대규모 언어모델의 의료관련감염 감시에서 드러나지 않는 실패 양상: NHSN 정의를 활용한 구조화된 평가

Hidden failure modes of large language models in healthcare-associated infection surveillance: a structured evaluation using NHSN definitions

Infect Control Hosp Epidemiol. 2026

원문 PDF: https://static.potados.com/papers-ko/2026-06-04/214904-alzyood-2026-llm-hai-failure-modes-nhsn_original.pdf
번역 생성일: 2026-06-04
※ 기계 보조 번역(자동 검수 1회 포함)입니다. 인용·임상 판단 시 반드시 원문을 확인하십시오. 참고문헌은 원문 그대로 두었습니다.

초록

배경(Background): 대규모 언어모델(large language models, LLM)은 의료관련감염(healthcare-associated infection, HAI) 감시에 점점 더 활용되고 있으나, 공식적인 국가 의료안전 네트워크(National Healthcare Safety Network, NHSN) 정의를 적용할 때의 신뢰성은 충분히 규명되어 있지 않습니다. 본 연구는 NHSN으로 정의된 감염을 분류할 때 GPT-5.1 Thinking의 정확도와 그 판단 근거를 평가합니다.

방법(Methods): 5개 NHSN 감염 유형을 대표하고 복잡한 엣지 케이스를 포함하며, 완전하고 정리된 임상 데이터를 담은 70개의 합성 증례 비네트(case vignette)를 2025년 NHSN 감시 정의를 기준으로 평가했습니다. GPT-5.1 Thinking은 표준(standard), 구조화(structured), 제약(constrained)의 세 가지 프롬프트 전략으로 각 증례를 분류했습니다. 정량적 정확도 지표 분석과 함께, 판단 근거 및 실패 양상에 대한 질적 귀납적 내용분석(qualitative inductive content analysis)을 수행했습니다.

결과(Results): 210건의 분류 전체에 걸친 정확도는 표준 프롬프트 78.6%에서 구조화 프롬프트 88.6%, 제약 프롬프트 95.7%로 향상되었습니다. 명확한 해부학적 또는 영상학적 기준을 가진 감염(수술부위감염[surgical site infection, SSI], 인공호흡기관련폐렴[ventilator-associated pneumonia, VAP])에서 성능이 가장 높았고, 복잡한 제외 규칙이 관여하는 감염(중심정맥관관련혈류감염[central line-associated bloodstream infection, CLABSI], *Clostridioides difficile* 감염[CDI])에서 가장 낮았습니다. 제약형 프롬프트는 NHSN 규칙 준수를 향상시켰으나, 위계적 제외(hierarchical exclusion)에서의 오류를 완전히 제거하지는 못했습니다. 내용분석은 반복적으로 나타나는 세 가지 실패 범주를 확인했습니다: 감시 논리보다 임상적 개연성을 우선하는 경향, 정량적·시간적 역치 적용 실패, 위계적 출처 귀속(hierarchical source attribution)의 오류.

결론(Conclusion): GPT-5.1 Thinking은 엄격한 제약 하에서 감염 감시를 보조할 잠재력을 보이지만, 임상적 직관에 대한 과의존과 복잡한 제외 경로 처리의 어려움을 포함한 체계적 한계를 나타냅니다. 현재로서 LLM은 자율적 NHSN 분류에는 부적합하며, 견고한 인간 감독을 동반한 감독형 의사결정 지원 도구로는 활용될 수 있습니다. 효과적인 HAI 감시에 핵심적인 감시 정의와 복잡한 상황적 특성을 통합하는 LLM의 능력을 향상시키기 위해서는 추가 개발이 필요하며, 완전한 자율적 배치를 위해서는 추가 검증이 요구됩니다. 본 결과는 합성 데이터에 기반한 것으로, 이러한 도구의 정확도를 과대평가하는 방향으로 실제 임상 데이터와 차이가 있을 수 있습니다.

(접수: 2025년 12월 5일; 게재 승인: 2026년 3월 8일; 전자 출판: 2026년 4월 6일)

서론

LLM, 특히 GPT-5.1과 같은 진보된 모델은 HAI 감시를 자동화하기 위해 탐구되고 있습니다.[1,2,3] 연구들은 LLM이 업무 흐름을 간소화하고 데이터 처리 시간을 단축하며 국가 감시 정의 적용의 일관성을 향상시킬 수 있음을 시사합니다.[4,5] 초기 연구들은 CLABSI, CAUTI, SSI, CDI, VAP를 식별하는 데서 유망한 결과를 보고하며 효율성과 정확도 향상의 가능성을 제시했습니다.[4,6,7]

그러나 NHSN의 엄격한 기준 하에서의 성능은 여전히 불분명합니다. 선행 연구들은 흔히 단순화된 증례를 사용하여, 데이터가 불완전하거나 시간적·정량적 역치가 적용되거나 정착(colonization)을 감염과 구분해야 할 때의 모델 거동에 대한 이해를 제한합니다.[8-10] NHSN 기준은 독소 기반 CDI 알고리즘, 집락형성단위(colony-forming unit, CFU) 역치, 기기-일(device-day) 윈도우, 위계적 이차 혈류감염(secondary bloodstream infection, BSI) 귀속과 같은 경직된 규칙에 의존합니다. 상충하는 진단 데이터에 대한 LLM의 반응을 평가한 연구는 거의 없으며,[11] 실제 임상 근거는 NHSN 요구사항의 빈번한 오적용을 보여줍니다. 예를 들어 GPT-4.0은 유망한 민감도를 보고했으나 CLABSI에 대해 제한된 특이도/귀속의 어려움을 보였고 이차 BSI를 잘못 귀속했습니다.[12] 경직성과 근거 없는 추론을 포함한 LLM 추론의 더 광범위한 한계도 지적되었으나,[10] 정의의 정밀성을 요구하는 감시 프레임워크 내에서 체계적으로 연구되지는 않았습니다.

본 연구는 완전하고 정리된 임상 정보를 담은 70개의 합성 비네트를 사용하여 5개 HAI 범주에 걸쳐 세 가지 프롬프트 전략 하에서 GPT-5.1 Thinking의 NHSN 감시 정의 적용을 평가하고, 정량적 정확도와 오분류에 대해 제시된 근거를 함께 평가합니다. 이러한 목표는 안전한 AI 통합을 위해 정밀성(precision), 협력(partnership), 실천(practice), 사람(people)을 강조하는, 감염 감시에서 책임 있는 AI를 위한 4Ps 프레임워크와 부합합니다.[13]

방법

연구 설계

이 횡단면 평가는 다섯 가지 주요 HAI 유형(CLABSI, CAUTI, CDI, SSI, VAP)에 대해 증례 비네트가 공식 NHSN 2025 감시 정의를 충족하는지 판정하는 GPT-5.1 Thinking의 성능을 평가했습니다.[14] 수렴적 혼합방법(convergent mixed methods) 설계가 채택되었습니다.[15] 정량적 분석축은 NHSN 규칙에 엄격히 기반한 사전 정의된 표준(gold-standard) 결과를 사용하여 HAI 범주 및 프롬프트 전략에 걸친 정확도를 평가했습니다. 질적 분석축은 정량적 오류 패턴을 설명하기 위해 오분류된 비네트(n = 26)의 판단 근거에 귀납적 질적 내용분석을 적용하여,[16] 모델의 정당화가 NHSN 규칙에서 벗어난 부분을 코딩했습니다. 코드는 데이터로부터 귀납적으로 도출된 후, 반복적 비교를 통해 범주로 그룹화되었습니다. 코드, 증례 식별자, 할당된 범주의 전체 목록은 보충 파일 1(Supplementary File 1)에 제공됩니다.

증례 비네트 개발

의사결정 지원 시스템 평가를 위한 확립된 방법론에 따라 70개의 시나리오 기반 증례 비네트가 개발되었습니다.[17] 비네트 세트는 NHSN 기준을 완전히 충족하는 양성(positive) 증례 28개와 기준을 충족하지 않는 음성(negative) 증례 42개로 구성되어, 민감도와 특이도 양쪽 모두에 대한 평가를 보장했습니다. 세트에는 CLABSI(n = 17; 양성 6, 음성 11), CAUTI(n = 14; 양성 4, 음성 10), CDI(n = 14; 양성 6, 음성 8), SSI(n = 12; 양성 6, 음성 6), VAP(n = 13; 양성 6, 음성 7)가 포함되었습니다. 첫 50개의 비네트(C1-C50)는 핵심 감시 상황을 반영했고, 나머지 20개(C51-C70)는 위계적 규칙, 기기-일 윈도우, 이차 BSI 귀속, 정량적 역치를 스트레스 테스트하도록 설계된 고난도 엣지 케이스였습니다.

SSI 비네트에는 객관적 소견(화농성 배액, 양성 배양, 영상으로 확인된 농양)을 가진 증례와 비감염성 수술 후 합병증(장액종[seroma], 혈종[hematoma])과의 감별이 필요한 증례가 포함되었습니다. 모든 비네트와 표준 분류는 보충 파일 2(Supplementary File 2)에 제공됩니다. 중요하게도, 모든 비네트는 구조화된 형식으로 제시된 완전하고 잘 정리된 임상

데이터를 담고 있었습니다. 실제 임상 문서는 일반적으로 더 단편적이고 불완전하며 서술적으로 복잡하므로, 이 비네트로부터 얻은 결과는 일상 임상 진료에서 달성 가능한 정확도를 과대평가할 가능성이 높습니다.

표준(Gold-standard) NHSN 분류

표준 분류는 주저자(MA, 감염예방관리 컨설턴트)가 부여하고 공중보건 및 감염 감시 전문성을 가진 두 명의 공저자가 독립적으로 검증했습니다. 불일치는 2025년 NHSN 환자안전요소 매뉴얼(Patient Safety Component Manual, PSCM) 기준을 참조한 합의 논의를 통해 해결되었습니다.[14] NHSN 감시는 엄격한 이분법적 기준을 적용합니다: 필요한 모든 기준이 명확히 충족될 때에만 감염으로 계수됩니다.

평가 대상 LLM

GPT-5.1 Thinking(OpenAI)은 진보된 추론 능력을 갖춘 현세대 모델로 선정되었습니다. OpenAI는 GPT-5.1 Thinking을 더 복잡한 작업/더 지속적인 추론을 위한 것으로 기술하며,[18] 이는 구조화된 분류 작업에 적합합니다. 모든 분류는 2025년 10월에 생성되었습니다.

프롬프트 전략

세 가지 프롬프트 전략이 사용되었습니다. 각 HAI 유형에 대한 2025 NHSN PSCM 정의가 증례 평가 전에 참조 문서로 GPT-5.1 Thinking에 업로드되었습니다. 표준(standard) 프롬프트는 간단한 정당화와 함께 직접적인 Yes/No 분류를 요청했습니다. 구조화(structured) 프롬프트는 단계적 추론을 요구했습니다: 핵심 임상 소견 요약, 충족 및 미충족 기준 식별, 그 후 최종 분류 제시. 제약(constrained) 프롬프트는 엄격한 Yes/No 판정과 함께 명시적 규칙 점검을 강제했습니다. 모든 프롬프트는 서면 판단 근거를 요청했습니다. 전체 프롬프트 텍스트는 보충 파일 2에 제공됩니다.

평가 지표

70개의 각 비네트는 각 프롬프트 전략 하에서 1회씩 제시되어 총 210건의 분류를 생성했습니다(그림 1은 평가 워크플로를 보여줍니다). 성능은 정확도(accuracy), 민감도(sensitivity), 특이도(specificity), 위양성률(false positive rate), 위음성률(false negative rate)을 사용하여 평가되었습니다.[19] Cochran's Q 검정으로 전략 간 차이를 평가했으며, 사후 쌍별 정확 McNemar 검정(exact McNemar test)을 수행했습니다. HAI별 비교는 셀 수가 제한적이어서 기술적으로 요약했습니다.

윤리적 고려

본 연구는 인간 참여자나 실제 환자 데이터를 포함하지 않았습니다. 모든 증례 비네트는 새로 창작되어 전적으로 합성된 것이며 식별 가능한 정보를 포함하지 않았습니다. 따라서 윤리 승인과 사전 동의가 필요하지 않았으며, 본 연구는 윤리 심의 면제로 간주되었습니다.

결과

NHSN 양성 28개와 NHSN 음성 42개로 구성된 70개의 합성 증례 비네트가 평가되었습니다. 비네트 세트는 다섯 가지 NHSN HAI 범주를 다루었습니다: CLABSI(n = 17; 양성 6, 음성 11), CAUTI(n = 14; 양성 4, 음성 10), CDI(n = 14; 양성 6, 음성 8), SSI(n = 12; 양성 6, 음성 6), VAP(n = 13; 양성 6, 음성 7). 각 비네트는 각 프롬프트 전략(표준, 구조화, 제약) 하에서 1회씩 평가되어 정량 분석을 위한 210건의 증례-프롬프트 분류를 산출했습니다(그림 1).

정량적 결과

전체 정확도는 프롬프트 구조에 따라 유의하게 달랐습니다(Cochran's Q(2) = 18.2, P < .001). 제약(95.7%) 및 구조화(88.6%) 프롬프트는 표준(78.6%) 프롬프트보다 높은 정확도를 보였습니다(McNemar 각각 P = .00049, P = .0156). NHSN 적격 HAI 사건 탐지에 대한 민감도는 모든 프롬프트에서 높게 유지되었습니다(96.4-100%). 정확도는 구조화 프롬프트와 제약 프롬프트 사이에서 유의하게 다르지 않았습니다(McNemar P = .0625). 위양성률은 표준 프롬프트의 33.3%에서 제약 프롬프트의 7.1%로 떨어져, NHSN 규칙 역치에 대한 점진적으로 더 엄격해지는 준수를 나타냈습니다.

감염 범주별 정확도

HAI별 분석은 정식 통계 검정에 대해 검정력이 부족하므로, 이러한 차이는 기술적으로 해석되어야 합니다. SSI가 가장 높은 관찰 정확도(97.2%)를 보였고, CAUTI(92.9%)와 VAP(89.7%)가 뒤를 이었습니다. 정확도는 CDI(81.0%)와 CLABSI(80.4%)에서 수치적으로 더 낮았으며, 그 오류들은 위계적 제외 적용과 정착을 감염과 구분하는 데서의 어려움을 반영했습니다(표 1).

HAI 유형	n	정확(Correct)	정확도	FP	FN	주요 오류 패턴
SSI	36	35	97.2%	1	0	비감염성 장액종을 표재성 SSI로 오분류
VAP	39	35	89.7%	4	0	모호한 영상소견으로 VAP를 과진단; 발관 후 시간 윈도우 무시
CAUTI	42	39	92.9%	3	0	CFU 역치(<10^5) 무시; 부적격 혼합균총에도 분류
CDI	42	34	81.0%	7	1	음성 독소 및 대체 원인에도 NAAT 양성을 감염과 동일시
CLABSI	51	41	80.4%	10	0	MBI-LCBI 제외 실패; 이차 BSI를 일차 CLABSI로 오귀속
합계	**210**	**184**	**87.6%**	**25**	**1**	

표 1. HAI 범주별 기술적 정확도. 주: n, 증례-프롬프트 관찰 수(증례 × 3개 프롬프트 전략); FP, 위양성(false positive); FN, 위음성(false negative).

그림 1. 증례 입력, 프롬프트 전략, 분류 출력을 보여주는 단계적 평가 워크플로. 각 비네트는 표준, 구조화, 제약 프롬프트 하에서 평가되었습니다. 출력은 표준 NHSN 2025 분류와 비교되어 정량적 정확도 지표를 생성했습니다. 오분류된 비네트는 귀납적 내용분석의 대상이 되었습니다. HAI, 의료관련감염; NHSN, 국가 의료안전 네트워크; CLABSI, 중심정맥관관련혈류감염; CAUTI, 카테터관련요로감염(catheter-associated urinary tract infection); CDI, *Clostridioides difficile* 감염; SSI, 수술부위감염; VAP, 인공호흡기관련폐렴. [원문 그림 참조]

질적 결과: 실패 양상에 대한 내용분석

오분류된 26개 비네트의 분석은 GPT-5.1 Thinking이 NHSN 기준 적용에 실패하는 반복적 방식을 포착하는 세 가지 범주를 확인했습니다. 추가로 두 개의 증례(C59, C68)는 주요 범주에 들어맞지 않는 오류를 산출했으며, 보충 파일 1에 상세히 기술됩니다.

범주 1: 임상적 개연성이 감시 규칙을 압도

모델은 필수적인 NHSN 기준이 충족되지 않았더라도 임상적으로 개연성이 있을 때 일관되게 감염을 추론했습니다. CDI 증례에서는 양성 핵산증폭검사(nucleic acid amplification test, NAAT) 결과가 독소 요건과 대체 설명을 압도했습니다. Case 58에서 모델은 음성 독소 검사와 고용량 마그네슘 사용이 설사에 대한 명확한 대체 설명을 제공함에도 NAAT 양성을 CDI와 동일시했습니다. 구조화 프롬프트 하에서는 마그네슘 사용을 인정하면서도 여전히 CDI로 분류했습니다. Case 67은 동일한 패턴을 보였습니다: 경장영양(enteral feed) 조정 후 묽은 변이 단 2회만(NHSN 역치인 3회 미만) 기록되었음에도, 표준 및 구조화 프롬프트 하에서 CDI가 부여되었습니다. VAP 증례에서도 모호한 영상소견이 유사하게 무시되었습니다. Case

60에서는 흉부 방사선 사진이 명확한 새로운 경화(consolidation) 없음을 보고했음에도, 모델은 표준 및 구조화 프롬프트에 걸쳐 VAP로 분류했으며, 모호한 소견을 인식하면서도 여전히 감염이 존재한다고 결론지었습니다. 가장 지속적인 오류는 점막장벽손상 검사실확인 혈류감염(mucosal barrier injury laboratory-confirmed bloodstream infection, MBI-LCBI) 제외와 관련되었습니다. Case 61에서는 호중구감소증(절대호중구수 300 cells/μL), 중증 점막염, *Enterococcus faecium* 균혈증이 명시적 MBI-LCBI 기준을 충족했음에도, 모델은 세 가지 프롬프트 전략 모두에서 이를 CLABSI로 분류하여, 관찰된 가장 프롬프트-저항적인 오류를 나타냈습니다.

범주 2: 정량적 역치 및 시간 규칙의 비일관적 적용

모델은 수치적 역치와 시간 윈도우를 고정된 정의 기준이 아니라 유연한 가이드라인으로 취급했습니다. CAUTI의 경우 Case 57은 소변 배양에서 8×10^4 CFU/mL(NHSN 최소 기준인 10^5 CFU/mL 미만)가 자랐음에도, 표준 프롬프트 하에서 CAUTI가 부여되었습니다. Case 65는 동일한 거동을 보여, 소변 배양 4×10^4 CFU/mL가 CAUTI로 분류되었으며, 이는 세균뇨·카테터·증상의 어떤 조합이든 정량적 역치와 무관하게 CAUTI와 동일시하는 패턴을 반영합니다. CLABSI의 경우 Case 64는 단일 양성 혈액배양 병에서 coagulase-negative staphylococci(NHSN 규칙상 두 건의 양성 배양이 필요한 알려진 피부 오염균)가 자란 사례였는데, 표준 프롬프트는 단일 병임에도 이를 CLABSI로 분류했습니다. 시간 규칙도 유사하게 문제가 되었습니다. Case 63에서는 균혈증 발생 3일 전에 중심정맥관이 제거되어 2일 제거 후 윈도우를 초과했는데, 모델은 표준 프롬프트 하에서 제거 시점을 간과했고, 구조화 프롬프트 하에서는 시간 순서를 기술하면서도 여전히 CLABSI로 결론지었습니다. Case 69에서는 발관 3일 후 폐렴이 발생하여 VAP의 2일 인공호흡기 윈도우를 벗어났는데, 모델은 시간 규칙을 적용하지 않고 임상적 근거로 이를 VAP로 분류했으며, 구조화 프롬프트는 발관 시점을 기술하면서도 VAP 분류를 유지했습니다. 이러한 오류들은 모델이 관련 시간적·정량적 정보를 인식하지만 이를 필수 제약으로 적용하지 않음을 시사합니다. 이 범주의 모든 오류는 제약 프롬프트에 의해 해결되었습니다.

범주 3: 위계적 귀속 오류

모델은 다른 곳에서 일치하는 균이 식별될 때 균혈증을 일차 감염 부위와 연결하도록 요구하는 NHSN의 이차 BSI 귀속 규칙을 적용하는 데 어려움을 겪었습니다. Case 62에서는 엽성 경화(lobar consolidation)가 있는 환자가 혈액 및 객담 배양 모두에서 *Streptococcus pneumoniae*를 보였습니다. NHSN 규칙은 호흡기 출처로의 귀속을 이차 BSI로 요구하며 CLABSI를 제외하지만, 모델은 중심정맥관이 존재하는 어떤 균혈증이든 정맥관 감염을 나타낸다는 휴리스틱을 적용하여 표준 프롬프트 하에서 이를 CLABSI로 분류했습니다. Case 51은 동일한 실패를 보였습니다: 복부 통증과 복강 내 출처로 귀속 가능한 영상 소견을 동반한 *Klebsiella pneumoniae* 균혈증이 모든 프롬프트에 걸쳐 CLABSI로 오분류되었습니다. Case 66은 관련 오류를 예시했습니다: 소변 배양에서 혼합균총(NHSN상 부적격)이 자라고 새로운 폐 침윤이 일차 출처로서 폐렴을 시사함에도, 농뇨(pyuria)와 세균뇨를 근거로 CAUTI가 부여되었습니다. MBI-LCBI 제외 경로는 가장 지속적인 실패를 보였으며, Case 61은 이미 범주 1에서 기술되었습니다. 이러한 결과들은 위계적 출처 귀속 논리가 검정된 프롬프트 전략 하에서 이 모델의 신뢰 가능한 능력 범위를 여전히 벗어남을 나타냅니다.

고찰

이 평가는 NHSN 감시에서 GPT-5.1 Thinking의 성능이 기저 정의의 복잡성에 따라 달라졌음을 시사합니다. 모델은 단순한 기준이 적용되는 HAI(SSI, VAP)에서 더 높은 정확도를 달성한 반면, 복잡한 위계적 출처 귀속이나 엄격한 시간적 역치를 요구하는 정의(CDI, CLABSI)에서는 신뢰성이 떨어졌습니다. 선행 연구와 일관되게, 이러한 결과는 LLM이 HAI를 과진단하는 경향이 있음을 나타냅니다: 지나치게 민감하고 충분히 특이적이지 않습니다. 이 패턴은 직접적인 임상적 함의를 갖는데, 인위적으로 부풀려진 HAI 발생률은 품질과 규제 양쪽 관점에서 병원에 매우 바람직하지 않기 때문입니다.

이러한 관찰은 더 큰 해석적 판단을 요구하는 감시 정의가 본질적으로 일관되게 적용하기 더 어렵다는 근거와 일치합니다.[20] 2024년 체계적 문헌고찰은 자연어처리(NLP) 모델이 폐렴 탐지에서 최대 99% 특이도를 달성한 반면, GPT-4는 BSI에 대해 단 59%의 진단 정확도만을 보였다고 보고했습니다.[21] 유사한 패턴이 LLM 성능이 작업 복잡성에 따라 달라짐을 보여주는 연구들에서 나타납니다.[22] 우리의 결과는 정확한 임상적 해석이, 분류가 역치, 시간 윈도우 또는 위계적 제외에 의존할 때 NHSN 규칙의 올바른 적용으로 이어지지 않음을 보여줍니다.

우리의 결과는 최근의 실제 환경 평가와 대비하여 맥락화될 필요가 있습니다. Rodriguez-Nava 등은 LLM이 Stanford의 임상 기록을 사용하여 높은 민감도로 CLABSI를 신속하게 평가할 수 있다고 보고하며, LLM을 인간 검토를 대체하기보다 보조하는 유망한 도구로 보았습니다.[12] Morgan 등은 일부 경계 증례는 어느 쪽으로든 합당하게 분류될 수 있어, 주관적 요소를 가진 감염에 대해 확정적 판정이 덜 명확함을 지적했습니다.[3] 우리 결과와의 외견상 불일치를 설명할 수 있는 몇 가지 방법론적 요인이 있습니다. 첫째, 우리 평가는 완전한 정보를 담은 합성 비네트를 사용한 반면, 실제 기록은 역설적으로 분류를 단순화하는 모호성을 포함할 수 있습니다. 둘째, 우리는 표준화된 프롬프트로 단일 모델을 검정한 반면, 기관별 구현은 미세조정된(fine-tuned) 접근을 포함할 수 있습니다. 셋째, 엣지 케이스에 대한 우리의 초점은 일상적 감시 모집단에서는 덜 빈번하게 발생할 수 있는 실패 양상을 의도적으로 스트레스 테스트했습니다.

CDI와 CLABSI 분류의 오류는 다단계 NHSN 논리 적용의 어려움을 시사합니다. 그러나 일부 NHSN 기준, 특히 CLABSI에서의 출처 귀속은 인간과 자동화된 판정 양쪽 모두에 도전이 되는 내재적 주관성을 포함합니다.[3] 미생물학적 신호가 존재할 때 감염을 과진단하는 지속적 경향이, 위계적 귀속과 시간적 논리에서의 실패와 결합되어, 감독되지 않은 배치가 HAI 발생률을 부풀릴 위험을 시사합니다.

실무 및 연구에 대한 함의

현재 GPT-5.1 Thinking은 자율적 HAI 감시 도구로 기능할 수 없습니다. 그러나 높은 민감도는 검토가 필요한 증례를 식별하는 데 잠재적 유용성을 나타내며, 감염예방 전문가(infection preventionist)는 최종 결정에 대한 책임을 유지합니다. 이는 4Ps 프레임워크의 협력(Partnership) 축과 부합합니다.[13] 세 가지 실무적 함의가 도출됩니다. 첫째, 제약 프롬프트로 정확도가 78.6%에서 95.7%로 증가했다는 점을 고려하면, 조직은 감시 규칙에 부합하는 프롬프트로부터 이익을 얻을 수 있습니다. 둘째, 감염예방 팀 내에서 AI 리터러시를 개발하면 직원들이 모델 출력을 효과적으로 해석하는 데 도움이 될 수 있습니다. 셋째, 거버넌스 프로세스는 성능 표류(performance drift)를 식별하기 위해 주기적 감사를 포함해야 합니다. 향후 연구는 이러한 오류 패턴이 서로 다른 LLM 아키텍처에 걸쳐 나타나는지 검토하고, 언어모델과 결정론적 규칙 엔진(deterministic rule engine)을 결합한 하이브리드 접근을 탐구해야 합니다.

강점과 한계

이는 2025 PSCM을 사용하여 NHSN 감시 정의의 전 범위에 걸쳐 GPT-5.1 Thinking을 평가한 최초의 체계적 평가입니다. 70개의 비네트는 이차 BSI 귀속, MBI-LCBI 제외, 정량적 역치를 포함한 정의의 미묘한 차이를 포착했습니다. 세 가지 프롬프트 전략을 평가함으로써 프롬프트 설계에 따른 성능 변화를 드러냈습니다.

한계로는 실제 환경 문서의 복잡성을 결여한 표준화된 비네트의 사용이 있습니다. 우리는 단일 시점에서 하나의 모델을 평가했으며, 결과가 다른 LLM이나 향후 버전으로 일반화되지 않을 수 있습니다. 모델은 실시간 전자건강기록(electronic health record)에 연결되지 않았으며, 자동화된 데이터 추출의 어려움은 평가되지 않았습니다. NHSN 정의가 진화함에 따라, 우리의 결과는 2025 PSCM 기준을 반영합니다. 더 나아가, 특정 HAI 유형을 탐지하는 능력은 임상 비네트에 제공된 정보의 완전하고 정리된 특성에 영향을 받을 수 있습니다; 실제 환경에서의 탐지는, 일반적으로 더 많은 정보가 누락되거나 모호하기 때문에, 더 낮을 가능성이 높습니다. 따라서 본 연구에서 제시된 정확도 추정치는 이러한 도구가 실무에서 달성할 수 있는

것의 상한선을 나타냅니다. 종합하면, 이러한 결과는 이러한 종류의 LLM이 인간 감독을 요구하며 HAI 감시에서 완전한 자율적 배치에 아직 준비되지 않았음을 시사합니다.

결론

GPT-5.1 Thinking은 NHSN 정의에 긴밀하게 부합하는 제약 프롬프트로 향상된 성능을 보였습니다. 분류는 명확하고 관찰 가능한 기준으로 정의된 감염에 대해 더 신뢰할 수 있었고, 복잡한 제외 규칙, 시간 윈도우, 위계적 출처 귀속에 의존하는 감염에 대해서는 덜 신뢰할 수 있었습니다. 수치적 역치와 출처 귀속 적용의 지속적 어려움은 이러한 약점이 단순한 프롬프트 설계 문제를 넘어 현재 LLM 추론의 한계를 반영할 수 있음을 시사합니다. 우리의 결과는 이러한 프롬프트 전략을 사용한 GPT-5.1 Thinking의 완전한 자율적 NHSN 분류가 감독 없는 배치 이전에 해결이 필요한 한계를 보였음을 나타냅니다. 서로 다른 모델, 프롬프트 접근, 또는 검증된 상용 제품이 수용 가능한 자율적 성능을 달성할 수 있을지는 여전히 열린 질문으로 남아 있습니다.

보충 자료 및 선언

보충 자료(Supplementary material). 본 논문의 보충 자료는 https://doi.org/10.1017/ice.2026.10444 에서 찾을 수 있습니다.

감사의 글(Acknowledgments). 본 연구 전반에 걸쳐 필수 자원과 지원에 대한 접근을 제공해 준 Oxford Brookes University에 감사드립니다. 또한 연구 중 통찰이나 도움을 제공해 준 모든 동료 및 직원에게 감사를 표합니다.

재정 지원(Financial support). 본 연구는 공공, 상업, 비영리 부문의 어떠한 기금 기관으로부터도 특정 지원금을 받지 않았습니다.

이해상충(Competing interests). 저자(들)는 이해상충이 없음을 선언합니다.

윤리 기준(Ethical standard). 해당 없음.

참고문헌 (원문)

1. Marra AR, Langford BJ, Nori P, Bearman G. Revolutionizing antimicrobial stewardship, infection prevention, and public health with artificial intelligence: The middle path. Antimicrob Steward Healthc Epidemiol 2023;3:e219.
2. Arnold A, McLellan S, Stokes JM. How AI can help us beat AMR. NPJ Antimicrob Resist 2025;3:18.
3. Morgan DJ, Goodman KE, Branch-Elliman W, et al. Using generative AI for healthcare-associated infection surveillance. Clin Infect Dis 2026;82:473-479.
4. Wiemken TL, Carrico RM. Assisting the infection preventionist: Use of artificial intelligence for health care-associated infection surveillance. Am J Infect Control 2024;52:625-629.
5. Mendes VIS, Mendes BMF, Moura RP, et al. Harnessing artificial intelligence for enhanced public health surveillance: A narrative review. Front Public Health 2025;13:1601151.
6. Antonie NI, Gheorghe G, Ionescu VA, Tiucă L-C, Diaconu CC. The role of ChatGPT and AI chatbots in optimizing antibiotic therapy: A comprehensive narrative review. Antibiotics (Basel) 2025;14:60.
7. He Y, Huang J, Li N, Zhou G, Liu J. Artificial intelligence-driven prediction and interpretation of central line-associated bloodstream infections in ICU: Insights from the MIMIC-IV database. Front Public Health 13:2025;1675077.
8. Ge J, Lai JC. Artificial intelligence-based text generators in hepatology: ChatGPT is just the beginning. Hepatol Commun 2023;7:e0097.
9. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology: A recent scoping review. Diagn Pathol 2024;19:43.

10. Kim J, Podlasek A, Shidara K, Liu F, Alaa A, Bernardo D. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. Sci Rep 2025;15:39426.
11. Shan G, Chen X, Wang C, et al. Comparing diagnostic accuracy of clinical professionals and large language models: Systematic review and meta-analysis. JMIR Med Inform 2025;13:e64963.
12. Rodriguez-Nava G, Egoryan G, Goodman KE, Morgan DJ, Salinas J. Utility of a large language model for identifying central line-associated bloodstream infections (CLABSI) using real clinical data at Stanford Health Care. Open Forum Infect Dis 2025;12:P-408.
13. Alzyood M. Artificial intelligence and the future of healthcare-associated infection surveillance: bridging precision, partnership, practice, and people. Clin Infect Dis 2026;82:480-483.
14. Centers for Disease Control and Prevention. National Healthcare Safety Network (NHSN) Patient Safety Component Manual. Atlanta, GA, 2025. Division of Healthcare Quality Promotion. Centers for Disease Control and Prevention website. https://www.cdc.gov/nhsn/pdfs/validation/2025/pcsmanual_2025.pdf. Accessed February 25, 2026.
15. Creswell JW, Clark VLP. Designing and Conducting Mixed Methods Research, 3rd edition. Thousand Oaks, CA: SAGE Publications; 2017.
16. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62:107-115.
17. Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: A prospective validation study of 3 methods for measuring quality. JAMA 2000;283:1715-1722.
18. OpenAI. GPT-5.1: A smarter, more conversational ChatGPT, 2025. OpenAI website. https://openai.com/index/gpt-5-1/. Published 2025. Accessed October 5, 2025.
19. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.
20. Uwishema O, Ghezzawi M, Charbel N, et al. Diagnostic performance of artificial intelligence for dermatological conditions: A systematic review focused on low- and middle-income countries to address resource constraints and improve access to specialist care. Int J Emerg Med 2025;18:172.
21. Shenoy ES, Branch-Elliman W. Automating surveillance for healthcare-associated infections: Rationale and current realities (Part I/III). Antimicrob Steward Healthc Epidemiol 2023;3:e25.
22. Omar M, Brin D, Glicksberg B, Klang E. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. Am J Infect Control 2024;52:992-1001.
23. Abosi OJ, Kobayashi T, Ross N, et al. A head-to-head comparison of the accuracy of commercially available large language models for infection prevention and control inquiries. Infect Control Hosp Epidemiol 2024;46:1-3.