基于NLP的骨科查房量化反馈对住院医师临床思维的干预研究

张元强[1],李 茜[1],王文卿[2],李 硕[3],张保良[1]

山东大学齐鲁医院:1.骨科;2.神经内科,济南 250000;3.爱尔兰国立科克大学凯瑟琳麦考利护理和助产学院,科克,爱尔兰 T12YN60

【摘要】目的 探讨一种基于自然语言处理(natural language processing,NLP)的教学查房量化评估与反馈模式,在提升骨科住院医师临床思维能力方面的有效性。**方法** 采用前瞻性非随机对照研究设计,将山东大学齐鲁医院60名骨科住院医师分为试验组(n=30)与对照组(n=30)。试验组在为期3个月的干预期间,每周接受基于NLP技术分析其查房发言后生成的个人量化报告,并由带教医师进行一对一精准反馈。对照组则接受常规的、基于主观印象的口头反馈。采用微型临床评估演练(Mini-CEX)总分的变化值为主要结局指标,辅以理论知识测试、自信心问卷及360度评估等次要结局指标,综合评价干预效果。**结果** 研究结束时,试验组的Mini-CEX得分变化值(1.73±0.72)分显著高于对照组的(0.70±0.61)分,差异具有统计学意义(P<0.001)。此外,试验组在自信心问卷得分与360度评估得分上的改善也均显著优于对照组(均P<0.05)。相关性分析证实,NLP模型提取的"逻辑关联词频率"等关键指标与Mini-CEX的"临床判断"分项得分呈显著正相关(r=0.68,P<0.001)。**结论** 基于NLP的教学查房量化评估与反馈模式,能有效提升骨科住院医师的临床思维综合能力与自信心。该模式通过将隐性的思维过程显性化,实现了对临床教学的客观评估与精准干预,为住院医师培养体系的优化提供了有价值的实证依据与技术路径。

【关键词】自然语言处理;临床思维;住院医师;教学查房;形成性评价

【中图分类号】R-4;C975 **【文献标志码】**A **【文章编号】**1004-5287(2026)04-0568-11

【DOI】10.13566/j.cnki.cmet.cn61-1317/g4.202604020

Intervention study of NLP-based quantitative feedback in orthopedic teaching rounds on residents' clinical thinking ability

ZHANG Yuanqiang[1], *LI Qian*[1], *WANG Wenqing*[2], *LI Shuo*[3], *ZHANG Baoliang*[1]

Qilu Hospital of Shandong University: 1. *Department of Orthopedics*; 2. *Department of Neurology*, *Jinan* 250000, *China*; 3. *Catherine McAuley School of Nursing and Midwifery*, *University College Cork*, *Cork* T12YN60, *Ireland*

【Abstract】Objective To evaluate the effectiveness of a natural language processing (NLP)-driven quantitative assessment and feedback mode applied during orthopedic teaching rounds in enhancing clinical reasoning skills among resident physicians. **Methods** A prospective, non-randomized controlled study was conducted with 60 orthopedic residents assigned to either the experimental group (n=30) or the control group (n=30). During the 3-month intervention period, the experimental group received personalized quantitative reports generated based on NLP technology analysis of their ward rounds speeches every week, and received one-on-one precise feedback from the supervising physician. The control group, on the other hand, received conventional, subjective impression-based oral feedback. The total score change value of the Mini-CEX was used as the primary outcome indicator, while theoretical knowledge tests, self-confidence questionnaires, and 360-degree evaluations were

收稿日期:2026-03-17

作者简介:张元强,主治医师,博士,研究方向为骨科临床医学,临床医学教育方向。

通信作者:张保良,主治医师,博士,研究方向为临床骨外科医学方向。E-mail:sduqilu_zbl@yeah.net

employed as secondary outcome indicators to comprehensively evaluate the intervention effect. **Results** At study completion, the experimental group showed a significantly greater mean increase in Mini-CEX score (1.73±0.72) compared with the control group [(0.70±0.61), $P<0.001$]. Improvements in self-confidence and 360-degree evaluation scores of the experimental group were also significantly higher than those of the control group (both $P<0.05$). Correlation analysis revealed that the key indicators extracted by the NLP model, such as "frequency of logical connectives", have a significant positive correlation with the "clinical judgment" sub-score of Mini-CEX. ($r=0.68$, $P<0.001$). **Conclusion** The NLP-based quantitative assessment and feedback mode for teaching rounds effectively enhances the comprehensive clinical thinking ability and self-confidence of orthopedic residents. By making implicit cognitive processes explicit and measurable, this mode achieves objective assessment and precise intervention in clinical teaching, offering valuable empirical evidence and technical approaches for the optimization of the resident training system.

【Keywords】 natural language processing; clinical reasoning; resident physicians; teaching rounds; formative assessment

教学查房是住院医师规范化培训的核心环节，是其临床思维、知识整合与沟通能力培养的关键场域。在这一动态交互过程中，带教医师通过引导和反馈，帮助住院医师将理论知识转化为临床实践能力[1-4]。临床思维作为医师的核心素养，其培养效果直接关系到医疗质量与患者安全，而教学查房正是淬炼此项能力的主阵地[5-7]。

然而，传统的教学查房评估与反馈模式在很大程度上依赖于带教医师的主观经验与即时印象[8]。这种模式普遍缺乏统一、客观的评价标准，导致评估结果可能存在偏倚，难以精准、全面地识别出住院医师在临床思维链路中的具体短板，例如是知识应用不足，还是逻辑推理欠缺[9-11]。同时，基于主观印象的反馈往往是概括性、非结构化的，这限制了教学指导的针对性与可操作性，进而影响了住院医师学习效能的进一步提升[12]。因此，探索一种能够客观、精细化捕捉并评估住院医师临床思维过程的方法，已成为临床教学领域备受关注的问题。

近年来，以自然语言处理(natural language processing，NLP)为代表的人工智能技术取得了长足进步，其在深度分析与挖掘非结构化文本数据方面的能力，为破解上述教学难题提供了新的可能[13-16]。有研究表明，语言是思维的外在表征，通过分析个体在特定场景下的语言学特征，可以有效推断其思维过程的质量[17-18]。将NLP技术引入教学查房场景，通过对查房过程中的师生对话进行量化分析，有望将模糊、隐性的教学过程转化为可度量、可分析的数据，从而为建立客观、多维度的住院医师能力评价体系，以及实施数据驱动的精准教学反馈开辟了全新路径[19-21]。

基于此，本研究旨在探索构建一种基于自然语言处理的骨科教学查房量化评估模型，并验证该模型驱动下的精准反馈模式对提升住院医师临床思维能力的实际效果。本研究假设，相较于接受传统教学反馈的住院医师，接受基于NLP量化评估与精准反馈的试验组住院医师，其临床思维综合能力，尤其是在病例分析的逻辑性、知识应用的深度等方面，将获得更显著的提升。本研究期望通过实证研究，为优化临床教学评估方法、改进教学反馈模式、促进住院医师高质量成长提供科学依据与实践参考。

1　研究对象与方法

1.1　研究设计

本研究采用前瞻性、单中心、非随机对照研究设计，旨在比较基于NLP的量化反馈模式与传统反馈模式对骨科住院医师临床思维能力的影响。研究于2025年3月至2025年9月在山东大学齐鲁医院(以下简称"我院")骨科开展，方案已通过我院伦理委员会审查批准，所有研究对象在入组前均充分了解研究内容并签署了书面知情同意书。录音数据采集前均获得在场人员口头同意，文件立即去标识化并加密存储，仅课题组核心成员可访问。转录完成后原始录音即刻删除。知情同意书明确说明数据仅用于教学研究，不用于考核，研究对象可随时退出。为遵循教育公平原则，研究结束后立即为对照组提供其完整量化评估报告及补充性一对一反馈(延迟干预设计)，确保所有参与者最终均能从新模式中受益。

1.2　研究对象

纳入标准：我院骨科在培的第一年或第二年(PGY1/PGY2)住院医师；本次骨科轮转周期不少于3个月；自愿参加本研究并签署书面知情同意书。排除标准：轮转时间不足3个月或中途退出轮转者；有严重听力或语言交流障碍，影响正常教学查房与录音者；既往已接受过类似的量化教学反馈培训者。

1.3　NLP量化评估模型的构建与验证

本研究基于开源Python库构建了一套模块化、

可审计的NLP分析流程。该模型核心组件包括spaCy 3.2（Honnibal et al., 2020）与jieba 0.42.1（Sun, 2012），各模块间通过标准化JSON接口实现数据传递，确保全流程可追溯与版本控制。spaCy 3.2是基于OntoNotes 5.0及Common Crawl的语料训练，能够在通用中文句法中较高程度地完成分析任务；jieba分词工具则采用基于Trie树的前缀词典扫描与动态规划切分策略，用户可自定义词典接口支持动态加载专业词库，能够满足批量快速处理的效率需求。

本研究采用了基于规则与专业词典的混合方法，而非端到端的深度学习神经网络模型。主要原因是考虑到本研究所涉及的骨科临床能力评估数据为有限样本场景（$n<500$），难以满足深度神经网络模型对大规模标注数据的训练需求，这可能会导致结果存在过拟合的风险。其次，临床能力的评估多涉及到多维度可解释性的指标，运用规则驱动的方法能够提供较为清晰透明的决策路径，便于教育者认识和理解模型输出的依据。同时，考虑到医学教育评价领域对评估结果的信度与效度一般具有较为严格的要求。因此，本研究选择基于词典与规则的方法，能够帮助我们在概念对齐与人工校验方面获得方法论的优势。医学垂直领域的预训练模型（如MC-BERT、PubMedBERT）虽可缓解数据稀缺问题，但其对骨科亚专科术语的覆盖度有限，且微调过程需大量标注资源，成本效益比劣于规则方法。

1.3.1 骨科专业词库的构建方法

骨科专业词库是本分析流程的核心基础资源，本研究对于专业词库的构建主要遵循着“权威来源提取、临床指南补充和专家人工校验”的阶段性质量控制流程。研究组成员首先从《坎贝尔骨科学》（Campbell's Operative Orthopaedics，第14版）及《骨科学》（人民卫生出版社，第9版）等权威教材中系统提取核心专业术语，经过去重与同义词归并后，纳入约3000个术语条目。第二阶段，参考中华医学会骨科学分会发布的《骨关节炎诊疗指南（2021年版）》《股骨颈骨折临床诊疗指南（2018年版）》等10部临床诊疗指南，补充疾病诊断分类、手术操作编码及治疗路径相关的标准术语约500个，确保词库与现行临床规范的一致性。第三阶段，邀请3位高年资骨科主任医师对词库进行独立审核与交叉验证，重点修正解剖术语的规范性表述、补充亚专科特色术语（如脊柱外科、运动医学、骨肿瘤学），并最终形成共识版本。

1.3.2 词库层级分类体系

将词库按照临床认知深度与知识应用层级进行了三级分类。基础解剖词汇层（Level 1）：涵盖骨骼、关节、韧带、肌肉等结构术语，如“股骨头”“髋臼”“前交叉韧带”等，主要反映受试者的解剖学知识储备；核心诊断词汇层（Level 2）：包括疾病名称、分型标准及病理机制术语，如“股骨颈骨折（Garden Ⅳ型）”“骨性关节炎（Kellgren-Lawrence Ⅲ级）”等，用于评估疾病识别与分类能力；高阶治疗决策词汇层（Level 3）：主要包含手术术式、保守治疗策略及预后评估术语，如“全髋关节置换术”“保髋治疗”“负重时间窗”等，表征临床决策与知识整合能力。该分级体系不仅服务于专业词汇密度的计算，更为后续知识应用深度的梯度评估提供了结构化框架。

1.3.3 各项量化指标的操作性定义与计算方法

本研究对所有NLP提取指标均给出明确的操作性定义与数学表达式，以确保量化评估的客观性与研究的可重复性。

专业词汇密度（professional vocabulary density，PVD）指发言文本中专业词库术语总频次与有效总词数之比，以百分比表示。计算公式为

$$\mathrm{PVD}=\sum_{i=1}^{3}\frac{n_i}{N_{\mathrm{valid}}}\times 100\% \tag{1}$$

式中：n_i为第i层级术语出现频次（同一术语多次出现分别计数）；N_{valid}为经预处理后有效词数（剔除填充词〈FILLER〉、语气词、重复修正片段）。

逻辑关联词频率（logical connector frequency，LCF）指逻辑关联词出现总次数与发言总句数之比，以每百句频次表示。逻辑关联词包括因果类（因为、所以、导致）、递进类（首先、其次、此外）、总结类（综上、总之）等，共计约50个词汇。计算公式为

$$\mathrm{LCF}=\frac{m}{S_{\mathrm{total}}}\times 100 \tag{2}$$

式中：m为逻辑关联词出现总次数；S_{total}为发言总句数。

鉴别诊断数量（differential diagnosis count，DDC）通过正则表达式触发识别后，提取窗口内疾病名称的去重计数。通过正则表达式匹配“鉴别诊断”“需要排除”“还要考虑”等关键短语后，提取其后出现的疾病名称数量。

人文关怀词汇频率（patient-centered communication frequency，PCCF）为人文关怀类词汇出现频次与有效总词数之比，以千分比表示。计算公式为

$$\mathrm{PCCF}=\frac{k}{N_{\mathrm{valid}}}\times 1000 \tag{3}$$

式中：k为共情与沟通类词汇出现次数；N_{valid}为发言总词数，词汇表包括“理解”担心”“沟通”“解释”等约30个词汇。

1.3.4　模型验证

针对鉴别诊断识别，经过人工标注50份文本后，对比模型识别的结果显示该模型的准确率为87.3%，召回率为82.1%，F_1值为84.6%（表1）。此外，为验证模型有效性，随机抽取30份转录文本，由2位骨科教学主任独立评分（参与度、知识深度、逻辑性、人文关怀各1～5分），两位专家评分一致性良好（ICC=0.82）。此外，NLP指标与专家评分呈显著正相关，其中专业词汇密度与知识深度（r=0.61，P=0.003），逻辑关联词频率与逻辑性（r=0.74，P<0.001）（表2）。

此外，为评估NLP指标在真实临床培训场景中的时间稳定性，本研究采用纵向追踪设计，对10名住院医师的临床考核文本进行连续12周的数据采集，计算各指标的组内相关系数（ICC，two-way random effects，absolute agreement）。

表1　鉴别诊断识别准确性验证结果　（n=50）

指标	数值
人工标注的鉴别诊断总数	156
NLP模型识别数	147
正确识别数（真阳性）	128
误识别数（假阳性）	19
漏识别数（假阴性）	28
准确率（Precision）	87.3%（128/147）
召回率（Recall）	82.1%（128/156）
F_1值	0.846
标注者间Kappa	0.88

表2　NLP指标纵向稳定性评估结果

指标	ICC	95% CI
专业词汇密度	0.78	[0.71，0.84]
逻辑关联词频率	0.76	[0.68，0.82]
鉴别诊断数量	0.72	[0.63，0.79]
人文关怀词汇频率	0.70	[0.61，0.77]

1.4　研究流程与干预方法

研究流程主要分为基线评估、干预实施与终末评估三个阶段，具体流程如图1所示。

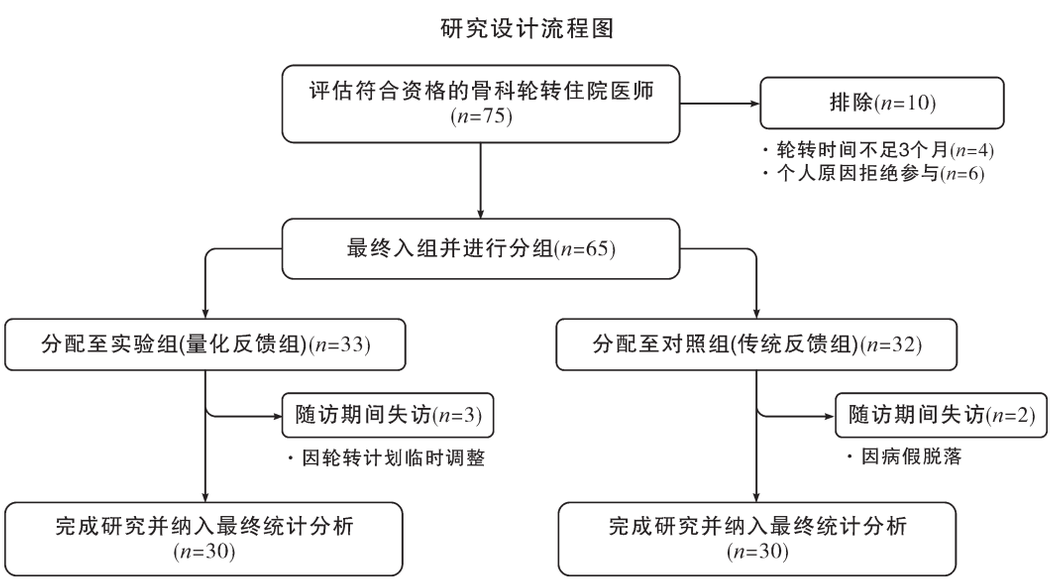


图1　研究设计流程图

1.4.1　基线评估与分组

所有符合标准的住院医师在研究开始时均完成一次基线评估，内容包括：微型临床评估演练（Mini-CEX）；理论知识测试；自信心与自我效能问卷。完成评估后，根据研究对象的入组顺序奇偶数，将其非随机分配至试验组（量化反馈组）或对照组（传统反馈组）。

1.4.2　数据采集与干预实施

研究为期3个月。在此期间，每周固定两次对科室教学查房的全过程进行高保真录音，以采集所有住院医师的发言数据。具体步骤和内容包括（图2）以下几个方面。试验组（量化反馈组）：每周末，带教主治医师会收到其负责的试验组住院医师的个人量化评估报告。在下一周初，带教医师会与每位住院医师进行一次约10～15 min的一对一反馈面谈，依据报告内容进行针对性指导和讨论。对照组（传统反馈组）：正常参与教学查房并被录音，但其本人及带教医师均不接触任何量化评估报告。带教医师仅依据个人主观印象，在查房过程中或结束后进行常规的、非结构化的口头点评与指导。数据处理：每周对录音文件进行

语音转文本处理，并分离出每位住院医师的发言片段。NLP模型分析：采用本研究构建的NLP量化评估模型，从参与度、知识深度、逻辑结构、人文关怀四个维度对发言文本进行自动化分析，提取“专业词汇密度”“逻辑关联词频率”等关键指标。生成报告：基于分析结果，为每位住院医师生成一份可视化的个人量化评估报告，直观展示其多维度能力表现及与同级平均水平的对比。精准反馈：每周带教医师依据此量化报告，与每位住院医师进行一次约10～15 min的“一对一”反馈面谈，进行针对性的指导与讨论。

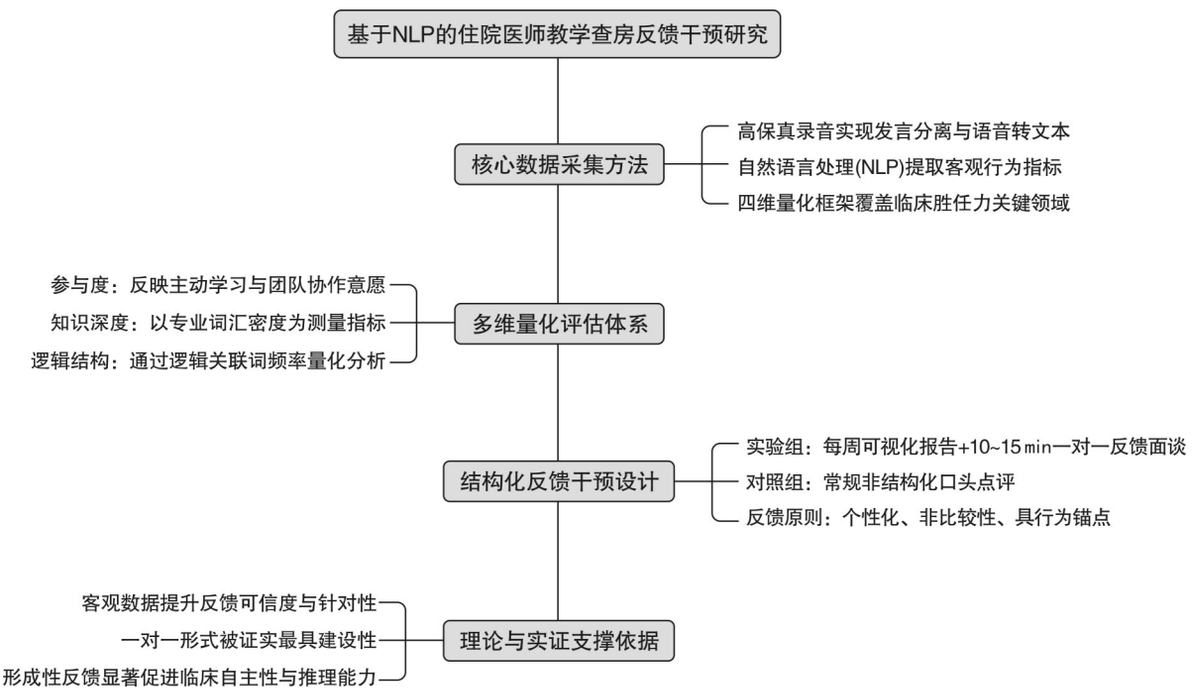


图2 干预与数据采集思维导图

1.4.3 终末评估

所有研究对象在3个月干预期结束时，再次完成一套与基线评估平行且难度相当的终末评估，内容同样包括Mini-CEX、理论知识测试和自信心问卷。此外，还需完成一次360度评估。

1.5 评估工具与结局指标

1.5.1 主要结局指标

Mini-CEX总分的变化值用于衡量研究前后住院医师临床思维综合能力的变化。该工具由经过标准化的带教医师进行评估，信效度良好。

1.5.2 次要结局指标

理论知识测试成绩的变化值：评估基础理论与临床知识水平的变化。自信心与自我效能问卷得分变化值：评估研究对象在临床思维方面的心理层面变化。360度评估得分：于研究结束时，由带教医师、护士长及同组同事从多角度对住院医师的综合能力进行评价。试验组NLP量化指标的变化趋势：分析试验组内部，如“专业词汇密度”“逻辑关联词频率”等指标随时间的变化情况。

1.6 质性研究方法

在研究结束时，采用目的性抽样，对两组部分住院医师及带教医师进行半结构化访谈，以深入探究他们对各自反馈模式的主观体验、感知效果及优化建议。所有访谈均进行录音、转录，并采用主题分析法(thematic analysis)对文本资料进行编码和主题提炼。

1.7 统计学方法

所有数据采用SPSS 26.0软件进行统计分析。计量资料以均数±标准差($\bar{x}\pm s$)或中位数(四分位距)[M(IQR)]表示，计数资料以例数(百分比)[n(%)]表示。基线资料比较中计量资料采用独立样本t检验，计数资料采用χ^2检验或Fisher精确检验。主要结局指标分析采用协方差分析(ANCOVA)比较两组终末Mini-CEX得分的差异，并以基线得分为协变量进行校正。次要结局指标分析计量资料的变化值比较采用独立样本t检验；等级资料(360度评估)采用Mann-Whitney U检验。组内趋势分析对试验组的NLP量化指标，采用重复测量方差分析检验其随时间的变化趋势。相关性分析采用Spearman相关性分析，探讨终末NLP量化指标与Mini-CEX各分项得分之间的相关性。所有检验均为双侧检验，以$P<0.05$为差异具有统计学意义。

2　结果

2.1　研究对象基本情况

本研究共初步评估了75名符合条件的骨科轮转住院医师，其中10名因轮转时间不足或个人原因未参与，最终65名住院医师签署知情同意书并入组。研究对象按入组顺序的奇偶数分配至试验组（n=33）与对照组（n=32）。在为期3个月的干预期间，试验组有3名、对照组有2名研究对象因故脱落。最终，共有60名住院医师（每组30名）完成了全部研究流程并被纳入最终统计分析。对两组最终纳入分析的住院医师的基线资料进行比较，结果如表3所示。两组研究对象在年龄、性别构成、培训年资（PGY1/PGY2比例）上均无显著差异（均P>0.05）。同时，两组在研究开始时的主要及次要结局指标，包括Mini-CEX得分、理论知识测试成绩以及自信心问卷得分，其差异亦无统计学意义（均P>0.05）。上述结果表明，两组研究对象在基线水平上具有良好的可比性。

表3　两组研究对象基线资料比较

特征	试验组（n=30）	对照组（n=30）	检验值（t/χ^2）	P
年龄[($\bar{x}\pm s$)，岁]	26.5±1.2	26.8±1.4	−0.88	0.38
性别（男/女，n）	18/12	17/13	0.14	0.71
培训年资（PGY1/PGY2，n）	15/15	15/15	-	>0.99
基线Mini-CEX得分[($\bar{x}\pm s$)，分]	6.12±0.81	6.05±0.89	0.34	0.73
基线理论知识成绩[($\bar{x}\pm s$)，分]	72.5±8.1	73.1±7.9	−0.31	0.76
基线自信心问卷得分[($\bar{x}\pm s$)，分]	3.51±0.62	3.44±0.70	0.44	0.66

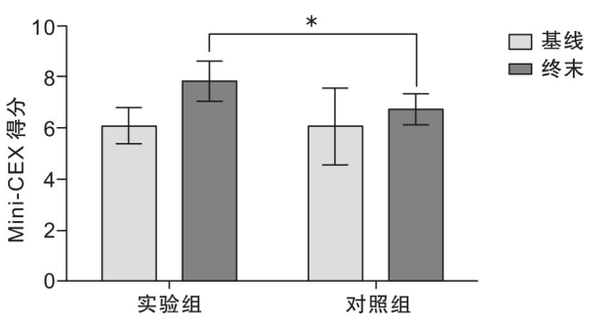
2.2　两组临床思维能力及相关指标变化比较

2.2.1　主要结局指标：Mini-CEX得分变化

如图3和表4所示，干预结束后，两组住院医师的Mini-CEX总分均有提高。试验组的Mini-CEX平均得分从基线的（6.12±0.81）分显著提升至（7.85±0.65）分，平均变化值为（1.73±0.72）分。对照组的平均得分则从（6.05±0.89）分提升至（6.75±0.78）分，平均变化值为（0.70±0.61）分。经协方差分析校正基线得分后，试验组的终末Mini-CEX得分显著高于对照组，差异具有统计学意义（P<0.001），表明量化反馈模式能更有效地提升住院医师的临床思维综合能力。

2.2.2　次要结局指标变化

次要结局指标的组间比较结果详见表4。在理论知识方面，虽然两组成绩在干预后均有增长，但其变化值的组间差异无统计学意义（P=0.40），提示新模式对理论知识的提升效果与传统模式相当。在心理层面，试验组的自信心问卷得分变化值（1.02±0.45）分显著高于对照组（0.31±0.58）分，差异具有高度统计学意义（P<0.001），表明量化反馈模式能显著增强住院医师在临床思维方面的自信心与自我效能感。在360度评估中，试验组获得的综合评价得分中位数（4.2，IQR：3.8～4.5）显著优于对照组（3.5，IQR：3.2～3.9），差异具有统计学意义（P=0.004），说明从同事、带教及护士等多角度观察，试验组住院医师的综合表现获得了更高的认可。



注：*为P<0.001

图3　两组干预前后Mini-CEX得分对比箱线图

表4　两组研究前后各项结局指标得分比较

指标	组别	基线得分[($\bar{x}\pm s$)，分]	终末得分[($\bar{x}\pm s$)，分]	变化值[($\bar{x}\pm s$)，分]	P
Mini-CEX总分	试验组	6.12±0.81	7.85±0.65	1.73±0.72	<0.001
	对照组	6.05±0.89	6.75±0.78	0.70±0.61	
理论知识成绩	试验组	72.5±8.1	80.1±7.5	7.6±4.2	0.40
	对照组	73.1±7.9	79.5±7.2	6.4±3.8	
自信心问卷得分	试验组	3.51±0.62	4.53±0.55	1.02±0.45	<0.001
	对照组	3.44±0.70	3.75±0.61	0.31±0.58	
360度评估得分（M，IQR）	试验组	-	4.2(3.8～4.5)	-	0.004
	对照组	-	3.5(3.2～3.9)	-	

2.3 试验组NLP量化指标变化趋势

为了探究量化反馈模式的作用过程，对试验组住院医师在10周干预期间由NLP模型提取的关键指标进行了纵向分析。如图4所示，“专业词汇密度”“逻辑关联词频率”两项核心指标均呈现出持续且稳定的增长趋势。重复测量方差分析结果显示，这两项指标随时间的增长均具有统计学意义（均$P<0.001$）。具体而言，“专业词汇密度”从研究初期的平均12.1%增长至末期的19.8%；“逻辑关联词频率”则从平均每百句发言使用4.5次增长至8.2次。这一趋势直观地反映出，在持续的量化反馈指导下，试验组住院医师在查房发言中的专业性和逻辑性得到了逐步且显著的改善。

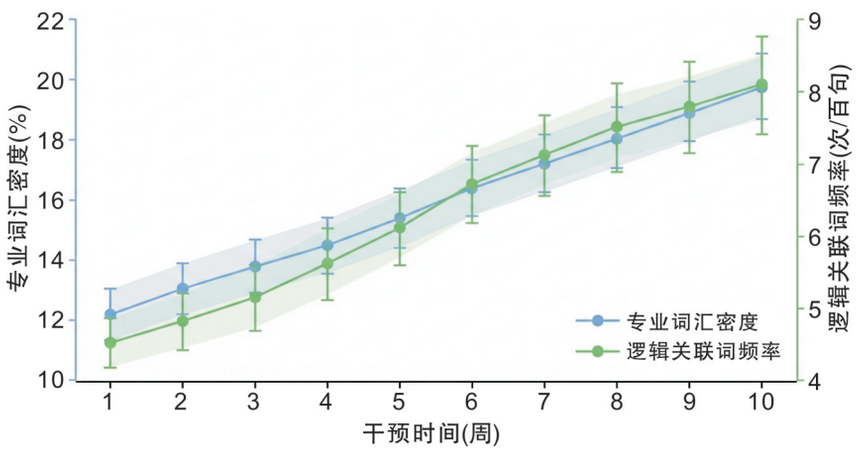


图4 试验组关键NLP量化指标随时间变化趋势折线图

2.4 NLP量化指标与临床能力评估的相关性

为了验证NLP量化模型的效度，对所有研究对象在终末评估时的NLP指标与Mini-CEX各分项得分进行了相关性分析，结果如表5所示。分析发现，NLP模型的核心指标与Mini-CEX评估结果存在显著的正相关关系。其中，“逻辑关联词频率”与Mini-CEX的“临床判断”分项得分呈现出强度正相关（$r=0.68$，$P<0.001$），同时“鉴别诊断数量”也与“临床判断”中度正相关（$r=0.55$，$P<0.001$）。这提示住院医师发言的逻辑性是其临床判断能力的重要外在体现。此外，模型提取的“人文关怀词汇频率”与Mini-CEX的“人文关怀”分项得分高度相关（$r=0.72$，$P<0.001$），直接证明了模型在该维度评估上的有效性。最后，代表综合能力的“整体表现”得分与“逻辑关联词频率”（$r=0.61$，$P<0.001$）、“专业词汇密度”（$r=0.51$，$P=0.003$）及“鉴别诊断数量”（$r=0.48$，$P=0.006$）等多个NLP指标均显著相关。

表5 终末NLP量化指标与Mini-CEX各维度得分的相关性矩阵

语言特征	Mini-CEX病史采集	Mini-CEX临床判断	Mini-CEX人文关怀	Mini-CEX整体表现
总发言时长	0.28（$P=0.15$）	0.35（$P=0.06$）	0.19（$P=0.29$）	0.42（$P=0.021$）*
专业词汇密度	0.38（$P=0.04$）	0.49（$P=0.005$）*	0.11（$P=0.56$）	0.51（$P=0.003$）*
逻辑关联词频率	0.41（$P=0.025$）	0.68（$P<0.001$）*	0.22（$P=0.21$）	0.61（$P<0.001$）*
鉴别诊断数量	0.25（$P=0.18$）	0.55（$P<0.001$）*	−0.08（$P=0.68$）	0.48（$P=0.006$）*
人文关怀词汇频率	0.31（$P=0.11$）	0.15（$P=0.41$）	0.72（$P<0.001$）*	0.33（$P=0.08$）

注：*为两者具有正相关关系。

2.5 半结构化访谈质性结果概述

对访谈资料进行主题分析后发现，质性结果为上述量化数据提供了有力的支持与解释。试验组住院医师普遍认为，量化评估报告（图5）具有诊断性价值，使其首次能够客观、清晰地看见自己在查房发言中的优势与短板。一位住院医师典型地提到：“报告让我意识到我总是在描述病情，但很少提出分析和鉴别诊断，这促使我查房前会多想一步。”这种具体、可量化的反馈被认为比传统口头点评更具针对性和可操作性。带教医师则反馈，量化报告为教学提供了客观依据和抓手，帮助他们发现了以往主观印象中容易忽略的问题（如某位学生发言积极但逻辑混乱），从而使一对一反馈的深度和效率得到显著提升。相比之下，对照组住院医师则表达了对自身进步评估依据模糊的困惑，并对量化反馈模式表现出浓厚的兴趣和期待。这些质性发现从主观体验层面，印证了量化反馈模式

在激发反思、促进精准教学方面的积极作用。

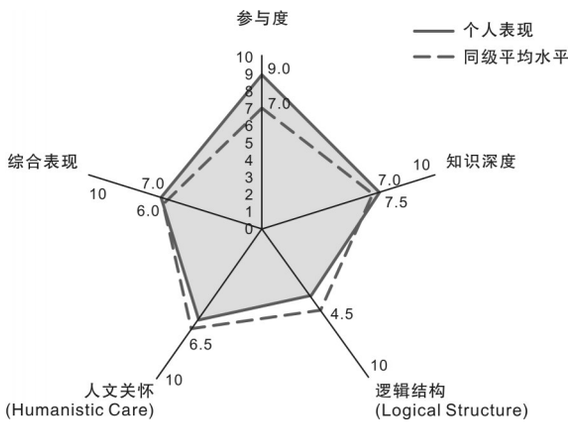


图5　量化评估报告示例雷达图

3　讨论

本研究提供了一项前瞻性的证据，证实了基于NLP的量化评估与反馈模式在提升骨科住院医师临床思维能力方面的有效性。研究结果明确显示，与接受传统反馈的对照组相比，接受新型量化反馈的试验组住院医师在Mini-CEX所衡量的临床思维综合能力上取得了更为显著的进步。同时，该模式还能有效提升住院医师的自信心与自我效能感，并使其综合表现获得更广泛的认可。本研究不仅验证了研究假设，也为数据驱动的临床教学改革提供了新的思路与实证支持。

3.1　主要发现

本研究的核心发现与“形成性评价”及“反思性实践”等教育理论紧密契合。其内在机制在于，量化反馈将隐性的思维过程显性化，从而激发了学习者的元认知。本研究通过量化反馈机制实现了住院医师临床推理过程的可视化，使其从隐性状态转化为可观察、可测量的显性指标，进而触发元认知监控并促进学习者的反思性学习状态。这一机制似乎与当前的人工智能技术在医学文本处理领域的最新进展形成跨领域呼应且相关。具体而言，如Saban等[22]利用自然语言处理技术从自由文本电子病历中提取出与糖尿病并发症有关的医疗信息，并展示了算法如何将非结构化的叙述性记录转化为结构化量化数据，其穿透文本、捕捉隐性临床判断的能力，与本研究中量化评估工具作为“认知透镜”的功能具有一定的相似性质。在此基础上，Wollek等[23]同样基于语言模型的放射报告自动标注系统进一步印证了思维显性化的价值，该团队研究通过将描述性语言转化为标准化标签实现诊断思维的可视化，揭示了当模糊的感觉被转化为清晰的数据时，专业人员的自我监控与自我调节能力便获得切实抓手。不仅如此，Zaidat等[24]揭示了量化反馈的另一重要维度，即减少主观变异、提升标准化水平，其利用大型语言模型将复杂手术记录转化为精确计费代码的实践，该过程和临床教学的实践意义进一步提示了量化反馈同样能够为医学教育提供客观、公正的进步参照。这些研究虽聚焦人工智能临床应用，但其核心机制仍多以隐性信息显性化、模糊描述量化、主观判断标准化为主要的功能设计，共同支持了量化反馈能够通过解构并重构专业人员的思维过程，不仅可以提升信息处理的效率，更关键的是激发使用者对自身认知活动的觉察与反思，最终促使学习者从无意识的能力向有意识的能力跃迁，从而为其终身专业发展奠定元认知的基础。传统口头反馈常因其模糊性而难以触及深层问题，而本研究中的量化报告通过多维度雷达图，将抽象的表现好坏转化为具体的、可视化的数据。如质性访谈中所述，住院医师首次得以客观地首次看见自己在知识深度、逻辑结构等方面的具体短板。这种清晰的自我认知是驱动行为改变的先决条件，促使他们从被动接受信息转向主动思考“如何改进”，从而实现了从“描述病情”到“分析与鉴别诊断”的认知飞跃，这与Ba、Jia、Liu、Luo及O’Sullivan等[25-28]的研究核心发现相一致，即结构化评估工具能够通过促进自我认知来驱动更为深度的学习。当学习者能够清晰识别自身与预期标准之间的差距时，便会产生内在动机去弥补这一差距，从而实现认知水平和实践能力的跃升。

同时，这种数据驱动的反馈也促进了刻意练习。带教医师反馈，量化报告为教学提供了客观抓手，能够发现主观观察中易被忽略的模式（如发言积极但逻辑混乱）。这使得每周的一对一面谈能够聚焦于具体问题，提出如“尝试使用更多逻辑关联词”或“在汇报时有意识地构建鉴别诊断”等可执行的建议，使住院医师的日常练习更具目标性[29]。试验组NLP指标随时间稳步提升的趋势正是这一精准干预效果的客观体现。数据驱动的反馈提升了教学指导的精准性与可操作性。将本研究置于更广阔的学术背景中，其价值在于将NLP技术应用于真实、动态、交互性强的教学查房场景。以往相关研究多集中于对模拟训练或书面病历分析等静态文本的分析，而教学查房的口语对话具有即时性、非结构化和高度情境依赖的特点，对其进行量化分析面临更大挑战。本研究的实践表明，通过分析住院医师在真实临床互动中的语言学特征，可以有效、客观地捕捉到临床思维过程的关键特征，这为现有文献提供了重要补充。此外，相关性分

析结果显示,模型的多个关键指标,特别是“逻辑关联词频率”和“鉴别诊断数量”,与公认的临床能力评估工具Mini-CEX中的临床判断分项高度相关,这为NLP模型作为形成性评价工具的效度提供了证据支持。

3.2 次要目标

本研究的次要结局指标进一步揭示了量化反馈模式的综合效应。试验组在自信心与自我效能感上的显著提升,可能源于反馈的赋能效应。当学习者明确知道自己的不足以及如何改进时,不确定性带来的焦虑感会降低,对自身能力的掌控感则会增强[30-32]。研究发现,当学习者通过量化反馈明确知晓自身不足及具体改进路径时,不确定性带来的焦虑感得到了显著降低,而对自身能力的掌控感则明显增强。这一心理机制与既往医学教育研究中关于结构化支持促进学习者自我效能形成的结论高度契合。例如,Ren等[30]通过探索监督教学门诊对妇产科教育的影响,进而揭示了从旁观到主动参与的渐进式过渡如何通过明确的角色定位与即时反馈降低学习焦虑、增强职业认同,其强调的渐进式责任转移与本研究中“明确不足-定向改进”的反馈逻辑相一致。此外,在Ren的研究基础上,Sui等[31]的团队将情景模拟与Mini-CEX相结合应用于妇产科住院医师教学,发现结构化的情景演练为学习者提供了安全的试错空间,而Mini-CEX的即时、具体反馈则使其清晰知晓错在何处与如何修正,双重机制共同作用显著提升了学习者对复杂临床情境的应对信心。相较而言,Urbina等[32]则是从评估工具验证的角度证实了标准化、可靠的反馈机制是降低模拟环境焦虑、促进沉浸学习的关键要素,其强调的评估维度具体化、学习目标可达成与本研究中量化反馈降低不确定性、增强掌控感的核心发现相互印证。对比分析上述的研究虽聚焦了不同的教学场景,但共同指向了当学习支持系统能够提供清晰的现状定位与明确的改进路径时,学习者的心理资源便从应对焦虑转向投入学习,其自我效能感与自我调节能力随之提升的教学原理。而本研究在具体的骨科实习临床教学的情境中验证并拓展了这一原理,为优化住院医师培训中的反馈设计提供了循证依据。与此形成对比,理论知识测试成绩的组间差异不显著,这一结果符合预期,也恰恰凸显了本干预措施的特异性。本研究的核心干预并非知识灌输,而是针对临床思维这一高阶认知技能的训练,即如何“应用”而非“记忆”知识[33-34]。此外,360度评估结果显示,试验组的进步获得了包括护士、同事在内的多方认可,这表明通过该模式培养的能力提升是真实且可外显的,能够转化为更优的团队协作与临床沟通表现。

3.3 研究创新与局限

本研究的创新之处在于,将前沿的人工智能技术成功应用于动态、交互式的教学查房场景,实现了对临床思维这一核心能力的非侵入式、客观化和纵向追踪评估,突破了传统评估方法的主观性与局限性。本研究的优势在于采用了前瞻性对照设计,并结合了定量与定性(混合方法)的研究方式,使得研究结论既有客观数据的支持,又有来自参与者主观体验的深度解释。

然而,为客观审视本研究的价值,其存在的局限性亦不容忽视。首先,本研究为单中心、非随机对照设计。尽管我们对两组的基线资料进行了详尽比较且未发现统计学差异,从而在一定程度上降低了选择偏倚的风险,但非随机分配仍无法完全排除未知混杂因素的干扰,这可能影响结论在外推至其他医疗中心或专业时的普适性。其次,60例的样本量作为单中心探索性研究相对有限。虽然主要结局指标的差异达到了高度统计学意义(均$P<0.001$),提示干预效果较为稳健,但较小的样本量仍可能限制结论的外部有效性。再者,试验组接受的一对一反馈使其获得了更多关注,可能存在霍桑效应。然而,这种由量化报告驱动的精准关注本身正是本教学模式的核心优势之一,它将传统的、泛化的关注转变为有据可依、聚焦于具体学习目标的结构化指导。尽管如此,我们承认无法完全剥离纯粹被关注所带来的效应。未来的研究可设计更为复杂的对照组(如提供同等时长的非结构化常规反馈),以进一步分离技术工具与教学互动各自的效应。最后,本研究主要关注为期3个月的短期教学效果。这一时长与典型的临床轮转周期相符,但量化反馈模式对住院医师远期临床执业能力、职业发展乃至对患者的最终临床结局所产生的长期影响仍是未知数。开展纵向追踪研究对于全面评价本教学模式的价值至关重要。

此外,本研究面临干预措施特异性的问题。试验组同时接受了量化报告和结构化一对一反馈,而对照组仅接受常规口头反馈。因此,结果差异可能源于量化数据本身、结构化反馈形式,或两者协同作用,我们无法完全分离各要素的独立贡献。本研究评估的实质上是基于NLP的量化反馈模式作为整体教学策略的有效性,而非单一技术要素的效应。质性访谈提示量化数据确实具有独特价值(如“第一次清楚看到自己的问题”),但未来研究应采用多臂对照设计(如增

设“无量化＋结构化反馈”组）来分离各要素效应，深入理解作用机制。

4　结论

综上所述，本研究证实了基于自然语言处理的教学查房量化评估与反馈模式是一种有效提升骨科住院医师临床思维能力、自信心及综合表现的教学策略。该模式通过将隐性思维显性化和提供精准、可操作的反馈，成功地将临床教学从依赖主观经验的艺术向数据驱动的科学转变。未来的研究应致力于开展多中心、大样本的随机对照试验，以增强证据等级。同时，可进一步优化NLP模型，融合更深度的语义分析技术，并探索该模式在不同临床专科及不同年资学习者中的适用性与可迁移性，最终为构建智能化、个性化的临床医学教育体系贡献力量。

参考文献

[1] 陈燕，朱唯一，王维，等. 基于CiteSpace的国内外护理教学查房研究热点与趋势分析[J]. 上海护理，2025，25(11)：39-43.

[2] 阮莉莉，叶彩凤，曾文娟，等. 医护一体化教学查房对基层医务人员理论与操作技能的影响研究[J]. 现代医院，2025，25(8)：1309-1312.

[3] 董奕君，侯玮，尚进. CBL与PAL结合模式在肾内科住院医师规范化培训中的应用探索[J]. 中国病案，2025，26(8)：91-94.

[4] 刘博，陈歆，彭平，等. MDT结合Mini-CEX评分量表在口腔肿瘤临床实习中的教学研究[J]. 临床口腔医学杂志，2025，41(2)：106-109.

[5] Chen F，Li H，Liang X，et al. Use of the Delphi method in the construction of the temperature chain management scheme for da vinci robot-assisted urological surgical patients[J]. Therapeutic Hypothermia and Temperature Management，2025，15(4)：174-183.

[6] Johnson B，Bennett C，Carrier J，et al. Introducing multidisciplinary ward rounds in Malawi：A best practice implementation project[J]. JBI Evidence Implementation，2025，23(S1)：S35-S43.

[7] Kates J，Brown C，Campolieto J，et al. Health professions students’ reflections about principles of interprofessional collaboration after shadowing interprofessional palliative care rounds[J]. The American Journal of Hospice & Palliative Care，2025，42(11)：1102-1108.

[8] Mastalerz K A，Jordan S R，Connors S C. A qualitative study of patient and interprofessional healthcare team member experiences of bedside interdisciplinary rounds at a VA：Language，teamwork，and trust[J]. Journal of General Internal Medicine，2025，40(3)：538-546.

[9] Moro F，Chiarini V，Scquizzato T，et al. Variables for reporting studies on extended-focused assessment with sonography for trauma (E-FAST)：An international Delphi consensus study[J]. Injury，2025，56(1)：111931.

[10] Quade A，Jarrett E S，Rizvi R，et al. Impact of a novel virtual rounding queue software on nurse and family presence for rounds in the neonatal intensive care unit：A pilot study[J]. Academic Pediatrics，2025，25(2)：102584.

[11] Sharma S，Huang R S，Malik A，et al. Insights from North American radiology grand rounds：Determining patterns of gender bias in professional introductions[J]. Current Problems in Diagnostic Radiology，2025，54(1)：81-86.

[12] Song A，Safdieh J E，Robbins M S. Group text messaging as a residency teaching tool in outpatient neurology and headache：A mixed-methods observational study[J]. Headache，2025，65(4)：539-544.

[13] Conway A，Li J，Rad M G，et al. Automating sedation state assessments using natural language processing[J]. Journal of Nursing Scholarship，2025，57(1)：17-27.

[14] Dalal T C，Liang L B，Silva A M，et al. Speech based natural language profile before，during and after the onset of psychosis：A cluster analysis[J]. Acta Psychiatrica Scandinavica，2025，151(3)：332-347.

[15] Demilie M M，Aschale A A，Salau A O，et al. Speaker-based language identification for Ethio-Semitic languages using CRNN and hybrid features[J]. Network：Computation in Neural Systems，2025，36(4)：1313-1335.

[16] Dougherty R F，Clarke P，Atli M，et al. Psilocybin therapy for treatment resistant depression：Prediction of clinical outcome by natural language processing[J]. Psychopharmacology，2025，242(7)：1553-1561.

[17] Goldberg S B，Tanana M，Stewart S H，et al. Automating the assessment of multicultural orientation through machine learning and natural language processing[J]. Psychotherapy，2025，62(3)：292-300.

[18] Graham L A，Illarmo S S，Wren S M，et al. Use of natural language processing method to identify regional anesthesia from clinical notes[J]. Regional Anesthesia and Pain Medicine，2025，50(3)：271-275.

[19] Kaźmierczak I，Jakubowska A，Pietraszkiewicz A，et al. Natural language sentiment as an indicator of depres-

sion and anxiety symptoms: A longitudinal mixed methods study[J]. Cognition & Emotion, 2025, 39(7): 1693-1702.

[20] Li Z B, Zhao H Y, Zhu G H, et al. Classification method of traditional Chinese medicine compound decoction duration based on multi-dimensional feature weighted fusion[J]. Computer Methods in Biomechanics and Biomedical Engineering, 2025, 28(6): 867-881.

[21] Robinson F E, Grimm D, Horning D, et al. An automated content-based measure of closed loop communication among critical care air transport teams[J]. Military Psychology, 2025, 37(2): 168-174.

[22] Saban M, Lutski M, Zucker I, et al. Identifying diabetes related-complications in a real-world free-text electronic medical records in Hebrew using natural language processing techniques[J]. Journal of Diabetes Science and Technology, 2025, 19(4): 999-1007.

[23] Wollek A, Haitzer P, Sedlmeyr T, et al. Language model-based labeling of German thoracic radiology reports[J]. RöFo - Fortschritte Auf Dem Gebiet der Röntgenstrahlen und der Bildgebenden Verfahren, 2025, 197(1): 55-64.

[24] Zaidat B, Lahoti Y S, Yu A, et al. Artificially intelligent billing in spine surgery: An analysis of a large language model[J]. Global Spine Journal, 2025, 15(2): 1113-1120.

[25] Ba H, Xu L, Gu Y, et al. Comparative study of problem-based learning and traditional teaching methods on medical students' outcomes in pediatrics clerkships[J]. Advances in Medical Education and Practice, 2025, 16: 615-624.

[26] Jia F, Wang M, Lü Q, et al. Application of the BOPPPS-CBL model in emergency chest pain management teaching for resident physicians: A randomized comparison[J]. BMC Medical Education, 2025, 25(1): 378.

[27] Liu C, Ren M, Luo C E, et al. Application of mini-CEX evaluation combined with segmented teaching approach in the training of surgical clinical practice skills for resident physicians[J]. Journal of Surgical Education, 2025, 82(5): 103497.

[28] Luo F L, Cai J, Ma H X, et al. Application of a training program system centered on job competency in the standardized training of new nurses[J]. BMC Nursing, 2025, 24(1): 358.

[29] O'Sullivan O E, Daykan Y, Rotem R. Advancing obstetrics and gynecology training: The shift toward outcome and based education, a narrative review of literature[J]. International Journal of Gynaecology and Obstetrics, 2025, 171(3): 1156-1161.

[30] Ren S S, Zhan H, Fernando A, et al. From shadowing to active learning: Exploring the impact of supervised teaching clinics on gynecology education[J]. Frontiers in Medicine, 2024, 11: 1498393.

[31] Sui X X, Cui Z, Li Q X. Application of scenario simulation combined Mini-CEX teaching model in gynecology and obstetrics resident teaching: A case study of paclitaxel and cisplatin use[J]. Pakistan Journal of Pharmaceutical Sciences, 2025, 38(1): 157-164.

[32] Urbina J, Monks S M. Validating assessment tools in simulation[M]. Treasure Island (FL): StatPearls Publishing, 2025.

[33] Xu T, Cui L P, Li C. Effect evaluation of the 7E teaching model based on real clinical cases in nursing students' clinical rounds: A quasi-experimental study[J]. Nurse Education in Practice, 2025, 84: 104316.

[34] Yao Y, Yang G, Li X, et al. An assessment of clinical reception training using standard patient and dental simulator in prosthodontic dentistry for dental undergraduates: A historical control trial[J]. BMC Medical Education, 2024, 24(1): 1537.