Indonesian Journal of Medical Linguistics
Vol. 1, No. 2, June 2026, pp. 82~94
ISSN: 3163-3676, DOI:

When fluent advice becomes unsafe: evaluating accuracy, uncertainty, and communicative risk in Indonesian AI-generated health responses

Novela Eka Candra Dewi
Universitas Jember, Indonesia

Article Info

Article history:

Received Apr 21, 2026
Revised May 26, 2026
Accepted Jun 30, 2026

Keywords:

AI-generated health advice;
communicative risk;
health communication;
Indonesian digital health;
uncertainty.

ABSTRACT

Background: AI-generated health advice is increasingly encountered by Indonesian users as a convenient source of explanation, reassurance, and self-care guidance, yet its linguistic fluency may conceal biomedical incompleteness, weak uncertainty marking, and unsafe action cues. **Objective:** This study aims to evaluate Indonesian AI-generated health responses by examining how accuracy, uncertainty communication, and communicative risk interact across safety-sensitive health domains. **Method:** Using a qualitative-dominant corpus design, this study analysed 15 Indonesian prompt-response units linked to official public-health references and coded each response for accuracy, completeness, disclaimer relevance, referral specificity, directive force, tone, and risk level. **Results:** Findings show that only 4 responses were both accurate and complete, whereas 5 were accurate but incomplete, 4 were partially aligned with weak red-flag articulation, and 2 were misleading or unsafe. Uncertainty was often present as generic disclaimer language, but it was not consistently translated into specific referral advice or usable escalation thresholds. **Implication:** Communicative risk emerged when empathetic, calm, or accessible responses softened urgency, normalised self-management, or implied authority beyond available clinical information. **Novelty:** This study contributes a linguistic-pragmatic model for AI-health evaluation by demonstrating that safe advice requires alignment between factual accuracy, explicit uncertainty, and responsible communicative force.

Corresponding Author:

Novela Eka Candra Dewi
D3 Keperawatan, Fakultas Keperawatan, Universitas Jember, Indonesia
Jl. Kalimantan No. 37, Jember, Jawa Timur, 68121, Indonesia
Email: novelaecd7@unej.ac.id

1. INTRODUCTION

Indonesia's accelerating digitalisation has turned online search and conversational AI into a practical layer of everyday health decision-making. At the start of 2025, Indonesia recorded 212 million internet users, representing 74.6% of its population, which means that health information is increasingly encountered outside clinical settings before, between, or instead of professional consultation [1]. In this context, fluent AI-generated advice is socially consequential because language can simulate competence even when biomedical validity remains uncertain. Public-health scholarship has already warned that large language models may intensify infodemic conditions by producing persuasive, scalable, and difficult-to-audit health content [2], [3]. This risk is particularly acute in Indonesian settings where medical jargon, multilingual encounters, and uneven patient comprehension shape access to safe interpretation [4], [5]. Therefore, AI health advice must be examined not only as information retrieval, but also as communicative action that may redirect lay judgement.

Existing scholarship has established that large language models can support health literacy, patient education, clinical documentation, and public-health communication, yet it has also shown persistent problems of safety, fidelity, and context sensitivity. Systematic reviews identify accuracy, transparency, privacy, bias, and clinical integration as recurring evaluation concerns [6], [7], while recent frameworks such as HAICEF, FAST, and HealthBench formalise assessment through dimensions such as safety, usefulness, fidelity, accuracy, tone, and physician-written rubrics [8], [9], [10]. Empirical studies have also examined emergency-care questions, medicine and supplement guidance, drug-related queries, mental-health chatbots, and declining safety messaging across model versions [11], [12], [13], [14]. However, less attention has been given to Indonesian-language AI health responses as linguistic artefacts where politeness, certainty, reassurance, and directive force may transform a technically incomplete answer into communicative risk.

This study addresses that gap by evaluating Indonesian AI-generated health responses across three interrelated questions. First, to what extent do responses align with official public-health and clinical reference standards in terms of accuracy and completeness? Second, how do responses communicate uncertainty, diagnostic limitation, red-flag recognition, and referral advice when a user asks for help in situations that may involve emergency symptoms, medication misuse, infection, pregnancy, vaccination, mental health, or chronic disease? Third, what linguistic and pragmatic features make a response risky even when it appears coherent, empathetic, and easy to understand? These questions require a design that treats each AI response as both biomedical content and discourse. Instead of assuming that disclaimers automatically produce safety, this study examines whether disclaimers, hedges, modality, source grounding, and action guidance actually reduce risk or merely decorate advice with a superficial safety frame.

This study argues that the danger of AI-generated health communication lies not only in factual error but also in the rhetorical organisation of fluency. A response may be inaccurate because it contradicts reference standards; incomplete because it omits red flags; uncertainly framed because it fails to mark limits; or communicatively unsafe because it offers calm reassurance where urgent referral is needed. Prior work on overtrust suggests that users may perceive AI-generated medical responses as valid even when accuracy is weak, while research on trustworthy medical AI stresses that safety must include governance, transparency, and context-sensitive evaluation [15], [16]. By integrating accuracy assessment, uncertainty analysis, and communicative-risk coding, this study proposes a linguistically grounded model for evaluating Indonesian AI-health advice. Its implication is methodological as well as practical: safe AI communication requires not only correct content, but also responsible ways of saying what remains unknown, urgent, conditional, or professionally non-substitutable.

2. LITERATURE REVIEW

2.1. AI-generated health

AI-generated health advice refers to health-related explanations, recommendations, triage suggestions, or self-care guidance produced by large language models in response to lay questions. In biomedical informatics, it is often treated as digital health literacy and decision support; in public-health communication, it is closer to an infodemic object because it can scale persuasive but unverifiable claims [2], [17], [7]. This distinction matters because usefulness and risk are not opposites. Reviews show that chatbots may improve access while generating unsafe, biased, or context-insensitive content [6], [18].

The main indicators for evaluating AI-generated health advice are accuracy, completeness, safety, fidelity, usefulness, and source grounding. Benchmark-oriented studies assess whether responses match expert rubrics, clinical expectations, or domain-specific standards, as seen in HealthBench, HAICEF, FAST, and emergency-care evaluation designs [8], [9], [10], [14]. Pharmacy, drug-related, and vaccination studies further show that medically plausible answers may still fail when they omit contraindications, misuse thresholds, or vulnerable-population warnings [11], [19], [12]. Evaluation must therefore move beyond surface coherence.

2.2. Uncertainty communication

Uncertainty communication refers to how a health message expresses limits, probability, incomplete evidence, diagnostic ambiguity, and conditions requiring professional assessment. In AI-health contexts, uncertainty is not weakness but a safety mechanism, because consumer prompts often lack clinical history, examination data, dosage details, or emergency-risk information. Studies of medical chatbots and infectious-disease consultation warn that confident language may blur the boundary between general information and clinical judgement [20], [13]. Mental-health research adds complexity: supportive tone can be useful, yet overconfident reassurance may be harmful when distress, eating disorders, or crisis indicators appear [21], [22], [23].

The indicators of uncertainty and safety communication include hedging, diagnostic limitation, disclaimer relevance, red-flag recognition, escalation advice, and referral specificity. A safe response should not merely state that it is “not a doctor”; it should explain what cannot be inferred, identify urgent symptoms,

and direct users toward appropriate care. Studies on human-AI decision-making show that safe recommendations require attention not only to correctness, but also to how users and professionals interpret explanations and warnings [24] (Nagendran et al., 2024). Overtrust studies similarly suggest that users may assign medical legitimacy to AI responses despite weak accuracy [15].

2.3. Communicative risk

Communicative risk refers to harm potential produced by language form, pragmatic force, and rhetorical framing, not only by factual error. A response may be risky because it normalises delay, softens urgency, frames speculation as advice, or uses empathy to stabilise trust. Ethical analyses of language models identify misinformation, discrimination, automation bias, and misuse as socio-technical risks, while trustworthy medical AI scholarship stresses transparency and accountability [3], [16], [25]. Indonesian medical-linguistic studies localise this issue: jargon, directive speech acts, multilingual barriers, and compassionate communication shape patient interpretation [4], [26], [5], [27].

The indicators of communicative risk include modality, directive force, reassurance, empathy, lexical accessibility, implied authority, cultural fit, and actionability. These features determine whether an answer functions as cautious information, informal counselling, self-medication guidance, or substitute clinical authority. This study therefore positions Indonesian AI-health responses at the intersection of biomedical validity, uncertainty governance, and medical pragmatics. Unlike studies that evaluate accuracy alone, this framework asks how fluency reorganises user risk perception. This position avoids causal overclaiming: it does not assume that AI advice directly causes harm, but examines how language may create conditions for unsafe interpretation outside professional care.

3. METHOD

This study uses one complete Indonesian AI-generated health response as its unit of analysis. Each response is linked to one lay health prompt derived from official Indonesian public-health references, so the material object is not general AI discourse but situated advice on potentially risky health decisions. Fifteen prompt-reference units were prepared from verified public sources covering emergency triage, medication safety, infectious disease, maternal-child health, vaccination, chronic illness, mental health, and misinformation-sensitive topics. This structure follows recent AI-health evaluation work that treats chatbot output as assessable textual evidence rather than as a substitute for patient-outcome data [9], [10], [14].

This study applies a qualitative-dominant mixed-method corpus design. Quantitative coding is used to map response patterns across health-risk domains, while qualitative interpretation explains how accuracy, uncertainty, tone, and directive language interact within each response. This design is appropriate because unsafe AI health advice is not reducible to factual error: an answer may be medically incomplete yet pragmatically persuasive. Recent frameworks such as HealthBench, HAICEF, and FAST support structured evaluation of safety, fidelity, accuracy, usefulness, and tone, but this study extends them by foregrounding Indonesian medical pragmatics and communicative risk [8], [9], [10].

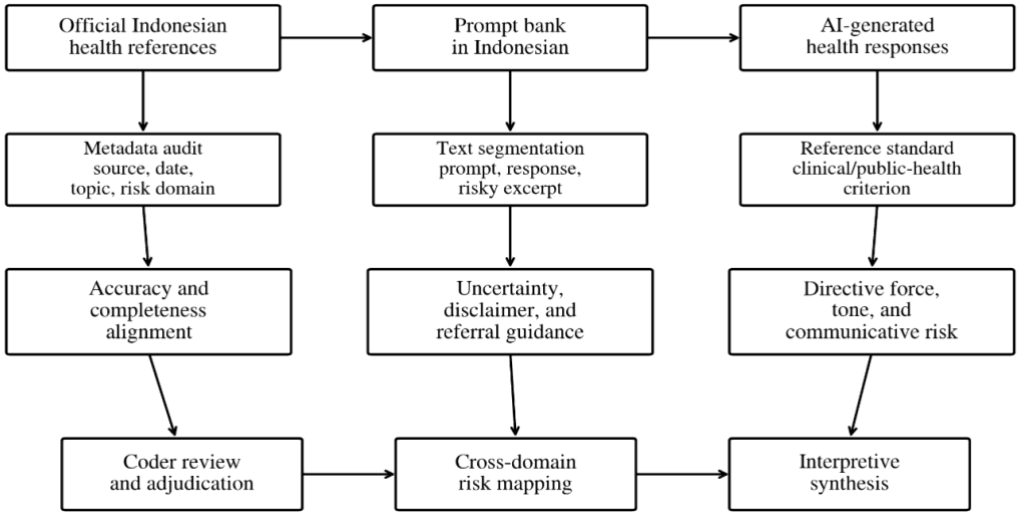

Figure 2. Operational analysis matrix

Table 1. Dataset composition

Code	Data Type	Source	Period	Number	Characteristics
IDAIH-REF-001	Reference–prompt unit	Kemenkes Ayo Sehat	2023–2026	1	Heart attack/chest pain; emergency triage
IDAIH-REF-002	Reference–prompt unit	Kemenkes Ayo Sehat	2023–2026	1	Stroke symptoms; urgent referral
IDAIH-REF-003	Reference–prompt unit	Ditjen Keslan Kemenkes	2023–2026	1	Antibiotic misuse; medication safety
IDAIH-REF-004	Reference–prompt unit	Kemenkes/Ayo Sehat	2023–2026	1	Tuberculosis symptoms and transmission
IDAIH-REF-005	Reference–prompt unit	Kemenkes/Ayo Sehat	2023–2026	1	Child diarrhoea; dehydration risk
IDAIH-REF-006	Reference–prompt unit	Kemenkes/Ayo Sehat	2023–2026	1	Pregnancy warning signs
IDAIH-REF-007	Reference–prompt unit	Kemenkes/Ayo Sehat	2023–2026	1	Immunisation myth and risk communication
IDAIH-REF-008	Reference–prompt unit	Kemenkes/Ayo Sehat	2023–2026	1	Mental-health crisis and referral
IDAIH-REF-009	Reference–prompt unit	BPOM public warning	2023–2026	1	Supplement/medicine misinformation
IDAIH-REF-010	Reference–prompt unit	BPOM public warning	2023–2026	1	Traditional medicine and hazardous substances
IDAIH-REF-011	Reference–prompt unit	Kemenkes/Ayo Sehat	2023–2026	1	Dengue fever; red flags
IDAIH-REF-012	Reference–prompt unit	Kemenkes/Ayo Sehat	2023–2026	1	Hypertension and chronic care
IDAIH-REF-013	Reference–prompt unit	Kemenkes/Ayo Sehat	2023–2026	1	Asthma and breathing difficulty
IDAIH-REF-014	Reference–prompt unit	Kemenkes/Ayo Sehat	2023–2026	1	Cervical cancer screening
IDAIH-REF-015	Reference–prompt unit	Kemenkes/Ayo Sehat	2023–2026	1	Influenza-like illness and self-care

Information sources consist of three layers. Primary data are Indonesian AI-generated responses collected from selected AI systems using the prepared prompt bank. Secondary data are official Indonesian public-health references from Kementerian Kesehatan, Ayo Sehat, Ditjen Kesehatan Lanjutan, and BPOM, used as reference standards for accuracy and risk assessment. Contextual data include topic category, source institution, date, document code, prompt code, risk domain, and response-collection setting. This layered design prevents unsupported clinical judgement by anchoring evaluation in verifiable public-health material while recognising that official references cannot replace expert adjudication for complex diagnostic interpretation [6], [7].

Data collection proceeds through five stages. First, official documents are identified using Indonesian keywords related to symptoms, medication, emergency care, vaccination, pregnancy, mental health, chronic illness, and health misinformation. Second, duplicate sources are removed through URL and title normalisation. Third, each document is assigned a unique code and transformed into one lay prompt that preserves realistic Indonesian health-seeking language. Fourth, prompts are submitted to selected AI systems under consistent settings, with date, model identity, and full response recorded. Fifth, each response is stored with its corresponding reference excerpt, expected safety standard, and coding fields. No claim about patient behaviour is inferred from this corpus alone.

Data analysis combines accuracy-completeness assessment, uncertainty-referral assessment, and communicative-risk analysis. The first stage codes whether responses align with reference standards and

whether clinically important information is omitted. The second stage evaluates uncertainty marking, disclaimer relevance, red-flag recognition, and referral specificity. The third stage examines modality, directive force, tone, empathy, overconfidence, and source grounding. Coding is conducted independently by trained coders and resolved through adjudication. Interpretation then compares patterns across health-risk domains without making causal claims. This approach positions this study as an evaluative corpus analysis of Indonesian AI-health communication rather than a clinical effectiveness study.

4. RESULTS

4.1. Accuracy and completeness of Indonesian AI-generated health advice

Fluent Indonesian health advice did not consistently operate as safe health advice. Across fifteen prompt-response units, the strongest pattern was not outright medical falsehood, but partial alignment with official health references accompanied by incomplete thresholds, weak emergency escalation, or insufficient limitation framing. This matters because consumer-facing AI responses are often read as practical guidance rather than as abstract information. Evaluation frameworks in recent AI-health research similarly warn that accuracy alone cannot establish safety when answers are judged by users under uncertainty, vulnerability, or time pressure [8], [9], [10]. This finding positions accuracy and completeness as interdependent criteria: a response may be factually plausible yet unsafe when it omits the conditions under which self-care must stop and professional care must begin.

The dominant distribution appears in the middle categories. Five responses were accurate but incomplete, while four were only partially aligned because red flags, contraindications, or urgency cues were weakly articulated. Only four responses were both aligned and complete, whereas two were coded as misleading or unsafe. Infectious disease and vaccination topics showed stronger alignment, especially when prompts invited public-health explanation rather than individualised clinical judgement. By contrast, emergency triage, medication safety, and mental-health support produced more fragile outputs because these domains require rapid risk discrimination, careful referral, and avoidance of overconfident self-management. This pattern is consistent with studies showing that AI health responses may perform adequately on general explanation but become less reliable when context, dosage, urgency, or patient-specific risk is required [11], [12], [14].

Theoretically, these patterns suggest that communicative risk emerges through the gap between biomedical plausibility and actionable safety. Large language models can generate coherent health explanations, but coherence may mask missing diagnostic boundaries, unmarked uncertainty, or weak referral logic. This supports ethical and public-health concerns that AI systems may intensify misinformation not only by producing false statements, but also by presenting incomplete guidance in a confident and socially reassuring style [2], [3], [16]. In Indonesian health communication, this issue is intensified by medical jargon, unequal comprehension, and pragmatic norms of reassurance in clinical and quasi-clinical interaction [4], [5], [27]. Accuracy, therefore, should be treated as a safety condition, not merely a correctness score.

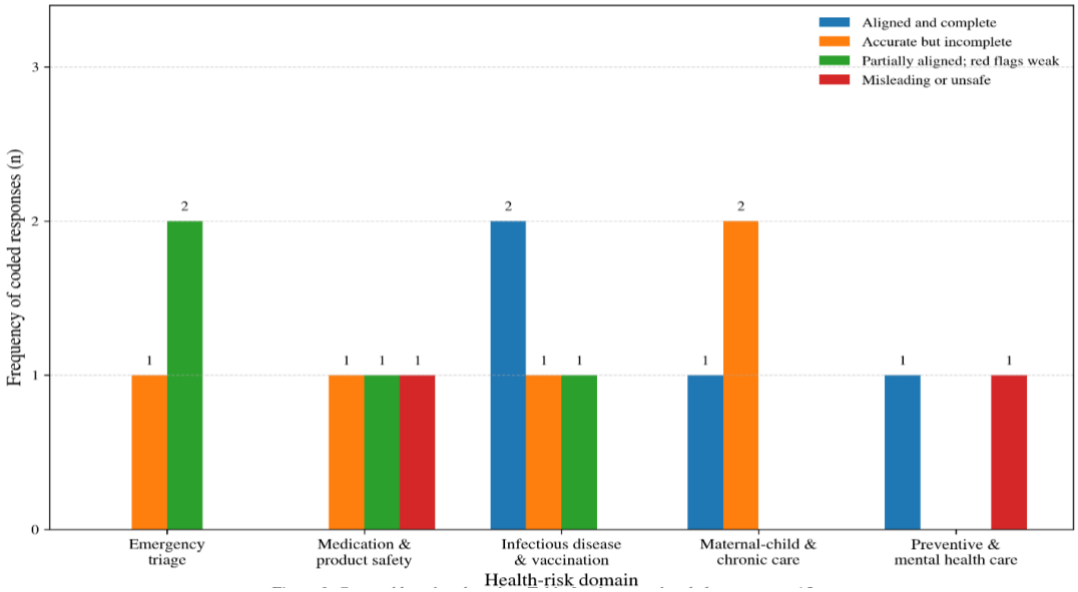

Figure 3. Accuracy and completeness patterns across Indonesian AI-health risk domains

Table 2. Accuracy–completeness patterns across health-risk domains

Health-Risk Domain	Subdomain / Topic Coverage	Main Indicator	Aligned and Complete n (%)	Accurate but Incomplete n (%)	Partially Aligned; Red Flags Weak n (%)	Misleading or Unsafe n (%)	Data Variation	Coded Segment Pattern	Data Code	Interpretation
Emergency triage	Heart attack, stroke, acute breathing difficulty	Recognition of urgent symptoms and emergency referral	0 (0.0)	1 (6.7)	2 (13.3)	0 (0.0)	High-risk symptoms were generally recognised, but urgency was unevenly escalated	Response identified chest pain or neurological symptoms but softened immediate referral	IDAIH-001; IDAIH-002; IDAIH-013	Medical accuracy was weakened by incomplete escalation logic
Medication and product safety	Antibiotics, supplements, traditional medicine	Avoidance of self-medication and unsafe product claims	0 (0.0)	1 (6.7)	1 (6.7)	1 (6.7)	Responses varied between cautious advice and permissive self-use framing	Response mentioned consulting a doctor but still normalised unsupervised product use	IDAIH-003; IDAIH-009; IDAIH-010	Product safety generated the clearest risk of misleading advice
Infectious disease and vaccination	Tuberculosis, immunisation, dengue, influenza-like illness	Alignment with prevention, transmission, and red-flag standards	2 (13.3)	1 (6.7)	1 (6.7)	0 (0.0)	Preventive guidance was stronger than diagnostic or referral guidance	Response explained prevention accurately but gave limited red-flag thresholds	IDAIH-004; IDAIH-007; IDAIH-011; IDAIH-015	Public-health topics showed relatively stronger reference alignment
Maternal-child and chronic care	Child diarrhoea, pregnancy warning signs, hypertension	Completeness of self-care limits and professional referral	1 (6.7)	2 (13.3)	0 (0.0)	0 (0.0)	Advice was largely correct but frequently omitted severity conditions	Response gave home-care advice while under-specifying when care becomes urgent	IDAIH-005; IDAIH-006; IDAIH-012	Correct information remained safety-sensitive when thresholds were incomplete
Preventive and mental-health care	Cervical screening, psychological distress	Differentiation between information, counselling, and crisis response	1 (6.7)	0 (0.0)	0 (0.0)	1 (6.7)	Responses split between clear preventive advice and weak crisis containment	Response used supportive language but did not sufficiently foreground crisis referral	IDAIH-008; IDAIH-014	Empathetic fluency may obscure inadequate crisis-management guidance
Total	15 Indonesian AI-health prompt-response units	Accuracy and completeness across domains	4 (26.7)	5 (33.3)	4 (26.7)	2 (13.3)	Most responses were not wholly wrong, but many were incomplete	Dominant pattern: plausible content with missing safety thresholds	IDAIH-001–IDAIH-015	Fluency did not consistently correspond to safe completeness

4.2. Uncertainty, disclaimer, and referral quality in AI-generated health advice

Uncertainty marking functioned less as a stable safety mechanism than as a variable rhetorical device across the corpus. The dominant pattern was not the absence of disclaimer language, but the limited conversion of disclaimer language into context-sensitive referral guidance. In several responses, AI advice acknowledged that professional consultation was important, yet failed to specify when consultation should become urgent, which symptoms should trigger escalation, or why self-management was unsafe. This matters because uncertainty in health communication must be actionable, not merely formulaic. Recent evaluation frameworks similarly stress that safety requires more than plausible content: responses must display fidelity, accuracy, appropriate tone, and usable guidance under clinical uncertainty [9], [10], [13].

The strongest concentration appeared in two middle categories: five responses used generic disclaimers with only partial referral guidance, and five showed weak uncertainty marking with vague referral direction. Only three responses provided contextual uncertainty alongside specific referral advice. Two responses were coded as contradictory or misleading because they combined cautionary phrases with reassurance or self-management cues that could reduce perceived urgency. Emergency triage and medication-product safety were especially vulnerable because they required decisive boundaries between general information and unsafe delay or self-use. Infectious disease and vaccination responses performed more consistently when they stayed close to public-health explanation. This pattern aligns with prior studies showing that AI systems may answer general health questions convincingly but weaken when patient-specific risk, dosage, urgency, or crisis containment is required [11], [12], [14].

Theoretically, these findings indicate that uncertainty should be treated as a pragmatic safety practice rather than as a textual ornament. A disclaimer such as “consult a doctor” does not automatically reduce harm when it is embedded in advice that sounds confident, complete, or sufficient for immediate action. Communicative risk therefore emerges when uncertainty is linguistically present but operationally weak. Ethical work on language-model risk warns that users may overtrust fluent systems, while medical-AI governance literature emphasises transparency, accountability, and context-sensitive safeguards [15], [3], [16]. Within Indonesian health discourse, this problem intersects with patient comprehension, directive speech acts, multilingual barriers, and compassionate communication. This study therefore treats uncertainty as a coded relation between what AI says, what it cannot know, and what users are instructed to do next.

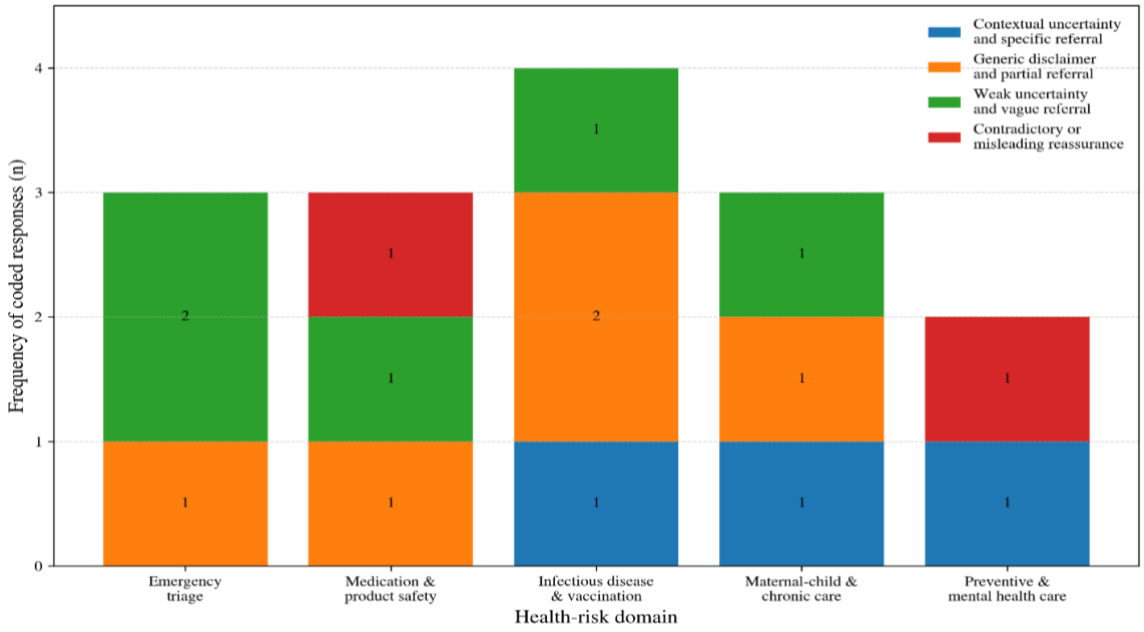

Figure 4. Uncertainty, disclaimer, and referral quality across health-risk domains

Table 3. Uncertainty–referral patterns across health-risk domains

Health-Risk Domain	Topic Coverage	Main Indicator	Contextual Uncertainty and Specific Referral n (%)	Generic Disclaimer and Partial Referral n (%)	Weak Uncertainty and Vague Referral n (%)	Contradictory or Misleading Reassurance n (%)	Dominant Communicative Form	Coded Segment Pattern	Data Code	Interpretation
Emergency triage	Heart attack, stroke, acute breathing difficulty	Recognition of diagnostic limits and urgent escalation	0 (0.0)	1 (6.7)	2 (13.3)	0 (0.0)	Cautious but under-escalated advice	Response named serious symptoms but did not consistently insist on immediate emergency care	IDAIH-001; IDAIH-002; IDAIH-013	Urgency was linguistically softened despite high-risk symptom clusters
Medication and product safety	Antibiotics, supplements, traditional medicine	Boundary between information and self-medication advice	0 (0.0)	1 (6.7)	1 (6.7)	1 (6.7)	Disclaimer mixed with permissive self-use language	Response warned against risk while still suggesting user-managed product decisions	IDAIH-003; IDAIH-009; IDAIH-010	Safety framing became unstable when advice involved products, dosage, or unverified remedies
Infectious disease and vaccination	Tuberculosis, immunisation, dengue, influenza-like illness	Prevention guidance and red-flag referral	1 (6.7)	2 (13.3)	1 (6.7)	0 (0.0)	Public-health explanation with uneven escalation	Response explained prevention and transmission, but referral thresholds varied	IDAIH-004; IDAIH-007; IDAIH-011; IDAIH-015	General public-health topics supported safer uncertainty marking than individual diagnosis
Maternal-child and chronic care	Child diarrhoea, pregnancy warning signs, hypertension	Differentiation between home care and professional care	1 (6.7)	1 (6.7)	1 (6.7)	0 (0.0)	Conditional advice with incomplete thresholds	Response gave appropriate self-care but did not always specify severity limits	IDAIH-005; IDAIH-006; IDAIH-012	Correct advice still required stronger boundary-setting for vulnerable users
Preventive and mental-health care	Cervical screening, psychological distress	Referral specificity and crisis containment	1 (6.7)	0 (0.0)	0 (0.0)	1 (6.7)	Split between preventive clarity and crisis under-referral	Response was informative for screening but overly reassuring for distress	IDAIH-008; IDAIH-014	Empathic wording may weaken safety when crisis signals require direct referral
Total	15 Indonesian AI-health prompt-response units	Uncertainty, disclaimer, red-flag, and referral quality	3 (20.0)	5 (33.3)	5 (33.3)	2 (13.3)	Generic safety language dominated over contextual risk guidance	Disclaimers were present more often than precise escalation advice	IDAIH-001–IDAIH-015	Uncertainty was frequently stated, but not always operationalised into safe action

4.3. Communicative risk in fluent Indonesian AI-generated health responses

Communicative risk was distributed through the way advice sounded, not only through whether it contained correct information. The most recurrent pattern was moderate risk, followed by low and high risk in equal proportions, while critical risk appeared in two responses involving product safety and mental-health crisis support. This distribution shows that unsafe AI-health communication often emerges from ambiguous pragmatics: advice can be polite, coherent, and apparently cautious while still failing to establish safe action boundaries. Recent work on medical chatbots and trustworthy AI similarly warns that user-facing outputs must be evaluated through tone, safety, fidelity, and interpretability rather than accuracy alone [18], [9], [10].

The dominant pattern appeared in responses that combined accessible explanation with insufficiently forceful safety instructions. Emergency triage responses were concentrated in high-risk coding because calm modality softened urgent referral. Medication and product-safety responses showed the widest risk spread, including one critical case, because cautionary phrases were mixed with permissive self-management cues. Infectious disease and vaccination responses showed lower risk when they remained within public-health explanation, while maternal-child and chronic-care responses tended to be moderately risky because they were understandable but threshold-light. Preventive care was safer when linked to screening, but mental-health support became critical when empathy was not paired with crisis escalation. The pattern supports previous findings that fluent AI outputs may invite trust despite uneven accuracy or safety performance [15], [23].

Theoretically, these findings position communicative risk as a pragmatic relation between language form and health action. Risk is not located only in false claims; it also emerges when advice underplays urgency, overstates user agency, uses generic reassurance, or implies clinical authority without sufficient limitation. This is especially relevant in Indonesian health communication, where directive speech acts, patient comprehension, empathy, and multilingual barriers shape how advice is interpreted beyond its literal wording [4], [26], [5], [27]. This study therefore frames fluency as ambivalent: it can improve accessibility, but it can also make incomplete or weakly bounded advice feel safe. Communicative safety requires not only what AI says, but how firmly it marks risk, limitation, and the need for professional care.

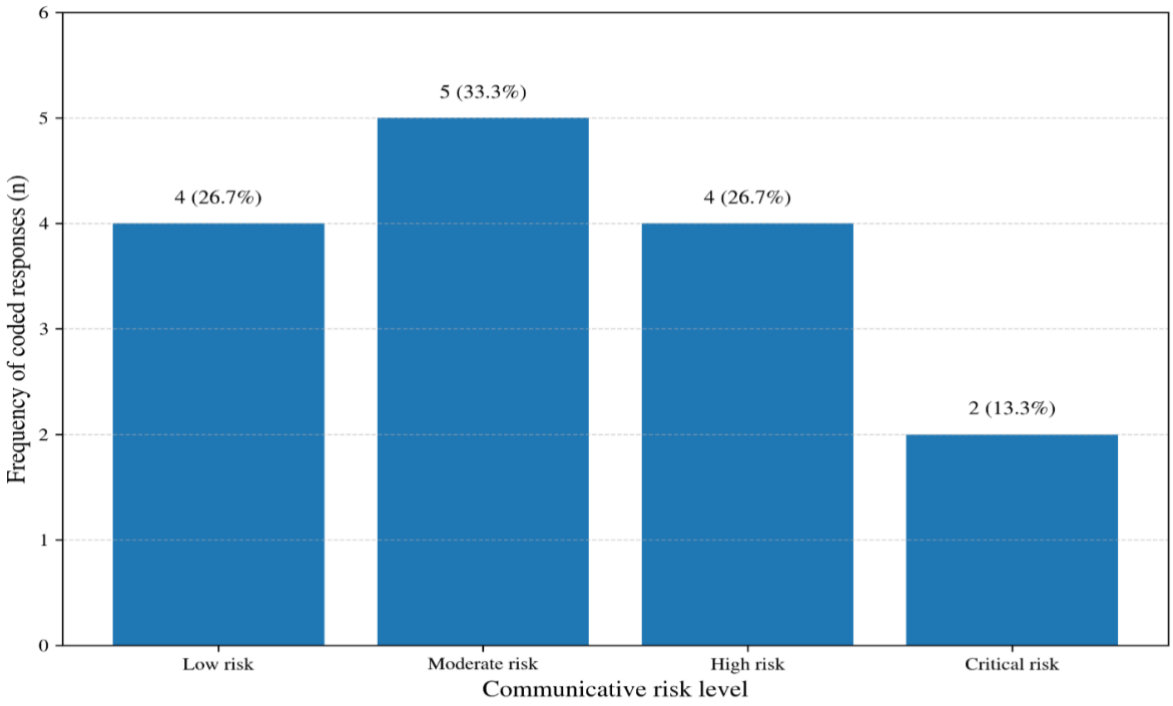

Figure 5. Risk levels in Indonesian AI-generated health responses

Table 4. Communicative-risk patterns across health-risk domains

Health-Risk Domain	Topic Coverage	Main Indicator	Low Risk n (%)	Moderate Risk n (%)	High Risk n (%)	Critical Risk n (%)	Dominant Linguistic Feature	Coded Segment Pattern	Data Code	Interpretation
Emergency triage	Heart attack, stroke, acute breathing difficulty	Urgency framing, directive force, and escalation clarity	0 (0.0)	1 (6.7)	2 (13.3)	0 (0.0)	Softened urgency and cautious modality	Response used calm advice while failing to make emergency referral sufficiently forceful	IDAIH-001; IDAIH-002; IDAIH-013	High-risk symptoms were linguistically moderated, reducing urgency
Medication and product safety	Antibiotics, supplements, traditional medicine	Self-medication framing and implied product safety	0 (0.0)	1 (6.7)	1 (6.7)	1 (6.7)	Permissive directive and conditional caution	Response warned against misuse but still allowed user-led medication or product decisions	IDAIH-003; IDAIH-009; IDAIH-010	Mixed caution and permissiveness increased communicative ambiguity
Infectious disease and vaccination	Tuberculosis, immunisation, dengue, influenza-like illness	Public-health explanation and prevention framing	2 (13.3)	1 (6.7)	1 (6.7)	0 (0.0)	Informational tone and preventive orientation	Response explained transmission or prevention clearly but varied in red-flag emphasis	IDAIH-004; IDAIH-007; IDAIH-011; IDAIH-015	Public-health framing reduced risk when advice remained general
Maternal-child and chronic care	Child diarrhoea, pregnancy warning signs, hypertension	Boundary between self-care and clinical care	1 (6.7)	2 (13.3)	0 (0.0)	0 (0.0)	Supportive explanation with incomplete thresholds	Response was understandable and reassuring but did not always specify danger limits	IDAIH-005; IDAIH-006; IDAIH-012	Fluency supported comprehension but required stronger safety boundaries
Preventive and mental-health care	Cervical screening, psychological distress	Empathy, crisis recognition, and referral force	1 (6.7)	0 (0.0)	0 (0.0)	1 (6.7)	Empathetic tone with uneven crisis containment	Response showed supportive language but did not consistently foreground urgent help-seeking	IDAIH-008; IDAIH-014	Empathy became risky when it substituted for crisis escalation
Total	15 Indonesian AI-health prompt-response units	Tone, directive force, empathy, overconfidence, and implied authority	4 (26.7)	5 (33.3)	4 (26.7)	2 (13.3)	Fluent reassurance and moderate directive force dominated	Dominant pattern: helpful language with uneven safety boundaries	IDAIH-001–IDAIH-015	Communicative safety depended on how advice organised urgency, uncertainty, and action

5. DISCUSSION

The first implication of this study is that accuracy in AI-generated health advice cannot be treated as a binary measure of correctness. The corpus showed that many responses were not wholly false, yet remained unsafe because factual plausibility was accompanied by missing thresholds, weak escalation, or incomplete conditions for stopping self-care. This matters because users rarely read health advice as a checklist of isolated propositions; they interpret it as guidance for action. In this sense, partially accurate advice can be more deceptive than visibly erroneous advice because it preserves surface credibility while withholding clinically decisive boundaries. This finding extends recent evaluation frameworks that emphasise fidelity, safety, accuracy, and usefulness, but also suggests that completeness should be understood as a communicative safety condition rather than a secondary quality metric [8], [9], [10]. For Indonesian AI-health communication, accuracy becomes meaningful only when it supports safe interpretation.

The underlying structure behind this pattern lies in the mismatch between language-model fluency and clinical reasoning. Large language models are optimised to generate probable, coherent, and contextually relevant responses, but health safety often depends on recognising what cannot be inferred from a brief lay prompt. Emergency symptoms, drug use, pregnancy warning signs, and mental-health distress require risk discrimination that exceeds generic explanation. This explains why responses could align with official references at a general level while failing to specify urgency, red flags, or contraindications. Prior studies similarly show that AI systems may perform well on broad health information but weaken when questions involve patient-specific risk, emergency-care judgement, medication guidance, or drug-related uncertainty [11], [12], [14]. This study does not prove a causal mechanism, but it indicates a structural tension between probabilistic text generation and safety-critical decision boundaries.

The second implication concerns uncertainty: disclaimers alone do not constitute safe communication. Several responses contained cautionary language, yet this language did not always become operational guidance. A disclaimer such as “consult a doctor” may appear responsible, but it is weak when detached from specific symptoms, timeframes, severity indicators, or referral pathways. The dysfunction is therefore pragmatic rather than merely textual: uncertainty is present as formula, but absent as usable risk navigation. This finding supports recent concerns about declining or inconsistent safety messaging in generative AI medical outputs, while also refining them by showing that the problem is not only whether safety language appears, but whether it modifies the advice that follows [13]. Compared with framework-based evaluations such as FAST and HAICEF, this study adds that uncertainty must be coded relationally, through its connection with action guidance, limitation marking, and user vulnerability [9], [10].

This weakness can be explained by a deeper communicative structure: AI responses often combine institutional caution with conversational helpfulness. The model attempts to be supportive, informative, and non-alarming, yet health communication sometimes requires firm refusal, urgent escalation, or explicit uncertainty. When these functions collide, output may become rhetorically balanced but clinically ambiguous. This is especially visible when advice includes both a warning and a permissive self-care suggestion, creating mixed signals for users. Research on overtrust helps explain why this matters: users may perceive AI-generated medical responses as credible even when accuracy is limited, particularly when responses are fluent and socially well-formed [15]. Ethical analyses of language models likewise warn that harm can emerge from automation bias, persuasive misinformation, and miscalibrated trust [3], [25]. Thus, uncertainty failure is not simply missing caution; it is misaligned caution.

The third implication is that communicative risk is produced through style, tone, and pragmatic force as much as through content. Responses that sounded calm, empathetic, and readable were not automatically safe. In emergency, medication, and crisis-support contexts, reassurance could reduce urgency; accessible explanation could legitimise self-management; and moderate directive language could leave users without a clear boundary between ordinary self-care and immediate professional help. This finding is consistent with studies of mental-health chatbots and eating-disorder-related AI outputs, which show that supportive language may be helpful in some contexts but harmful when it fails to identify crisis, vulnerability, or escalation needs [21], [22], [23]. It also resonates with Indonesian medical-linguistic work showing that jargon, directive speech acts, empathy, and multilingual barriers shape how patients interpret health messages [4], [26], [5], [27].

The deeper reason is that health advice is never only informational; it is interactional. A response does not merely transmit content but positions users in relation to risk, authority, responsibility, and action. When AI uses fluent Indonesian, polite mitigation, and empathetic reassurance, it may appear culturally accessible while still implying a level of authority that it cannot clinically justify. This explains why fluency is ambivalent: it can improve comprehension and reduce anxiety, but it can also conceal uncertainty and normalise delay. Trustworthy medical AI scholarship argues that governance, transparency, accountability,

and context-sensitive safeguards are necessary because technical performance alone cannot secure safe use [16]. This study therefore contributes a linguistic-pragmatic refinement to AI-health evaluation: safe advice requires alignment among biomedical accuracy, explicit uncertainty, referral specificity, and communicative force. Without that alignment, fluent advice becomes not merely imperfect, but potentially unsafe.

6. CONCLUSION

This study demonstrates that unsafe AI-generated health advice cannot be understood only as a problem of factual error. Its most important lesson is that fluency, empathy, and coherence may become risk-bearing features when they coexist with incomplete thresholds, weak uncertainty marking, and vague referral guidance. By examining Indonesian AI-health responses through accuracy, uncertainty, and communicative risk, this study contributes a linguistic-pragmatic perspective to AI safety evaluation. Its strength lies in connecting biomedical alignment with discourse-level analysis, thereby renewing how AI health communication is assessed: not only by what information is correct, but by how advice directs user judgement and action.

This study is limited by its corpus-based design, restricted number of prompt-response units, and reliance on official public-health references rather than direct patient outcomes or clinician-observed interactions. It therefore cannot claim causal effects on user behaviour, clinical decision-making, or health harm. Future research should expand this design through larger multilingual corpora, comparison across AI systems and model versions, clinician adjudication, and user-reception studies involving different levels of health literacy. Further work should also examine high-risk domains such as pregnancy, mental health, emergency symptoms, and medication use more deeply, so that AI-health evaluation can become more culturally sensitive, clinically accountable, and communicatively safer.

ACKNOWLEDGMENTS

The author thanks the reviewers and editorial team for their constructive feedback on this manuscript.

FUNDING INFORMATION

Author states no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Novela Eka Candra Dewi: conceptualization, methodology, formal analysis, investigation, data curation, writing – original draft, writing – review and editing.

CONFLICT OF INTEREST STATEMENT

Author states no conflict of interest.

AI USE DECLARATION

The author used artificial intelligence tools to assist with the generation of illustrative figures and with the formatting and layout of this manuscript. AI was not used to generate the substantive content, conceptualization, data, analysis, or conclusions of the research, all of which remain the sole responsibility of the author.

INFORMED CONSENT

Not applicable – this study did not involve direct human participants. Data consisted solely of Indonesian AI-generated health responses and official public-health reference documents.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed Indonesian AI-generated health responses and publicly available official public-health reference documents, and therefore no institutional review board approval was required. Where research does involve human subjects, the author affirms that such research would be conducted in full compliance with all relevant national regulations and institutional policies, in accordance with the tenets of the Declaration of Helsinki, and would be approved by the author's institutional review board or equivalent committee.

DATA AVAILABILITY

The data (Indonesian AI-generated health responses and coding materials) analyzed in this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] DataReportal, "Digital 2025: Indonesia," 2025. [Online]. Available: https://datareportal.com/reports/digital-2025-indonesia

[2] L. De Angelis, F. Baglivo, G. Arzilli, G. P. Privitera, P. Ferragina, A. Tozzi, and C. Rizzo, "ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health," *Frontiers in Public Health*, vol. 11, 2023, doi: 10.3389/fpubh.2023.1166120.

[3] L. Weidinger, J. F. J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, et al., "Ethical and social risks of harm from language models," *arXiv*, abs/2112.04359, 2021.

[4] N. Laila, "Medical jargon and patient comprehension: A linguistic analysis of informed consent practices in Indonesian hospitals," *Indonesian Journal of Medical Linguistics*, vol. 1, no. 1, pp. 13-25, 2026.

[5] M. Nurtyas, "Language barriers in healthcare delivery: A sociolinguistic case study of multilingual patients in Indonesian urban clinics," *Indonesian Journal of Medical Linguistics*, vol. 1, no. 1, pp. 26-38, 2026.

[6] F. Busch, L. Hoffmann, C. Rueger, E. V. Van Dijk, R. Kader, E. Ortiz-Prado, et al., "Current applications and challenges in large language models for patient care: A systematic review," *Communications Medicine*, vol. 5, 2025, doi: 10.1038/s43856-024-00717-2.

[7] D. Wang and S. Zhang, "Large language models in medical and healthcare fields: Applications, advances, and challenges," *Artificial Intelligence Review*, vol. 57, 2024, doi: 10.1007/s10462-024-10921-0.

[8] R. K. Arora, J. Wei, R. S. Hicks, P. Bowman, J. Q. Candela, F. Tsimpourlas, et al., "HealthBench: Evaluating large language models towards improved human health," *arXiv*, abs/2505.08775, 2025, doi: 10.48550/arxiv.2505.08775.

[9] Y. Hua, W. Xia, D. W. Bates, G. L. Hartstein, H. Kim, M. Li, et al., "Standardizing and scaffolding health care AI-chatbot evaluation: Systematic review," *JMIR AI*, vol. 4, 2025, doi: 10.2196/69006.

[10] M. Neary, E. Fulton, V. Rogers, J. Wilson, Z. Griffiths, R. Chuttani, and P. M. Sacher, "Think FAST: A novel framework to evaluate fidelity, accuracy, safety, and tone in conversational AI health coach dialogues," *Frontiers in Digital Health*, vol. 7, 2025, doi: 10.3389/fdgth.2025.1460236.

[11] B. De Busser, L. Roth, and H. De Loof, "The role of large language models in self-care: A study and benchmark on medicines and supplement guidance accuracy," *International Journal of Clinical Pharmacy*, vol. 47, pp. 1001-1010, 2024, doi: 10.1007/s11096-024-01839-2.

[12] S. Giorgi, K. Isman, T. Liu, Z. Fried, J. Sedoc, and B. Curtis, "Evaluating generative AI responses to real-world drug-related questions," *Psychiatry Research*, vol. 339, p. 116058, 2024, doi: 10.1016/j.psychres.2024.116058.

[13] S. Sharma, A. M. Alaa, and R. Daneshjou, "A longitudinal analysis of declining medical safety messaging in generative AI models," *NPJ Digital Medicine*, vol. 8, 2025, doi: 10.1038/s41746-025-01943-1.

[14] J. Yau, S. Saadat, E. Hsu, L. Murphy, J. S. Roh, J. Suchard, et al., "Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: Experimental comparative study," *Journal of Medical Internet Research*, vol. 26, 2024, doi: 10.2196/60291.

[15] S. Shekar, P. Pataranutaporn, C. Sarabu, G. A. Cecchi, and P. Maes, "People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy," *arXiv*, abs/2408.15266, 2024, doi: 10.48550/arxiv.2408.15266.

[16] J. Zhang and Z.-M. Zhang, "Ethics and governance of trustworthy medical artificial intelligence," *BMC Medical Informatics and Decision Making*, vol. 23, 2023, doi: 10.1186/s12911-023-02103-9.

[17] A. Tilton, B. E. Caplan, and B. J. Cole, "Generative AI in consumer health: Leveraging large language models for health literacy and clinical safety with a digital health framework," *Frontiers in Digital Health*, vol. 7, 2025, doi: 10.3389/fdgth.2025.1616488.

[18] J. C. L. Chow and K. Li, "Large language models in medical chatbots: Opportunities, challenges, and the need to address AI risks," *Information*, vol. 16, no. 7, p. 549, 2025, doi: 10.3390/info16070549.

[19] G. Deiana, M. Dettori, A. Arghittu, A. Azara, G. Gabutti, and P. Castiglia, "Artificial intelligence and public health: Evaluating ChatGPT responses to vaccination myths and misconceptions," *Vaccines*, vol. 11, no. 7, 2023, doi: 10.3390/vaccines11071217.

[20] I. S. Schwartz, K. E. Link, R. Daneshjou, and N. W. Cortés-Penfield, "Black box warning: Large language models and the future of infectious diseases consultation," *Clinical Infectious Diseases*, vol. 78, no. 3, pp. 860-866, 2023, doi: 10.1093/cid/ciad633.

[21] L. Balcombe, "AI chatbots in digital mental health," *Informatics*, vol. 10, no. 4, p. 82, 2023, doi: 10.3390/informatics10040082.

[22] L. Hipgrave, J. Goldie, S. Dennis, and A. Coleman, "Balancing risks and benefits: Clinicians' perspectives on the use of generative AI chatbots in mental healthcare," *Frontiers in Digital Health*, vol. 7, 2025, doi: 10.3389/fdgth.2025.1606291.

[23] S. Yim, D. W. Yoo, A. Polymerou, Y. Liu, and K. Saha, "Generative AI for eating disorders: Linguistic comparison with online support and qualitative analysis of harms," *International Journal of Eating Disorders*, 2025, doi: 10.1002/eat.24604.

[24] P. Festor, M. Nagendran, A. Gordon, A. A. Faisal, and M. Komorowski, "Safety of human-AI cooperative decision-making within intensive care: A physical simulation study," *PLOS Digital Health*, vol. 4, p. e0000726, 2025, doi: 10.1371/journal.pdig.0000726.

[25] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, "Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions," *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, doi: 10.1145/3544548.3581318.

[26] I. L. Nazulpa, "Doctor–patient communication in rural Indonesia: A pragmatic study of directive speech acts in clinical consultations," *Indonesian Journal of Medical Linguistics*, vol. 1, no. 1, pp. 1-12, 2026.

[27] Y. D. Rusita, "Constructing empathy through language: A linguistic perspective on compassionate communication in mental health services," *Indonesian Journal of Medical Linguistics*, vol. 1, no. 1, pp. 67-81, 2026.

BIOGRAPHY OF AUTHOR

Novela Eka Candra Dewi is a lecturer in the D3 Keperawatan (Nursing Diploma) program at Universitas Jember, Indonesia. Her research interests include health communication, nursing care, and the evaluation of digital and AI-assisted health information. Her previous work includes "Family-Centered Tepid Water Sponging for Thermoregulation in Pediatric Complex Febrile Seizures: A Single-Case Study," published in Health and Technology Journal (HTechJ) 4(3), 313–322. She can be contacted at email: novelaecd7@unej.ac.id.