ABSTRACT

Doğan T. Comparison of Generative Artificial Intelligence Tools in Terms of Information Provision in the Genetic Counseling of Hereditary Cancer Syndromes. Istanbul University, Institute of Health Sciences, Department of Medical Genetics, Genetic Counseling Program. Master's Thesis. Istanbul. 2026.

This study aimed to comparatively evaluate the potential contribution of generative artificial intelligence tools (GAI) as supportive information sources for genetic counselors during preparation for genetic counseling (GC) sessions in the context of hereditary cancer syndromes (HCS). As of March 2026, the six most widely used GAI tools worldwide, namely ChatGPT, Gemini, Claude, Grok, DeepSeek, and Copilot, were evaluated across three separate analyses using the latest models available in their free plans. In the first stage, 60 open-ended questions based on current guidelines were submitted to the GAI tools using standardized prompts. In the second stage, 60 case narratives adapted from PubMed-indexed case reports were submitted in the same manner. The responses were scored by two independent doctoral-level genetics specialists. In the third stage, Human Phenotype Ontology (HPO) IDs derived from the case narratives using Doc2HPO were provided to PubCaseFinder and the GAI tools, and whether each tool reached the correct final diagnosis in its first response was recorded as a binary outcome. In the statistical analyses, combined scores from the open-ended question and case narrative responses were compared using the Friedman test. Pairwise comparisons were performed using the Wilcoxon signed-rank test, and multiple comparisons were corrected using the Holm correction. Inter-rater agreement was assessed using quadratic weighted Cohen's kappa. In the HPO-based diagnostic analysis, the overall difference among tools was evaluated using Cochran's Q test, while pairwise comparisons between PubCaseFinder and the GAI tools were assessed using exact McNemar tests. In the open-ended question analysis, Gemini and DeepSeek achieved the highest median scores, both with a median of 8. Grok and Claude showed intermediate performance with a median of 7, whereas Copilot and ChatGPT had the lowest median scores with a median of 6. The difference among GAI tools was statistically significant (Friedman test, chi-square = 252.59, df = 5, $p < 0.0001$), and inter-rater agreement was moderate (kappa = 0.484). In the case narrative analysis, Claude clearly stood out with the highest scores, reaching a median of 8. Grok and DeepSeek followed Claude, while ChatGPT, Gemini, and Copilot showed lower performance. The overall difference among GAI tools was also statistically significant in this analysis (Friedman test, chi-square = 140.90, df = 5, $p < 0.0001$), but inter-rater agreement was markedly low (kappa = -0.096). This finding suggested that the evaluation of case-based responses was more open to interpretation. In the HPO-based diagnostic analysis, the highest accuracy was observed for PubCaseFinder (20%). Claude followed with an accuracy of 16.7%. ChatGPT reached an accuracy of 10%, whereas Copilot, DeepSeek, Grok, and Gemini remained below 10%. The overall difference among tools was statistically significant (Cochran's Q test, Q = 23.17, df = 6, $p = 0.0007$). However, after multiple-comparison correction, no pairwise difference between PubCaseFinder and any GAI tool reached the predefined significance threshold of $p < 0.005$. Overall, the findings showed that no single GAI tool was superior across all tasks and that performance varied substantially according to task type, namely information provision, clinical case interpretation, and ontology-based diagnostic prediction. The main limitations of this study include the limited sample size, evaluation of only first responses, the fact that the findings reflect the model versions available during the study period due to the rapidly evolving nature of GAI tools, and the exclusion of empathy, ethics, and autonomy-related dimensions of real counseling encounters. GAI tools have high potential to provide reliable and efficient informational support to genetic counselors. However, these tools should not be regarded as independent decision-makers, but rather as supportive information resources that require expert supervision.

Key Words: Genetic counseling, hereditary cancer syndromes, generative artificial intelligence tools, Human Phenotype Ontology, PubCaseFinder.