Can Artificial Intelligence Produce Safe Infection Information Leaflets for Maternity and Neonatal Patients?

Shideh Kiafar[1,2], Susan Knowles[1,2] | [1]University College Dublin [2]National Maternity Hospital, Dublin

Key Messages

- AI can produce simple, readable, and coherent patient information leaflet rapidly. However, it is not suitable without human evaluation and input.
- NMH leaflets had higher mean scores for accuracy, safety, completeness and actionability, while AI scored higher for understandability, human-centred tone and fairness/inclusion. No domain remained statistically significant after correction for multiple comparisons.
- AI-generated patient information requires substantial specialist clinical input, multidisciplinary review and approval before use.

Background

Artificial Intelligence (AI) can produce easy to read patient information rapidly. However, readability does not necessarily mean that information is clinically accurate, complete or safe. This is particularly important in maternity and neonatal care, where missed precautions, incorrect advice or oversimplification could influence decisions made by women, parents or healthcare providers.

This project explored whether AI-generated infection prevention and control information leaflets were comparable to existing human-made National Maternity Hospital (NMH) patient information leaflets.

Aim

To evaluate and compare AI-generated infection prevention and control leaflets with the existing human made NMH leaflets using a structured scoring tool assessing nine domains: accuracy, safety, completeness, understandability, actionability, human-centred tone, fairness and inclusion, transparency, and overall suitability.

Leaflets Evaluated

Six NMH leaflets were selected:

- Neonatal MRSA
- Group B Streptococcus (GBS)
- Mumps
- Zika virus
- Neonatal multidrug-resistant organisms (MDROs)
- Postnatal wound care

For each topic, an AI-generated leaflet was compared with the NMH leaflet. A total of 12 leaflets were evaluated.

Generation and Evaluation of AI leaflets

All AI leaflets were generated using the same predefined prompt. Only the clinical topic was changed on each search. This reduced variation caused by prompt differences. The AI was not provided with the corresponding NMH leaflets.

Predefined Prompt

The AI was given the following prompt: Create a patient information leaflet about (Topic) for use in a maternity or neonatal hospital setting. Use clear, plain English language with non-judgmental professional tone. Explain the condition, infection or colonisation, transmission risk, implications for the

mother and the baby, prevention measures, signs or symptoms, and when urgent medical assistance is needed. Avoid unnecessary medical terminology and define them where inclusion is required. Use bullet points where appropriate.

Using this predefined prompt for all leaflets could minimise result variation caused by different prompts.

Selection of the AI tool

A preliminary testing showed that Microsoft Copilot produced very brief leaflets, around half an A4 page, and did not provide sufficient detail for a meaningful comparison with the existing NMH leaflets, whereas ChatGPT generated more comprehensive leaflets with clear structure and greater clinical detail. ChatGPT was therefore selected to generate the six AI leaflets used in the final evaluation. The preliminary Copilot outputs were not included in the formal scoring or analysis. This was not intended as a direct comparison between Copilot and ChatGPT, but the comparison was necessary at the initial development phase to choose better leaflet options.

Evaluation Process

Four independent reviewers - microbiology registrar, patient representative, midwife, information officer – scored each leaflet across nine domains using a five-point scale: 1 = poor; 2 = below standard; 3 = adequate; 4 = good; 5 = excellent

Of 432 possible ratings (4 reviewers × 12 leaflets × 9 domains) 10 ratings were missing and 422 valid ratings (213 AI, 209 NMH) were analysed. Paired AI and NMH scores within each domain were compared using the Wilcoxon signed-rank test. Holm adjustment was applied to account for multiple testing across the nine domains.

Results

The overall mean scores were:

- AI-generated leaflets: **4.42/5**
- NMH leaflets: **4.52/5**

Although the overall scores were close, when individual domains were examined, important differences observed. AI performed particularly well in the use of simple language, human-centred tone and fairness. reviewers frequently noted a clear and simple language, logical layout, supportive and human-centred tone, inclusive wording and good transparency. However, the main concerns were clinically significant omissions and lack of details required for decision-making, oversimplification of the information and lack of local information or hospital pathway instructions. NMH leaflets performed better in clinically important domains, particularly accuracy, clinical safety, completeness, actionability. Accuracy and completeness showed nominal differences before adjustment for multiple comparisons (p=0.014 and p=0.032, respectively). However, neither remained statistically significant after Holm correction (adjusted p=0.129 and p=0.256, respectively), and no domain showed a statistically significant difference after adjustment. Overall suitability was almost similar (4.12 vs 4.14). Given the small exploratory sample, the absence of statistically significant adjusted differences should not be interpreted as evidence of equivalence between AI and NMH leaflets.

These findings suggest that evaluating the AI leaflets based on readability may results in false reassurance regarding their clinical suitability.

Full Results

Domain	What Was Assessed	AI Mean	NMH Mean	Unadjusted p	Holm-adjusted p
Accuracy	Information in line with current national and international guidelines	**4.04**	*4.58*	0.014	0.129
Safety	No advice/omission likely to delay care, increase transmission or cause harm	**4.36**	**4.61**	0.107	0.669
Completeness	Essential explanations, precautions and escalation advice are included	**3.98**	**4.50**	0.032	0.256
Understandability	Simple language, correct grammar/dictation, unnecessary use of medical terms and logical order	**4.60**	**4.29**	0.128	0.669
Actionability	Readers can identify what to do and when to seek help	**4.31**	**4.58**	0.098	0.669
Human-centred tone	Compassionate, non-stigmatising and supportive of clinical judgement	**4.75**	**4.54**	0.096	0.669
Fairness & inclusion	Inclusive language and assumptions that do not disadvantage groups	**4.98**	**4.88**	0.285	0.771
Transparency	Limits, source currency and need for individual clinical advice are clear.	**4.58**	**4.60**	0.257	0.771
Overall suitability	Suitable for use only after normal hospital approval processes.	**4.12**	**4.14**	0.823	0.823

Limitations

This was a small exploratory evaluation involving four reviewers, six leaflet topics from a single setting and subjective scoring. Reviews were partly unblinded, and the small sample limited statistical power to detect differences between AI and NMH leaflets. Multiple statistical comparisons were

therefore adjusted using the Holm method. AI models are also evolving rapidly, which may limit the generalisability and reproducibility of these findings.

Implications

AI could be used in producing the first draft of the patient-information leaflets or improve a human written information by enhancing the tone. adopting the plain language and restructuring the difficult subjects in a more understandable language. The AI leaflets should not be used without the specialist review and multidisciplinary approval. The clinical experts should maintain the professional accountability for clinical accuracy while using the AI communication advantages. As AI-generated material are more commonly used in many areas of medical practices, healthcare professionals need skills in a safe usage of AI and methods in critically evaluating the contents. Nurses and midwives should be able to identify inaccurate or missing information, oversimplification, outdated data, recommendations without robust evidence and omission of local context. Some structured evaluation tools may assist in assessing AI-generated materials.

Conclusion

AI-generated patient information leaflets were clear, easy to read and human-centred, with higher mean scores in several communication-focused domains. NMH leaflets had higher mean scores for accuracy, safety, completeness and actionability. Although nominal differences were observed for accuracy and completeness, no domain remained statistically significant after correction for multiple comparisons.

AI should therefore be used as a tool to support information development rather than an alternative to clinical expertise. Safe AI usage needs expert input, multidisciplinary review and staff AI literacy.