不同人工智能工具自动生成外科疾病护理诊断的质量分析

李鹏[1,2]，张源慧[3]，唐龙[4]，刘杉[5]，郑晓妮[1]，Ji Jianchun[6]，刘坚[7]

摘要：目的 评估不同人工智能工具自动生成外科疾病护理诊断的准确性，测试其生成护理诊断的潜力，扩宽人工智能在护理领域的应用范围。**方法** 2024年12月，选择《外科护理学》中10个典型案例，以统一指令分别输入豆包、ChatGPT-4、Kimi、文心一言、通义千问5种人工智能工具，令其自动生成外科疾病护理诊断。邀请11名专家就生成的护理诊断质量进行评价。**结果** 5种人工智能工具生成护理诊断质量评分按高到低排序为豆包，ChatGPT-4，Kimi，文心一言，通义千问；不同人工智能工具生成的外科疾病护理诊断质量评分比较差异有统计学意义（$P<0.05$）。11名专家就生成的10个护理诊断质量评价之间的Fleiss′s Kappa值为0.445（$P<0.05$）。**结论** 人工智能工具自动生成外科疾病护理诊断的质量的认可度较高，但需要在实际使用过程中结合临床经验进行进一步的准确性判断。

关键词：外科疾病；　护理诊断；　人工智能；　专家函询；　质量评价；　外科护理；　数智化护理
中图分类号：R473.6；TP18　**DOI：**10.3870/j.issn.1001-4152.2026.03.110

Quality analysis of automatic generation of surgical disease nursing diagnoses by different artificial intelligence tools

Li Peng, Zhang Yuanhui, Tang Long, Liu Shan, Zheng Xiaoni, Ji Jianchun, Liu Jian. Scool of Nursing, Yiyang Medical College, Yiyang 413000, China

Abstract: Objective To evaluate the accuracy of automatic generation of surgical disease nursing diagnoses by different artificial intelligence (AI) tools, test their potential in generating nursing diagnoses, and expand the application scope of AI in the nursing field. **Methods** In December 2024, 10 typical cases from Surgical Nursing were selected. Unified instructions were input into 5 AI tools (Doubao, ChatGPT-4, Kimi, ERNIE Bot, and Qwen) respectively to generate surgical disease nursing diagnoses automatically. Eleven experts were invited to evaluate the quality of the generated nursing diagnoses. **Results** The quality scores of nursing diagnoses generated by the 5 AI tools, ranked from highest to lowest, were Doubao, ChatGPT-4, Kimi, ERNIE Bot, and Qwen; there was a statistically significant difference in the quality scores of surgical disease nursing diagnoses generated by different AI tools ($P<0.05$). The Fleiss′s Kappa value among the 11 experts′ evaluations of the quality of 10 generated nursing diagnoses was 0.445 ($P<0.05$). **Conclusion** The quality of surgical disease nursing diagnoses automatically generated by AI tools has high recognition, but further accuracy judgment combined with clinical experience is needed in practical use.

Keywords: surgical diseases; nursing diagnosis; artificial intelligence; expert consultation; quality evaluation; surgical nursing; digital-intelligent nursing

护理诊断（Nursing Diagnosis，NP）作为一种方法工具能促使护士系统识别、理解、评估、预测患者的健康问题，其目的是通过系统化过程改善患者健康状况并实现以患者为中心的护理，最大限度减少错误或失误[1-2]。护理诊断作为一种陈述，反映护理人员初始评估阶段是否能很好识别患者的体征、症状以及其相关风险因素。如果护理诊断是正确的，将引导正确的护理干预和评价，反之，则无法准确判断患者健康问题，从而导致错误的干预及评价，降低护理质量[3]。人工智能（Artificial Intelligence，AI）作为互联网主要组成部分，通过复杂的人机技术和人机交换来增强干预与交互，分析用户输入并根据历史记录和偏好，提供量身订制的建议和支持[4]。因此，AI工具受到越来越多的认可，但其在护理教育和临床护理领域应用仍处于起步阶段，如语音识别、数据挖掘、身体恶化预测等[5-7]，其潜在用途、好处、挑战和道德考虑仍有待深度挖掘。目前未见有基于AI的护理诊断输出的研究。本研究通过对比分析AI工具生成护理诊断的准确性，揭示AI技术在护理诊断中的现存问题与改进空间，筛选符合我国护理实践需求的智能工具提供参考。

1　资料与方法

1.1　一般资料　2024年12月，以人民卫生出版社十四五规划教材《外科护理学》（第七版）[8]为蓝本，经研究小组讨论，按照外科疾病大类，择选主要章节中具有典型代表性的思考题案例作为研究对象，最终纳入失血性休克、大肠癌、颅内血肿、损伤性气胸、腹股沟斜疝、急性腹膜炎、急性梗阻性化脓性胆管炎、尿道

作者单位：1. 益阳医学高等专科学校护理学院（湖南 益阳，413000）；2. 忻州现代康养职业学院；桂林医科大学附属医院 3. 急诊医学科 4. 重症医学科；5. 益阳市中心医院体检中心；6. 墨尔本澳大利亚联盟护理学院；7. 益阳医学高等专科学校附属医院放射科
通信作者：张源慧，170153896@qq.com
李鹏：男，博士，副教授，护理学院副院长，451792921@qq.com
科研项目：2024年湖南省教育厅教学改革项目（ZJGB2024324）；2024年度湖南省社会科学成果评审委员会课题（XSP24YBC210）
收稿：2025-09-05；修回：2025-11-02

损伤、Colles 骨折、急性乳腺炎 10 个案例。每个案例统一描述现病史、既往史、体征和检查结果 4 部分。以急性梗阻性化脓性胆管炎为例。现病史：患者，女性，83 岁，因“突发右上腹疼痛、伴皮肤巩膜黄染 2 d，加重 12 h”入院，诊断为急性梗阻性化脓性胆管炎。既往史：患者既往有肝内胆管结石，未予治疗；否认高血压、心脏病、糖尿病等慢性病史，无药物过敏史及手术外伤史。体征：T 39.5℃，P 126 次/min，R 26 次/min，BP 82/54 mmHg；神志不清、烦躁不安，口唇发绀，皮肤巩膜黄染，皮下有瘀斑，右上腹及剑突下压痛，轻度反跳痛及肌紧张。检查结果：血常规示 WBC 20.8×10^9/L；腹部超声示肝左外叶有多个强回声团，呈串排列，最大直径约 0.5 cm。

1.2 方法

基于工具可及性（公开访问无门槛）、功能适配性（支持长文本分析，匹配护理案例场景输入）、用户普及度（护理人员及大众认知度、使用频率较高）及技术代表性（能较为全面地进行文本内容分析）选取当前主流 AI 工具。经研究小组商议，决定选取 5 个 AI 工具：豆包（https://www.doubao.com/chat/）、通义千问（https://tongyi.aliyun.com/qianwen/）、文心一言（https://yiyan.baidu.com/）、Kimi（https://kimi.moonshot.cn/）以及 ChatGPT-4（https://chatgpt.com/）进行护理诊断的生成，具体方法如下。

1.2.1 输入前训练 目前 AI 工具均不是医学专用 AI，因此在输入指令前，事先对 AI 工具进行相关知识训练，使其对所需分析的相关护理知识有所掌握。本研究在进行指令输入前，对 AI 工具上传《NANDA-I 护理诊断：定义与分类（2021－2023）》[9]、人民卫生出版社十四五规划教材《外科护理学》（第七版）[8]，并输入指令如下：我现在将给你 2 个文件，请你认真学习，学习完成后通知我，等候我下一步指令。

1.2.2 输入内容及自动生成内容输出 为保证输入信息的准确性，输入前由 2 名研究小组成员核对输入信息，确认无误后采用 4 段式格式输入 AI 工具（ChatGPT-4 信息输入与导出由国外研究小组成员负责）。①角色：你现在是一位资深的护理专家；②背景：正在指导护生学习或临床护士的护理工作，需要针对临床护理病例分析其目前存在的护理问题/诊断；③任务：从专业角度出发，请对以下病例进行全面分析；④目标：要求在训练学习内容的基础上，重点关注北美护理诊断与案例相关护理诊断的逻辑性、精准性、严谨性，并遵循“首优”原则（即首先关注并解决患者当前最紧迫、最重要的护理问题）；输出至少 3 个护理诊断及其依据，并按重要性先后排序，其中排序第 1 位为“首优”护理诊断；诊断符合 NANDA 护理问题/诊断的要求，并列出信息来源或依据；输出格式为护理问题/诊断的 PES 三段式，即问题（problem）、病因（etiology）、症状和体征（signs and symptoms）。之后导出 AI 自动生成的文本，并将文本整理成 Excel 文档，按序号排列。

1.2.3 专家函询 自行设计专家函询表，包括专家的一般信息、AI 自动生成护理诊断和依据评价两部分。其中 AI 自动生成护理诊断和依据的评价包括文本的准确性（首优护理诊断、护理诊断重要性排序、诊断正确性 3 个条目）、专业性（使用学术用语、解释诊断的准确性 2 个条目）2 个方面，采取 Likert 5 级评分法计分，0～4 分分别表示“极差、较差、一般、良好、优秀”，总分 0～20 分，其中 0～＜6、6～＜11、11～＜16、16～20 分分别表示为“质量较差、一般、良好、优秀”。邀请专家根据理论、经验和参考文献对 10 个案例的输出内容进行评价。最终纳入来自中国湖南省、广西壮族自治区和澳大利亚的 11 名护理专家，男 4 名，女 7 名；年龄 36～54（42.55±3.59）岁；工作年限 15～36（19.73±5.80）年；均为副高级以上职称、研究生学历。本研究共进行 1 轮专家函询发放问卷共 11 份，回收有效问卷 11 份，有效回收率为 100%，其中专家判断系数（Ca）为 0.982，专家熟悉程度为 0.946，专家权威系数为 0.964。

1.2.4 统计学方法 采用 SPSS27.0 软件进行统计分析。本研究数据不服从正态分布，以中位数 M（P_{25}，P_{75}）表示，采用非参数检验。采用 Cronbach′s α 系数法评估专家评分的可靠性，采用 Fleiss′s Kappa 检验来评估专家评分的一致性（以＞0.4 表示有一定的可接受一致性）。检验水准 $\alpha=0.05$。

2 结果

2.1 专家对不同 AI 工具自动生成的外科疾病护理诊断质量评价的一致性 11 名专家对 5 种 AI 工具自动生成的 10 个外科疾病的护理诊断的质量评分 Cronbach′s α 系数范围为 0.948～0.996，总 Cronbach′s α 系数为 0.982。Fleiss′s Kappa 检验结果显示，Kappa 值为 0.445（$P<0.05$），专家对 10 个外科疾病的护理诊断评分的一致性可接受。

2.2 不同 AI 工具自动生成的外科疾病护理诊断的质量比较 通义千问、文心一言、Kimi1、ChatGPT-4、豆包 5 种 AI 工具生成外科疾病护理诊断的质量评分为分别为：9～14、12～18、13～18、10～18、12～19 分。≥11 分的评分占所有评分的 98.18%。不同 AI 工具自动生成的外科疾病护理诊断的质量评分比较，见表 1。

2.3 AI 工具生成不同案例护理诊断质量评分 见表 2。

3 讨论

3.1 AI 工具可以生成较为准确的护理诊断内容 本研究经研究小组充分讨论，选择 10 个典型案例，并对豆包、ChatGPT-4、Kimi、文心一言、通义千问 5 种 AI 工具进行专业化指令训练，要求这 5 种 AI 工具根据输入

的案例自动生成外科疾病护理诊断，邀请 11 名专家对 5 种 AI 工具自动生成的外科疾病护理诊断的准确定和专业性进行评分，其中专家权威系数为 0.964，总 Cronbach′s α 系数为 0.982，Kappa 值为 0.445（$P<0.05$），说明专家评分结果权威可信，一致程度较好。5 种 AI 工具自动生成的外科疾病护理诊断的质量评分总分为 9～19 分，≥11 分的评分占所有评分的 98.18%，表明总体质量良好。以上结果表明，AI 工具可以生成较为准确的护理诊断内容，这与徐文博等[10]的研究结论基本一致，可见 AI 展现出可作为护理教学和临床护理辅助工具的巨大潜力。自从 ChatGPT 发布以来，其以出色的对话交互能力和广泛的知识储备，成为 AI 领域的标志性产品[11]。目前国内 AI 也有了突飞猛进的发展，对于生成内容的准确性提升迅速[12]。但目前个别内容可能还会有准确性较差的问题，如本研究中案例 10 得分最低，分析原因，案例 10 并存高血压、糖尿病等多种并发症，而 AI 生成的护理诊断在细节上存在偏差，未能全面覆盖所有症状，也未能全面覆盖所有并发症潜在风险，提示在使用 AI 工具时需结合专业知识和临床经验进行综合判断。

表 1　不同 AI 工具自动生成的外科疾病护理诊断的质量评分比较　　分，$M(P_{25},P_{75})$

项目	准确性			专业性		总分
	首优护理诊断	护理诊断重要性排序	诊断准确性	使用学术术语	护理诊断完整性	
通义千问	3.0(2.0,3.0)	2.0(2.0,3.0)	2.0(2.0,3.0)	3.0(2.0,3.0)	3.0(3.0,3.0)	13.0(12.0,14.0)
文心一言	3.0(3.0,4.0)	3.0(3.0,3.0)	3.0(3.0,4.0)	3.0(3.0,4.0)	3.0(3.0,3.0)	15.0(14.0,17.0)
Kimi	3.0(3.0,3.0)	3.0(3.0,3.0)	3.0(3.0,4.0)	3.0(3.0,4.0)	3.0(3.0,3.0)	15.0(14.0,16.0)
ChatGPT-4	3.0(3.0,3.0)	3.0(3.0,3.0)	3.0(3.0,4.0)	3.0(3.0,4.0)	3.0(3.0,3.0)	16.0(15.0,17.0)
豆包	4.0(3.0,4.0)[abcd]	3.0(3.0,4.0)[abc]	4.0(3.0,4.0)[abcd]	4.0(3.0,4.0)[abcd]	4.0(3.0,4.0)[abcd]	18.0(17.0,19.0)[abcd]
Hc	112.721	216.949	161.149	143.156	157.664	332.360

注：均 $P<0.001$。与通义千问比较，[a]$P<0.05$；与文心一言比较，[b]$P<0.05$；与 Kimi 比较，[c]$P<0.05$；与 ChatGPT-4 比较，[d]$P<0.05$。

表 2　AI 工具生成不同案例护理诊断质量评分

分，$M(P_{25},P_{75})$

案例	准确性	专业性	总分
1	10.0(8.0,11.0)	7.0(6.0,7.0)	17.0(13.0,18.0)
2	9.0(8.0,11.0)	7.0(6.0,8.0)	17.0(15.0,18.0)
3	8.0(8.0,10.0)	6.0(6.0,7.0)	15.0(13.0,16.0)
4	6.0(8.0,11.0)	6.0(6.0,7.0)	16.0(14.0,17.0)
5	9.0(8.0,10.0)	6.0(6.0,7.0)	15.0(14.0,17.0)
6	10.0(9.0,11.0)	6.0(6.0,7.0)	17.0(16.0,17.0)
7	10.0(9.0,11.0)	6.0(6.0,7.0)	16.0(15.0,17.0)
8	10.0(8.0,11.0)	7.0(6.0,7.0)	16.0(14.0,17.0)
9	10.0(8.0,11.0)	6.0(6.0,7.0)	16.0(14.0,17.0)
10	8.0(7.0,9.0)	6.0(5.0,7.0)	14.0(12.0,15.0)

3.2　不同 AI 工具生成的内容在准确性和专业性有一定的差异性，需要鉴别判断　从本研究所针对的 5 种 AI 工具的研究情况来看，豆包的准确性和专业性评分最高，其次为 ChatGPT-4，通义千问的准确性和专业性评分最低，表明不同 AI 工具之间在对问题分析的准确性、专业性方面有一定的差异，这与相关研究结果一致[13-14]。提示在选择和使用 AI 工具时，应充分考虑其适用性和局限性，避免盲目依赖。通过对比分析，发现 AI 在处理复杂案例时，仍需人工审核和补充，以确保诊断的全面性和精确性。这可能与不同 AI 的开源性、自我学习能力等有关，不能盲目地相信 AI 工具生成的信息[15]。同时本研究结果显示，开创 AI 的划时代产物 ChatGPT 生成护理诊断质量并不是得分最高的 AI 工具，一方面可能与国外的网络资源与国内有一定的区别有关；另一方面，也可能与国内 AI 快速发展有关。因此，建议护理人员在日常工作与学习中使用 AI 工具时，可以对多个 AI 工具进行比较，选择较为准确的生成内容。AI 工具功能正呈指数级跃迁[16-18]，并迅速渗透到护理全场景，成为变革核心驱动力。护理人员及护理教育者应紧跟数字时代步伐，积极把握机遇，推动护理实践、护理教学与 AI 技术的深度融合，以实现护理领域的创新与突破，为护理事业的高质量发展注入新动力。

3.3　正确使用及强化训练可以增加 AI 工具生成内容的准确性　在使用 AI 工具生成信息的过程中，正确输入指令、反复强化训练能够在一定程度上提高 AI 工具生成内容的准确性。此外，精准提问，清晰准确地表达需求，能引导 AI 工具给出更契合期望的答案[19]。如在提供病例的基础上，准确描述要求：重点关注北美护理诊断与案例相关护理诊断的逻辑性、精准性、严谨性。按照“首优”原则（即首先关注并解决患者当前最紧迫、最重要的护理问题），拟出至少 3 个护理诊断及其依据，并按重要性先后排序，其中排序第一位为“首优”护理诊断，可让人工智能聚焦于准确的北美护理诊断，而非宽泛模糊作答。同时做适当引导，对于复杂问题，通过分步骤提问或给予提示信息引导 AI 工具，逐步引导能使生成内容更精准、有条理。反复训练也是提高 AI 工具生成内容准确性的重要方法[20]。如本研究在输出内容前，将人民卫生社出版的《外科护理学》（第 7 版）教材的相关内容先传给 AI 工具学习，规范了外科护理学相关知识，增加大

量包含临床表现、辅助检查、治疗原则、护理问题、护理措施等信息，可拓宽模型知识面，使其在回答医疗问题时更准确专业，有效防止输出内容的随意性。目前的 AI 工具多为通用型，缺乏医学专用 AI 工具，提示在使用 AI 工具时，应针对外科护理学特定领域进行强化训练，使人工智能掌握专业知识和表达习惯或实时检索最新和权威的医学文献，将文献与相关医学知识图谱供 AI 学习，以增强生成模型的知识背景[21]。此外，使用外科护理学教科书等专业资料对 AI 工具进行训练，使其在回答护理问题如“分析当前患者存在的主要护理问题”时，给出符合专业逻辑的准确分析。

4 结论

本研究显示，利用不同 AI 工具生成护理诊断的准确性较好，可以为护理教育和临床护理的信息化智能化发展提供有力支持。但仍需注意 AI 工具的局限性，尤其在复杂病例分析中，需结合专业判断进行验证和调整，以确保诊断的可靠性和安全性。目前国内 AI 工具及网址众多，本研究只涉及到常见的几种，只对护理问题/诊断进行了验证，且本研究仅分析教材内案例，而临床疾病更为复杂和多变。因此，未来研究可针对真实案例的输出进行质量分析，并且需要纳入更符合中国临床环境和语言环境的人工智能工具，为临床决策做出更为精准和个性化的诊疗服务。

参考文献：

[1] American Nurses Association. The nursing process[EB/OL]. [2025-01-08]. https://www.nursingworld.org/practice-policy/workforce/what-is-nursing/the-nursing-process/.

[2] Leenara Bezerra da Silva C, Lopes de Moura E, Nayara do Nascimento Dantas T, et al. Nursing diagnoses in patients with COVID-19 admitted to the intensive care unit: cross-mapping[J]. Heliyon, 2024, 10(5): e27088.

[3] Park J, Jeong S. The analysis of nursing diagnoses determined by students for patients in rehabilitation units[J]. J Exerc Rehabil, 2022, 18(5): 299-307.

[4] Rony M K K, Kayesh I, Bala S D, et al. Artificial intelligence in future nursing care: exploring perspectives of nursing professionals: a descriptive qualitative study[J]. Heliyon, 2024, 10(4): 25718-25726.

[5] 王珍妮，须月萍，夏开建，等. 基于 YOLO 神经网络构建压力性损伤自动检测和分期的人工智能模型[J]. 中国全科医学，2024，27(36)：4582-4590.

[6] Hannaford L, Cheng X, Kunes-Connell M. Predicting nursing baccalaureate program graduates using machine learning models: a quantitative research study[J]. Nurse Educ Today, 2021, 99: 1047-1055.

[7] Liao C T, Tsay S F, Chen H C. Exploring generative AI's role in alleviating nursing workload and burnout in Taiwan[J]. J Formos Med Assoc, 2024, 123(7): 736-737.

[8] 李乐之，路潜. 外科护理学[M]. 7 版. 北京：人民卫生出版社，2021.

[9] Herdman T H, Kamitsuru S, Lopes C T. NANDA-I 护理诊断：定义与分类(2021－2023)[M]. 李小妹，周凯娜，译. 北京：世界图书出版公司，2023：20-23.

[10] 徐文博，陈凤敏，王超，等. GPT-4 大语言模型对护理知识理解的测试研究[J]. 护理学杂志，2024，39(19)：93-96.

[11] 卢宇，余京蕾，陈鹏鹤，等. 生成式人工智能的教育应用与展望：以 ChatGPT 系统为例[J]. 中国远程教育，2023，43(4)：24-31，51.

[12] 李艳艳，李洁，钟珍童，等. AI 语言模拟聊天机器人在孕产妇尿失禁健康教育中的应用[J]. 护理学杂志，2025，40(4)：1-5.

[13] Cheong R, Unadkat S, Mcneillis V, et al. Artificial intelligence chatbots as sources of patient education material for obstructive sleep apnoea: ChatGPT versus Google Bard[J]. Eur Arch Otorhinolaryngol, 2024, 281(2): 985-993.

[14] Srinivasan N, Samaan J S, Rajeev N D, et al. Large language models and bariatric surgery patient education: a comparative readability analysis of GPT-3. 5, GPT-4, Bard, and online institutional resources[J]. Surg Endosc, 2024, 38(5): 2522-2532.

[15] 金云波，龚盼盼，包莹莹，等. 强人工智能时代大学生自主学习能力发展面临的机遇与挑战：基于 ChatGPT 的审思[J]. 信阳师范学院学报(哲学社会科学版)，2024，44(1)：105-111.

[16] Ghanem Y K, Rouhi A D, Al-Houssan A, et al. Google to Dr. ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis[J]. Surg Endosc, 2024, 38(5): 2887-2893.

[17] 冯俊霞. 人工智能在企业人力资源招聘中的运用研究[J]. 商展经济，2025(1)：165-168.

[18] 杨静，张萌，朱亮，等. 人工智能在环境工程中的应用：热点演化与未来趋势[J]. 环境工程学报，2024(11)：3049-3058.

[19] 张汇典. 人工智能浪潮下新闻传播领域的风险及治理思考[J]. 西部广播电视，2024，45(5)：71-74.

[20] 王绍源，杨东航，任宇东. 大语言模型在护理领域的应用场景与伦理探讨[J]. 护理学杂志，2025，40(5)：108-113.

[21] 吴金玉，陈曦，黎慧，等. 大语言模型在护理领域的应用进展[J]. 护理学杂志，2024，39(17)：26-29.

(本文编辑　黄辉，吴红艳)