

인공지능 기반 공유 의사결정 교육 주제범위 문헌고찰과 가상 표준화 환자 설계의 윤리적 함의

김준혁*

연세대학교 치과대학 치의학교육학교실, 부교수

Artificial Intelligence in Shared Decision-Making Education: A Scoping Review with Ethical Implications for Designing Virtual Standardized Patients

Junhewk Kim*

Associate Professor, Department of Dental Education, College of Dentistry, Yonsei University, Seoul, Korea

Received: Aug 4, 2026
Revised: Sep 2, 2026
Accepted: Sep 14, 2026

*Corresponding author

Junhewk Kim
Associate Professor, Department of
Dental Education, College of Dentistry,
Yonsei University, Seoul, Korea
Tel: +82-2-2228-3022
E-mail: MOLE0619@yuhs.ac

Copyright © 2026 Korean Society for Medical Ethics. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Conflict of interests

The author is engaged in the development of an educational virtual standardized patient system of the general kind described in Section IV.3. No specific product is named, evaluated, or endorsed in this review, and no claim is made that such a system has been built, deployed, or shown to be effective. The kind of warrant carried by each design principle is declared in Table 1 so that a reader may judge which of them the reviewed literature supports and which follow from the ethical argument. Junhewk Kim is an editorial board member of this journal. However, the author was not involved in the reviewer

Abstract

Shared decision-making (SDM) operationalizes respect for patient autonomy, yet health professions education often evaluates it solely through self-reported confidence. This scoping review synthesized how artificial intelligence (AI) is used to teach, assess, or support SDM competency among clinical trainees and practitioners. A search of four databases in August 2026 identified 1,008 unique records, with 13 meeting inclusion criteria. AI served as an instructional environment (n=5), an assessment infrastructure (n=6), or catalyst for curricular improvement (n=2). Only one randomized educational trial was identified, and no study demonstrated durable transfer to clinical practice. Eleven sources explicitly targeted patients' values and preferences, yet only one addressed ethical reasoning, and only two of five simulated encounter models articulated clinical options. Current approaches therefore emphasize the informational dimension of consent at the expense of collaborative deliberation. Accordingly, we propose five ethical design principles for virtual standardized patients, formulated as a design specification warranting empirical evaluation.

Keywords: decision making, shared; artificial intelligence; education, medical; ethics, medical; patient simulation

I. 서론

공유 의사결정(shared decision-making, SDM)은 임상 의사와 환자가 의사결정이 필요함을 인식하고, 가능한 선택지와 근거를 함께 검토한 뒤, 환자의 가치와 선호를 탐색하여 이를 반영한 선택에 이르는 협력적 과정이다[1,2]. 실제 진료에선 주로 선택지 검토, 대안 및 장단점 비교, 환자에게 문제가 되는 것들을 결합한 결정 도출로 구성된다[3]. 그러나, SDM을 검토해야 하는 것

selection, editorial evaluation, or decision-making process for this article. Otherwise, no potential conflicts of interest relevant to this article were reported.

Acknowledgements

A GPT-5-based OpenAI agent was used for title and abstract triage during screening in both the initial and the revised search, as described in Section II.4; all triage outputs were treated as decision support, no record was excluded by the agent without subsequent human review of the retained set, and the recall recheck was configured so that it could only promote records. Generative AI tools were also used to assist with language editing and English revision of this manuscript. All eligibility decisions, data charting, verification against source texts, interpretation, and the design principles proposed in Section IV.3 were made and verified by the author, who takes full responsibility for the content.

Funding information

This research was supported by a grant from the Korea Health Technology R&D Project through the Patient-Doctor Shared Decision Making Research Center (PDSDM), funded by the Ministry of Health and Welfare, Republic of Korea (grant number: RS-2023-KH142275).

Data availability

Upon reasonable request, the datasets of this study can be available from the corresponding author.

ORCID

Junhewk Kim
https://orcid.org/0000-0002-9109-270X

Author contributions

The article is prepared by a single author.

Ethics approval

Not applicable.

은 그것이 뛰어난 의사소통 기술이어서라기보다는, 임상에서 환자 자율성 존중을 구현할 수 있는 실천적 방법이기 때문이다. 결정에서 실행되지 않고 원칙의 이름으로만 남는 자율성은 오히려 환자를 방해할 수 있다.

SDM은 환자-의료진 관계 모형의 한 위치를 차지한다. 후견인으로 의료진을 설정하고 환자의 자율성을 무시하는 온정주의, 이를 역전하여 고객으로서 환자와 판매자로서 의료진을 설정하고 전문가적 판단을 선호에 복속시키는 소비자주의는 명확한 한계를 보인다[4]. 환자를 최종 의사결정자로, 의료진을 정보 공개자로 두는 충분한 설명에 의한 동의는 두 모형으로부터 명백한 진전이였다. 그러나, 이를 정보 공개와 동의 획득의 절차로만 이해할 경우, 의료진은 선택지의 의미를 환자와 함께 해석하거나 결정의 형성을 지원할 책임을 회피할 수 있다. 반면 SDM은 양측을 모두 숙고의 참여자로 둔다. 여기에서 양측 모두 응답책임의 주체로서 각각 서로 다른 지식과 책임을 가지고 결정 과정에 기여할 것을 요청받는다.¹⁾

단, SDM이 모든 임상적 의사결정에서 동일한 정도로 요구되는 것은 아니라는 점을 검토할 필요가 있다. Whitney et al.은 적절한 의사결정은 문제의 리스크와 확실성에 의존한다고 주장한 바 있다[5]. 임상적 증거가 단일의 합리적 경로를 제안하는 경우 단순 동의 또는 충분한 설명에 의한 동의로 충분하지만, 그 증거가 선택을 과소결정(underdetermine)하며 결과가 환자의 가치에 달려있는 경우, SDM은 그저 선택이 아니라 윤리적 의무가 된다는 것이다. 따라서, SDM의 본령은 단순히 여러 선택지를 제시하는 것이 아니라, 임상적 평형 또는 선호 민감성이 존재하는 상황에서 환자에게 중요한 고려사항이 무엇인지 함께 해석하는 데 있다. 반면, Beauchamp & Childress가 제시한 충분한 설명에 의한 동의의 분석틀은 문턱 요건인 동의능력과 자발성, 정보 요건인 공개·권고·이해, 그리고 동의 요건인 결정과 승인으로 구성된다[6]. 이 요건들은 동의가 충족해야 할 도덕적 조건이지만, 그 자체가 함께 숙의하는 구체적인 방법을 제시하지는 않는다. 또한, Sandman & Munthe가 지적한 것처럼, ‘공유’ 의사결정은 온정주의의 잔여로부터 환자 선택 일반까지 다양한 실천으로 구현될 수 있다[7]. 그러므로 SDM의 실천을 위해선 특정한 대화 절차나 의사소통 기술을 훈련하는 것을 넘어, 임상적 근거와 환자의 가치, 구체적인 삶의 맥락을 함께 해석하고, 이들 사이의 긴장과 우선순위를 검토하여 결정에 이르는 숙고의 과정을 교육할 필요가 있다. 이러한 숙고는 무엇이 이 결정에서 중요하며, 어떤 이유가 정당한 고려사항이 되는지를 판단하는 과정이라는 점에서 본질적으로 윤리적이다.

그러나 현재 교육에서 환자의 가치를 이끌어내고 해석하는 과정은 명확히 과소평가되고 있으며, 흔히 추가 문진 항목 정도로만 여겨지고 있다. 그러나, 가치 탐색에는 의사결정이 가치 판단을 내포하고 있음에 대한 도덕적 지각력과, 어떤 고려사항이 선택에 정당하게 영향을 미칠 수 있는지에 대한 도덕적 판단력이 요구된다. 또한, 가치 탐색은 인식적 부정의에 취약함을 살필 필요가 있다. 환자의 인구학적 특성이 그 경험의 진술에 대한 신뢰성에 영향을 미치는 경우, 어떤 환자의 가치는 다른 환자의 가치보다 덜 고려할 만한 것이 된다[8].

이 때문에 임상에서 의료진이 존중하거나 침해하는 대상으로 자율성을 이해하는 것보다, 의료진의 실천을 통해 지지되거나 훼손될 수 있는 관계적 개념으로 접근할 필요가 있다[9].

1) 이때 응답책임(answerability)이란 결정 과정에 참여한 각자가 자신의 기여에 대하여 상대방에게 설명하고 응답할 의무를 뜻하며, 잘못된 결과에 대해 사후적으로 부과되는 법적 책임(liability)과는 구별된다. 의료진은 어떤 임상적 근거에서 무엇을 왜 권고하는지를 설명할 책임을 지고, 환자는 자신에게 무엇이 중요한지를 표현하고 그 이해가 정확한지 확인해 줄 책임을 진다. 어느 쪽도 상대의 몫을 대신 수행할 수 없다는 점에서 이 책임은 상호적이며, 아래 IV.3의 셋째 및 넷째 원칙은 이 개념을 학습자와 평가자의 관계에 적용한 것이다.

특히 한국에서는 중요한 의사결정에 가족이 참여하는 것이 일반적이며, 환자의 참여 정도가 곧 바로 환자의 이익으로 환원되지도 않는다[10]. 즉, 자율성을 순전히 개인주의적으로 이해할 경우, 학습자는 실제 진료에서 마주하게 될 가족 관계, 의존, 책임 분배를 제대로 다루기 어렵다.

이것을 숙고 역량이라고 한다면, 개념이나 의사소통 원칙을 설명하는 것만으로는 충분히 교육하기 어려울 것이다. 학습자는 불완전한 임상 근거와 서로 충돌할 수 있는 가치가 제시되는 상황에서 환자의 관점을 이끌어내고, 자신의 판단을 설명하며, 새로운 정보에 따라 대화를 조정하는 과정을 직접 수행해 보아야 한다. 따라서 SDM 교육은 상호작용 상황에서 숙고를 반복하고, 그 수행을 관찰·평가하며, 피드백을 바탕으로 다시 시도할 수 있는 교육 환경을 요청한다.

보건의료 전문직 학습자와 임상의를 대상으로 하는 SDM 교육 프로그램은 빠르게 확산되고 있으며, 일반적으로 개념 교육에 역할극, 사례 토의, 표준화 환자(standardized patient, SP)를 활용한 실습을 결합한다[11,12]. 그러나 이러한 프로그램에 대한 평가는 여전히 이질적이고 단기적인 수준에 머물러 있으며, 검증된 수행평가나 임상 실무의 지속적인 변화보다는 주로 학습자가 스스로 보고한 자신감과 만족도에 의존한다. 그러나, 충분한 설명에 의한 동의 역량을 위한 시뮬레이션 교육의 무작위 대조시험과 같이 수행 중심 설계가 적용된 연구는 해당 기술이 교육 가능하고 측정 가능함을 보인다[13].

객관구조화진료시험(objective structured clinical examination, OSCE)과 임상수행시험(clinical performance examination, CPX)을 비롯한 SP 기반 방법을 활용하면 학습자는 안전하고 표준화된 환경에서 면담을 실시하고, 선택지를 설명하며, 가치를 탐색하고, 의사결정을 지원할 수 있다. 또한, 관찰자와 SP는 이러한 수행을 평가하고 피드백을 제공할 수 있다. 이런 SP는 의사소통 교육 전반에 폭넓게 활용되고 있으나, 다른 시뮬레이션 방법보다 일관되게 우수하다는 근거는 아직 제한적이다[14].

SDM 교육과 관련하여 SP가 활용되고 있으나, 이미 SP가 부여한 전반적인 의사소통 평점만으로는 SDM에 특화된 수행 수준을 추론할 수 없음이 제시된 바 있다[15]. 따라서, SDM에 특화된 평가도구가 필요하다. 또한, 대규모 학습자 집단을 대상으로 여러 스테이션을 운영하면서 SP와 평가자를 동시에 배치하려면 상당한 인력과 기반시설이 필요하다[16]. 이러한 제약 때문에 학습자가 자신의 속도에 맞추어 반복적으로 연습하고 즉각적인 개별 피드백을 받는 교육을 대규모로 제공하는 것은 어렵다. 진술한 바와 같이 SDM 교육이 윤리적 특성을 지닌다고 할 때, 윤리적 성향은 습관화를 통해 형성되며, 따라서, 이런 반복 가능성의 제한은 교육의 문제이기도 하다. 또한, 다양한 연령, 문화적 배경, 취약성의 경험을 안전하고 충실하게 재현하려면 SP 모집과 설계, 교육, SP 보호에 이르는 설계가 검토되어야 하며, 이는 배역을 맡기고 보수를 지급하는 문제를 넘어 참여자에 대한 정의의 문제로 확장된다[17]. 더불어, 고령자 및 복합질환 환자를 대상으로 하는 SDM에는 관점을 조율해야 하는 가족 돌봄제공자와 여러 보건의료 전문가가 함께 참여하는 경우가 많으므로, 단일 임상-환자 면담의 범위를 넘어서는 시나리오 및 스테이션이 요청된다[18]. 즉, SP가 SDM 교육에서 널리 활용되고 있다고 해도, 그 제한점 또한 명확하다.

여기서 이 제약들이 서로 같은 성격을 지니지 않는다는 점을 구분할 필요가 있다. 반복 연습의 횟수, 일정 조정과 비용, 사례의 다양성 확보, 즉각적인 개별 피드백의 제공은 인력과 기반시설의 한계에서 비롯되므로, 원리상 계산 자원으로 대체할 수 있는 종류의 제약이다. 반면 표정·침묵·자세와 같은 체현된 단서, 실제 사람을 마주할 때 발생하는 관계적 긴장은 그렇지 않다. SP로 참여하는 사람에 대한 정의의 문제 역시 가상 환경으로 옮긴다고 사라지지 않

으며, 어떤 환자 집단을 누구의 관점에서 어떻게 표상할 것인가라는 문제로 형태를 바꿀 뿐이다. 따라서 가상 환경은 전자의 제약군을 겨냥한 보완 수단으로 검토될 수 있다.

생성형 인공지능(Artificial Intelligence, AI)이 널리 보급되기 훨씬 전부터 가상 환자²⁾는 이러한 보완 수단 가운데 하나로 제시되어 왔다. 미리 작성된 응답 선택지를 바탕으로 구축된 SDM 가상 환자는 학습자가 진료 상황을 탐색할 수 있도록 했지만, 제한된 응답 선택지 때문에 한계를 노정했다[19]. 그러나, 오늘날 대규모 언어모형에 기반을 둔 가상 SP는 개방형 자연어 대화, 다양한 사례의 생성, 반복 연습, 자동화된 피드백을 가능하게 하며, 이에 따라 면담 및 의사소통 교육에 대한 접근성을 확대할 잠재력을 지닌다[20,21]. 그러나, 가상 환경에서의 한계 또한 명확하다. 감정적·비언어적 단서를 제대로 재현하기 어렵고, 임상 실무로의 학습 전이가 아직 확인되지 않았으며, 평가도구도 표준화되어 있지 않다. 또한, 부정확한 내용 생성, 편향, 개인정보 노출 위험 또한 고려되어야 한다[22]. 이에 보건의료 전문직 교육에서 AI를 윤리적으로 활용하기 위한 지침은 투명성과 감독을 유지하고 무비판적 도입을 경계할 것을 강조해 왔다[23].

따라서 SDM 교육에서 AI의 역할을 가상 SP에만 한정할 필요는 없으며, AI를 통해 진료 전 준비와 가치 탐색을 위한 발판을 제공하고, 학습자와 임상과의 대화를 수집·분석하여 SDM 행동을 평가하며, 이러한 평가를 피드백 및 교육과정 개선과 연계할 수 있다[24,25]. 이에 기존의 한 주제범위 문헌고찰은 SDM에서 AI가 활용되는 양상을 체계적으로 제시하였다[26]. 그러나 해당 문헌은 SDM을 수행해야 하는 전문가의 교육·훈련·평가보다는 임상 실무에서 AI가 SDM을 촉진하는 방식을 탐색하는 데 초점을 두었다. 즉, 전문직 학습의 관점에서 SDM 교육에 관한 문헌과 AI 기반 가상 환자의 문헌을 통합적으로 검토한 연구는 아직 없다. 또한, 실제 AI가 SDM 학습을 지원하기 위해 어떤 기능을 수행해 왔으며, 그 활용이 어떻게 평가되었는지 또한 근거 기반으로 통합되지 않았다.

이에 본 연구는 두 부분으로 구성된다. 첫째, 보건의료 전문직 학습자와 임상의를 대상으로 SDM의 교육, 훈련, 평가 또는 학습 지원을 위해 AI를 활용한 문헌을 대상으로 주제범위 문헌고찰을 수행하였다. 둘째, 이 고찰에서 확인된 공백을 전제로 삼아, SDM을 윤리적 실천으로 지원할 수 있는 가상 SP 기반 시스템이 갖추어야 할 설계 원칙을 규범적으로 논증하였다. 전자는 문헌에서 확인된 것에 관한 기술이고, 후자는 전자에 기초하여 별도의 윤리적 논증을 거친 제안이다.³⁾ 두 부분의 절차는 각각 II.3-II.6과 II.7에 기술하였으며, 다섯 원칙 각각이 문헌 근거와 윤리적 논증 가운데 무엇에 의존하는지는 Table 1에 구분하여 제시하였다.

연구 질문: 보건의료 전문직 학습자와 임상의를 대상으로 SDM을 교육하거나, 훈련하거나, 평가하거나, 그 학습을 지원하기 위해 AI는 어떻게 활용되고 있는가?

2) 본 연구에서 가상 환자(virtual patient)는 컴퓨터로 구현된 환자 사례 일반을 가리키고, 가상 표준화 환자(virtual standardized patient, 이하 가상 SP)는 그 가운데 표준화된 교육과 평가에서 인간 SP가 맡아온 역할을 수행하도록 설계된 것을 가리킨다. 즉 가상 SP는 가상 환자의 하위 범주이며, 이하에서는 이 구분에 따라 용어를 사용한다.

3) 본 고찰의 대상 문헌은 가상 SP 연구에 한정되지 않는다. 연구 질문이 정한 범위는 AI가 보건의료 전문직의 SDM 교육·훈련·평가·학습 지원에 활용된 문헌 전체이며, 가상 SP는 고찰의 대상이 아니라 후반부의 규범적 논증이 향하는 설계 대상이다. 이 때문에 포함 문헌에는 훈련 중재뿐 아니라 임상 상담을 자동으로 평가한 연구가 함께 포함되어 있으며, 이들은 동일한 근거 수준에서 통합되지 않고 III에서 유형별로 구분하여 기술하였다.

Table 1. Proposed ethical design principles for a virtual standardized patient system supporting shared decision-making as ethical practice

ID	Design principle	Evidential and analytical basis	Ethical warrant
R1	Preference-sensitive and staged encounters: scenarios should present at least two clinically defensible options, including non-treatment or deferral where appropriate, and proceed from preference exploration to tailored information, joint deliberation, and decision or deferral.	Only 2 of 5 sources modeling a teaching encounter specified the available options and explicitly allowed the decision to be deferred [27,28]; a third specified them only partially [29]. Deferral was built into the scenario in one of the two [27] and scored as an instrument item rather than designed into the encounter in the other [28]. Staged SDM models appeared in the same two sources, although the preference-first ordering proposed here follows from the ethical analysis rather than directly from the evidence.	SDM is required where evidence does not determine a single choice [5]. Exploring what matters before framing the options allows the patient's values to shape, rather than merely react to, the decision.
R2	Relational and representative encounters: the system should support family members, surrogates, and interprofessional participants, and include diverse patient personas developed through co-design and reviewed for stereotyping.	Family-centered decision-making was an explicit target in 1 of 13 sources. Twelve sources modeled or assumed a clinician-patient dyad or no encounter, and none evaluated a multi-party consultation. The requirement that personas be developed through co-design and reviewed for stereotyping follows from the ethical analysis.	Relational autonomy [9], the importance of family participation in clinical decisions [10], and justice toward represented patient groups and standardized patients [8,17].
R3	Traceable, calibrated, and contestable assessment: scores should be linked to transcript-verified evidence and observable behavior, use instruments with clearly stated provenance, be calibrated against multiple human raters, and permit human review and learner appeal.	One automated scorer credited learners with information volunteered by the simulated patient and scored higher than the human rater in more than two-thirds of encounters [28]. Agreement was more interpretable where human-human reliability was also reported [30]. The included studies also showed that instrument scores cannot be converted uncritically into conventional percentage grades. The learner's entitlement to inspect the evidence and to appeal follows from the ethical analysis.	Veracity: a judgment recorded as the learner's achievement must correspond to what the learner actually did, and a record that does not is a false statement about that learner. Answerability (Section I): whoever fixes such a judgment owes the person judged an account of its grounds, which entails inspectability, appeal, and human review. Validity is a property of the interpretation and use of scores, not of the instrument, so agreement coefficients do not by themselves license the use [51].
R4	Transparent governance and safeguards against deskilling: learners should be informed about data retention, model and prompt provenance, and the origin of feedback; records should remain under institutional control where possible; scaffolding should be withdrawn progressively; human debriefing should be retained; and AI should not finalize a high-stakes judgment about a learner without human confirmation, the requirement being human accountability for the judgment rather than independent human re-derivation of every automated output.	Privacy, explainability, automation bias, and automation complacency were identified as concerns but rarely evaluated directly [27-31]. None of the safeguards prescribed here was itself evaluated in the reviewed sources, so the prescription as a whole follows from the ethical analysis.	Learners are persons whose data and performance are being judged. Human professionals retain responsibility for clinical and educational decisions, and AI use should preserve rather than displace independent judgment [32].
R5	Longitudinal, multi-session learning: the system should enable repeated practice, feedback-informed retry, delayed reassessment, comparison across attempts, and linkage to workplace-based observation. It should supplement rather than replace human standardized patients and supervised clinical experience.	Three of the four training interventions evaluated a single encounter with outcomes measured immediately afterward [27,28,33], and no source demonstrated sustained clinical behavior or patient benefit. A virtual system can reduce the logistical constraints that often limit human standardized-patient education to one or a few sessions. The premise that competence forms through habituation rather than in a single successful performance follows from the ethical analysis.	SDM competence requires repeated practice and habituation rather than a single successful performance. Human encounters remain necessary for embodied, relational, and nonverbal dimensions that virtual environments cannot fully reproduce.

AI: artificial intelligence, SDM: shared decision-making.

II. 연구 방법

1. 연구 설계 및 보고 지침

본 연구는 보건의료 전문직을 대상으로 한 SDM 교육에서 AI가 활용된 방식과 이를 뒷받침하는 근거의 범위를 파악하기 위해 수행한 주제범위 문헌고찰 및 그 결과를 전제로 삼는 규범적 분석을 함께 수행하였다. 연구 질문과 검토 범위는 PCC(population, concept, context) 요소와 데이터베이스 검색 기록을 함께 활용하여 설정하였다. 연구 방법의 전체 틀은 JBI(Joanna Briggs Institute)의 주제범위 문헌고찰 지침을 따랐으며[34], 보고는

PRISMA-ScR(preferred reporting items for systematic reviews and meta-analyses extension for scoping reviews)에 따라 이루어졌다[35]. 단, 문헌 선별과 자료 도표화에선 복수의 연구자가 모든 단계를 독립적으로 중복 수행하는 JBI의 기본 절차를 적용하는 대신, Cochrane Rapid Reviews Methods Group의 축약 절차를 참고하는 한편[36], 제외 문헌의 검토와 추출 자료의 확인을 AI 에이전트의 도움을 받아 단일 연구자가 수행하는 방식으로 수정하여 진행하였다. 이러한 축약이 적용된 지점과 그 한계는 이후 해당 절에 명시하였다.

본 연구의 절차는 세 부분으로 구분된다. 문헌의 확인과 선별 및 도표화(II.3-II.5), 도표화 자료의 기술적 요약과 내용분석(II.6), 그리고 이를 전제로 한 규범적 논증(II.7)이다. 앞의 두 부분은 PRISMA-ScR과 JBI 지침의 적용을 받으나, 마지막 부분은 문헌고찰의 산출이 아니라 별도의 논증이므로 그 절차를 따로 기술하였다.⁴⁾

본 검토는 두 차례에 걸쳐 수행되었다. 최초 검색은 2026년 6월 18일에 시행하였다. 이후 해당 검색식의 구성을 PRESS(peer review of electronic search strategies) 방식에 준하여 동료 검토하였다[37]. 결과에 따라 PCC 요소와 네 개 데이터베이스의 검색식을 수정하고, 2026년 8월 1일에 검색을 다시 시행하였다. 수정 후 검색을 본 검토의 공식 검색으로 삼았으며, 이하에 제시된 모든 검색 건수와 Figure 1의 흐름도는 이 검색 결과를 기준으로 한다.

2. 연구 질문 및 문헌 선정 기준

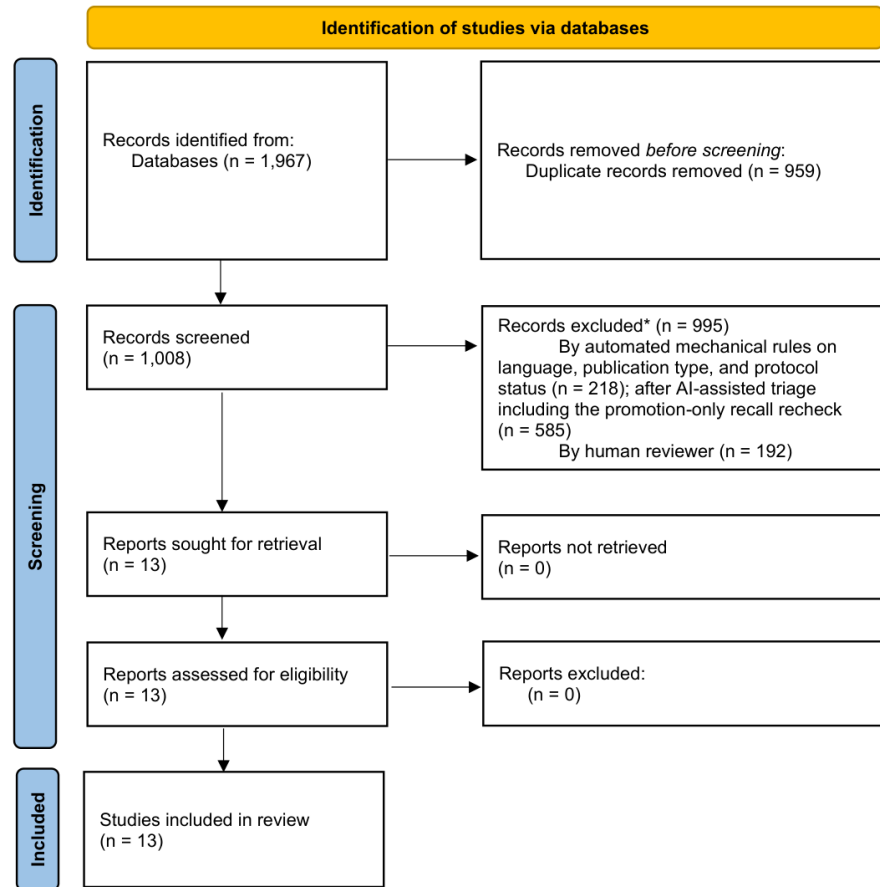
선정 기준은 JBI가 권고하는 PCC 틀에 따라 구성하였다. 수정된 기준에서 population은 학생, 전공의, 펠로우 및 임상 현장에서 활동하는 보건의료인을 포함한 보건의료 분야의 학습자와 임상의로 정의하였다. Concept는 SDM을 교육·훈련하거나, 시뮬레이션·코칭·평가하고, 피드백을 제공하는 등 SDM의 학습과 수행을 지원하기 위한 AI의 활용으로 정의하였다. Context는 학부 및 대학원 교육, 전문직 평생교육, 시뮬레이션 기반 학습 및 임상 현장 기반 학습을 포함한 보건의료 전문직 교육과 전문직 학습으로 정의하였다.

다음 네 조건을 모두 충족하는 것을 적격 문헌으로 선정하였다. 첫째, AI(기계학습, 딥러닝, 대규모 언어모형, 자연어처리, 생성형 AI, 알고리즘 예측 모형, 대화형 에이전트, AI로 구동되는 가상 환자를 포함한다)가 사용되었을 것. 둘째, 대상 주제가 SDM 또는 이에 준하는 환자-임상의 의사결정 의사소통일 것. 셋째, AI가 SDM을 교육·훈련·평가하거나 그 학습을 지원하는 목적에 사용되었을 것. 넷째, 그 대상이 환자나 일반인이 아니라 보건의료 전문직 학습자 또는 임상의일 것. 이에 따라 환자 대상 의사결정 도구와 환자용 챗봇, SDM 교육과 연결되지 않은 임상 예측·진단 모형, AI 요소가 없는 가상 환자 및 시뮬레이션은 제외하였다. 단, 임상에서 의료진의 SDM 수행을 지원하되 교육 목적이 분명하지 않은 문헌은 곧바로 제외하는 대신, 보류로 두고 후속 단계에서 다시 판단하였다.

최초 제목·초록 선별은 재현율을 확보하기 위해 의도적으로 폭넓게 수행하였다. AI가 SDM을 직접 교육·훈련하거나 평가한 문헌은 일차 문헌으로 분류하였다. 명확한 교육 목적 없이 임상에서 의료진의 SDM 수행을 지원한 문헌은 이차 문헌으로 보류하였으며, 정보가 불충분한 문헌은 불명확한 것으로 분류하였다. 한편, 실증연구뿐 아니라 교육 중재 또는 평가도구의 설계·방법 개발 연구와 최종 기준을 직접 충족하는 개념적 문헌고찰도 포함하였다.

두 차례의 검색 사이에서 적격성 기준은 변경하지 않았다. PRESS 방식의 동료 검토에서 수

4) 본 검토의 프로토콜은 연구의 신속성으로 인해 사전에 등록하지 못한 채로 진행하였다(PRISMA-ScR 항목 5).



*If automation tools were used, indicate how many records were excluded by a human and how many were excluded by automation tools.

Figure 1. PRISMA-ScR flow diagram of the source selection process. PRISMA-ScR; preferred reporting items for systematic reviews and meta-analyses extension for scoping reviews.

정된 것은 검색식과 PCC 요소의 서술이며, 실제 선별에 적용된 판정 기준은 최초 검색 당시 작성한 선별 지침 파일을 수정 없이 그대로 사용하였다(Supplementary S1).⁵⁾ 수정 후 검색에서 발견한 1,008건 가운데 최초 검색에서 이미 선별한 842건에 대해서는 당시의 판정을 그대로 수용하고, 새로 발견된 166건만을 새로 선별하였다.

본 결정의 타당성은 두 가지 방식으로 확인하였다. 첫째, 최초 검색에서 최종 포함되었던 13편을 표지 문헌으로 삼아 수정 후 검색이 이들을 다시 회수하는지 확인한 결과 12편이 회수되었다. 회수되지 않은 1편은 이후 원문 확인에서 AI 요소가 전혀 없음이 확인되어 적격성 기준에 따라 독립적으로 제외된 문헌으로서, 이 누락은 최종 결과와 무관하다. 둘째, 최초 검색에서 제목·초록 단계를 통과했으나 수정 후 검색에서 발견되지 않은 문헌은 20건이었으며, 이들은 모두 최초 검색 당시의 연구자 평가에서 이미 제외된 문헌이었다.

5) Population 관련 용어를 교육 블록에서 제거한 것은 해당 용어가 검색식에서 검색 대상을 좁히는 역할을 하고 있었기 때문이다. Population 적격성 자체는 최초 검색에서도 선별 단계에서 판정되었으므로, 이 변경으로 기존 판정의 근거가 달라지지는 않았다. 다만 PCC 요소의 서술에서 'educators'라는 표현이 빠지고 학습자와 임상자의 범위가 열거 형식으로 구체화되었는데, 선별 지침은 두 차례 모두 교수진(faculty)을 population에 포함하고 있었으므로 이 역시 서술의 정밀화에 속한다.

3. 정보 출처 및 검색 전략

PubMed, Web of Science Core Collection, Scopus 및 CINAHL, CINAHL (EBSCOhost)를 검색하였다. 각 검색식은 SDM 관련 용어, AI 관련 용어, 교육·훈련·시뮬레이션·평가 관련 용어의 세 블록을 AND로 결합하였으며, 각 블록 내에서는 동의어를 OR로 연결하였다. 검색 단계에서는 출판연도 제한을 적용하지 않았고, 언어 기준은 문헌 선별 단계에서 적용하였다. 두 차례 검색 각각에 대하여 네 데이터베이스의 완전한 검색식과 데이터베이스별 검색 건수는 Supplementary S2에 제시하였다.

PRESS 방식에 준한 동료 검토를 통해 세 블록을 조작화하는 과정에서의 문제점이 확인되었다. 통제가 사용되지 않았고, AI 블록에는 그 자체로 AI를 의미하지 않는 양식 관련 용어가 포함되어 있었으며, 그중 두 용어는 교육 블록에도 중복되어 있었다. 또한 population 관련 용어가 교육 블록 안에 포함되어 있었고, 네 데이터베이스 간 블록의 개념적 범위가 서로 동등하지 않았다.

이차 검색에서는 이러한 문제를 모두 수정하였다. PubMed와 CINAHL에서는 자유어와 함께 통제를 추가하였으며, SDM, AI 기계학습, 자연어처리, 전문직교육 및 계속교육, 교수, 시뮬레이션 훈련, 환자 시뮬레이션, 교육평가, 교육과정 및 임상역량에 해당하는 Medical Subject Headings와 CINAHL Headings를 사용하였다. AI 블록은 AI의 사용을 명확히 나타내는 표현으로 제한하였고, 최초 검색에서 일부 데이터베이스에 포함되지 않았던 용어를 보완하였다. Population 관련 용어는 교육 블록에서 제거하고, population 적격성은 문헌 선별 과정에서 평가하였다. 이후 네 데이터베이스에서 동일한 개념을 검색하도록 SDM 및 AI 블록의 범위를 조정하였다.

4. 문헌 선별

문헌 선별과 자료 도표화는 2026년 6월부터 8월에 걸쳐 수행되었다.

검색 결과는 자체 제작한 Python 스크립트로 통합하였다. 중복 문헌은 digital object identifier를 우선 기준으로, 식별자가 없거나 일치하지 않는 경우에는 정규화한 제목의 완전 일치를 기준으로 판정하였다. 내보내기 형식, 정규화 규칙 및 서지정보 병합 우선순위는 Supplementary S2에 제시하였다.

문헌 선별은 다음 다섯 단계로 진행하였다. 각 단계에서 AI가 관여한 지점, 사용한 도구와 버전, 사용 일자, 사용 목적, 검증 방식은 Supplementary S1에 항목별로 보고하였다.

1. 자동화 규칙. 규칙 기반 스크립트로 영어가 아닌 문헌, 비원저 출판 유형(사설, 서한, 회초록, 정오표 등) 및 연구계획서를 제거하였다. 이 단계는 서지 필드와 정규표현식만을 사용하므로 결정론적이므로, 스크립트를 다시 실행하면 동일한 결과가 재현되도록 하였다.
2. 생성형 AI 보조 선별. 문헌의 제목·초록·키워드를 GPT-5 기반 에이전트에 제공하고, 사전에 작성한 선별 지침에 따라 각 문헌을 일차, 이차, 불명확 또는 제외로 분류하도록 하였다. 에이전트는 각 판단의 근거와 신뢰도(high/med/low)를 정해진 JSONL 형식으로 함께 기록하였다. 신뢰도는 판정을 자동으로 좌우하는 임계값으로 사용하지 않았고, 낮은 신뢰도로 기록된 판정은 연구자 검토 대상으로 표시하는 데 사용하였다. 두 차례 검

- 색 모두 동일한 선별 지침 파일을 수정 없이 사용하였으며, 지침 전문과 출력 형식, 배치 구성은 Supplementary S1에 그대로 수록하였다.
3. 재현을 보호를 위한 재검토. 제외된 문헌 가운데 AI, SDM 및 교육 관련 키워드 신호가 동시에 나타난 문헌을 동일한 에이전트가 다시 검토하였다. 이 재검토는 상위 범주로의 승격만 허용하도록 스크립트 수준에서 강제하였으며, 재검토에서 새로운 제외 판정을 만들거나 유지된 문헌을 강등하는 것은 코드상 불가능하도록 설계하였다.
 4. 연구자의 적격성 평가. 위 단계에서 일차, 이차 또는 불명확으로 남은 문헌은 연구자가 적격성 기준에 대조하여 평가하였다. 이 단계의 판단 근거는 제목과 초록이었으며, 초록이 색인되지 않은 26건은 서지정보만으로 평가하였다. 환자 대상 도구, 임상 의사결정지원에만 사용된 도구, AI 요소가 없는 가상 시뮬레이션 등 경계선상에 있는 문헌은 기준에 따라 다시 검토하였다.
 5. 원문 평가. 위 단계를 통과한 문헌은 원문을 확보하여 적격성 기준 전체에 대하여 평가하였다. 이후 연구와 무관한 외부 연구자가 최종 포함된 각 문헌을 적격성 기준과 대조하여 확인하였다. 이 확인에서 이견은 제기되지 않았다.

기계학습 분류기를 이용한 선별 보조 자체는 근거합성에서 이미 확립된 절차로서, 본 연구가 새로 도입한 방식이 아니다. Cochrane은 무작위대조시험 분류기를 등록 시스템에 상시 운용하여 선별 부담을 줄이고 있고[38], ASReview[39]와 Rayyan[40]을 비롯한 도구도 널리 사용되고 있다. Cochrane, Campbell Collaboration, JBI 및 Collaboration for Environmental Evidence의 공동 입장문 또한 인간의 감독이 유지되고 판정에 관여한 AI의 사용이 투명하게 보고되는 한 이러한 활용을 허용하며, 본 연구는 이 입장문이 요구하는 보고 항목에 따라 Supplementary S1을 작성하였다[41]. 다만 대규모 언어모형을 이용한 선별은 특정 과제에 대해 사전 훈련되고 성능이 공표된 분류기와 달리 그 성능이 개별 검토마다 확인되어야 하므로, 아래와 같이 별도의 검증을 시행하였다.

이 절차의 설계에 있어서 선별 과정에서 어떤 오류가 더 문제가 되는지를 고려하였다. 단일 연구자의 선별은 이중 선별에 비해 더 많은 문헌을 놓칠 수 있고[42], 대규모 언어모형을 이용한 선별 역시 특이도에 비해 민감도가 낮은 경향이 보고되었다[43]. 다시 말해, AI 지원 단일 선별의 경우 적격 문헌을 제외하는 오류를 일으킬 가능성이 더 크므로, 후속 검증은 AI가 제외한 문헌 가운데 실제 포함 대상이 섞여 있는지를 확인하는 데 초점을 맞췄다.

에이전트의 분류는 그 자체로 확정된 판정이 아니라 의사결정 보조 자료로 취급하였다. 연구자는 제외로 분류된 문헌 전체에 대하여 에이전트가 기록한 제외 근거를 검토하였고, 근거가 기준에 부합하지 않거나 판단이 의심스러운 문헌은 원자료를 다시 확인하여 상위 범주로 되돌렸다. 다만 이 단계에서 제외 문헌의 제목과 초록을 연구자가 개별적으로 다시 읽지는 않았다.

이러한 검토 방식이 적격 문헌을 놓쳤을 가능성을 정량적으로 확인하기 위하여 별도의 재선별을 시행하였다. 고정된 난수 시드를 사용하여 AI 보조 선별에서 제외된 585건 가운데 100건을 무작위로 추출하고, 여기에 초록이 색인되지 않아 서지정보만으로 제외된 26건을 더한 126건을 대상으로 하였다. 재선별에서는 에이전트의 분류와 근거를 모두 제거하고 서지정보와 초록만을 제시하였으며, 연구자가 동일한 선별 지침에 따라 다시 판정하였다. 위임 성률은 이 재선별에서 연구자가 상위 범주로 되돌린 문헌의 비율로 정의하고 Wilson 95% 신뢰구간과 함께 제시하였다. 재선별에서 상위 범주로 되돌려진 문헌은 원문 확보와 적격성 평

가를 포함한 이후 절차를 그대로 거치도록 사전에 정하였으며, 그 결과 적절한 문헌이 확인될 경우 포함 문헌 수를 갱신하기로 하였다. 재선별에 사용한 표본과 판정 결과는 Supplementary S1에 제시하였다.

수정된 검색의 문헌 선정 과정과 단계별 문헌 수는 PRISMA-ScR 도표에 제시하였다 (Figure 1). 제목·초록 선별 단계와 원문 평가 단계는 도표에서 구분하였다.

5. 자료 정리 및 평가

자료는 본 검토의 연구 질문에 맞추어 개발한 자체 양식을 사용하여 도표화하였다. 본 보고서에서는 해당 양식의 항목을 JBI가 권고하는 문헌 특성 및 결과 항목에 대응시켜 제시하였다. 자료 추출은 한 명의 연구자가 13편의 원문 전체에 대하여 수행하였다. 추출한 값이 원문과 어긋나지 않도록 각 항목을 원문과 대조하는 확인 작업을 AI 에이전트의 보조를 받아 수행하였고, 보고된 빈도와 도표의 값이 도표화 자료와 일치하는지는 스크립트를 통해 기계적으로 대조하였다. 도표화한 항목은 서지정보, 논문 유형 및 연구 설계, 국가 및 임상 분야, 학습자 또는 목표 사용자, 표본 또는 자료원, AI 양식·입력·출력, 교육적 기능, 목표로 한 SDM 역량, 중재 또는 도구에 대한 설명, 비교군 또는 참조표준, 결과지표, Kirkpatrick 평가 수준, 주요 결과 및 보고된 한계, 실행 관련 고려사항, 윤리·형평성·안전 문제, 연구자의 해석이었다.

‘목표로 한 SDM 역량’ 항목의 범주는 각 문헌이 스스로 교육 목표 또는 평가 대상으로 명시한 표현에서 귀납적으로 도출한 뒤 동일한 의미의 표현을 통합한 것이다. 도출된 범주는 이후 Makoul & Clayman의 통합 모형[1] 및 Elwyn et al.의 three-talk 모형[3]과의 대응 관계를 확인하여, 대응이 분명한 경우 해당 모형의 명칭을 사용하였다. 예컨대 team talk, option talk, decision talk는 observer OPTION(Observing Patient Involvement)-5를 사용한 문헌이 그 도구의 항목 구조를 따른 것이다. 두 모형 어디에도 대응하지 않는 범주, 즉 윤리적 추론, 형평성 및 편향 인식, AI 설명, 문서화 등은 해당 문헌의 표현을 그대로 유지하였다.

각 범주는 다음 원칙에 따라 코딩하였다. 어떤 범주는 해당 문헌이 그것을 교육 중재의 목표, 평가도구가 측정하는 대상, 또는 보고된 결과지표로 명시적으로 지목한 경우에만 코딩하였다. 한편, 문헌의 내용으로 미루어 그 구성개념이 암묵적으로 관여하였으리라고 독자가 추론할 수 있다는 이유에서 코딩을 추가하지는 않았다. 따라서 Table 2의 수치는 문헌이 그것을 명시적 목표 또는 측정 대상으로 삼았는지를 나타낸다. 범주별 조작적 정의와 범주별로 코딩된 문헌은 Supplementary S3에 제시하였다.

이 가운데 의사결정 조건은 각 문헌이 교육하거나 분석한 상황에서 선택 가능한 선택지를 명시하였는지를, 당사자 구성은 그 상황에 참여하는 당사자가 누구인지를 뜻한다. 이 두 항목은 13편의 원문에 대하여 판단마다 근거가 되는 원문을 그대로 기록하는 방식으로 도표화하였다. 13편 전체의 도표화 자료는 두 항목의 근거 원문과 함께 Supplementary S4에 제시하였다.

자료 추출은 한 명의 연구자가 수행하였으므로, 해석이 개입하는 항목에 대해서는 판단의 근거를 공개하여 제3자가 검증할 수 있도록 하였다. 목표로 한 SDM 역량은 각 범주의 조작적 정의를 Supplementary S3에 제시하여 동일한 기준으로 코딩을 재현할 수 있게 하였고, 의사결정 조건과 당사자 구성은 판단마다 근거가 된 원문을 Supplementary S4에 함께 제시하였다. 이로써 각 판정이 어떤 문장에 근거하는지를 원문과 대조하여 확인할 수 있다. 단, 이 방식은 제2의 연구자에 의한 독립적 이중 추출을 대체하지 못하므로, 한계는 IV.4에 기술하였다.

www.kci.go.kr

Table 2. Shared decision-making competencies targeted

Competency targeted	n	Sources
Values and preferences	11	27–31, 33, 44, 47–50
Option talk	9	27–30, 33, 44–47
Decision facilitation	8	27–29, 33, 44–46, 49
Risk communication	6	27–29, 45, 46, 50
Empathy and person-centeredness	4	29, 33, 48, 50
Communication quality	2	47, 48
Goals of care	2	31, 49
Patient-specific information	2	45, 46
Adherence support	1	47
AI explanation	1	27
Conflict resolution	1	31
Contextualized care	1	44
Decision awareness	1	30
Decision talk	1	30
Documentation	1	49
Equity and bias awareness	1	31
Ethical reasoning	1	31
Family-centered decision-making	1	31
Information provision	1	50
Management reasoning	1	44
Patient activation	1	47
Patient-centered communication	1	33
Patient-centered outcomes	1	49
Question elicitation	1	50
Team talk	1	30
Trust building	1	48

Operational definitions for each category are given in Supplementary S3.
AI: artificial intelligence.

6. 자료 분석 및 제시

도표화한 자료는 연구 유형, AI 양식, 교육적 기능, 목표로 한 SDM 역량 및 Kirkpatrick 수준에 따라 빈도와 문헌 식별자를 제시하는 방식으로 기술적으로 요약하였다.⁶⁾ 이후 도표화한 자료를 반복적으로 비교하는 귀납적 내용분석을 수행하여 AI의 교육적 역할에서 반복적으로 나타나는 양상을 확인하였다. 그 결과, AI를 훈련 환경으로 활용하는 경우, 평가 인프라로 활용하는 경우, 교육과정 또는 실무 개선의 촉매로 활용하는 경우의 세 가지 주제를 도출하였다.

직접적인 교육 중재에서 보고된 학습자 성과는 AI 기반 측정도구의 타당화 결과 및 교육과정 설계·개념적 문헌의 결과와 구분하여 해석하였다. 결과는 빈도표와 서술적 요약의 형태로 제시하였다.

단, 개별 문헌에 대한 공식적인 편향 위험 평가나 방법론적 질 평가는 수행하지 않았다. 이는 주제범위 문헌고찰에서 선택적으로 적용되는 항목이며, 본 검토의 목적이 근거의 강도를 판정하는 데 있지 않기 때문이다(PRISMA-ScR 항목 12). 같은 이유에서 메타분석과 근거 확실성 평가도 수행하지 않았다. 포함 문헌은 연구 목적과 설계가 서로 이질적이므로 하나의 근거 수준에서 통합하지 않았으며, 훈련 중재, 자동 평가도구의 타당화, 실무 분석, 개념적 문헌을 구분하여 기술하였다.

6) 이때 Kirkpatrick 수준은 교육 중재를 시행하고 그 효과를 학습자에게서 측정할 문헌에만 적용하였다. 교육 중재 없이 실제 임상 상담이나 진료기록을 분석한 문헌, 학습자 성과를 평가하지 않은 방법 개발 연구 및 개념적 문헌에는 이 틀을 적용하지 않고, 무엇을 측정하였는지를 별도로 기술하였다.

7. 설계 원칙의 도출

IV.3에 제시한 다섯 가지 설계 원칙은 문헌고찰의 결과를 전제로 하여 수행한 규범적 논증의 산물이다. 도출은 세 단계로 진행하였다.

첫째, III의 기술적 요약과 내용분석에서 확인된 공백을 열거하였다. 여기에는 숙고 관련 역할을 명시적 목표로 삼은 문헌이 드물다는 점, 상담에서 무엇이 선택의 대상인지가 대체로 명시되지 않는다는 점, 거의 모든 문헌이 임상-환자 이자 관계를 전제한다는 점, 자동 채점의 근거가 학습자의 실제 수행과 연결되지 않는 사례가 확인된다는 점, 훈련 연구 대부분이 단일 회차 직후 결과만을 측정한다는 점이 포함된다.

둘째, 각 공백에 대하여 그것이 왜 단지 연구상의 미비가 아니라 규범적으로 문제가 되는지를 서론에서 제시한 전제들에 비추어 검토하였다. 사용한 전제는 다음과 같다. (1) 증거가 선택을 과소결정하고 결과가 환자의 가치에 달려 있을 때 SDM은 선택이 아니라 윤리적 의무가 된다는 조건[5]. (2) 자율성을 관계적으로 이해할 필요와[9,10], 가치 진술의 신뢰성이 환자의 인구학적 특성에 따라 불균등하게 배분될 수 있다는 인식적 부정의의 문제[8]. (3) 학습자의 성취로 기록되는 판단은 학습자가 실제로 수행한 바에 대응해야 한다는 진실성의 요구와, 판단을 내리는 쪽이 그 근거를 설명할 응답책임(I 참조). (4) 윤리적 성향이 설명이 아니라 습관화와 사회화를 통해 형성된다는 논의(IV.2 참조).

셋째, 각 검토의 결론을 설계상 확인 가능한 진술로 다시 쓰고, 그 진술 가운데 문헌 근거에서 직접 따라 나오는 부분과 윤리적 논증에서 따라 나오는 부분을 구분하여 Table 1의 두 열에 각각 제시하였다.

이 절차는 전문가 합의나 델파이 조사와 같은 외부 타당화를 거치지 않았다. 따라서 다섯 원칙은 합의된 기준이 아니라 하나의 설계 명세로 제안하는 것이다.

III. 결과

1. 선정된 문헌

2026년 8월 1일 네 개 데이터베이스를 검색한 결과 총 1,967건이 확인되었다. 데이터베이스별로는 PubMed 476건, Web of Science 480건, Scopus 883건, CINAHL 128건이었다. 중복 문헌 959건을 제거하였다. 남은 1,008건 가운데 995건을 제목-초록 단계에서 제외하였다. 언어, 출판 유형 및 연구계획서 여부에 관한 자동화된 규칙을 통해 218건을 제외하였고, 재현율을 보호하기 위해 상위 범주로의 재분류만 허용한 재검토를 포함한 AI 보조 선별을 통해 585건을 제외하였으며, 남은 205건 가운데 192건은 연구자의 적격성 평가에서 제외하였다.

연구자 단계에서 제외한 192건의 사유는 임상 수행 지원에만 해당하고 교육이 아님 104건, 판정에 필요한 정보가 부족함 24건, 임상의 또는 환자를 대상으로 하는 도구로 전문직 교육에 해당하지 않음 22건, 전문직의 학습이나 평가에 해당하지 않음 19건, 환자 또는 환자용 교육자료를 대상으로 함 18건, 연구 질문에 답하지 않는 포괄적 문헌고찰 4건, SDM 교육이 아님 1건이었다. 개별 문헌의 목록과 사유는 Supplementary S2에 제시하였다.

제목-초록 단계를 통과한 13편은 모두 원문을 확보하였으며, 원문 단계에서 추가로 제외된

문헌은 없었다. 이는 사전 단계에서 연구자가 205건 전체에 대하여 적격성 기준을 이미 적용하고 경계선상의 문헌을 그 단계에서 판정하였기 때문이다. 따라서 PRISMA-ScR이 요구하는 원문 단계 제외 문헌의 목록은 해당 사항이 없다. 제목·초록 선별과 원문 평가의 두 단계는 Figure 1에 구분하여 표시하였다.

AI 보조 선별이 적격 문헌을 놓쳤는지를 확인하기 위한 눈가림 재선별에서는, AI 보조 선별로 제외된 585건에서 무작위 추출한 100건과 초록 없이 서지정보만으로 제외된 26건을 합한 126건을 다시 판정하였다. 두 표본 모두에서 상위 범주로 되돌려진 문헌은 없었다(무작위 표본 0/100, 서지정보 표본 0/26). 무작위 표본에 근거한 위음성률은 0%였고, Wilson 95% 신뢰구간은 0.0%–3.7%였다. 이 상한을 AI 보조 선별로 제외된 585건 전체에 적용하면, 재선별에서 상위 범주로 되돌려질 문헌은 최대 약 22건에 해당한다. 되돌려진 문헌이 없었으므로 사전에 정한 처리 규칙에 따른 후속 절차는 발생하지 않았다. 재선별의 표본 구성과 판정 결과는 Supplementary S1에 제시하였다.

2. 포함된 문헌의 특성

최종 포함된 13편의 문헌 특성은 Table 3에 제시하였다[27–31,33,44–50]. 출판연도는 2022년부터 2026년까지였으며, 13편 가운데 10편이 2025년 또는 2026년에 발표되었다. 국가별로는 미국이 7편, 독일이 2편, 한국과 중국이 각각 1편이었고, 다국가 연구가 2편이었다. 논문 유형별로는 AI 평가 연구가 7편, 훈련 중재 연구가 4편이었으며, 방법 개발 논문과 개념적 문헌고찰이 각각 1편이었다. 이 가운데 1편은 동료평가를 거치지 않은 사전출판 문헌이었다[46].

Kirkpatrick 수준은 교육 중재를 시행하고 그 효과를 학습자에게서 측정한 4편에만 적용하였다[27–29,33]. 이 가운데 3편은 학습자의 반응과 자기평가 학습성고를 보고하였고[27,28,33], 1편은 반응, 학습 및 시뮬레이션 상황에서의 행동까지 보고하였다[29].

나머지 9편에는 이 틀을 적용하지 않고, 각 문헌이 실제로 무엇을 측정하였는지를 별도로 기술하였다. 교육 중재 없이 실제 임상 상담이나 진료기록을 자동으로 측정한 6편은 임상 현장에서 이루어진 의료진의 SDM 관련 행동 또는 그 문서화를 측정한 것이며[30,45–49], 이는 측정 대상 집단의 수행 수준을 나타낸다. 1편은 연구진이 수행한 상담에 대하여 환자 성과의 대리지표를 보고하였으나 교육 연구가 아닌 실무 분석이었다[50]. 방법 개발 연구 1편[44]과 개념적 문헌고찰 1편[31]은 학습자 성과를 평가하지 않았다. 교육 중재를 거친 학습자에게서 지속적인 임상 행동 변화나 환자 편익 수준의 성과를 입증한 문헌은 없었다.

논문에 나타난 AI 양식과 교육적 기능의 분포를 확인하였다(Table 4). 가장 빈번하게 나타난 기능적 요소는 평가 점수 산출로 8편에서 나타났으며, 다음은 질 향상 7편, 피드백 또는 성찰과 시뮬레이션 또는 역할극이 각각 6편이었다. 직접 훈련은 5편에서 확인되었다.

또한, 각 문헌이 명시적으로 목표 또는 측정 대상으로 삼은 SDM 역량을 검토하였다(Table 2). 13편 가운데 11편이 환자의 가치와 선호를, 9편이 선택지 논의를 명시하였다. 의사결정 촉진은 8편, 리스크 커뮤니케이션은 6편에서 다루어졌다. 반면 윤리적 추론과 가족 중심 의사결정은 각각 1편에서만 명시되었으며, 두 항목 모두 동일한 개념적 문헌고찰에서 나타났다[31]. 팀 논의와 결정 논의 역시 각각 1편에서만 나타났으며, 이는 하나의 자동화 점수 산출 연구에 국한되었다[30]. Table 2의 수치는 문헌이 그것을 명시적 목표 또는 측정 대상으로 삼았는지를 나타내며, 범주별 조작적 정의는 Supplementary S3에 제시하였다.

Table 3. Characteristics of the 13 included sources

#	Source	Country	Article type / design	Learners or target users	AI modality	SDM competencies targeted	Kirkpatrick level
27	Kim et al. 2026	South Korea	Training intervention; pilot mixed-methods course evaluation	Final-year medical students	Clinical diagnostic AI in simulation	Risk communication; option talk; values/preferences; decision facilitation; AI explanation	1, 2
29	Gao et al. 2026	China	Training intervention; randomized controlled educational trial	Surgical residents (PGY 1–5)	LLM preparation scaffold	Values/preferences; risk communication; option talk; decision facilitation; empathy	1, 2, 3 (simulated encounter)
33	Müller et al. 2026	Germany	Training intervention; quasi-experimental pre-post	Medical students (semesters 2–10)	LLM chatbot simulated patient	Empathy; patient-centered communication; option talk; values/preferences; decision facilitation	1, 2
44	Cook 2025	United States	Method development	Health professions learners and educators	LLM virtual patient	Management reasoning; values/preferences; contextualized care; option talk; decision facilitation	-
28	Herschbach et al. 2026	Germany	Training intervention; feasibility mixed-methods pre-post	Medical students in the practical year; licensed physicians	LLM chatbot simulated patient; LLM scorer	Option talk; values/preferences; decision facilitation; risk communication	1, 2
30	Selvaraj et al. 2026	United States	AI assessment; validation against trained human raters	Clinicians; SDM educators; QI teams	LLM scorer	Decision awareness; team talk; option talk; values/preferences; decision talk	-
45	Zheng et al. 2026	United States	AI assessment; NLP model development and validation	Physicians; oncology/urology educators	ML/NLP classifier	Risk communication; option talk; patient-specific information; decision facilitation	-
46	López-García et al. 2025	United States	AI assessment; preprint, LLM scoring framework	Physicians; oncology educators; QI teams	LLM scorer; transformer classifier	Risk communication; patient-specific information; option talk; decision facilitation	-
47	Topaz et al. 2022	United States	AI assessment; secondary analysis of recorded encounters	Primary care clinicians; QI teams	Speech-to-text; ML classifier	Patient activation; option talk; values/preferences; adherence support; communication quality	-
48	Khaikin et al. 2026	United States, United Kingdom	AI assessment; feasibility/face-validity research letter	Clinicians; equity-focused SDM quality teams	Audiovisual digital phenotyping	Trust building; empathy; values/preferences; communication quality	-
49	Streid et al. 2023	United States	AI assessment; retrospective EHR review	Surgeons; perioperative teams	Rule-based NLP	Goals of care; values/preferences; patient-centered outcomes; documentation; decision facilitation	-
31	Sullivan et al. 2026	United States	Conceptual review	Neonatal clinicians, fellows, residents, faculty	Generative AI; conversational agent; simulation; NLP feedback	Family-centered decision-making; values/preferences; goals of care; ethical reasoning; conflict resolution; equity/bias awareness	-
50	Lambrakos et al. 2024	Australia, Spain, United Kingdom, United States	AI assessment; four-stage interview/survey with video analytics	Electrophysiologists; material developers	AI video analytics; sentiment analysis; speech-to-text	Risk communication; information provision; question elicitation; empathy; values/preferences	-

Kirkpatrick levels: 1: reaction, 2: learning, 3: behavior, 4: results; -: not applicable (no educational intervention; see Section III.2).

AI: artificial intelligence, SDM: shared decision-making, PGY: postgraduate year, LLM: large language model, QI: quality improvement, NLP: natural language processing, ML: machine learning, EHR: electronic health record.

3. SDM 교육에서 AI의 역할

포함된 13편의 문헌에서 AI의 교육적 역할은 훈련 환경(n=5), 평가 인프라(n=6), 교육과정 또는 실무 개선의 촉매(n=2)로 구분되었다.

훈련 환경 연구들은 상담 준비, 가상환자, 시뮬레이션 및 자동 피드백의 가능성을 보여주었다[27–29,33,44]. Kim et al.은 영상의학 AI 결과를 우연히 발견된 영상 소견에 관한 모의 상담에 통합하였다. 의과대학 본과 4학년 학생의 SDM 이해도와 자신감($t=6.51, 7.56$; 모두 $p<.001$), AI 보조 SDM 이해도와 자신감($t=5.72, 5.80$; 모두 $p<.001$)이 향상되었다[27]. 그러나 소규모 단일군 연구로 자기보고 결과만 측정했으며, 학습자들은 AI 결과의 설명가능성 부족을 한계로 지적하였다.

Gao et al.의 무작위시험에서는 외과 전공의 40명을 가치 중심 대규모 언어모형군과 자기 주도 학습군에 배정하였다[29]. AI 군은 SDM-Q-9 점수가 더 높았고(84.7 대 71.3, $p<.01$),

Table 4. Educational functions of AI across the included sources

Educational function	n	Sources
Assessment scoring	8	28, 30, 45-50
Quality improvement	7	30, 45-50
Feedback and reflection	6	27, 29-31, 45, 46
Simulation or role play	6	27-29, 31, 33, 44
Training delivery	5	27-29, 33, 44
Automated feedback generation	3	28, 33, 44
Curriculum guidance	2	31, 49
Assessment development	1	44
Assessment guidance	1	31
Case generation	1	31
Educational material development	1	50
Equity monitoring	1	48
Preparation scaffold	1	29
Risk prediction	1	47

AI: artificial intelligence.

의사결정 갈등은 더 낮았으며(19.5 대 32.1, $p < .01$), 관찰자 평가 의사소통 점수(17.5 대 14.8, $p < .01$)와 상담 개요 점수(13.8 대 7.5, $p < .01$)도 높았다. 단, 단일기관의 소규모 연구이고 비교군이 비구조화되어 있으며 모든 평가는 중재 직후 이루어졌다.

Müller et al.은 GPT-4o 챗봇을 모의환자와 자동 피드백 도구로 사용하였다[33]. 약 13 회의 대화 후 자기평가 의사소통 역량은 5.89점에서 6.83점으로 증가했으며(평균차 0.94, 95% CI 0.69-1.20, Cohen $d=0.58$), 피드백 유용성은 7.92/10이었다. 그러나 결과는 관찰된 수행이 아니라 자기평가였고, 피드백의 적절성과 역량 변화 사이에는 상관관계가 없었다. Cook은 GPT-4 가상환자가 GPT-3.5보다 실제적이고 선호에 부합한 대화를 생성한다고 보고했다. 그러나 해당 연구는 학습자 성과를 평가하지 않았다[44].

Herschbach et al.의 연구는 AI가 모의환자와 평가자 역할을 동시에 수행한 유일한 사례였다[28]. 의과대학생 22명, 의사 34명, 소속을 밝히지 않은 2명 등 총 58명이 SDM 6단계 모형에 대한 설명을 받은 뒤, AI 기반 가상 환자와 상담을 진행하고, Observer OPTION-12를 사용하여 자동 평가를 받았다. 참여자는 평가에서 평균 25.23점(48점 만점, 범위 7-37)을 기록하였다. AI와 인간 전문가가 평가한 Observer OPTION-12 총점의 일치도는 높았지만(ICC=0.86, $p < .001$), AI는 61건 중 68.85%에서 인간보다 높은 점수를 부여했다. AI와 인간이 정확히 일치한 경우는 9.84%였다. 자동 평가에 대한 질적 검사는 AI가 학습자가 묻지 않았는데도 환자가 자발적으로 제공한 정보를 학습자가 탐색한 것으로 간주하는 오류가 나타났음을 확인하였다. 의과대학생은 의사보다 높은 점수를 보였고(28.41 대 22.79, $p=.002$), 자기효능감은 4.02점에서 3.77점으로 감소하였다($p=.002$). 이는 구조화된 피드백이 과제의 어려움에 대한 자기평가를 재조정했을 가능성을 시사한다. 단, 이 연구의 인간 참조표준은 신뢰도가 확인되지 않은 단일 평가자였다.

평가 인프라 연구들은 상담 기록, 음성, 영상 및 의무기록에서 SDM 수행을 자동으로 식별하고 점수화하였다[30,45-49]. Selvaraj et al.은 유방암 수술 상담 287건에서 GPT-4o와 Gemini의 Observer OPTION-5 점수가 인간 평가와 약 $r=0.60$ 의 상관을 보였다고 보고하였다. 이는 인간 평가자 간 상관인 $r=0.77$ 의 약 75%-83% 수준이었다[30]. 전립선암 상담을 분석한 자연어처리 모형은 위험소통 개념을 AUC 0.84-0.99로 식별했고, 상담 단위 평가 정확도는 80%-100%였다[45]. 사전 출판 문헌에서 GPT-4o 기반 평가의 미시 평균(micro-averaged) F1은 85.00-92.00로 보고되었다[46].

Topaz et al.은 일차의료 상담 80건에서 전사 정확도 0.90, 환자가 인식한 SDM과 약물 순응도 분류 F-score 0.88과 0.87을 보고하였다[47]. Khaikin et al.은 긍정적인 의사소통 경험을 보고한 참여자 4명 중 3명에게서 AI가 부정적 또는 중립적 정서를 탐지하였다[48]. Streid et al.은 수술 전 기록 50,307건을 96.3%의 정확도로 분석했지만, 치료목표 기록의 필수 요소가 하나 이상 포함된 수술은 11.7%, 세 요소가 모두 포함된 경우는 0.2%였다[49]. 이러한 연구들은 대규모 평가와 피드백의 기반을 제공할 수 있으나, 대부분 특정 임상 영역과 제한된 자료에 기반하며 수동 주석과 인간 검증에 의존하였다. 또한 이들은 교육 중재를 시행한 연구가 아니라 이미 이루어진 임상 상담을 대상으로 한 측정 연구이므로, 그 결과는 학습의 성과가 아니라 측정 대상 집단의 수행 수준으로 읽어야 한다.

교육과정 또는 실무 개선을 다룬 두 문헌은 AI가 교육 내용과 우선순위를 형성할 가능성을 보여주었다[31,50]. 신생아 윤리교육에 관한 개념적 문헌고찰은 AI 기반 사례, 가상 튜터, 표준화 부모, 윤리 코칭 및 자연어처리 피드백을 제안하면서 환각, 편향, 고정관념 강화와 복잡한 윤리 문제의 단순화를 위험으로 제시하였다[31]. 이 문헌은 윤리적 추론을 명시적 목표로 다룬 유일한 문헌이었다. Lambrakos et al.은 심장기기 삽입을 거부한 환자 27명의 상담에 AI 영상·감정 분석을 적용하였다[50]. 기기 삽입 의향은 상담 후 5.0/7로, 기존 상담에 대한 회상 평가 3.1과 상담 전 4.4보다 높았다($p < .001$). 단, 이는 교육 효과가 아니라 실무 분석 결과였다.

훈련, 평가, 실무의 세 역할에 걸쳐, 문헌이 모형화하거나 분석한 상담 자체의 구성을 두 가지 측면에서 확인하였다. 첫째, 상담에서 무엇이 선택의 대상이 되는지가 명시되었는지를 보았다. 교육 상황을 모형화한 5편 가운데 선택 가능한 선택지를 밝힌 것은 2편이었고, 두 문헌 모두 결정을 유보하는 경로를 함께 두었다[27,28]. 나머지 3편은 임상 주제만 제시하였을 뿐, 학습자가 무엇을 두고 숙고해야 하는지는 명시하지 않았다. 즉, 선호 민감성이 성립하는 조건 자체는 대체로 설계의 대상이 아니라 주어진 배경으로 다루어졌다. 둘째, 상담에 누가 참여하는지를 보았다. 13편 가운데 12편은 임상사와 환자의 이자 관계를 모형화하거나 전제하였으며, 임상 상담을 다루지 않은 문헌도 여기에 해당하였다. 유일한 예외인 개념적 문헌고찰은 신생아 진료에서 부모를 의사결정자로 다루고 표준화 가족 및 AI 기반 표준화 부모와의 실습을 제안하였으나, 이를 실제로 평가하지는 않았다[31]. 다자간 상담을 평가한 문헌은 없었다.

종합하면, AI는 SDM 연습, 상담 준비, 자동 피드백 및 대규모 수행평가를 지원할 가능성을 보였으나, 그 근거는 단기 시뮬레이션, 측정도구 타당화 및 후향적 분석에 집중되어 있었다.

IV. 고찰

1. 근거 문헌의 한계

본 검토에서 확인된 직접적인 교육 근거는 무작위 교육시험 1편[29], 대조군이 없는 사전-사후 연구 3편[27,28,33], 방법 개발 연구 1편[44]이 확인되었다. 나머지 연구는 교육보다는 의사소통이나 문서화에 집중했다. 실행 가능성, 만족도, 지각된 역량, 시뮬레이션 수행은 초기 연구에서 의미 있는 결과지표로 사용될 수 있으나, 서로 다른 환경과 학습자 집단에서의 효과를 입증하지는 못한다는 단점을 지닌다. 특히, 대조군이 없는 연구들은 대부분 한 차례의 교육 직후 결과만을 측정하여 학습 효과의 지속을 보여주지 못했다. 평가 연구 또한 교

육 효과를 입증하는 데 한계를 보였다. 6편은 SDM 관련 행동 또는 문서화를 측정하였지만 [30,45-49], 이러한 측정에서 도출된 피드백이 이후 의료진의 행동을 변화시키지는 평가하지 않았다. 환자 수준의 결과를 보고한 연구⁷⁾는 연구진이 수행한 상담을 분석한 것으로 교육 효과를 평가한 연구는 아니었다[50].

따라서, 현재 근거만으로는 AI 기반 SDM 교육이 일상 진료에서 환자의 지식, 가치에 부합하는 결정, 의사결정 갈등, 신뢰 또는 치료 순응도를 향상시킨다고 결론 내리기 어렵다.

이러한 결과는 자동 채점과 평가자 일치도를 교육 효과나 도구의 타당화 근거로 확장 해석해서는 안 됨을 보여준다. 자동 채점은 향후 피드백 체계를 위한 기반일 뿐이며, 인간 평가와의 상관이나 높은 분류 성능만으로 평가도구의 타당성과 교육적 유효성이 확립되는 것은 아니다[30,46]. 타당성은 도구 자체가 아니라 그 점수를 특정한 목적으로 해석하고 사용하는 방식에 대해 성립하는 것이므로, 일치도 계수는 그 해석을 뒷받침하는 여러 근거 가운데 하나일 뿐 그 자체로 타당성을 확립하지 않는다[51]. 특히 한 연구는 인간 평가자 간 일치도를 기준으로 모델 성능을 해석했지만[30], 다른 연구는 신뢰도가 확인되지 않은 단일 전문가를 참조표준으로 삼았다[28]. 참조표준의 신뢰도가 확인되지 않으면 불일치가 모델에서 어느 만큼이나 비롯되었는지 알 수 없으므로, 높은 일치도 계수만으로 모델의 정확성을 판단하기는 어렵다.

이상을 다음 네 가지로 요약할 수 있다. 첫째, 무작위시험을 포함한 엄격한 비교연구가 부족했다. 둘째, SDM을 명시적 교육 목표로 삼은 교육과정이 드물었다. 셋째, AI 기반 교육이 장기적인 임상 행동 변화나 환자 결과로 이어지는지를 보여주는 근거가 없었다. 넷째, 편향, 프라이버시, 설명가능성은 대체로 연구의 한계나 안전장치로 언급되었을 뿐, 이를 독립적인 평가 대상으로 검증한 연구는 없었다.

2. 윤리학적 검토

문헌 검토에서 드러난 윤리적 이슈는 다음과 같다. 첫째, AI 기반 SDM 교육은 정보 전달과 행동 측정에만 주목하고, 숙고의 과정은 충분히 다루지 않는다. 이 판단은 서로 독립적인 두 가지 도표화 결과에 근거한다. 하나는 명시적 목표의 분포이다. 가치와 선호는 11편, 선택지 논의는 9편에서 명시적 목표 또는 측정 대상이었던 반면, 윤리적 추론과 가족 중심 의사결정을 명시한 문헌은 각각 1편뿐이었다[31]. 다른 하나는 도표화 항목 가운데 의사결정 조건, 즉 각 문헌이 모형화한 상황에서 무엇이 선택의 대상인지가 명시되었는가이다. 교육 상황을 모형화한 5편 가운데 이를 명시한 것은 2편이었고[27,28], 1편은 부분적으로만 명시하였다[29]. 이 두 번째 근거는 13편의 부분집합인 5편에 한정된 관찰이므로 그 범위가 작다는 점을 감안할 필요가 있다.

단, 어떤 문헌이 ‘윤리적 추론’이라는 용어를 사용하지 않았다는 사실이, 그 문헌의 교육이나 평가에서 숙고적 판단이 전혀 이루어지지 않았음을 뜻하지는 않는다. 가치 탐색이나 선택지 비교를 목표로 삼은 교육에도 숙고의 요소가 포함되어 있을 수 있다. 여기서 확인된 것은 숙고가 실제로 이루어졌는지 여부가 아니라, 숙고와 관련된 역량이 교육의 명시적 목표나 평가의 대상으로 지목되었는지 여부이다. 그러나 명시적 목표가 되지 않은 것은 설계되지

7) 또한, 포함된 13편은 연구 목적과 설계가 서로 다르다. 교육 중재를 시행하고 학습자의 변화를 측정한 연구, 이미 이루어진 임상 상담을 자동으로 채점한 연구, 도구를 제안한 방법 연구, 교육 방향을 논한 개념적 문헌이 함께 포함되어 있으며, 이들은 같은 질문에 답하는 서로 다른 크기의 근거가 아니라 서로 다른 질문에 답하는 문헌이다. 따라서 본 검토는 이들을 하나의 근거 수준에서 통합하지 않았으며, 각 유형이 무엇을 보여줄 수 있는 연구인지를 구분한 상태에서 종합하였다.

도 측정되지도 않는다는 점에서, 이 결과는 여전히 설계상의 문제를 가리킨다. 다시 말해, 설명과 선택지 제시는 적극적으로 측정했지만, 그 선택지가 특정 환자에게 갖는 의미를 함께 해석하고 판단하는 과정은 거의 평가하지 않았다. 이러한 불균형은 논문 한 편의 수행 자료에서도 관찰된다. 단일기관에서 58명을 대상으로 수행한 실험 가능성 연구[28]의 Observer OPTION-12 문항별 점수를 보면, 가장 높은 점수를 받은 항목은 복수의 선택지가 있음을 알리는 것이었고, 가장 낮은 점수를 받은 항목은 환자가 의사결정에 어떻게 참여하고 싶은지를 확인하는 것이었다. 이는 단일 연구의 결과이므로 포함 문헌 전체에서 확인된 경향으로 일반화할 수 없으나, 위의 두 도표화 결과와 같은 방향을 가리키는 사례로 읽을 수 있다.

둘째, SDM 역량이 단순한 대화 기술이 아니라 무엇이 중요한지 알아차리고 판단하는 성향이라면, 한 차례의 교육만으로는 이를 형성할 수 없다. 전문직 정체성은 설명을 듣는 것보다 사회화와 반복적 실천을 통해 형성되며[52], 덕이 습관화를 통해 획득된다는 의료윤리학의 오랜 논의도 같은 지점을 지적한다[53]. 그러나 네 편의 훈련 연구 가운데 세 편은 한 번의 상담과 직후 결과만을 평가하였다[27,28,33]. 한 연구에서는 자기효능감이 오히려 감소했으며, 이는 피드백을 통해 과제의 어려움을 더 정확히 인식한 결과로 해석되었다[28]. 또한 같은 연구에서 의과대학생($n=22$)이 면허를 가진 의사($n=34$)보다 높은 수행 점수를 보였다(나머지 2명은 소속 미기재)[28]. 이는 단일 연구의 소규모 하위집단 비교이므로 그 자체로 일반화하기 어려우나, 임상 경험의 축적이나 SDM 원칙에 대한 동기가 곧바로 수행으로 이어지지는 않을 수 있음을 시사하는 관찰로 볼 수 있다. 따라서 반복적 연습과 피드백이 가능한 시뮬레이션 환경이 필요하다.

또한, AI를 활용해 SDM을 교육할 때에는 의사소통 기술뿐만 아니라 AI의 판단을 비판적으로 다루는 능력도 함께 다룰 필요가 있다. AI에게 일부 과업을 맡기더라도 사용자는 독립적으로 판단하고 그 결과에 대한 최종 책임을 져야 한다[29,31,32]. 이러한 역량에 대한 고려 없이 AI를 활용한 대화 훈련만 반복한다면, 겉으로 드러나는 SDM 수행은 향상될 수 있으나 이를 윤리적으로 책임지는 능력은 오히려 약화될 수 있다.

셋째, 자동 채점은 입증된 타당성 이상으로 권위 있게 받아들여질 수 있다. 앞서 언급한 실험 가능성 연구에서 AI는 3분의 2 이상의 상담에서 인간 평가자보다 높은 점수를 부여했고, 환자가 자발적으로 제공한 정보를 학습자가 탐색한 것으로 잘못 인정하기도 하였다[28]. 이는 단순한 측정 오차를 넘어 학습자가 하지 않은 행동을 성취로 귀속하는 문제이다. 즉 여기서 먼저 문제가 되는 것은 점수의 정확도가 아니라 진실성이다. 학습자의 성취로 기록되는 판단은 학습자가 실제로 수행한 바에 대응해야 하며, 그렇지 않은 기록은 그 학습자에 관한 사실을 잘못 진술하는 것이 된다.

3. 가상 SP 시스템의 설계

앞서 확인한 공백은 연구상의 한계인 동시에 설계상의 과제이기도 하다. 현재 문헌이 정보 제공과 행동 측정에 집중하고 숙고를 충분히 다루지 못하며, SDM 역량의 형성에는 반복적 연습이 필요하고 자동 평가는 절차적으로 정당화되어야 한다면, 가상 SP 시스템은 다음과 같은 규범적 원칙을 갖추어야 한다.

첫째, 시나리오는 임상적으로 방어 가능한 복수의 선택지가 존재하고 환자의 가치에 따라 결정이 달라지는 선호 민감 상황을 중심으로 구성해야 한다. 가능한 경우 치료하지 않거나 결정을 유보하는 선택도 포함해야 한다. 또한 대화는 환자의 관심과 선호 탐색, 그에 맞춘 정보

제공, 공동 숙고, 결정 또는 유보의 순서로 진행되어야 한다. 이는 선택지를 일방적으로 설명한 뒤 선호를 묻는 방식보다 환자의 가치가 결정의 틀을 형성하도록 한다.

둘째, 시스템은 단순한 의료진-환자 이자 관계를 넘어 가족, 보호자, 대리결정자 및 보건의료팀이 참여하는 다자간 상황을 지원해야 한다. 낮은 건강문해력, 청각장애, 인지 변화, 소아 환자와 보호자 등 다양한 사례도 당사자와의 공동설계와 고정관념 검토를 거쳐 포함할 필요가 있다. 현재 문헌 대부분은 이자 관계만을 다루고 있어 실제 임상에서 흔한 관계적 의사결정이 충분히 반영되지 못하고 있다.

셋째, 자동 평가는 추적 가능하고 인간 평가와 교정되어야 한다. 점수는 학습자가 실제로 말하거나 수행한 근거와 연결되어야 하며, 각 평가 기준을 뒷받침하는 원문을 확인할 수 있어야 한다. 인간 평가와의 비교는 신뢰도가 확인된 복수 평가자를 기준으로 시행해야 한다. 또한 기존 평가도구를 사용할 때에는 원도구인지, 수정한 도구인지, 해당 환경에서 타당화되었는지를 명시하고, 척도 특성과 무관한 관습적 합격선을 적용해서는 안 된다[54]. 관찰자 기반 SDM 평가도구의 점수는 실제 진료에서 척도 상한에 크게 못 미치는 범위에 분포하기 때문이다.⁸⁾ 본 검토에 포함된 문헌에서도 Observer OPTION-12를 사용한 자동 평가의 평균은 48점 만점에 25.23점(범위 7-37)이었다[28]. 아울러 일반적인 의사소통 평가들[55]과 SDM 특화 도구를 함께 사용하는 경우, 각 도구가 무엇을 보증하는지 구분하여 제시해야 한다. 나아가 신뢰도 지표나 인간 평가와의 일치도는 그 점수를 학습자의 역량에 관한 판단으로 사용하는 것을 정당화하는 근거의 일부일 뿐이며, 그 사용 목적에 대한 타당성 논증을 대신할 수 없다[51]. 학습자는 평가 근거를 열람하고 이의를 제기하며 인간의 재검토를 요청할 수 있어야 한다.

넷째, 학습자 데이터와 AI 사용은 투명하게 관리되어야 한다. 저장되는 자료, 보존기간, 사용한 모델과 프롬프트, 피드백의 생성 주체를 학습자에게 알려야 하며, 대화 기록은 가능한 한 기관의 통제 아래 두어야 한다. AI는 학습자에 대한 고부담 판단을 단독으로 확정해서는 안 된다. AI의 산출은 판단의 근거로 활용될 수 있으나, 판단의 확정과 그에 대한 응답책임은 인간 교수자에게 있어야 하며, 학습자는 그 근거를 열람하고 인간의 재검토를 요청할 수 있어야 한다. 여기서 요구되는 것은 인간이 모든 자동 산출을 독립적으로 다시 산출하는 것이 아니라, 학습자의 성취로 확정되는 판단이 인간의 책임 아래 이루어지는 것이다. 또한 인간 디브리핑과 감독을 유지해야 한다. 교육과정이 진행됨에 따라 AI의 도움을 점차 줄여 학습자가 독립적으로 판단하고 수행할 수 있는지도 확인해야 한다.

다섯째, 가상 SP 시스템은 한 차례의 실습에 그치지 않고 반복 연습, 피드백 후 재시도, 후속 재평가로 이어지는 다회차 학습을 지원해야 한다. 인간 SP를 통한 교육은 교육적 가치가 높지만, 인력, 비용, 일정 조정의 부담으로 인해 실습 횟수를 충분히 확보하기 어렵다. 반면, 가상 SP는 동일한 또는 변형된 사례를 여러 시점에 반복하고, 수행의 변화를 종단적으로 추적할 수 있다는 장점이 있다. 이는 한 번의 성공적 수행보다 숙고와 판단 능력이 반복을 거치며 형성되는지를 살피는 데 적합하다. 단, 서론에서 구분한 바와 같이 가상 환경은 환자의 침묵, 자세, 망설임과 같은 비언어적 단서와 실제 관계에서 발생하는 긴장을 재현하기 어려우므로, 임상 경험을 완전히 대체하기보다 학습자의 반복 연습을 확장하는 보완적 환경으로 활용될 수 있다.

8) 단, Observer OPTION-12와 Observer OPTION-5는 문항 수와 채점 구조가 다른 별개의 도구이므로 두 도구의 점수를 하나의 분포로 다룰 수 없다. 여기서 말하는 것은 특정 점수대가 아니라, 관찰자 기반 도구의 점수를 백분율 성적으로 환산하여 관습적 합격선을 적용하는 방식이 도구의 척도 특성과 맞지 않는다는 점이다.

4. 연구의 한계

본 검토의 방법론적 한계는 다음과 같다. 첫째, 문헌 선별과 자료 도표화를 복수의 연구자가 독립적으로 중복 수행하지 않았으며, 본 논문은 Cochrane의 축약 절차를 AI 기반으로 변용한 것이므로 이를 충족했다고 말할 수 없다. 본 연구는 제외 오류가 발생할 가능성이 더 크다고 판단하여 이를 집중적으로 확인하였다. 그러나, 이러한 확인 방식이 이중 선별을 대체할 수는 없다. 둘째, 제목·초록 단계에서 연구자가 직접 확인한 대상은 각 문헌 전체가 아니라 에이전트가 제시한 제외 근거였다. 연구자는 제외된 모든 문헌에 대하여 해당 근거가 선별 기준에 부합하는지 검토했으나, 각 문헌의 제목과 초록을 다시 읽지는 않았다. 따라서 에이전트가 문헌의 내용을 잘못 파악하고도 표면적으로 타당해 보이는 근거를 기록한 경우에는 오류를 발견하지 못했을 수 있다. 이를 보완하기 위해 AI, SDM 및 교육 관련 용어가 함께 나타난 제외 문헌은 중복 검토하였고, 이 과정에서는 제외 결정 유지 또는 상위 범주로 상승(허용/불확실)하는 것만 허용하였다. 더하여 II.4에 기술한 눈가림 재선별로 위음성률을 추정하였으나, 이는 표본에 근거한 추정이므로 개별 누락 문헌의 존재를 배제하지 못한다. 셋째, 도표화한 자료를 제2의 연구자가 독립적으로 재추출하여 대조하지 않았으므로, 개별 항목의 추출 오류가 남아 있을 수 있다. 해석이 개입하는 항목에 대해서는 조작적 정의와 판단 근거 원문을 공개하여 검증이 가능하도록 하였으나, 이는 판정을 확인할 수 있게 할 뿐 독립적으로 재산출한 것은 아니다. 발표된 문헌의 재현가능한 검색 결과에 충실하고자 하였으므로, 회색문헌 검색, 참고문헌 목록 검색, 저자 연락을 수행하지 않았고, JBI 지침에서 권고하는 제3단계 참고문헌 검색도 시행하지 않았다.

넷째, 선별 기준을 영어 문헌으로 한정하였다. 그런데 본 연구는 서론과 설계 원칙에서 한국의 가족 참여 관행과 관계적 자율성을 중요한 논거로 사용하였다. 이러한 맥락을 다룬 비영어권 문헌, 특히 한국어와 그 밖의 동아시아 언어로 발표된 SDM 교육 문헌이 검토에서 배제되었을 가능성이 있으며, 이는 다자간·가족 참여 상황을 다룬 문헌이 본 검토에서 드물게 확인된 결과에도 영향을 주었을 수 있다.

다섯째, 목표로 한 SDM 역량의 분류체계는 포함 문헌이 스스로 사용한 표현에서 귀납적으로 도출한 것이다. 따라서 다른 분류틀을 적용하였다면 범주의 경계와 빈도가 달라졌을 수 있다.

포함 문헌은 13편으로 적고 연구 설계와 내용도 이질적이므로, 이를 바탕으로 해당 분야 전체에 관한 결론을 내리기는 어렵다. 특히 토의에서 중요한 근거로 활용한 관찰 가운데 SDM 수행의 문항별 분포, 자동 평가와 인간 평가 간 차이의 방향, 관찰자 평가도구 점수가 척도 내에서 차지하는 위치는 모두 단일기관에서 58명을 대상으로 수행한 하나의 실험 가능성 연구에 근거한다[28]. 이러한 결과는 전체 문헌에서 종합적으로 확인된 양상과 일치하기 때문에 제시하였지만, 그 양상을 독립적으로 확인한 근거라기보다 이를 보여주는 사례로 해석할 필요가 있다.

마지막으로 IV.3의 설계 원칙은 규범적 논증이다. 각 설계 원칙의 도출 근거는 Table 1에 제시하였다. 다섯 가지 설계 원칙은 포함 문헌에서 확인된 결과와 공백을 출발점으로 하되, 그 구체적인 구성은 IV.2의 윤리적 분석을 통해 제안하였다. 이 설계 원칙들은 향후 시스템을 평가하기 위한 하나의 설계 명세로 제안되는 것이며, 그 교육적 효과를 주장하기에 앞서 원칙 자체에 대한 실증적 평가가 필요하다.

V. 결론

현재 인공지능은 임상 의사 SDM을 더 잘 수행하도록 만드는 것보다, 그 수행을 더 정교하게 측정하는 방향으로 훨씬 빠르게 발전하고 있다. 포함된 13편의 문헌의 가장 두드러진 기여는 전사문, 음성, 영상 및 의무기록을 자동으로 분석하고 점수화하는 평가 인프라였다. 반면 직접적인 교육 효과에 관한 근거는 무작위시험 1편과 대조군 없는 소수의 연구에 의존하였으며, 임상 실무로의 전이나 환자와 가족의 편익을 입증한 연구는 없었다.

무엇보다 포함된 문헌은 동의의 정보적 측면을 주로 조작화한 반면, 숙의적 측면은 거의 다루지 않았다. 그 결과 SDM이 교정하고자 했던 문제, 즉 팀의 도덕적 사고 및 행위를 정보 전달의 기록으로 축소하는 문제를 수치화된 점수의 권위와 함께 재생산할 위험이 있다. 환자의 가치와 선호는 거의 모든 문헌에서 명시적 목표로 언급되었지만, 윤리적 추론은 단 한 편에서만 명시되었다. 교육 상황을 모형화한 5편 가운데 무엇이 선택의 대상인지를 명시한 것이 2편에 그쳤다는 점도 같은 방향을 가리킨다.

따라서 숙고를 반복적으로 연습하고, 측정도구가 무엇을 평가하는지 명확히 밝히며, 학습자가 평가 근거를 확인하고 이의를 제기할 수 있는 절차 안에서 수행을 판단하는 교육 환경을 구축하고 검증하는 일이 향후 과제로 제시된다. 본 연구는 이러한 환경이 갖추어야 할 조건에 대한 다섯 가지 설계 원칙을 제시하였다. 단 이는 문헌에서 확인된 한계를 바탕으로 구성한 규범적 제안이므로, 실제 교육 효과 입증을 위한 비교 연구와 장기 추적 연구를 시행할 필요가 있다. 

SUPPLEMENTARY MATERIALS

Supplementary materials are only available online from: <https://doi.org/10.35301/ksme.2026.29.3.163>

REFERENCES

1. Makoul G, Clayman ML. An integrative model of shared decision making in medical encounters. *Patient Educ Couns* 2006;60(3):301-312.
<https://doi.org/10.1016/j.pec.2005.06.010>
2. Elwyn G, Frosch D, Thomson R, et al. Shared decision making: a model for clinical practice. *J Gen Intern Med* 2012;27(10):1361-1367.
<https://doi.org/10.1007/s11606-012-2077-6>
3. Elwyn G, Durand MA, Song J, et al. A three-talk model for shared decision making: multistage consultation process. *BMJ* 2017;359:j4891.
<https://doi.org/10.1136/bmj.j4891>
4. Emanuel EJ, Emanuel LL. Four models of the physician-patient relationship. *JAMA* 1992;267(16):2221-2226.
<https://doi.org/10.1001/jama.1992.03480160079038>
5. Whitney SN, McGuire AL, McCullough LB. A typology of shared decision making, informed consent, and simple consent. *Ann Intern Med* 2004;140(1):54-59.
<https://doi.org/10.7326/0003-4819-140-1-200401060-00012>
6. Beauchamp TL, Childress JF. Principles of biomedical ethics. 7th ed. Oxford University Press; 2013.

7. Sandman L, Munthe C. Shared decision making, paternalism and patient choice. *Health Care Anal* 2010;18(1):60-84.
<https://doi.org/10.1007/s10728-008-0108-6>
8. Carel H, Kidd IJ. Epistemic injustice in healthcare: a philosophical analysis. *Med Health Care Philos* 2014;17(4):529-540.
<https://doi.org/10.1007/s11019-014-9560-2>
9. Entwistle VA, Carter SM, Cribb A, et al. Supporting patient autonomy: the importance of clinician-patient relationships. *J Gen Intern Med* 2010;25(7):741-745.
<https://doi.org/10.1007/s11606-010-1292-2>
10. Mo HN, Shin DW, Woo JH, et al. Is patient autonomy a critical determinant of quality of life in Korea? End-of-life decision making from the perspective of the patient. *Palliat Med* 2012;26(3):222-231.
<https://doi.org/10.1177/0269216311405089>
11. Coates D, Clerke T. Training interventions to equip health care professionals with shared decision-making skills: a systematic scoping review. *J Contin Educ Health Prof* 2020;40(2):100-119.
<https://doi.org/10.1097/CEH.0000000000000289>
12. Ørom ASK, Musaeus P. Shared decision-making in medical education: a systematic review. *Dan Med J* 2026;73(4):A04250257.
<https://doi.org/10.61409/A04250257>
13. McCarrick CA, Moynihan A, McEntee PD, et al. Impact of simulation training on communication skills and informed consent practices in medical students: a randomised controlled trial. *BMC Med Educ* 2025;25(1):1078.
<https://doi.org/10.1186/s12909-025-07671-0>
14. Rutherford-Hemming T, Herrington A, Ngo TP. The use of standardized patients to teach communication skills: a systematic review. *Simul Healthc* 2024;19(1S):S122-S128.
<https://doi.org/10.1097/SIH.0000000000000766>
15. Hauer KE, Fernandez A, Teherani A, et al. Assessment of medical students' shared decision-making in standardized patient encounters. *J Gen Intern Med* 2011;26(4):367-372.
<https://doi.org/10.1007/s11606-010-1567-7>
16. Colanhese M, Lima SF, Negri EC, et al. Cost analysis of the OSCE: scoping review. *Simul Healthc* Forthcoming 2026.
<https://doi.org/10.1097/SIH.0000000000000910>
17. Sibbald M, Last N, Keuhl A, et al. Challenges facing standardised patients representing equity-deserving groups: insights from health care educators. *Med Educ* 2023;57(6):516-522.
<https://doi.org/10.1111/medu.15085>
18. Pel-Littel RE, Snaterse M, Teppich NM, et al. Barriers and facilitators for shared decision making in older patients with multiple chronic conditions: a systematic review. *BMC Geriatr* 2021;21(1):112.
<https://doi.org/10.1186/s12877-021-02050-y>
19. Jacklin S, Maskrey N, Chapman S. Shared decision-making with a virtual patient in medical education: mixed methods evaluation study. *JMIR Med Educ* 2021;7(2):e22745.
<https://doi.org/10.2196/22745>
20. Gilbert V, Philip P, Llorca PM, et al. Use of artificial intelligence-enhanced virtual patients in educational approaches to medical interview training: a systematic review. *BMC Med Educ* 2026;26(1):588.
<https://doi.org/10.1186/s12909-026-08804-9>
21. Jiang J, Ye MZ, Kwok TTO, et al. GenAI-supported virtual patients in health care education: systematic review. *J Med Internet Res* 2026;28:e82756.
<https://doi.org/10.2196/82756>

22. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR Med Educ* 2023;9:e48785.
<https://doi.org/10.2196/48785>
23. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. *Med Teach* 2023;45(6):574-584.
<https://doi.org/10.1080/0142159X.2023.2186203>
24. Hui LM, Yu E, Chung A, et al. Artificial intelligence for assessment in competency-based medical education: current practices and future directions. *Postgrad Med J* 2026;102(1211):806-816.
<https://doi.org/10.1093/postmj/qgaf195>
25. Masters K, MacNeil H, Benjamin J, et al. Artificial intelligence in health professions education assessment: AMEE Guide No. 178. *Med Teach* 2025;47(9):1410-1424.
<https://doi.org/10.1080/0142159X.2024.2445037>
26. Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, et al. Application of artificial intelligence in shared decision making: scoping review. *JMIR Med Inform* 2022;10(8):e36199.
<https://doi.org/10.2196/36199>
27. Kim YM, Lee YM, Kim DH, et al. Artificial intelligence-assisted shared decision-making training for medical students transitioning to residency. *Acad Med* 2026;101(1):48-53.
<https://doi.org/10.1093/acamed/wvaf006>
28. Herschbach L, Festl-Wietek T, Sonanini A, et al. AI-simulated patients for training shared decision-making: feasibility study in medical education. *JMIR Med Educ* 2026;12:e100467.
<https://doi.org/10.2196/100467>
29. Gao C, Fan Y, Yao F, et al. The VALUE of AI-guided communication: enhancing shared decision-making in metabolic bariatric surgery consultations through a metacognitive framework. *Obes Surg* 2026;36(7):3497-3510.
<https://doi.org/10.1007/s11695-026-08760-9>
30. Selvaraj SP, Yen RW, Forcino R, et al. Automating the Observer OPTION-5 measure of shared decision making: assessing validity by comparing large language models to human ratings. *Patient Educ Couns* 2026;142:109362.
<https://doi.org/10.1016/j.pec.2025.109362>
31. Sullivan A, Geis G, Cummings C. Ethics education in neonatology: integrating theory, multimodal methods, and AI innovation. *Neoreviews* 2026;27(2):e73-e83.
<https://doi.org/10.1542/neo.27-2-099>
32. World Medical Association. WMA statement on artificial and augmented intelligence in medical care [internet]. World Medical Association; 2025 [cited 2026 Aug 3].
<https://www.wma.net/policies-post/wma-statement-on-artificial-and-augmented-intelligence-in-medical-care/>
33. Müller Y, Lemberg H, Gellert P, et al. Training empathetic communication skills in medical students with a role-prompted GPT-4o chatbot: quasi-experimental intervention study. *JMIR Med Educ* 2026;12:e88092.
<https://doi.org/10.2196/88092>
34. Peters MDJ, Godfrey C, McInerney P, et al. Scoping reviews. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. *JBIMES* manual for evidence synthesis. JBI; 2024.
<https://doi.org/10.46658/JBIMES-24-09>
35. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018;169(7):467-473.
<https://doi.org/10.7326/M18-0850>

36. Garritty C, Gartlehner G, Nussbaumer-Streit B, et al. Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews. *J Clin Epidemiol* 2021;130:13-22.
<https://doi.org/10.1016/j.jclinepi.2020.10.007>
37. McGowan J, Sampson M, Salzwedel DM, et al. PRESS peer review of electronic search strategies: 2015 guideline statement. *J Clin Epidemiol* 2016;75:40-46.
<https://doi.org/10.1016/j.jclinepi.2016.01.021>
38. Thomas J, McDonald S, Noel-Storr A, et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *J Clin Epidemiol* 2021;133:140-151.
<https://doi.org/10.1016/j.jclinepi.2020.11.003>
39. van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nat Mach Intell* 2021;3(2):125-133.
<https://doi.org/10.1038/s42256-020-00287-7>
40. Ouzzani M, Hammady H, Fedorowicz Z, et al. Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 2016;5(1):210.
<https://doi.org/10.1186/s13643-016-0384-4>
41. Flemyng E, Noel-Storr A, Macura B, et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence 2025. *JBIEvid Synth* 2025;23(11):2162-2166.
<https://doi.org/10.11124/JBIES-25-00480>
42. Waffenschmidt S, Knelangen M, Sieben W, et al. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Med Res Methodol* 2019;19(1):132.
<https://doi.org/10.1186/s12874-019-0782-0>
43. Oami T, Okada Y, Nakada TA. Performance of a large language model in screening citations. *JAMA Netw Open* 2024;7(7):e2420496.
<https://doi.org/10.1001/jamanetworkopen.2024.20496>
44. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. *Med Teach* 2025;47(1):40-42.
<https://doi.org/10.1080/0142159X.2024.2376879>
45. Zheng R, Friedrich NA, Luu M, et al. Development and validation of a natural language processing system to assess quality of physician communication in prostate cancer consultations. *Prostate Cancer Prostatic Dis* 2026;29(2):338-345.
<https://doi.org/10.1038/s41391-025-01011-5>
46. Lopez-Garcia G, Xu D, Luu M, et al. Scoring physician risk communication in prostate cancer using large language models. *medRxiv* Forthcoming 2025.
<https://doi.org/10.1101/2025.08.07.25333034>
47. Topaz M, Zolnoori M, Norful AA, et al. Speech recognition can help evaluate shared decision making and predict medication adherence in primary care setting. *PLOS ONE* 2022;17(8):e0271884.
<https://doi.org/10.1371/journal.pone.0271884>
48. Khaikin S, Tiruvadi V, Brooks J, et al. Machine learning-based audiovisual phenotyping for measuring communication, shared decision-making, and trust. *JMIR Form Res* 2026;10:e85906.
<https://doi.org/10.2196/85906>
49. Streid JL, Lee KC, Bader AM, et al. Shared decision making in the Geriatric Surgery Verification program: assessing baseline performance. *J Pain Symptom Manage* 2023;65(6):510-520.e3.
<https://doi.org/10.1016/j.jpainsymman.2023.01.018>

50. Lambrakos LK, Feigofsky SA, Wang Y, et al. Enhancing patient acceptance of ICD implantation through structured shared decision making: conversation is key. *J Interv Card Electrophysiol* 2024;67(9):2077-2087.
<https://doi.org/10.1007/s10840-024-01850-7>
51. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas* 2013;50(1):1-73.
<https://doi.org/10.1111/jedm.12000>
52. Cruess RL, Cruess SR, Boudreau JD, et al. A schematic representation of the professional identity formation and socialization of medical students and residents: a guide for medical educators. *Acad Med* 2015;90(6):718-725.
<https://doi.org/10.1097/ACM.0000000000000700>
53. Pellegrino ED, Thomasma DC. *The virtues in medical practice*. Oxford University Press; 1993.
54. Norcini J, Anderson MB, Bollela V, et al. 2018 consensus framework for good assessment. *Med Teach* 2018;40(11):1102-1109.
<https://doi.org/10.1080/0142159X.2018.1500016>
55. Makoul G. Essential elements of communication in medical encounters: the Kalamazoo consensus statement. *Acad Med* 2001;76(4):390-393.
<https://doi.org/10.1097/00001888-200104000-00021>