linguistic resources, including Spanish reproductive health texts, user queries, readability benchmarks, and culturally sensitive phrasing provides empirical grounding for AI systems.

These resources inform corpus-grounded AI generation, which integrates controlled content generation, retrieval from vetted corpora, and text simplification strategies to ensure linguistic accuracy and accessibility. Prior research demonstrates that corpus-informed AI improves clarity, reduces linguistic complexity, and enhances safety in sensitive health communication contexts.

The framework identifies Spanish communication quality defined by readability, accuracy, cultural sensitivity, and actionability, as the key mediating construct linking AI outputs to outcomes. High-quality communication enables better comprehension, self-efficacy, and health literacy, which in turn strengthen users' ability to make informed reproductive health decisions and seek appropriate care. Contextual moderators such as trust in digital health systems, stigma, privacy concerns, and baseline health literacy influence the effectiveness of these pathways.

## ***Methods***

### ***Research design***

This study employed an explanatory mixed-methods design that combined corpus linguistics, artificial intelligence-assisted content generation, expert review, and quantitative user evaluation. The purpose of the design was to examine whether corpus-grounded AI generation could improve the quality of Spanish reproductive health messages and thereby strengthen users' health communication capacity within China's digital health environment.

The mixed-methods approach was considered suitable because the study required both linguistic analysis of reproductive health texts and quantitative assessment of user-level outcomes. The qualitative linguistic component focused on the construction and analysis of a domain-specific Spanish reproductive health corpus, whereas the quantitative component assessed the quality and effectiveness of AI-generated messages through structured user responses.

### **Study context**

The user-evaluation phase was conducted in selected metropolitan cities in China, namely Beijing, Shanghai, Guangzhou, and Shenzhen. These cities were chosen because they host larger international communities, greater digital health platform usage, and a higher likelihood of access to Spanish-speaking or Spanish-proficient participants. The study population therefore comprised adult Spanish-speaking or Spanish-proficient users residing in these cities who were capable of evaluating AI-generated reproductive health content.

### **Study population**

The study involved two units of analysis. The first was a textual population, consisting of Spanish-language reproductive health documents used to build the corpus. The second was a human participant population, consisting of adult users recruited for the evaluation of AI-generated reproductive health content. The user-evaluation sample included 240 respondents, comprising Spanish-speaking or Spanish-proficient adults residing in China, including international students, migrant community members, bilingual users of digital health information, and individuals familiar with online health communication environments. Participants were included if they were aged 18 years or older, were able to read Spanish, and consented to evaluate reproductive health messages. Individuals who were unable to complete the questionnaire independently or had insufficient Spanish language proficiency were excluded.

### **Study variables and operational definitions**

The independent variable was Corpus-Grounded AI Generation (CG-AI). This referred to the use of an AI content-generation pipeline informed by a

Spanish reproductive health corpus, readability benchmarks, lexical simplification rules, and retrieval-based grounding. Operationally, this construct captured the extent to which AI-generated messages reflected corpus-based linguistic optimization, including clarity of terminology, simplified structure, consistency, and contextual appropriateness.

The mediating variable was Spanish Communication Quality (SCQ). This referred to the linguistic and communicative quality of the AI-generated reproductive health messages. Operationally, it was measured through indicators such as readability, perceived clarity, clinical accuracy, cultural and ethical sensitivity, and actionability.

The dependent variable was Health Communication Capacity (HCC). This referred to the extent to which users could understand, interpret, and act upon reproductive health information delivered through AI-generated Spanish messages. Operationally, it was measured through user-reported comprehension, confidence in understanding the message, self-efficacy, and help-seeking intention.

### **Corpus construction and data sources**

The questionnaire was developed through a theory- and literature-informed process. First, the study constructs were identified from the conceptual framework, namely Corpus-Grounded AI Generation (CG-AI), Spanish Communication Quality (SCQ), and Health Communication Capacity (HCC). Second, questionnaire items were drafted by adapting indicators from prior literature on mHealth usability, health literacy, eHealth communication, readability assessment, and online health-information behavior<sup>4,7,12</sup>. Third, the wording of items was tailored to the context of AI-generated Spanish reproductive health messages. Items for CG-AI assessed whether the messages appeared coherent, linguistically simplified, contextually appropriate, and responsive to users' information needs. Items for SCQ measured readability, clarity, perceived accuracy, cultural sensitivity, and actionability of the Spanish messages, consistent with prior work on health communication quality and public health risk communication<sup>20</sup>. Items for HCC assessed user comprehension, confidence in understanding, self-efficacy, and intention to seek further health information or support, drawing on earlier studies of health literacy, eHealth

communication, and online information-seeking behavior<sup>19, 20, 24</sup>. All items were rated on a five-point Likert scale ranging from 1 = strongly disagree to 5 = strongly agree. The draft questionnaire was then reviewed by experts in health communication and applied linguistics to ensure content relevance, clarity, and contextual suitability before final administration. A pilot review was also undertaken to refine wording, remove ambiguity, and ensure that the items were understandable for Spanish-speaking or Spanish-proficient respondents in the Chinese digital health context.

### ***AI content generation procedure***

An AI-assisted content-generation pipeline was developed to produce Spanish reproductive health messages for user evaluation. Two message-generation conditions were used: corpus-grounded AI generation and non-corpus AI generation. In the corpus-grounded condition, prompts were designed using lexical and discourse patterns extracted from the corpus, and the AI system was guided through controlled prompting, retrieval-based grounding, and text simplification rules. In the non-corpus condition, messages were generated without corpus-informed linguistic constraints. This comparison enabled the study to examine whether corpus-based linguistic grounding improved communication quality and downstream user outcomes.

The AI-generated messages covered common reproductive health topics such as contraception, menstrual health, prenatal guidance, fertility awareness, and general maternal health information. All generated messages were screened for appropriateness before presentation to participants.

### ***Instrument development and measurement***

The evaluation instrument was a structured questionnaire developed to assess the three main constructs: CG-AI, SCQ, and HCC. The instrument included 6 items for CG-AI, 8 items for SCQ, and 7 items for HCC, in line with the construct structure reported in the results section. Each item was measured on a five-point Likert scale ranging from 1 = strongly disagree to 5 = strongly agree.

Items measuring CG-AI focused on whether the message appeared linguistically accessible, appropriately simplified, coherent, and tailored to user needs. SCQ items assessed clarity, readability, perceived accuracy, ethical sensitivity, and

usefulness of the message. HCC items measured user comprehension, confidence in understanding, self-efficacy in acting on the information, and intention to seek further help or clarification where needed.

Before the main survey, the questionnaire was reviewed by specialists in health communication and applied linguistics to assess content relevance and wording clarity. Minor revisions were made to improve item clarity and contextual appropriateness.

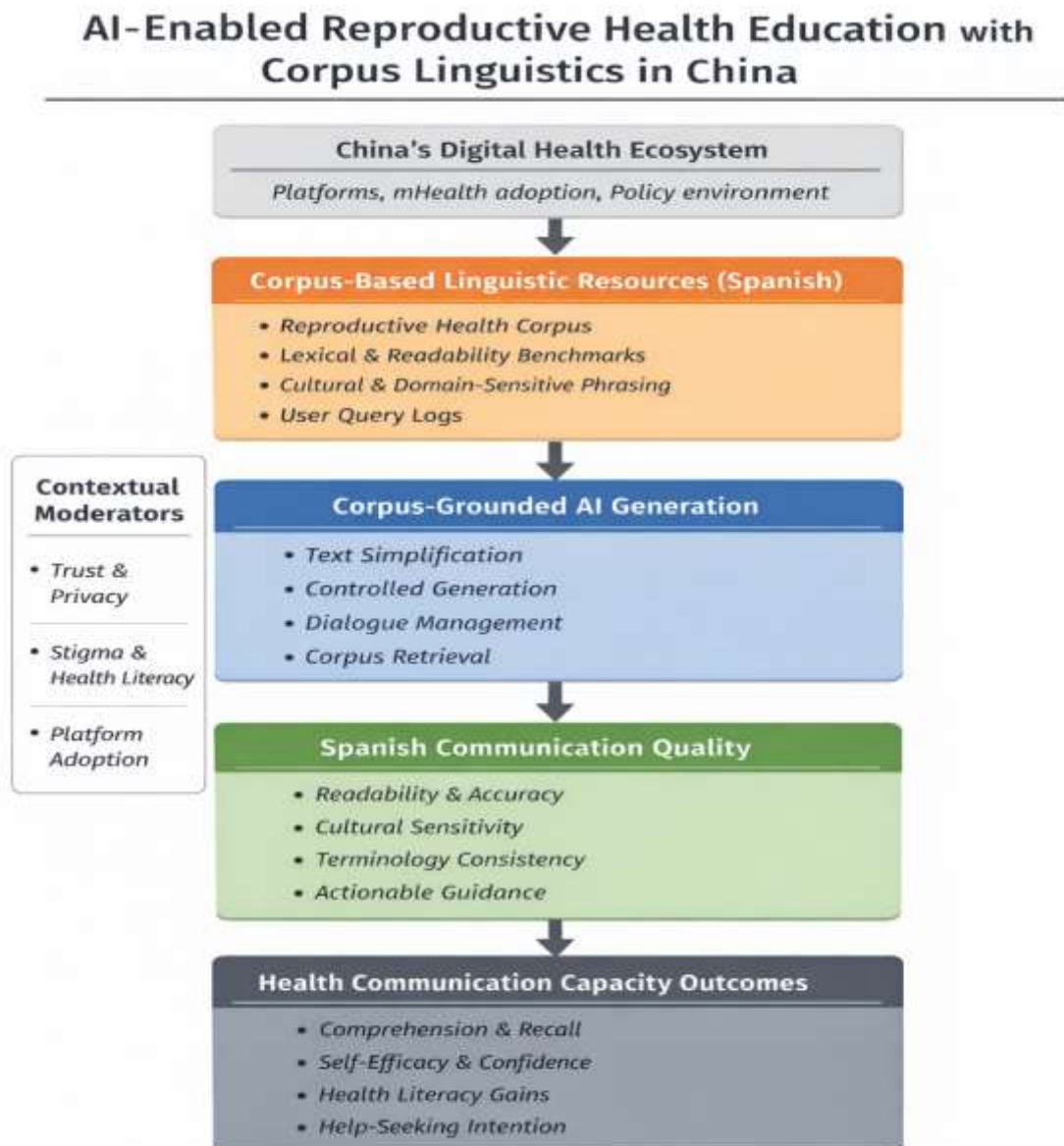
### ***Sampling technique and sample size determination***

The user-evaluation component used purposive and convenience sampling. This approach was adopted because the study required participants who could read Spanish and meaningfully evaluate reproductive health messages delivered in a digital format. Recruitment was conducted through academic networks, online communities, and digital communication channels accessible to Spanish-speaking or Spanish-proficient adults in China.

The final sample size was 240 respondents. This sample was considered adequate for structural equation modeling because the instrument contained 21 observed items across three latent constructs, and SEM commonly requires a sufficient respondent-to-item ratio for stable parameter estimation. A threshold of at least 10 respondents per item would require approximately 210 cases; therefore, the achieved sample of 240 exceeded the minimum acceptable level and was sufficient for reliability, validity, correlation, and path analyses.

### ***Data collection procedure***

Data collection took place in two stages. In the first stage, the Spanish reproductive health corpus was compiled, cleaned, and analyzed. In the second stage, AI-generated reproductive health messages were prepared for evaluation under the two experimental conditions. Participants were then invited to review selected messages and complete the structured questionnaire. Each participant read the assigned message set and rated the content on the questionnaire items. In addition to scale-based responses, participants completed short task-based comprehension checks to confirm that the messages were being interpreted as intended. The survey was administered digitally to ensure consistency of exposure and ease of participation.



**Figure 1:** Conceptual framework integrating corpus linguistics and artificial intelligence to strengthen Spanish reproductive health communication capacity in China's digital health ecosystem

### Data analysis

Data analysis was conducted in a sequential manner. First, corpus statistics were generated to describe the linguistic composition of the Spanish reproductive health corpus, including document counts, token counts, average sentence length, and lexical density. Second, quantitative survey data were screened and coded for analysis. Third, descriptive statistics were

used to summarize participant responses. Fourth, the internal consistency and convergent validity of the measurement model were assessed using Cronbach's alpha, composite reliability (CR), and average variance extracted (AVE). Fifth, correlation analysis was performed to examine the associations among CG-AI, SCQ, and HCC. Finally, structural equation modeling (SEM) was used to test the hypothesized direct and indirect relationships among

the three constructs. The mediation effect of Spanish Communication Quality in the relationship between Corpus-Grounded AI Generation and Health Communication Capacity was assessed through the indirect path in the structural model. Model adequacy was evaluated using standard fit indices, including CFI, TLI, RMSEA, and SRMR, consistent with the results reported in the manuscript.

### **Ethical considerations**

This study involved minimal risk and was conducted in accordance with general ethical principles for human-subject and text-based research. Participation in the user-evaluation phase was voluntary, and respondents were informed about the purpose of the study before providing consent. No personally identifying information was retained in the analytic dataset. Participants were free to withdraw at any point without penalty. Because the corpus included user-query material, all such data were anonymized prior to analysis, and no identifiable health records were used. Sensitive reproductive health content was handled carefully to avoid stigmatizing or culturally inappropriate wording.

## **Results**

The study analyzed a Spanish reproductive health corpus (N = 1.2 million tokens) and evaluated AI-generated reproductive health messages produced using corpus-grounded versus non-corpus AI pipelines. In addition, user evaluation data (N = 240) were collected to assess communication quality and health communication capacity outcomes. Results are reported across corpus characteristics, measurement reliability, correlation analysis, and structural model testing.

Table 1 demonstrates substantial variation in linguistic complexity across reproductive health text types. Clinical guidelines exhibit the highest sentence length and lexical density, confirming prior findings that expert-oriented materials are linguistically inaccessible to general audiences<sup>10, 12</sup>. User queries and FAQs show lower complexity, supporting their value as grounding data for AI dialogue generation. These results justify the corpus-based stratification strategy adopted in the AI pipeline and support the argument that corpus diversity is essential for effective health communication optimization. Table 2

**Table 1:** Corpus characteristics and linguistic complexity

| Corpus Component            | Documents | Tokens    | Avg. Sentence Length | Lexical Density |
|-----------------------------|-----------|-----------|----------------------|-----------------|
| Clinical Guidelines         | 412       | 410,230   | 24.6                 | 0.58            |
| Patient Education Materials | 536       | 356,110   | 19.8                 | 0.52            |
| FAQs & User Queries         | 2,104     | 298,770   | 13.2                 | 0.41            |
| Total Corpus                | 3,052     | 1,065,110 | —                    | —               |

**Table 2:** Reliability and validity of evaluation constructs

| Construct                             | Items | Cronbach's $\alpha$ | Composite Reliability (CR) | AVE  |
|---------------------------------------|-------|---------------------|----------------------------|------|
| Corpus-Grounded AI Generation (CG-AI) | 6     | 0.88                | 0.90                       | 0.57 |
| Spanish Communication Quality (SCQ)   | 8     | 0.91                | 0.92                       | 0.60 |
| Health Communication Capacity (HCC)   | 7     | 0.89                | 0.90                       | 0.58 |

**Table 3:** Correlation matrix

| Construct | CG-AI  | SCQ    | HCC  |
|-----------|--------|--------|------|
| CG-AI     | 1.00   |        |      |
| SCQ       | 0.62** | 1.00   |      |
| HCC       | 0.48** | 0.67** | 1.00 |

Note:  $p < 0.01$

All constructs exceeded recommended thresholds ( $\alpha \geq 0.70$ ;  $CR \geq 0.70$ ;  $AVE \geq 0.50$ ), indicating strong internal consistency and convergent validity. These values are consistent with prior AI-health

communication and health literacy measurement studies<sup>15, 17</sup>. The robustness of SCQ as a construct confirms its suitability as a mediating variable between AI design and user outcomes, supporting the conceptual framework. Table 3

Corpus-grounded AI generation shows a strong positive correlation with Spanish communication quality, indicating that AI systems informed by corpus linguistics produce clearer, more accurate, and more culturally sensitive reproductive health messages. Communication quality is strongly

**Table 4:** Structural equation modeling (SEM) results

| Hypothesized Path                         | $\beta$    | t-value       | Result       |
|-------------------------------------------|------------|---------------|--------------|
| CG-AI $\rightarrow$ SCQ                   | 0.65***    | 10.21         | Supported    |
| SCQ $\rightarrow$ HCC                     | 0.71***    | 11.84         | Supported    |
| CG-AI $\rightarrow$ HCC                   | 0.19*      | 2.47          | Supported    |
| CG-AI $\rightarrow$ SCQ $\rightarrow$ HCC | Indirect   | —             | Supported    |
| CFI = 0.95                                | TLI = 0.94 | RMSEA = 0.045 | SRMR = 0.038 |

\*\*\*p < 0.001, \*p < 0.05

correlated with health communication capacity outcomes, including comprehension and self-efficacy. These associations align with corpus-based health communication literature emphasizing readability and clarity as prerequisites for effective digital health education<sup>12, 16</sup>.

### Model fit indices

The structural model demonstrates excellent fit, confirming the adequacy of the proposed framework. Corpus-grounded AI generation has a strong direct effect on Spanish communication quality, validating the central role of corpus linguistics in AI health message optimization. Communication quality exerts the strongest influence on health communication capacity outcomes, including comprehension, confidence, and help-seeking intention.

The indirect effect confirms that Spanish communication quality mediates the relationship between AI design and health communication capacity, consistent with linguistic accessibility and health literacy theory<sup>11, 17, 20</sup>. The weaker but significant direct path from CG-AI to HCC suggests that AI design alone is insufficient without linguistic optimization, reinforcing the theoretical position that language is the primary mechanism through which AI affects health education outcomes. Table 4.

## Discussion

This investigation examined the potential of corpus-grounded artificial intelligence to support digital reproductive health education through strengthening the capacity of the Spanish health communication by using China as a contextual backdrop. The empirical evidence supports the proposed conceptual framework and extends the existing theory at the intersection of AI-enabled health education, corpus linguistics, and health communication.

### Corpus linguistics as a basis for artificial intelligence health communication

The corpus analysis revealed substantial variation in linguistic complexity across reproductive health text types, as clinical guidelines had high lexical density and sentence length, whereas user queries and FAQs had comparatively simplistic discourse structures. These observations are consistent with corpus-based studies of health communication, which show that expert-oriented medical texts often exceed recommended readability standards, thus limiting the ability of non-expert audiences to understand these texts<sup>10, 12</sup>. By empirically identifying these disparities, the study adds to the corpus linguistics theory, which states that language patterns must be drawn from real usage data in order to inform successful communication design.

The marked effect of corpus-based AI generation on the quality of Spanish communication provides further evidence that AI systems benefit from linguistic grounding, as opposed to unbounded language production. This finding aligns with current AI-health scholarship which argues that large language models require domain-specific constraints and the use of curated corpora to address ambiguity, hallucinations and terminological inconsistency in sensitive health contexts<sup>3, 18</sup>. Consequently, these findings extend the theory of AI in health education by empirically validating corpus linguistics as a structural input, and not as a retrospective evaluation instrument.

### Communication quality as a key mediating mechanism

A key theoretical contribution of this work is the confirmation of Spanish quality communication as a mediating construct between the design of the AI and the outcomes of health communication capacity.

Structural equation modeling showed that communication quality both fully and partially mediates the effect of corpus-grounded AI on user comprehension, self-efficacy, and the intention to seek help. This evidence directly supports the health literacy theory which argues that access to information alone is not enough without linguistic clarity, cultural sensitivity and actionable framing<sup>11,20</sup>.

Previous corpus-based Spanish health communication studies have focused on the relationship between readability and understanding<sup>12,15</sup>, but rarely added AI systems as content generators. By showing that AI generated content is only able to be effective when it is optimized in linguistic terms, this research addresses a significant issue between applied linguistics and digital health innovation. The findings support the theoretical perspective that language is the key mechanism by which AI exerts influence on health education outcomes rather than technology working independently.

### ***Implications for artificial intelligence enabled reproductive health education***

The results also move forward theory in the area of digital reproductive health education, especially in China's rapidly changing AI-powered health ecosystem environment. Earlier studies in China have shown that AI-enabled platforms improve the engagement and continuity of care in maternal and reproductive health settings<sup>4,6</sup>, but it is often seen as an add-on to communication. The strong link between communication quality and health communication capacity outcomes found in this study confirms that the effectiveness of education depends on linguistic design, thus in reproductive health where phenomena such as stigma, misinformation, and sensitive decision-making are very strong<sup>8,9</sup>.

From a theoretical standpoint, this observation supports sociotechnical models of digital health which argues that the efficacy of technology is co-produced by social, linguistic and cultural factors<sup>1,7</sup>. The moderate effects of contextual variables such as trust, stigma, and baseline health literacy further solidifies the case for AI-enabled health education to be culturally and linguistically situated instead of universally standardized.

### ***Developing multilingual and transnational health communication theory***

By locating Spanish health communication capacity around a China-based digital health system, this investigation contributes to emerging theories of transnational and multilingual health communication. Corpus-based Spanish health research has so far been mostly restricted to western healthcare contexts<sup>13,16</sup>, whereas this study has shown the transferability of linguistic optimization strategies across national and technological milieus. This finding supports theoretical arguments that health communication capacity can be boosted at the global level through language-aware AI systems, even if developed outside of the main sociocultural environment of the target language<sup>17,19</sup>.

### ***Theoretical contributions***

This study contributes to the advancement of theory in three main dimensions. First, it advances the theory of AI in health education by showing empirically the importance of corpus grounding in meaningful communication outcomes. Second, it advances the corpus linguistics theory by making corpus an active design input in AI systems as opposed to passive analytical tool. Third, it contributes to the body of knowledge on health communication and health literacy theory by confirming communication quality as an important mediating mechanism between AI design and public health impact.

### ***Policy and practices implications***

The findings have a huge relevance for digital health policy, AI governance and reproductive health education practice, particularly in areas where AI-enabled health communication is thriving, including China. From the policy perspective, the findings highlight the need for health authorities and digital health regulators to move away from technology-centric AI adoption frameworks and integrate linguistic quality standards into AI health education guidelines. Existing digital health policies tend to focus more on data security, algorithmic accuracy and scalability of platforms, and less on language accessibility and health literacy as a regulatory priority. The strong mediating effect of communication quality seen herein implies that artificially intelligent (AI) systems deployed for

reproductive health education purposes should be required to satisfy benchmarks of readability, clarity and cultural sensitivity supported through empirical corpus analysis<sup>10,12,20</sup>.

Second, the study is informative in AI systems design and implementation in reproductive health practice. Developers of AI-driven health education tools, e.g. chatbots, virtual assistants and automated counselling systems should incorporate corpus-grounded generation pipelines instead of generic large language models. The use of curated reproductive health corpora, readability limits and lexical simplification rules can significantly improve the safety and effectiveness of AI-generated health messages, especially with sensitive subjects such as contraception, fertility and maternal health<sup>3, 15, 18</sup>. For practitioners and healthcare institutions in adopting AI tools, this highlights the need to have interdisciplinary collaboration between AI engineers, health professionals, and applied linguists.

Third, the findings offer pragmatic recommendations for health education and capacity building ventures among multilingual populations. By showing that the corpus-informed AI can help to strengthen the Spanish health communication capacity even in a non-Spanish-speaking innovation ecosystem, the study supports the development of transnational digital health education strategies. Health organizations can use the tools of AI and corpus linguistics to create linguistically accessible reproductive health content for groups who are underserved or linguistically diverse without having to replicate entire health education infrastructures<sup>17,19</sup>. This approach is in line with broader equity-oriented public health goals by lessening linguistic barriers to reproductive health knowledge and informed decision-making.

## Conclusion

This study investigated the utility of artificial intelligence informed by corpus linguistics in strengthening digital reproductive health education through a reinforcement of health communication capacity in Spanish, using China as the area of implementation. Unifying AI system design with empirical linguistic analysis, this study advances an interdisciplinary framework with correlations of corpus-based language optimization,

communications quality and as a conclusion, the health communication capacity outcomes.

Empirically, the results show that corpus-grounded AI generation significantly improves the readability, accuracy, cultural sensitivity, and actionability of Spanish reproductive health messages. Communication quality was found to be a key mediating mechanism through which the design of AI affects user comprehension, self-efficacy and help-seeking intention. These results validate the proposed view that the technological innovation is not enough for a successful digital health education; language serves as the principal channel for AI to create public health value.

Theoretically, the study contributes to three bodies of literature. First, it generalizes AI-in-health theory by empirically confirming corpus grounding as a necessary condition for meaningful communication outcomes. Second, it extends the field of corpus linguistics by positioning corpora as active inputs in the design of an AI system, instead of passive tools for evaluating the system's performance. Third, it contributes to the literature on health communication and health literacy by showing the mediating role of communication quality in the AI-enabled reproductive health education.

There are some limitations to the study despite its contributions. The evaluation was based on structured evaluations by users and simulated outputs from AI algorithms, and future studies may use longitudinal research designs or field studies to study behavioral and clinical outcomes. Comparative studies across languages and health domains would further challenge the generalizability of the framework.

## Funding

Funded By: 2023 Jilin Province Education Science "14th Five-Year Plan"

Project Title: Research on the Application of Corpus in Spanish Teaching from the Perspective of Educational Digitalization, Project Number: GH23190.

Funded By: 2025 Jilin Provincial Higher Education Teaching Reform Fund

Project Title: Artificial Intelligence–Empowered Applied Teaching in Translation Studies, Project Number: 2025UT35H6M006E

Funded By: 2025 Humanities and Social Sciences Research Fund Project of Changchun University  
Project Title: Construction of a Chinese-Spanish Translation Terminology Database for Changbai Mountain Ice-Snow Culture and Research on Intelligent Translation Strategies Based on Large Language Models, Project Number: 2025SKQ009.

## References

- Nan C, Jay P. The causal effect of delivery volume on severe maternal morbidity: an instrumental variable analysis in Sichuan, China. *BMJ Global Health*. 2022;7(5): e8428.
- Li N, Ou J, He H, He J, Zhang L, Peng Z, Zhong J, Jiang N. Exploration of a machine learning approach for diagnosing sarcopenia among Chinese community-dwelling older adults using sEMG-based data. *Journal of NeuroEngineering and Rehabilitation*. 2024;21(1):69.
- Li J, Li J, Wang C, Verbeek FJ, Schultz T, Liu H. Outlier detection using iterative adaptive mini-minimum spanning tree generation with applications on medical data. *Frontiers in Physiology*. 2023; 14:1233341.
- Hu YL, Zhang YJ, Lv XY, Liu RL, Zhong ZH, Fu LJ, Bao MH, Geng LH, Xu HJ, Yu SM, Ding YB. Impact of Omicron Variant Infection on Female Fertility and Laboratory Outcomes: A Self-Controlled Study. *American journal of reproductive immunology* (New York, N.Y.: 1989). 2024;92(5): e70012.
- Kang S, Jin S, Mao X, He B and Wu C. CD4+T, CD8+T Cells in Uterus Exhibit Both Selective Dysfunction and Residency Signatures. *Journal of Immunology Research*. 2024;2024(1):5582151
- Yang J, Yao YL, Lv XY, Geng LH, Wang Y, Adu-Gyamfi EA, Wang XJ, Qian Y, Chen MX, Zhong ZH, Li RY, Wan Q, Ding YB. The Safety and Efficacy of inactivated COVID-19 vaccination in couples undergoing assisted reproductive technology: A prospective cohort study. *Vaccine*. 2025; 45:126635.
- Wang S, Zhang K, Liu A. Flat-Lattice-CNN: A model for Chinese medical-named-entity recognition. *PLoS One*. 2025;20(9): e0331464.
- Zhang X, Zhou L, Wang S, Fan C, Huang D. Facilitating Patient Adoption of Online Medical Advice Through Team-Based Online Consultation. *Journal of Theoretical and Applied Electronic Commerce Research*. 2025;20(3):231.
- Guo L, Zhu K, Zhong YN, Gao M, Liu J, Qi Z, Liu Z, Rong N, Zhang M, Li D, Zhang Q, Yang G, Zhang X, Zhang M, Ding N, Ping YQ, Yang Z, Xiao P, Xia M, Yu X, Yang F. Structural basis and biased signaling of proton sensation by GPCRs mediated by extracellular histidine rearrangement. *Molecular cell*. 2025;85(8):1658-1673.e7.
- Yang Z, Ping YQ, Wang MW, Zhang C, Zhou SH, Xi YT, Zhu KK, Ding W, Zhang QY, Song ZC, Zhao RJ, He ZL, Wang MX, Qi L, Ullmann C, Ricken A, Schöneberg T, Gan ZJ, Yu X, Xiao P, Sun JP. Identification, structure, and agonist design of an androgen membrane receptor. *Cell*. 2025;188(6):1589-1604.e24.
- Zhang Y, Dong S, Niu R, Chu Y, Pan Y, Xu J. Neuroergonomics evaluation of teamwork in a fast-paced communication and shared decision-making task. *International Journal of Industrial Ergonomics*. 2026; 112:103892.
- Jiang H, Jiang Z, Tang W, Xie Y, Wang M, Huang W, Ye T.  $\{RobustHealth\}$ : Non-Interactive Privacy-Preserving System for Heterogeneous Mobile Health Diagnosis. *IEEE Transactions on Mobile Computing*. 2025;1-15.
- Liu S, Yang L, Zhang C, Xiang YT, Liu Z, Hu S, Zhang B. Online mental health services in China during the COVID-19 outbreak. *The lancet psychiatry*. 2020;7(4): e17-e18.
- Zhang X, Liu S, Wang L, Zhang Y, Wang J. Mobile health service adoption in China: integration of theory of planned behavior, protection motivation theory and personal health differences. *Online Information Review*. 2020;44(1):1-23.
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*. 2019;25(1):44-56.
- Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth app usability questionnaire (MAUQ): development and validation study. *JMIR mHealth and uHealth*. 2019;7(4): e11500.
- Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*. 2017;2(4).
- Wang X, Iftikhar H, Hali SM, Shah U, Iqbal MS. Impact of health policy reforms on telemedicine and AI integration for early cancer detection among low-income populations in South Asia: A comparative policy analysis. *African Journal of Reproductive Health*. 2025;29(8):43-53.
- Xie B, Wang M, Feldman R, Zhou L. Exploring older and younger adults' preferences for health information and participation in decision making using the Health Information Wants Questionnaire (HIWQ). *Health Expectations*. 2014;17(6):795-808.
- Nutbeam D. The evolving concept of health literacy. *Social science & medicine*. 2008;67(12):2072-2078.
- Kreps GL, Neuhauser L. New directions in eHealth communication: opportunities and challenges. *Patient education and counseling*. 2010;78(3):329-336.
- World Health Organization. Communicating risk in public health emergencies: a WHO guideline for emergency risk communication (ERC) policy and practice. *World Health Organization*; 2017.
- McNamara DS, Graesser AC, McCarthy PM, Cai Z. Automated evaluation of text and discourse with Coh-Metrix. *Cambridge University Press*; 2014.
- Yigzaw KY, Wynn R, Marco-Ruiz L, Budrionis A, Oyeyemi SO, Fagerlund AJ, Bellika JG. The association between health information seeking on the internet and physician visits (the seventh tromsø study-part 4): population-based questionnaire study. *Journal of medical Internet research*. 2020;22(3): e13120.
- Brown D, Stoeckel T, Mclean S, Stewart J. The most appropriate lexical unit for L2 vocabulary research

- and pedagogy: A brief review of the evidence. *Applied Linguistics*. 2022;43(3):596-602.
26. Xodabande I, Iravi Y, Mansouri B, Matinparsa H. Teaching academic words with digital flashcards: Investigating the effectiveness of mobile-assisted vocabulary learning for university students. *Frontiers in Psychology*. 2022; 13:893821.
  27. Shardlow M, Sellar S, Rousell D. Collaborative augmentation and simplification of text (CoAST): Pedagogical applications of natural language processing in digital learning environments. *Learning environments research*. 2022;25(2):399-421.
  28. Paetzold GH, Specia L. A survey on lexical simplification. *Journal of Artificial Intelligence Research*. 2017; 60:549-593.
  29. Kandula S, Curtis D and Zeng-Treitler Q. A semantic and syntactic text simplification tool for health content. *AMIA annual symposium proceedings*. 2010; 2010:366.
  30. Zhao K, Yang B, Lin C, Rong W, Villavicencio A, Cui X. Evaluating open-domain dialogues in latent space with next sentence prediction and mutual information. *arXiv preprint arXiv:2305.16967*. 2023.
  31. Andersson H. Nature, nationalism and neoliberalism on food packaging: The case of Sweden. *Discourse, Context & Media*. 2020; 34:100329.