Revista Bioética
Print version ISSN 1983-8042 | On-line version ISSN 1983-8034

Rev. Bioét. vol.34 Brasília 2026

Bioethics and artificial intelligence: digital content analysis on labiaplasty

Vera Lucia Mota da Fonseca [1], *Jorge Artur Peçanha de Miranda Coelho* [2], *Gustavo Henrique Silva de Souza* [3], *Emanuel Duarte de Almeida Cordeiro* [4], *Luiza Andrade Botelho* [4], *Isabel Moreira Santos* [4], *José Humberto Belmino Chaves* [2], *Rui Nunes* [1]

1. Universidade do Porto, Porto, Portugal. **2.** Universidade Federal de Alagoas, Maceió/AL, Brasil. **3.** Instituto Federal do Norte de Minas Gerais, Teófilo Otoni/MG, Brasil. **4.** Universidade Estadual do Sudoeste da Bahia, Vitória da Conquista/BA, Brasil.

Abstract

Digital marketing for labiaplasty medicalizes natural anatomical variations, necessitating large-scale bioethical screening tools. This study validates a natural language processing framework by evaluating 145 websites. The GPT-4 model, guided by bioethical directives, was compared against independent expert judgment (gold standard). The artificial intelligence demonstrated sensitivity up to 100% in identifying pathologization but null specificity (0.00%), revealing systematic disagreement with humans (Kappa between -0.06 and -0.80). A visual "blind spot" (sensitivity <40%) was identified due to the absence of descriptive metadata on websites. In conclusion, natural language processing is an impartial screening tool yet limited by informational gaps and risks of allocative harm. These results reinforce the need for multimodal approaches and mandatory metadata to safeguard patient autonomy and ethics in digital medical marketing.

Keywords: Natural language processing. Bioethics. Body image. Artificial intelligence.

Resumo

Bioética e inteligência artificial: análise de conteúdo digital sobre labioplastia

O marketing digital da labioplastia frequentemente medicaliza variações anatômicas naturais, exigindo ferramentas de triagem bioética em larga escala. Este estudo valida uma estrutura de processamento de linguagem natural na avaliação de 145 websites. Utilizou-se o modelo GPT-4 guiado por diretivas bioéticas, comparando-o ao julgamento de especialistas independentes (padrão-ouro). A inteligência artificial demonstrou sensibilidade de até 100% na identificação da patologização, mas especificidade nula (0,00%), revelando discordância sistemática com os humanos (Kappa entre -0,06 e -0,80). Identificou-se um "ponto cego" visual (sensibilidade <40%) devido à ausência de metadados descritivos nos sites. Conclui-se que o processamento de linguagem natural é uma ferramenta de triagem imparcial, porém limitada por lacunas informacionais e riscos de dano alocativo. Os resultados reforçam a necessidade de abordagens multimodais e metadados obrigatórios para salvaguardar a autonomia da paciente e a ética no marketing médico digital.

Palavras-chave: Processamento de linguagem natural. Bioética. Imagem corporal. Inteligência artificial.

Resumen

Bioética e inteligencia artificial: análisis de contenido digital sobre labioplastia

El marketing digital de la labioplastia medicaliza las variaciones anatómicas naturales, lo que requiere herramientas de detección bioética a gran escala. Este estudio valida un marco de procesamiento de lenguaje natural evaluando 145 sitios web. El modelo GPT-4, guiado por directivas bioéticas, se comparó con el juicio de expertos independientes (estándar de oro). La inteligencia artificial mostró una sensibilidad de hasta el 100% en la identificación de la patologización, pero una especificidad nula (0,00%), revelando un desacuerdo sistemático con los humanos (Kappa entre -0,06 y -0,80). Se identificó un "punto ciego" visual (sensibilidad <40%) debido a la falta de metadatos descriptivos en los sitios. Se concluye que el procesamiento de lenguaje natural es una herramienta de detección imparcial, aunque limitada por brechas informativas y riesgos de daño alocativo. Los resultados refuerzan la necesidad de enfoques multimodales y metadatos obligatorios para salvaguardar la autonomía de la paciente y la ética en el marketing médico digital.

Palabras clave: Procesamiento de lenguaje natural. Bioética. Imagen corporal. Inteligencia artificial.

The authors declare no conflict of interest.

Research

In recent years, a progressive growth in female genital cosmetic surgeries, particularly labiaplasty, has been observed. According to the most recent survey by the *International Society of Aesthetic Plastic Surgery* [1], 210,633 labiaplasties were performed in 2023, representing 1.2% of all aesthetic surgical procedures globally and ranking as the 21st most performed aesthetic surgery worldwide. Compared to 2020, when 142,119 procedures were recorded, there has been a cumulative increase of 48.2% over four years. While the growth relative to 2022 was more modest (3.7%), this upward trend highlights the consolidation of labiaplasty within the context of female genital cosmetic surgeries (FGCS), interventions aimed exclusively at the aesthetic appearance of the female genitalia.

Such a scenario prompts ethical, clinical, and sociocultural reflections, especially considering that the procedure's popularity occurs alongside a scarcity of robust scientific evidence regarding its safety, efficacy, and precise indications. As labiaplasty becomes widespread in medical practice and collective imagination, questions regarding its actual necessity, potential risks, and implications for women's physical and psychological health have intensified.

Gynecologists-obstetricians must receive adequate training to identify women with sexual function disorders and psychiatric conditions such as depression and anxiety. Patients should be informed that surgical interventions to alter sexual appearance or function lack medical justification, carry substantial risks, and have not been proven safe or effective [2,3].

Ethicists and professional groups have long raised concerns about female cosmetic genital surgery. In evaluating the ethicality of elective vulvar cosmetic surgery, principles such as autonomy, beneficence, nonmaleficence, justice, and truthfulness must be critically considered [4]. Despite reported subjective success rates of 80-90%, these studies are often limited by their retrospective nature, short-term follow-up periods, superficial treatment of body image issues, and lack of control groups [5].

"Subjective success rates" are primarily based on patient satisfaction, which, in medical literature, does not necessarily equate to procedural efficacy or long-term outcomes. This discrepancy underscores the need for more rigorous studies to evaluate the true effectiveness and safety of these procedures. Concerns are further heightened when labiaplasty is performed on minors, as adolescents may not have the maturity or knowledge to make informed decisions about irreversible body modifications [6].

Female genital cosmetic surgery (FGCS) has become a significant and expanding sector within cosmetic surgery, driven by cultural, commercial, and technological forces that uphold and perpetuate rigid beauty standards. Feminist theory and bioethics frame FGCS not only as a medical intervention but as part of the broader medicalization of women's bodies, where natural variations are problematized and surgically altered under the guise of aesthetic improvement [7]. Media, including pornography, promotes the ideal of hairless, "neat" vulvas, intensifying societal pressures on women to conform [8]. This environment, coupled with aggressive marketing by the cosmetic surgery industry, creates a market that encourages women to undergo irreversible procedures to meet narrowly defined beauty standards [9].

Feminist critics argue that FGCS commodifies women's bodies, presenting conformity to beauty standards as empowerment [7]. These ethical concerns, particularly regarding informed consent, coercion, and the promotion of surgical solutions to socially constructed problems, necessitate a critical examination of the cultural, medical, and ethical dimensions of FGCS [8]. The ethical implications extend beyond individual autonomy, as FGCS often exploits women's insecurities and reinforces damaging gender norms [10].

The internet has played a pivotal role in shaping distorted body representations, particularly through the widespread depiction of hairless genitalia in pornographic media and the reinforcement of idealized female aesthetics across social media platforms [6,11,12]. Aggressive marketing and the increasing normalization of cosmetic surgery have further fueled the demand for FGCS [13-15]. Such pervasive exposure functions as a powerful normalization agent; as familiarity with these idealized digital representations increases, individual confidence in pursuing surgical modification grows, subsequently fostering a broader social acceptance of these procedures [16]. Additionally, heteronormative and androcentric depictions of sexual pleasure,

which prioritize heterosexual vaginal intercourse, intensify the pressure on women to conform to specific vulvar aesthetics [10,14,17].

Within this context, language emerges as both a reflection and an agent of influence, functioning as a political tool that shapes societal norms and individual perceptions. This interplay between language and society generates significant impacts within natural language processing, such as the reinforcement of cognitive biases resulting from excessive exposure [18]. For instance, the availability of heuristics leads individuals to overestimate the prevalence or significance of specific aesthetic standards based on their ease of recall, a process frequently intensified by pervasive media coverage. Consequently, both media exposure and individual relationship status have been identified as key factors influencing women's decisions to undergo labiaplasty [19,20].

Labiaplasty websites often employ marketing techniques such as behavioral modeling and vicarious learning to encourage women to seek aesthetic surgery through before-and-after photos, videos, testimonials, and images of satisfied individuals [21]. Internet content normalizes female genital cosmetic surgery by promoting behavioral norms that encourage surgical interventions. Such advertising, which reinforces societal standards regarding body image, demonstrates the effectiveness of behavioral modeling in shaping attitudes and behaviors [14]. Artificial intelligence and machine learning techniques, including natural language processing and computer vision algorithms, can be employed to analyze the content of medical websites [18,22-25].

This research presents a study that employs web scraping and natural language processing techniques to analyze online content related to labiaplasty. The objective of this study is to validate a natural language processing (NLP) framework for the automated bioethical appraisal of websites promoting labiaplasty, investigating how promotional narratives medicalize anatomical variations and reinforce rigid aesthetic standards.

Method

The analysis was conducted using a combination of web scraping and artificial intelligence (AI) techniques. A list of target websites was generated via a standardized Google search using the command: "List websites that offer labiaplasty". A standardized Google search identified an initial pool of 152 websites offering labiaplasty services. Following the screening process, seven websites were excluded as they were inaccessible or returned server errors at the time of data collection. Consequently, the final analytical sample consisted of 145 accessed websites. Data collection and automated processing were conducted between June 26 and July 30, 2025.

Data acquisition and web scraping

The data collection was designed to capture the primary communicative elements of labiaplasty-related websites. The procedure involved:

a. Content extraction: An automated scraping pipeline was implemented to parse the HTML structure of each target website. We prioritized the extraction of textual content from the main body, systematically excluding non-content-related nodes such as scripts, CSS styles, and commercial advertisements to ensure data integrity.

b. Visual metadata harvesting: Since visual marketing is a cornerstone of aesthetic surgery promotion, the scraper targeted the "alt" attributes of all image tags (img). This process converted visual stimuli into textual metadata, enabling the language model to analyze the bioethical context of imagery that would otherwise be inaccessible to text-only processing.

AI-driven bioethical appraisal

This study used a large language model (LLM) to perform an automated bioethical evaluation of the scraped content. This process ensured a standardized and replicable assessment of the promotional narratives. The evaluation was guided by three specific directives designed to identify the medicalization of aesthetics and the pathologization of anatomical variations.

a. Natural language processing: The analysis utilized the GPT-4 engine (gpt-4-0613 or latest available version) [26] engine via a dedicated API interface. This model was chosen for its advanced capability in processing complex

ethical nuances and generating narrative-style evaluations.

b. Instructional directives (prompts): The evaluation was guided by three specific bioethical directives designed to identify: (1) the promotion of harmful appearance norms; (2) the conflation of surgical aesthetics with sexual functionality or youthfulness; and (3) the pathologization of normal anatomical variations. The complete text for each instructional directive and representative samples of AI-generated narratives is provided in Appendix 1.

c. Processing pipeline: To maintain consistency across the 145 websites, the input was normalized through filtering and truncation before being processed by the LLM, ensuring the analysis focused on the core promotional narrative of each site.

Categorization of ai narratives

Since the GPT-4 model was instructed to provide narrative-style bioethical evaluations rather than simple binary outputs, a systematic post-hoc coding process was required. To ensure maximum methodological rigor and prevent confirmation bias, these textual narratives were independently categorized into binary classifications (yes/no) by a third researcher who was not involved in the initial human validation phase. This coder remained fully blinded to the results provided by the independent expert judges, ensuring that the interpretation of the AI's evaluative content was conducted without prior knowledge of the criterion-reference scores.

Human validation (criterion-reference)

To ensure the reliability of the AI-generated analyses and address potential concerns regarding algorithmic bias or inaccuracy, a human validation phase was conducted. Two independent judges, experienced in the field, were recruited to provide a gold standard (criterion-reference) for the evaluation of the 145 websites.

The validation framework established a strict separation of roles: Two independent raters provided the gold standard reference, while a third, unrelated researcher performed the blinded categorization of the AI's textual outputs. This triple-blinded approach was designed to uphold the psychometric integrity of the criterion-validity analysis.

Procedures and blinding

The experts performed a blind evaluation, meaning they had no access to the previous outputs generated by the OpenAI model. Each judge independently assessed the same 145 websites using a binary scale (yes/no) for three specific criteria that mirrored the prompts provided to the AI:

a. Criterion 1: Presence of language or imagery promoting unrealistic or harmful ideals of female genital appearance.

b. Criterion 2: Presence of before-and-after images suggesting enhanced appearance, sexual responsiveness, or pleasure.

c. Criterion 3: Presence of language misrepresenting natural vulvar variations as abnormal or promoting procedures as "empowerment".

Specific instructions were provided to ensure standardized assessment across the sample.

Statistical design and performance evaluation

The statistical analysis was designed to evaluate the criterion validity of the AI model through a multi-reference diagnostic framework. To ensure a robust assessment of the AI's performance in the face of potentially varied expert interpretations, the analytical plan was structured as follows:

a. Inter-rater reliability: To assess the stability of the human gold standard, Cohen's Kappa (kappa) was planned for each of the three evaluation criteria. This measure was intended to categorize the tasks as either objective (high agreement) or subjective (low agreement).

b. Multi-reference validation strategy: Given the interpretative nature of bioethical criteria, this study utilized a "divergent approach." Instead of forcing a consensus, the AI's performance was calculated individually against each expert judge. This strategy allows for an assessment of the AI's alignment with different professional frameworks.

c. Performance metrics: The AI's diagnostic accuracy was quantified using the following metrics, calculated for each judge-AI pair:

- Sensitivity (true positive rate): The model's capacity to correctly identify sites with infractions.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

- Specificity (true negative rate): The model's capacity to correctly identify ethical or educational sites.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- Overall accuracy: The total proportion of correct classifications.

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}}$$

Software and code availability

To ensure the transparency of the computational pipeline and the reproducibility of the findings, the following environment and tools were utilized:

a. Computational environment: All data collection, processing, and statistical analyses were performed using the R programming language (version 4.2.2) [27]. The analyses were conducted on a macOS Ventura operating system (version 13.2.1) [28], using RStudio as the Integrated Development Environment (IDE) [29].

b. Web scraping and AI integration: The extraction of textual content and visual metadata from the 145 websites was executed using the rvest package. The interface between R and the OpenAI GPT-4 API was managed via the reticulate package, which allowed for the direct application of the GPT-4 model to the scraped data.

c. Statistical software and metrics: Inter-rater reliability (Cohen's Kappa) was calculated using the irr package. Performance metrics for criterion validity, including sensitivity, specificity, and accuracy, were derived through custom-built scripts in base R to ensure full transparency in the diagnostic calculations.

d. Data stewardship: The complete dataset, including the raw scraped text, AI-generated evaluation narratives, and independent expert classifications, is documented and available for peer review to uphold Open Science standards. To ensure full transparency and reproducibility, the detailed prompt specifications, sample evaluations, and the comprehensive R programming script are hosted in a public GitHub repository [30].

Results

The analyzed websites' content frequently emphasized the importance of physical beauty and aesthetics, employing terms like "abnormal labia size", "abnormal vaginal shape", and "vaginal rejuvenation", and associating these procedures with luxury and prestige. This messaging frequently targets vulnerable populations, providing discounts to specific groups and employing potentially triggering language for those dealing with body image issues. Instead of procedure details, the websites emphasize before-and-after photos and doctor credentials, downplaying potential risks and complications. Phrases like "perfectly shaped labia", "restore your femininity", and "enhance self-confidence" reinforce these unattainable standards.

The qualitative analysis of the AI-generated narratives illustrates this pattern; for instance, when evaluating a website that featured imagery of social approval and concealing gestures, the model reasoned as follows: "The language used in this website promotes unrealistic and harmful ideals of female genital appearance by suggesting that women should have a certain type of genital appearance to be attractive or desirable. The imagery of a woman covering her private parts and a group of women laughing implies that there is anatomical deviance in the way a woman's genitalia looks and that it should be changed to fit a certain standard of beauty".

Furthermore, marketing language frequently employs behavioral modeling, implying that women must change their bodies to be attractive and desirable, and contributing to a body-shaming culture. Women who do not meet these standards may experience increased body dissatisfaction as a result. These findings highlight the significant role that marketing tactics play in shaping societal and individual perceptions of beauty, which may have implications for women's psychological health and self-perception.

In contrast, a small subset of the sample prioritized educational content over promotional narratives. Only four out of the 145 analyzed websites met these strictly educational criteria, such as the following example: "A labiaplasty is a surgical procedure used to reduce the size of the labia minora... It should not be performed on girls under the age of 18 because the labia continue to develop after puberty and into early adulthood. The labiaplasty is rarely and expensively available on the NHS".

Validation and inter-rater reliability

Inter-rater reliability varied significantly by criterion. Question 2 (objective imagery) showed near-perfect agreement (kappa=0.95; $p<0.001$). In contrast, Question 1 (kappa=-0.04) and Question 3 (kappa=0.13) demonstrated poor agreement. For these subjective criteria, the AI's performance was compared separately against each judge to identify patterns of alignment or systematic bias.

Criterion validity and AI performance

The diagnostic performance of the AI model relative to the human criterion-reference is summarized in Table 1. For the identification of before-and-after imagery (Question 2), the AI achieved an accuracy of 47.0% and a sensitivity of 39% when compared to Judge A. Similar results were observed for Judge B, with an accuracy of 47.8% and sensitivity of 39%.

In the assessment of bioethical criteria, the AI demonstrated high sensitivity across both references, ranging from 97.9% for Question 1 (Judge A) to 100% for Question 3 (both judges). Conversely, specificity for these criteria remained null (0.00%).

Table 1. AI diagnostic performance metrics against independent human experts

Criterion (Prompt)	Reference	Sensitivity	Specificity	Accuracy	Kappa
Q2 (Objective Imagery)	IA vs. Judge A	0.39	0.55	0.47	−0.06^{ns}
	IA vs. Judge B	0.39	0.55	0.48	−0.61^{ns}
	Judge A vs. Judge B	–	–	–	0.95***
Q1 (Harmful Ideals)	IA vs. Judge A	0.97	0.00	0.35	−0.48***
	IA vs. Judge B	1.00	0.00	0.10	−0.80***
	Judge A vs. Judge B	–	–	–	−0.04^{ns}
Q3 (Pathologization)	IA vs. Judge A	1.00	0.00	0.58	−0.26***
	IA vs. Judge B	1.00	0.00	0.24	−0.61***
	Judge A vs. Judge B	–	–	–	0.13^{ns}

****: p<001; **: p<01; *: p<05; ns: non-significant. P-values refer to the test of the null hypothesis that agreement is due to chance (Kappa=0).*

Discussion

General findings

The automated web analysis demonstrates that promotional content for labiaplasty systematically reinforces rigid beauty norms by problematizing natural anatomical variations. These narratives frame surgical intervention as a foundational requirement for self-confidence, effectively medicalizing aesthetic preferences and positioning the female body as a commodity subject to professional modification. This systematic commodification aligns with previous sociocultural critiques, illustrating how the digital marketing of female genital cosmetic surgery (FGCS) entrenches restrictive gender norms by presenting surgery as a predetermined response to socially constructed insecurities [7-9].

In parallel, the framing of FGCS as a surgical resolution for socially constructed concerns serves to entrench restrictive gender norms, effectively positioning body modification as a predetermined response to the very insecurities fostered by these normative standards [7].

In parallel, the frequent association of procedural outcomes with sexual responsiveness and youthfulness confirms the prevalence

of marketing strategies that conflict physical appearance with sexual desirability. By prioritizing "neat" and "youthful" aesthetics, these websites exert substantive pressure on women to conform to idealized standards frequently promoted through media portrayals [12,19,20]. The clinical implications of this marketing are profound, as the consistent downplaying of physical and psychological risks complicates the process of obtaining truly informed consent. The lack of transparency regarding procedural complications echoes the contentions of the medical community regarding the ethical justification of elective cosmetic genital surgery in the absence of robust safety evidence [10,31].

Ultimately, the application of natural language processing (NLP) in this research demonstrates the substantive potential of computational techniques for large-scale bioethical assessment. By serving as a standardized benchmark, the AI model provides a level of evaluative consistency that supplements the inherent subjectivity identified among human experts, offering a foundational framework for the future regulation of digital medical marketing.

Comparison with prior work

The findings of this research expand upon established qualitative and manual content analyses regarding the digital promotion of labiaplasty. The role of online content in normalizing FGCS and perpetuating rigid beauty standards has been well documented in previous systematic reviews [14,15,32]. However, traditional manual methodologies are often constrained by limited sample sizes and the significant human labor required for qualitative coding, which may hinder the ability to conduct continuous or large-scale surveillance.

In contrast, this study introduces a pivotal advancement by employing NLP to automate the bioethical assessment of a substantive dataset. While manual critical discourse analysis has been traditionally utilized to identify the sensationalism and medicalization of the female pudenda, the current model provides a standardized benchmark that processes these narratives with high sensitivity [33]. This transition to an automated evaluative framework allows for a more consistent identification of pathologization patterns that might otherwise be subject to individual interpretative bias in purely human-led evaluations.

Furthermore, the identification of a "blind spot" in text-based models offers a more nuanced understanding of the systemic limitations inherent in current AI technology, particularly regarding the ethical challenges of implementing machine learning in healthcare [22]. While earlier works focused primarily on the socio-cultural impact of promotional content, this research contributes a diagnostic performance analysis that bridges the gap between bioethical critique and computational validation.

The challenge of subjectivity in bioethical analysis

The marked disparity in inter-rater reliability reveals a central tension in digital bioethical assessment. While human experts reached near-perfect agreement on objective tasks, such as identifying the presence of before-and-after imagery (kappa=0.95; p<0.001), their consensus vanished when navigating the ethical nuances of the content. The poor agreement found for Question 1 (kappa=-0.04) and Question 3 (kappa=0.13), this evidence demonstrates that evaluating medical marketing is inherently interpretative, necessitating standardized benchmarks to supplement individual clinical judgment.

Furthermore, the impartiality of this benchmark was safeguarded by using a third independent researcher to categorize the AI's textual narratives. By ensuring this researcher remained blinded to the expert judges' scores, we mitigated the risk of confirmation bias, demonstrating that the AI's high sensitivity is a robust property of its linguistic reasoning rather than an artifact of interpretative alignment.

This subjectivity indicates that concepts such as "pathologization," "empowerment," and "unrealistic beauty standards" are not static benchmarks; they are filtered through the individual professional's framework, clinical experience, and sociocultural values. In this study, what one expert identified as a clear bioethical infraction regarding anatomical variation, another may have interpreted as acceptable clinical information. This "gold standard" problem, the absence of a universal consensus on what constitutes a digital

ethical violation, represents the primary hurdle in the regulation of medical marketing.

Within this context, the role of AI shifts from a mere automated substitute to a standardized benchmark. Although the AI demonstrated lower specificity, its sensitivity of up to 100% in identifying pathologization (Q3) suggests its utility as an impartial "first-pass" assessment. By applying the same instructional directives (prompts) across all 145 websites, the AI bypasses the individual biases of human judges, providing consistent, albeit rigid, ethical screening that serves as a foundation for subsequent human clinical judgment.

However, the premise of AI as a purely impartial benchmark must be approached with caution. Recent evidence indicates that LLMs systematically shift their ethical priorities, such as the balance between autonomy and justice, when presented with different socio-demographic cues [34]. This suggests that while AI bypasses individual human judge bias, it may introduce algorithmic biases that mirror or even amplify societal inequities, particularly in sensitive health-related decision-making.

The blind spot of text-based AI in bioethical surveillance

A critical finding of this study is the discrepancy between human expert consensus and AI performance in identifying visual marketing tactics. While the experts achieved near-perfect agreement in detecting before-and-after photos, the AI's sensitivity remained below 40%.

This "blind spot" highlights a fundamental limitation of current NLP models in bioethical assessment. Since the methodology relied on web scraping text and image descriptions (alt-text), the AI's failure suggests that many labiaplasty websites do not employ descriptive metadata for their visual content. Consequently, while a human visitor is immediately exposed to potentially harmful visual standards, a text-based automated assessment tool remains unaware of these infractions. This underscores the necessity for multi-modal AI approaches that combine computer vision with NLP for effective regulatory oversight of medical marketing.

The observed performance gap in imagery detection (<40% sensitivity) highlights a fundamental limitation of text-based NLP models. As the AI's diagnostic ability was constrained by its reliance on textual metadata (alt-text), it remained "blind" to visual adverts that were readily identified by human experts. This finding suggests that many labiaplasty websites do not employ descriptive metadata, rendering visual marketing infractions invisible to automated textual evaluative models. Beyond the technical blind spot regarding visual metadata, there exists a second, more nuanced limitation: socio-demographic bias. If promotional content for labiaplasty subtly targets specific economic or ethnic groups, current LLMs may alter their bioethical assessments based on these identifiers, potentially directing more utilitarian or pathologizing narratives toward certain populations [34].

Recent evidence highlights that while LLMs like GPT-4 can offer sophisticated linguistic reasoning, they exhibit systematic limitations when faced with complex medical ethical dilemmas, particularly regarding the prioritization of patient autonomy [35]. These findings align with the performance gaps observed in our study, where the model demonstrated high sensitivity but was constrained by a lack of multi-modal data integration.

Implications for policy and practice

The findings of this study offer substantive contributions to the regulation of digital healthcare environments. First, the high sensitivity demonstrated by the LLM in identifying the pathologization of anatomical variations suggests that automated bioethical assessment can serve as a pivotal tool for health surveillance agencies. Regulatory bodies, such as medical councils, could implement these AI-driven frameworks to conduct continuous, large-scale monitoring of online medical marketing, providing a standardized benchmark that addresses the ethical challenges of implementing machine learning in healthcare [22].

Second, the identification of a blind spot in text-based AI highlights a systemic requirement for improved transparency in medical web design. It is imperative that future digital advertising guidelines for surgical procedures mandate the inclusion of descriptive metadata (alt-text) for all clinical and promotional imagery. Such a requirement would not only facilitate more effective automated

oversight but also uphold the principles of digital accessibility and transparency [10].

Finally, for clinical practitioners, these results underscore the necessity of incorporating a critique of digital beauty standards into preoperative consultation. By acknowledging how online narratives often medicalize natural variations, clinicians can better assist patients in navigating the psychological dimensions of genital surgery, ensuring that the decision-making process is grounded in clinical reality and professional ethics [2].

Strengths and limitations

The robustness of this research is grounded in several methodological pillars. First, the application of NLP for the bioethical assessment of medical marketing represents a pivotal innovation in digital health surveillance. By employing an LLM as a standardized evaluative tool, this study demonstrates the utility of AI in identifying nuanced ethical infractions across a substantive dataset of 145 websites. Second, the use of blinded independent experts to establish a gold standard provided a rigorous framework for assessing the AI's criterion validity. The subsequent divergence approach utilized to handle inter-rater variability ensured that the inherent subjectivity of bioethical analysis was treated with psychometric precision. Furthermore, the study upholds Open Science standards by detailing the complete computational environment and providing the datasets for peer review, ensuring the reproducibility of the findings.

A significant strength of this study lies in its triple-blinded methodological structure. Beyond the use of independent expert judges, the final categorization of AI-generated narratives was performed by a third, unrelated researcher with no access to the human criterion-reference. This design ensures that the transition from qualitative AI outputs to quantitative performance metrics follows the highest standards of psychometric integrity.

Despite these strengths, certain limitations must be acknowledged:

- Scope and generalizability. While the analysis included 145 websites, this sample size may limit the generalizability of the findings to the broader global market of labiaplasty promotion.
- Search engine and selection bias. The sample was derived from a standardized Google search, which potentially prioritizes websites with high search engine optimization (SEO), possibly excluding less visible clinical platforms or independent forums.
- Modal constraint (the blind spot). The AI's diagnostic accuracy for visual content remained limited, with sensitivity falling below 40%. This performance gap is attributed to the model's reliance on textual metadata (alt-text), indicating that text-based automated assessments remain "blind" to potentially harmful visual marketing that lacks descriptive descriptors.
- Technological context. The results reflect the capabilities of the GPT-4 architecture as utilized during the data collection period in 2025. While GPT-4 represents a significant advancement over previous iterations (like GPT-3.5), performance metrics and ethical detection capabilities continue to evolve with the emergence of multi-modal and reasoning-heavy models.
- Algorithmic fairness and measurement bias. From a psychometric perspective, the disparity between the high sensitivity and extremely low specificity observed in our results (Table 1) suggests a challenge of algorithmic fairness and measurement bias. The model's tendency toward 'over-identification' indicates that it may be functioning with a rigid evaluative template that lacks the necessary calibration to distinguish between legitimate clinical information and unethical pathologization. This identifies a risk of allocative harm, where educational or neutral content is unfairly penalized by the AI due to linguistic patterns rather than actual bioethical infractions. Recent evidence highlights that while clinical NLP models exhibit systematic limitations in prioritizing patient autonomy [35], the path forward involves the development of Fairness Certification frameworks. These frameworks aim to provide formal guarantees that LLMs meet specific equity criteria before being deployed in sensitive healthcare assessments, mitigating the risks of biased pathologization identified in our study [36,37].

Final considerations

This research clarifies the intricate socio-cultural and ethical dynamics propelling the demand for labiaplasty. The evidence indicates that online promotional content frequently perpetuates unrealistic appearance standards, often presenting surgical modification as a primary solution for socially constructed insecurities.

A primary takeaway is the identification of a "blind spot" in current text-based AI monitoring; while these models are effective for large-scale textual screening, they remain limited by a lack of visual metadata on medical websites, allowing harmful visual marketing to persist undetected. Additionally, the divergence among human experts regarding ethical infractions highlights that bioethical screening is a nuanced, interpretative task that necessitates standardized benchmarks to supplement individual clinical judgment.

These findings necessitate a shift toward more robust, multi-modal regulatory oversight of medical marketing. For practitioners and policymakers, the results emphasize the requirement for enhanced education on the psychological dimensions of genital cosmetic surgery. Moving forward, the integration of AI-driven tools with human ethical oversight provides a foundational framework for safeguarding patient autonomy and ensuring transparency in the digital healthcare environment.

References

1. International Society of Aesthetic Plastic Surgery. ISAPS International Survey on Aesthetic/Cosmetic Procedures Performed in 2023 [Internet]. Munich: ISAPS; 2023 [acesso 20 fev 2026]. Disponível: https://bit.ly/45T9Lax
2. Elective female genital cosmetic surgery: ACOG Committee Opinion, no. 795. Obstet Gynecol [Internet]. 2020 [acesso 20 fev 2026];135:e36-e42. DOI: 10.1097/AOG.0000000000003616
3. Committee on Gynecologic Practice, American College of Obstetricians and Gynecologists. ACOG Committee Opinion no. 378: vaginal "rejuvenation" and cosmetic vaginal procedures. Obstet Gynecol [Internet]. 2007 [acesso 20 fev 2026];110(3):737-8. DOI: 10.1097/01.AOG.0000263927.82639.9b
4. Goodman MP. Female cosmetic genital surgery [Internet]. Obstet Gynecol [Internet]. 2009 [acesso 20 fev 2026];113(1):154-9. DOI: 10.1097/AOG.0b013e318190c0ea
5. Goodman MP. Female genital cosmetic and plastic surgery: a review. J Sex Med [Internet]. 2011 [acesso 20 fev 2026];8(6):1813-25. DOI: 10.1111/j.1743-6109.2011.02254.x
6. Kalampalikis A, Michala L. Cosmetic labiaplasty on minors: a review of current trends and evidence. Int J Impot Res [Internet]. 2023 [acesso 20 fev 2026];35(3):192-5. DOI: 10.1038/s41443-021-00480-1
7. Tiefer L. Female genital cosmetic surgery: freakish or inevitable? Analysis from medical marketing, bioethics, and feminist theory. Feminism & Psychology [Internet]. 2008 [acesso 20 fev 2026];18(4):466-79. DOI: 10.1177/0959353508095529
8. Braun V. Female genital cosmetic surgery: a critical review of current knowledge and contemporary debates. J Womens Health (Larchmt) [Internet]. 2010 [acesso 20 fev 2026];19(7):1393-407. DOI: 10.1089/jwh.2009.1728
9. Braun V. Selling a perfect vulva?: selling a "normal" vulva!. In: Creighton SM, Liao L-M, editores. Female Genital Cosmetic Surgery: Solution to What Problem? Cambridge: Cambridge University Press 2019. p. 23-32
10. Barbara G, Facchin F, Buggio L, Alberico D, Pina M, Kustermann A. Vaginal rejuvenation: current perspectives. Dovepress [Internet]. 2017 [acesso 20 fev 2026];(9):513-9. DOI: 10.2147/IJWH.S99700
11. Piña-García CA, Cruz G, Espinoza A. The impact of social media on sexual self-concept: a qualitative analysis of Instagram in Mexico. Sexes [Internet]. 2025 [acesso 20 fev 2026];6(1):9. DOI: 10.3390/sexes6010009
12. Veras EM, Román Ledesma S, Acosta Matos JA, Cortorreal MEC, Goncharova I, Bonilla RBR et al. Influence of social media filters on plastic surgery: a surgeon's perspective on evolving patient demands. Cureus [Internet]. 2025 [acesso 20 fev 2026];17(3):e80483. DOI: 10.7759/cureus.80483
13. Ashong AC, Batta HE. Sensationalising the female pudenda: an examination of public communication of aesthetic genital surgery. Glob J Health Sci [Internet]. 2013 [acesso 20 fev 2026];5(2):153-65. DOI: 10.5539/gjhs.v5n2p153

14. Mowat H, McDonald K, Dobson AS, Fisher J, Kirkman M. The contribution of online content to the promotion and normalisation of female genital cosmetic surgery: a systematic review of the literature. BMC Women's Health [Internet]. 2015 [acesso 20 fev 2026];15(1):110. DOI: 10.1186/s12905-015-0271-5
15. Ortega-Sánchez I, Lucha-López MO, Monti-Ballano S. Labiaplasty: a qualitative analysis of online discourse on labia minora. Sexes [Internet]. 2024 [acesso 20 fev 2026];5(4):721-38. DOI: 10.3390/sexes5040046
16. Blake AB, Nazarian M, Castel AD. Rapid Communication: The Apple of the mind's eye: everyday attention, metamemory, and reconstructive memory for the Apple logo. Sage Journals [Internet]. 2015 [acesso 23 jan 2026];68(5):858-65. DOI: 10.1080/17470218.2014.1002798
17. Boddy J. Re-thinking the zero-tolerance approach to FGM/C: the debate around female genital cosmetic surgery. Curr Sex Health Rep [Internet]. 2020 [acesso 20 fev 2026];12(4):302-13. DOI: 10.1007/s11930-020-00293-1
18. Hovy D, Spruit SL. The social impact of natural language processing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (v. 2: Short Papers) [Internet]. Berlin: ACL; 2016 [acesso 20 fev 2026]. p. 591-8. DOI: 10.18653/v1/P16-2096
19. Mohammed GF. The impact of social media on female genital self-image. Sex Med Rev [Internet]. 2025 [acesso 20 fev 2026];13(4):698-705. DOI: 10.1093/sxmrev/qeaf031
20. Sharp G, Mattiske J, Vale KI. Motivations, expectations, and experiences of labiaplasty: a qualitative study. Aesthet Surg J [Internet]. 2016 [acesso 20 fev 2026];36(8):920-8. DOI: 10.1093/asj/sjw014
21. Lee M, Lee HH. Do parasocial interactions and vicarious experiences in the beauty YouTube channels promote consumer purchase intention? Int J Consum Stud [Internet]. 2022 [acesso 20 fev 2026];46(1):235-48. DOI: 10.1111/ijcs.12667
22. Char DS, Shah NH, Magnus D. Implementing machine learning in health care: addressing ethical challenges. N Engl J Med [Internet]. 2018 [acesso 20 fev 2026];378(11):981-3. DOI: 10.1056/NEJMp1714229
23. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editores. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, v. 1 (Long and Short Papers) [Internet]. Berlin: ACL; 2019 [acesso 20 fev 2026]. p. 4171-86. DOI: 10.18653/v1/N19-1423
24. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM [Internet]. 2017 [acesso 20 fev 2026];60(6):84-90. DOI: 10.1145/3065386
25. Lipton ZC, Steinhardt J. Troubling trends in machine learning scholarship. Queue [Internet]. 2019 [acesso 20 fev 2026];17(1):45-77. DOI: 10.1145/3317287.3328534
26. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I et al. GPT-4 Technical Report. Computer Science [Internet]. 2023 [acesso 20 fev 2026];arXiv:2303.08774v6 [cs.CL]. DOI: 10.48550/arXiv.2303.08774
27. R: The R Project for Statistical Computing [Internet]. Vienna: The R Foundation; 2024 [acesso 20 fev 2026]. Disponível: https://www.r-project.org
28. Apple Inc. macOS Ventura [sistema operacional]. Versão 13.2.1. Cupertino: Apple Inc.; 2023.
29. RStudio Team. RStudio: Integrated Development Environment for R [programa de computador]. Boston: RStudio, PBC [Internet]; 2021 [acesso 20 fev 2026]. Disponível: http://www.rstudio.com/
30. Cordeiro EDA. Supplementary material: bioethics and artificial intelligence: digital content analysis on labiaplasty [Internet]. San Francisco: GitHub; 2026 [acesso 25 fev 2026]. https://github.com/Eddac/Bioetica_article
31. Halder GE, Iglesia CB, Rogers RG. Controversies in female genital cosmetic surgeries. Clin Obstet Gynecol [Internet]. 2020 [acesso 25 fev 2026];63(2):277. DOI: 10.1097/GRF.0000000000000519
32. Moran C, Lee C. 'Everyone wants a vagina that looks less like a vagina': Australian women's views on dissatisfaction with genital appearance. J Health Psychol [Internet]. 2018 [acesso 25 fev 2026];23(2):229-39. DOI: 10.1177/1359105316637588
33. Syed SA, Anwar A. The globally rising tide of cosmetic gynaecology: are providers aware of the ethical aspects? Cureus [Internet]. 2025 [acesso 25 fev 2026];17(11):e96327. DOI: 10.7759/cureus.96327
34. Sorin V, Korfiatis P, Collins JD, Apakama D, Omar M, Glicksberg BS et al. Socio-demographic modifiers shape large language models' ethical decisions. J Healthc Inform Res [Internet]. 2025 [acesso 20 fev 2026];9(4):567-86. DOI: 10.1007/s41666-025-00211-x

35. Xiong YT, Zeng YM, Liu HN, Sun YN, Tang W, Liu C. Exploring the medical ethical limitations of GPT-4 in clinical decision-making scenarios: a pilot survey. Front Public Health [Internet]. 2025 [acesso 20 fev 2026];13. DOI: 10.3389/fpubh.2025.1582377

36. Anadria D, Giachanou A, Kernahan J, Dobbe R, Oberski D. Algorithmic fairness in clinical natural language processing: challenges and opportunities [Internet]. In: Proceedings of the 3rd European Workshop on Algorithmic Fairness (EWAF 2024); 1-3 jul 2025; Mainz. Aachen: CEUR-WS; 2025 [acesso 20 fev 2026]. https://ceur-ws.org/Vol-3908/paper_50.pdf

37. Freiberger V, Buchmann E. Fairness certification for natural language processing and large language models [Internet]. In: Arai K, editor. Intelligent systems and applications. Cham: Springer Nature Switzerland; 2024 [acesso 20 fev 2026]. p. 606-24. DOI: 10.1007/978-3-031-66329-1_39

Vera Lucia Mota da Fonseca – Master – veralmfonseca@gmail.com
0009-0008-3636-0784

Jorge Artur Peçanha de Miranda Coelho – PhD – jorge.coelho@famed.ufal.br
0000-0002-0021-5963

Gustavo Henrique Silva de Souza – Master – gustavo.souza@ifnmg.edu.br
0000-0002-4046-9669

Emanuel Duarte de Almeida Cordeiro – PhD – emanuel.cordeiro@uesb.edu.br
0000-0001-6437-3197

Luiza Andrade Botelho – Undergraduate – luiza.andrade.bot@gmail.com
0009-0002-4843-5858

Isabel Moreira Santos – Undergraduate – isabelmoreirasantos2002@gmail.com
0009-0008-8183-6104

José Humberto Belmino Chaves – PhD – jose.chaves@famed.ufal.br
0000-0003-2704-6538

Rui Nunes – PhD – ruinunes@med.up.pt
0000-0002-1377-9899

Correspondence
Jorge Artur Peçanha de Miranda Coelho. Universidade Federal de Alagoas. Av. Lourival Melo Mota, s/n, Tabuleiro do Martins. CEP 57072-900. Maceió/AL, Brasil.

Contribution of the authors

Vera Lucia Mota da Fonseca: conceptualization; methodology; writing – original draft. Jorge Artur Peçanha de Miranda Coelho: conceptualization; formal analysis; writing – original draft; validation. Gustavo Henrique Silva de Souza: conceptualization; writing – original draft. Emanuel Duarte de Almeida Cordeiro: data curation; formal analysis; writing – original draft. Luiza Andrade Botelho: data curation. Isabel Moreira Santos: data curation. José Humberto Belmino Chaves: writing – original draft; validation. Rui Nunes: conceptualization; validation.

Data availability: All data used or generated in this study are described and presented in full in the body of the article.

Editor in charge: Dilza Teresinha Ambrós Ribeiro

Received: 26.2.2026
Revised: 8.4.2026
Approved: 17.4.2026

Research