

# AI와 의료의 대전환





# 차례

|                                               |           |
|-----------------------------------------------|-----------|
| <b>Part 1. 진료 기록과 임상 업무 자동화</b>               | <b>7</b>  |
| <b>제1장. AI 스크라이브와 기록 업무</b>                   | <b>9</b>  |
| 1.1 진료 음성 인식과 기록 시간 단축                        | 10        |
| 1.2 예일과 클리블랜드 클리닉의 도입 사례                      | 15        |
| 1.3 의료진 번아웃과 근무 외 기록 업무                       | 19        |
| <b>제2장. 임상 의사결정 지원(CDSS)과 진단</b>              | <b>24</b> |
| 2.1 거대언어모델(LLM) 기반 항생제 처방 지원                  | 25        |
| 2.2 의료 세계 모델과 U2-Med의 진단·검사 요청                | 31        |
| 2.3 AMIE와 MIRA의 진단 추론                         | 37        |
| <b>Part 2. AI 신약개발과 디지털 트윈</b>                | <b>44</b> |
| <b>제3장. 바이오 파운데이션 모델과 신약 표적</b>               | <b>46</b> |
| 3.1 의료 멀티모달 파운데이션 모델                          | 47        |
| 3.2 Octopus의 대장암 바이오마커(IGF2) 탐색               | 56        |
| 3.3 세포 세계 모델(Chreode)과 분자 설계                  | 62        |
| <b>제4장. 디지털 트윈(Health Digital Twin)과 임상시험</b> | <b>71</b> |
| 4.1 환자 특성을 반영한 디지털 트윈                         | 72        |

|                                                                 |            |
|-----------------------------------------------------------------|------------|
| 4.2 가상 대조군(Synthetic Control Arm) 임상시험 . . . . .                | 79         |
| 4.3 모델 기반 약물 개발과 동물실험 대체(FDA Modernization Act) . . . . .       | 85         |
| <b>Part 3. 해외 의료 AI 도입 사례</b>                                   | <b>94</b>  |
| <b>제5장. 미국 의료기관의 AI 도입 . . . . .</b>                            | <b>96</b>  |
| 5.1 메이요 클리닉의 VoiceCare AI와 원무 자동화 . . . . .                     | 97         |
| 5.2 Tempus의 정밀 종양학 지원 . . . . .                                 | 103        |
| <b>제6장. 중국의 의료 AI 도입 . . . . .</b>                              | <b>110</b> |
| 6.1 하북의대 부속병원의 AI 운영 성과 . . . . .                               | 111        |
| 6.2 WiNGPT 3.0과 엔트아푸 서비스 . . . . .                              | 116        |
| 6.3 마인드레이 启元과 1차 의료 디지털화 . . . . .                              | 122        |
| <b>Part 4. 의료 AI의 규제·안전·경제성</b>                                 | <b>128</b> |
| <b>제7장. AI 의료기기 규제와 PCCP(사전변경허가계획) . . . . .</b>                | <b>130</b> |
| 7.1 PCCP 지침과 AI 모델 변경 관리 . . . . .                              | 131        |
| 7.2 FDA 역량 평가(Competency Assessment)와 위험 평가 . . . . .           | 137        |
| 7.3 영국 AI Airlock과 유럽 AI Act . . . . .                          | 143        |
| <b>제8장. 환자 안전과 의료 AI의 경제성 . . . . .</b>                         | <b>149</b> |
| 8.1 통역 오류 전파(Interpreter Error Propagation)와 음성 인식 편향 . . . . . | 150        |
| 8.2 자동화 편향(Automation Bias)과 인간 검토(Human-in-the-loop) . . . . . | 155        |

|                                                    |            |
|----------------------------------------------------|------------|
| 8.3 의료 에이전틱 AI 시장과 CPT 수가 . . . . .                | 161        |
| <b>Part 5. 의료 AI 에이전트와 협진</b>                      | <b>168</b> |
| <b>제9장. 의료 에이전틱 AI(Agentic AI) . . . . .</b>       | <b>170</b> |
| 9.1 단일 예측 모델의 한계 . . . . .                         | 171        |
| 9.2 인지·사고·행동(Sense-Think-Act)과 메모리 . . . . .       | 177        |
| 9.3 의료 에이전틱 AI의 분류와 기술 성숙도 . . . . .               | 183        |
| <b>제10장. 다중 에이전트(Multi-Agent)와 AI 병원 . . . . .</b> | <b>188</b> |
| 10.1 다중 에이전트 협업 설계 . . . . .                       | 189        |
| 10.2 MAI-DxO의 역할 분담과 복합 진단 . . . . .               | 195        |
| 10.3 ClinVerse와 AI 병원의 진료 과정 . . . . .             | 200        |
| <b>Part 6. 미래 의료와 AI</b>                           | <b>207</b> |
| <b>제11장. 의사와 AI 에이전트의 협업 . . . . .</b>             | <b>209</b> |
| 11.1 AI 선별·인간 검토·종합 진단 . . . . .                   | 210        |
| 11.2 의사의 역할과 종합적 임상 판단 . . . . .                   | 217        |
| <b>제12장. 미래 의료 AI의 기술·윤리 과제 . . . . .</b>          | <b>223</b> |
| <b>참고문헌</b>                                        | <b>232</b> |
| <b>책 정보</b>                                        | <b>282</b> |



## Part 1. 진료 기록과 임상 업무 자동화



## 제1장. AI 스크라이브와 기록 업무



## 1.1 진료 음성 인식과 기록 시간 단축

TXP Medical의 제품 발표문에는 SpeechER가 음성을 어떤 기록으로 바꾸는지와 시간 절감 수치의 조사 주체가 나와 있다. 기능 설명과 효과 주장을 나누어 읽을 수 있는 자료다.

<https://txpmedical.jp/en/news/2jmiSPoM4bpIdl9GUZF0BI/>

진료 기록은 환자의 증상과 검사, 치료 계획을 전달하고 이후 진료를 이어가는 데 필요한 자료다. 그러나 진료 중 입력과 진료 후 문서 작성은 의료진의 업무 부담을 높일 수 있다. 미국 의료계에서 사용하는 파자마 타임(pajama time)은 퇴근 후에도 이어지는 전자의무기록(EHR, Electronic Health Record) 작성 업무를 가리킨다.

AI 스크라이브(ambient AI scribe)는 진료 대화를 인식하고 기록 초안을 만드는 기술이다. 스크라이브(scribe)는 기록 담당자를 뜻하며, 앰비언트(ambient)는 이 기술에서 주변의 대화를 별도 입력 없이 처리하는 방식을 나타낸다. 의료진은 환자와 대화하고, 시스템은 필요한 정보를 문서 형태로 정리한다.

음성 인식과 기록 작성은 서로 다른 작업이다. 음성 인식은 들린 말을 글자로 옮긴다. 기록 작성은 그중 진료에 필요한 내용을 고르고 알맞은 항목에 배치한다. 예를 들어 환자의 어제는 아팠지만 지금은 괜찮다는 말을 옮길 때에는 증상의 시간 변화가 남아 있어야 한다. 통증이 있다고만 줄이면 현재 상태를 잘못 전달할 수 있다. 이 예시는 기록 오류의 원리를 설명하기 위한 것이다.

## 진료 음성을 기록으로 바꾸는 과정

첫 단계는 대화의 수집이다. 마이크나 스마트폰 앱으로 음성을 기록한다. 환자 동의와 녹음 자료의 처리 방식은 해당 지역의 법과 의료기관 정책에 맞게 정해야 한다.

자동 음성 인식(ASR, Automatic Speech Recognition)은 수집한 음성을 텍스트로 변환한다. 이 과정에서는 의학 용어와 약물명, 화자의 구분, 주변 소음 등이 정확도에 영향을 줄 수 있다.

거대언어모델(LLM, Large Language Model)은 텍스트에서 증상과 복용 약물, 진찰 소견, 치료 계획 등을 정리한다. SOAP 형식에 따라 주관적 호소(Subjective), 객관적 소견(Objective), 평가(Assessment), 계획(Plan)으로 나누어 작성할 수도 있다. 제품에 따라 대화 중 처리하거나 진료 후 초안을 생성한다.

SOAP에서 환자가 느낀 증상과 의료진이 측정한 수치를 나누는 이유도 여기에 있다. 열이 나는 것 같다는 호소와 체온계로 측정한 결과는 출처가 다르다. AI가 둘을 섞으면 나중에 기록을 읽는 사람이 무엇을 실제로 확인했는지 알기 어렵다. 좋은 초안은 문장을 매끄럽게 만드는 데 더해 정보의 출처와 시점을 보존해야 한다.

마지막으로 의료진이 초안을 검토하고 수정해 확정한다. AI가 작성한 문서는 그 자체로 확인된 기록이 아니다. 사람이 처리 과정에 참여하는 휴먼인 더 루프(Human-in-the-loop) 구조에서 기록의 내용과 정확성을 확인해야 한다.

## 입력 시간과 검토 시간의 변화

AI 스크라이브는 직접 입력하는 시간을 줄이는 대신 초안을 확인하는 업무를 만든다. 따라서 전체 시간 절감은 입력 감소와 검토·수정 시간을 함께 고려해야 한다. 기록 형식과 진료과, 환자와의 대화 특성에 따라 효과도 달라질 수 있다.

예를 들어 직접 쓰던 시간이 10분에서 4분으로 줄어도 초안 확인에 새로 3분이 필요하면 전체 기록 업무는 7분이다. 이 숫자는 계산 방법을 보여 주는 가상 예다. 도입 전에는 입력 시간만 재고 도입 후에는 초안 생성 시간만 재는 식으로 비교하면 절감 효과가 부풀려진다. 환자 한 명의 기록을 시작해서 확정할 때까지 같은 범위로 시간을 재야 한다.

일본의 SpeechER와 간호 기록 지원 도구 Caretomo는 음성 입력을 문서로 정리하는 사례다.

SpeechER의 개발사는 일본 TXP Medical이다. 회사가 2025년 10월 15일 공개한 수상 발표문에는 이미 상용 앱으로 출시되었다고 적혀 있다. 이 날짜는 해당 발표의 날짜이며 최초 출시일로 읽어서는 안 된다.

Caretomo는 일본 SIND가 제공하는 간호 기록 지원 제품이다. 공식 제품 설명에서 음성 인식, SOAP 분류, 확인·수정 후 등록 순서를 확인할 수 있으나 해당 제품 소개에는 최초 출시일이 명시되어 있지 않다. 두 이름은 범용 AI 모델의 이름이 아니라 의료 업무에 맞게 구성한 제품 이름이다. 업체들은 자체 조사와 단기 시험에서 기록 시간 감소를 발표했다. 다만 이 결과를 모든 기관에서 기대할 수 있는 감소율로 일반화하기는 어렵다.

비교 연구에서도 제품과 지표에 따른 차이가 나타났다. UCLA 연구에서는 한 제품의 기록 작성 시간이 감소했지만 다른 제품은 대조군과 뚜렷한 차이가 없었다. 여러 의료기관의 운영 자료를 분석한 연구는 문서 작성 시간의 감소를 보고했으나, 근무 외 업무가 함께 줄어든 것은 아니었다.

따라서 특정한 70% 단축 수치를 기술 전체의 보편적 효과로 소개하는 것은 적절하지 않다. 기록 입력, 문서 작성, EHR 사용, 진료 의사결정은 서로 다른 업무다. 예를 들어 원격 심전도 플랫폼의 판단 시간 단축은 기록 작성 시간과 구분해야 한다.

### 시간 절감에 영향을 주는 조건

실제 효과는 사용 빈도와 업무 방식에 영향을 받는다. 초안을 자주 사용하는 의료진과 거의 사용하지 않는 의료진의 결과가 다를 수 있다. 서식과 용어가 진료과에 맞지 않으면 수정 부담이 커진다.

AI와 함께 진료 절차와 데이터 관리가 바뀌는 경우도 있다. 병원의 효율이 개선되더라도 변화 전체가 AI만으로 발생했다고 단정하기 어렵다. 도입 효과를 살필 때는 어떤 업무가 달라졌고 어떤 지원이 함께 제공되었는지 확인해야 한다.

### 기록 정확성과 안전

음성 인식과 내용 정리에서는 누락이나 잘못된 추가, 의미의 반전이 발생할 수 있다. 통역을 포함한 모의 진료 연구에서는 통역 오류가 AI 기록에 전달되는 문제가 나타났다. 환자가 말한 증상이 기록에서는 없는 것으로 정리된 사례도 있었다.

자동화 편향(automation bias)은 사용자가 AI 결과를 지나치게 신뢰하는 경향이다. 문장이 자연스럽게 작성되어 있으면 약물 용량과 알레르기, 증상의 유무 같은 오류를 놓칠 수 있다. 초안 검토를 형식적인 서명 절차로 운영해서는 안 된다.

AI 스크라이브의 도입 목적은 기록 작성의 부담을 줄이면서 필요한 정보의 정확성을 유지하는 데 있다. 의료기관은 작성 시간과 수정 부담, 환자와의 대화, 기록 품질을 함께 평가해야 한다. 반복 입력을 줄인 효과가 진료 개선으로 이어지도록 업무 절차와 검토 여건도 마련해야 한다.

## 1.2 예일과 클리블랜드 클리닉의 도입 사례

클리블랜드 클리닉의 2025년 운영 소개에는 선정한 제품의 개발사와 실제 기록 생성 과정이 설명되어 있다. 예일 연구진의 조사 결과는 다음 절 첫머리에 안내한 원논문에서 확인할 수 있다.

<https://consultqd.clevelandclinic.org/less-typing-more-talking-how-ambient-ai-is-reshaping-clinical-workflow-at-cleveland-clinic>

AI 스크라이브를 병원 전체에 적용하려면 제품의 성능과 함께 사용자 교육, 지원, 진료과별 서식을 준비해야 한다. 소규모 시험에서 사용할 수 있었던 기능도 많은 의료진이 서로 다른 업무에 적용하면 새로운 요구가 나타난다.

예일 뉴헤이븐 헬스(Yale New Haven Health)와 클리블랜드 클리닉(Cleveland Clinic)은 이러한 도입 과정을 보여 주는 사례다. 예일은 의료진의 업무 부담을 평가하는 연구에 참여했고, 클리블랜드 클리닉은 제품 비교와 대규모 보급 과정을 공개했다.

예일 연구에서 사용한 Abridge는 미국 Abridge가 개발한 서비스다. 회사의 공식 출시 발표는 2020년 10월 6일로, 당시에는 환자가 진료 대화를 기록하고 약 복용 안내 등을 다시 확인하는 서비스에 초점이 있었다. 뒤에 병원이 도입한 의료진용 기록 제품과 초기 환자용 서비스의 기능 범위는 같지 않다. 회사 설립, 서비스 공개, 특정 병원의 도입은 별개의 사건으로 구분해야 한다.

## 예일의 도입과 업무 부담 평가

예일 연구진은 여러 의료 시스템의 의료진을 대상으로 AI 스크라이브(Abridge) 사용 전후의 업무 경험을 조사했다. 번아웃과 기록 부담이 줄어드는 경향이 보고되었지만, 도입 전후 비교이므로 변화 전체를 AI의 효과로 단정할 수는 없다.

예일이 참여한 다른 다기관 연구에서도 문서 작성 시간의 감소가 나타났다. 다만 근무 외 업무 시간이 함께 줄어든 것은 아니었다. 이 결과는 제품 도입의 효과를 기록 속도와 실제 업무 환경 양쪽에서 평가해야 함을 보여준다.

예일은 앰비언트 AI를 진료에 확대하면서 수련 교육의 문제도 검토했다. 기록 초안에 의존하더라도 의대생과 수련의가 직접 내용을 해석하고 작성하는 능력을 유지해야 한다. 역량 약화(de-skilling)를 방지하는 교육은 도입 이후에도 필요한 과제다.

## 클리블랜드 클리닉의 제품 선정

클리블랜드 클리닉은 여러 AI 스크라이브를 시범 운영한 뒤 앰비언스 헬스케어(Ambience Healthcare)의 제품을 선택했다. 다양한 전문 분야의 의료진을 참여시켜 실제 업무에 맞는지 검토했다.

이 병원이 선택한 개발사는 미국 Ambience Healthcare다. 병원의 2025년 8월 14일 소개에 따르면 환자의 동의를 받고 스마트폰 앱으로 대화를 수집한 뒤 Epic 전자의무기록에 의료 문서와 진료 후 요약を 만든다. 2025년은 이 병원의 도입 경과를 확인하는 시점이며 제품 자체가 처음 개발된 해라는 뜻은 아니다. 병원은 치료와 기록의 책임을 맡고, 개발사는 그

업무를 지원하는 소프트웨어를 제공한다. 새로운 기술에 관심이 많은 사용자뿐 아니라 기록 업무의 부담이 큰 사용자도 포함했다.

시범 운영은 기능 비교와 사용자 경험을 확인하는 과정이었다. 이후 외래 진료 전반으로 보급을 확대했고, 비교적 짧은 기간에 수천 명의 의료진이 사용을 시작한 것으로 보고했다. 다만 기관과 공급기업의 운영 보고는 독립적인 임상 효과 평가와 구분해야 한다.

### 대규모 도입의 운영 체계

거버넌스(governance)는 의사결정과 안전 기준을 정하는 체계다. 클리블랜드 클리닉은 경영진과 임상 지도부, 의료정보, IT, 진료비 청구 등 관련 부서가 함께 참여하도록 구성했다. 기록 작성 도구라도 여러 업무에 영향을 주기 때문이다.

예를 들어 초안의 병명 표현이 바뀌면 다음 의사의 판단뿐 아니라 청구 담당자의 분류에도 영향을 줄 수 있다. 따라서 병원이 정할 사항에는 어떤 항목을 자동으로 채울지, 누가 확정할지, 오류를 발견하면 어디에 알릴지가 포함된다. 한 부서에서 편리한 기능이 다른 부서에서는 확인할 일을 늘릴 수 있으므로 전체 업무 흐름을 함께 살피는 것이다.

교육은 사용 전에 필요한 절차로 운영되었다. 진료 분야를 나누어 순차적으로 도입하고, 온라인 교육으로 기본 사용법과 검토 책임을 안내했다. 복잡한 진료과에서는 모델과 서식을 조정할 시간이 필요했다.

사용자 지원은 설정과 초안 수정, 병원 정책 등의 문제를 해결하는 데 사용되었다. 여러 의료진이 함께 진료하거나 특수한 서식이 필요한 경우에는 제품의 기본 설정만으로 충분하지 않을 수 있다.

운영 자료와 사용자 의견을 반영하는 개선 과정도 마련했다. 진료과별 교육과 사용 현황을 확인하고, 사용자 모임을 통해 기록 형식을 보완했다. 제품 제공 이후에도 실제 사용 조건을 계속 조정해 나갔다.

## 사용 범위와 만족도의 해석

기관 보고는 많은 진료에서 AI 초안이 생성되고 사용자가 편의를 경험했음을 보여 준다. 그러나 최초 사용과 지속적인 사용은 다르며, 만족도가 높다는 사실만으로 기록의 정확성이나 이직률 개선이 입증되지는 않는다.

진료과별 차이도 고려해야 한다. 대화의 길이와 복잡성, 필요한 기록 형식에 따라 초안 수정의 부담이 달라진다. 외래에서 확인한 효과가 응급실과 입원 병동, 간호 업무에 그대로 적용된다고 볼 수도 없다.

서터 헬스(Sutter Health) 역시 AI 스크라이브를 외래와 응급·입원 영역으로 확대하는 사례다. 이런 흐름은 사용 범위가 넓어지고 있음을 보여 주지만, 누적 기록 건수와 진료 품질의 개선은 별도로 평가해야 한다.

## 도입 사례의 공통 의미

예일의 사례는 의료진의 업무 경험을 측정하는 필요성을, 클리블랜드 클리닉의 사례는 대규모 도입을 지원하는 조직 체계의 중요성을 보여 준다. 기술의 기능뿐 아니라 제품 선정, 교육, 지원, 서식 개선이 실제 사용에 영향을 준다.

AI 스크라이브를 안정적으로 운영하려면 의료진이 초안을 확인할 수 있어야 하고, 문제가 생기면 지원받고 수정할 수 있어야 한다. 도입 규모를 늘리는 과정에서도 기록 품질과 검토 부담을 함께 살피는 것이 필요하다.

### 1.3 의료진 번아웃과 근무 외 기록 업무

예일 연구진 등이 발표한 JAMA Network Open 원논문은 도입 전후 설문과 번아웃 추정치를 함께 제시한다. 참여 인원, 분석에 포함된 인원, 대조군의 유무를 확인하며 읽으면 수치의 의미를 이해하기 쉽다.

<https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2839542>

AI 스크라이브의 도입 목적은 기록 시간을 줄이는 데 그치지 않는다. 기록 업무에 따른 부담을 낮춰 의료진이 환자 진료에 집중할 수 있도록 하는 것도 중요한 목적이다. 진료가 끝난 뒤까지 이어지는 문서 작업은 의료진의 피로를 높이고 일과 생활의 균형을 어렵게 한다.

번아웃(burnout)은 업무로 인한 소진 상태를 뜻한다. 의료진에게는 지속적인 피로, 업무에 대한 의욕 저하, 환자와의 정서적 거리감 등으로 나타날 수 있다. 원인은 근무시간, 인력 부족, 조직 환경 등 다양하며, 진료 기록 작성도 업무 부담의 한 요인이다.

#### 기록 부담과 번아웃의 변화

예일 연구진이 주도한 미국 6개 의료 시스템의 조사에서는 AI 스크라이브 도입 후 번아웃과 기록 업무 부담이 줄어드는 경향이 나타났다. 의료진은 같은 회사의 AI 스크라이브(Abridge)를 사용했고, 도입 전과 30일 사용 후에 업무 경험을 평가했다. 번아웃 추정 비율은 51.9%에서 38.8%로 낮아졌다.

이 값은 조사 참여자 전원을 그대로 나눈 비율이 아니라 여러 특성을 보정한 모형의 추정치다. 전후 설문을 마친 참여자는 263명이었지만 해당 보

정 분석에 포함된 사람은 184명이었다. 설문에 응답했다는 사실과 특정분석에 필요한 항목이 모두 있다는 사실이 다르기 때문이다. 다만 대조군없이 도입 전후를 비교한 설문 결과이므로, 변화 전체를 AI의 효과로 단정할 수는 없다.

이 조사의 의미는 의료 AI를 평가하는 기준이 기록 속도에서 의료진의 업무 환경으로 확대되고 있다는 데 있다. 기록을 얼마나 빨리 작성하는지뿐 아니라, 진료 중 집중하기 쉬워졌는지, 업무를 마친 뒤에도 문서 작업이 이어지는지 등을 함께 살펴보기 시작한 것이다.

위스콘신 대학교 병원(UW Health)과 UCLA 헬스의 연구에서도 AI 스크라이브 사용 후 기록 부담이나 번아웃이 개선되는 경향이 보고되었다. 다만 연구마다 도구, 사용 기간, 업무 환경이 달라 같은 크기의 효과를 기대하기는 어렵다.

## 업무 부담이 줄어드는 과정

AI가 대화를 기록하고 초안을 작성하면 의사는 환자의 설명을 듣는 동시에 입력해야 하는 부담을 덜 수 있다. 화면과 환자를 번갈아 보는 횟수가 줄면 대화에 집중하기도 쉬워진다.

진료 내용을 듣고 판단하는 일과 이를 기록하는 일을 반복해서 전환하는 과정도 부담을 준다. 이러한 작업 전환에는 문맥 전환(context switching) 비용이 따른다. AI가 기록 초안을 맡으면 의사는 환자의 증상과 치료 방향을 검토하는 데 집중할 여지가 생긴다. 다만 이런 변화가 번아웃 감소로 이어지는 구체적인 경로는 추가 연구가 필요하다.

퇴근 전에 기록을 마칠 수 있는지도 중요하다. 근무 외 문서 작업이 줄면 휴식 시간을 확보하기 쉬워진다. 그러나 진료 기록을 작성하는 시간이 줄었다고 해서 퇴근 후 업무가 같은 정도로 줄어드는 것은 아니다. 다른 행정 업무가 남아 있거나 AI 초안을 검토하는 데 시간이 필요할 수 있기 때문이다.

### 체감 부담과 실제 업무 시간

의료진이 느끼는 부담과 시스템에 기록된 업무 시간은 구분해서 볼 필요가 있다. 여러 의료기관의 전자의무기록(EHR) 사용 자료를 분석한 연구에서는 문서 작성 시간이 줄었지만, 근무 외 업무 시간은 뚜렷하게 감소하지 않았다. 보스턴의 병원 체계에서도 진료 기록을 작성하는 시간은 줄었으나 근무 외 기록 마감에는 차이가 없었다.

이 차이는 AI의 효과가 시간 감소만으로 설명되지 않음을 보여 준다. 반복적인 입력 부담이 줄어 업무가 수월해졌다고 느낄 수 있고, 반대로 AI 초안을 수정하는 일부 시간이 측정에 포함되지 않았을 수도 있다. 따라서 사용자 경험과 실제 업무 시간을 함께 살펴야 한다.

설문은 의료진이 느끼는 부담을 알 수 있지만 응답자의 기대나 당시의 근무 상황에 영향을 받는다. 전자의무기록의 접속 기록은 실제 사용 흔적을 보여 주지만 화면을 열어 둔 채 환자와 대화한 시간까지 정확히 구분하지 못할 수 있다. 어느 한 측정값을 완전한 정답으로 삼기보다 두 자료가 같은 방향을 가리키는지 확인하는 편이 낫다.

장기간 사용했을 때의 효과도 아직 분명하지 않다. 도입 후 여러 시점에 설문한 연구에서는 문서 작업 부담이 줄어도 번아웃이 함께 개선되지는 않

았다. 번아웃은 기록 업무 외에도 조직과 근무 환경의 영향을 받으므로, AI 도입만으로 해결하기 어렵다.

### 초안 검토라는 새로운 업무

AI 스크라이브를 도입하면 의료진의 기록 업무는 직접 작성하는 일에서 초안을 확인하고 수정하는 일로 일부 바뀐다. 초안이 자연스러운 문장으로 작성되어 있어도 약물 용량, 증상 유무, 검사 결과가 정확한지 확인해야 한다. 통역 과정의 오류가 그대로 기록에 반영될 가능성도 있다.

병원은 기록 시간 단축과 함께 검토에 필요한 업무량을 고려해야 한다. 진료과에 맞는 서식을 제공하고, 오류를 쉽게 수정할 수 있도록 하며, 검토 책임을 명확히 정하는 운영 방식이 필요하다.

검토 시간은 진료 일정 안에 들어가야 한다. 초안이 빨리 만들어진다는 이유로 그 시간까지 다른 환자 예약으로 채우면 기록 확인은 다시 퇴근 뒤로 밀릴 수 있다. 병원은 도입 뒤에도 기록이 마무리되는 시점과 수정량을 살펴야 한다. 의료진이 절약한 시간을 어디에 쓰게 되는지는 소프트웨어의 속도만으로 결정되지 않는다.

AI 스크라이브는 의료진의 기록 부담을 줄일 가능성을 보여 주고 있다. 이를 지속적인 업무 환경 개선으로 연결하려면 기술 도입과 함께 업무 절차, 교육, 인력 운영을 검토해야 한다. 기록의 정확성을 유지하면서 의료진이 환자 진료와 휴식에 쓸 시간을 확보하는 것이 도입의 목표다.



## 제2장. 임상 의사결정 지원(CDSS)과 진단



## 2.1 거대언어모델(LLM) 기반 항생제 처방 지원

중국 수도의과대학과 칭화대학 부속병원 등의 연구진이 발표한 원논문에  
는 환자 기록, 진료 지침, 처방 규칙을 연결한 과정이 담겨 있다. 실제 처방  
에 투입하기 전 과거 자료로 평가한 연구라는 점에 유의하며 읽을 수 있다.

<https://medinform.jmir.org/2026/1/e98207>

항생제 처방은 감염 원인과 환자 상태를 함께 고려해야 하는 임상 판단이  
다. 의사는 의심되는 세균, 감염 부위, 중증도, 지역의 항생제 내성 양상에  
따라 약을 선택한다. 간과 신장 기능, 나이, 병용 약물에 따라 용량이나 투  
여 간격을 조절해야 하는 경우도 있다.

임상 의사결정 지원 시스템(CDSS, Clinical Decision Support  
System)은 이러한 판단에 필요한 정보를 제공한다. 환자의 검사 결과와  
진료 지침을 확인하고, 약물 선택이나 용량에서 검토할 사항을 제시하는  
방식이다. AI의 역할은 의사가 여러 조건을 빠짐없이 검토하도록 돕는 데  
있다.

이때 시스템에 넣는 정보의 시점이 중요하다. 같은 환자의 검사라도 입원  
당시 결과와 며칠 뒤 결과는 의미가 다를 수 있다. 약물을 추천하는 화면에  
는 어떤 검사와 기록을 근거로 삼았는지가 드러나야 한다. 환자 이름이 맞  
는지만 확인해서는 충분하지 않다. 최근 결과가 빠져 있거나 다른 진료에  
서 사용한 약이 남아 있으면 그럴듯한 답변도 현재의 처방 판단과 어긋날  
수 있다.

## 규칙 기반 시스템에서 언어모델로

1970년대 스탠퍼드 대학교에서 개발한 MYCIN은 감염 진단과 항생제 추천을 위한 초기 전문가 시스템이다. 전문가가 정리한 규칙에 따라 결론을 제시했으므로 판단 과정을 설명하기 쉬웠다. 그러나 규칙을 계속 보완해야 했고, 자유로운 문장으로 작성된 진료 기록을 처리하는 데 한계가 있었다.

MYCIN은 스탠퍼드의 에드워드 쇼트리프 등을 중심으로 연구되었으며 1976년 저서에 체계적으로 정리되었다. 오늘날의 생성형 AI와 달리 전문가가 작성한 조건과 결론을 연결하는 방식이었다. 예를 들어 정해진 조건이 충족되면 특정 판단을 검토하도록 하는 식이다. MYCIN이라는 역사적 연구 시스템과 최근의 언어모델 제품은 같은 CDSS 범주 안에서도 기술적 출발점이 다르다.

이후 기계학습 모델은 환자 자료에서 패턴을 찾아 위험을 예측하는 데 사용되었다. 다만 예측 이유를 설명하기 어려운 경우가 있어, 처방 판단에 적용하려면 추가 검토가 필요했다.

거대언어모델(LLM)은 진료 기록을 읽고 추천 이유를 문장으로 설명할 수 있다. 의사가 자연어로 질문하고 답변의 근거를 검토할 수 있다는 점에서 기존 시스템과 차이가 있다. 그러나 설명이 자연스럽다고 해서 처방이 정확한 것은 아니다. 사실과 다른 내용을 생성하는 환각(hallucination), 용량 조절 조건의 누락, 질문 표현에 따른 답변 변화가 발생할 수 있다.

## 정확성과 일관성의 문제

이탈리아 감염내과 연구진은 여러 언어모델에 감염 사례를 제시하고 항생제 추천을 비교했다. 모델별 성능 차이가 컸으며, 후속 평가에서도 잘못되거나 불필요한 추천이 나타났다. 같은 사례를 반복해 질문했을 때 답이 달라지는 문제도 확인되었다.

문헌 근거를 제시하는 의료용 도구(OpenEvidence)와 범용 대화형 모델(ChatGPT-5.2)을 비교한 실험에서는, 더 정확한 답을 제시한 도구가 반드시 더 일관적인 것은 아니었다. 정확성은 답이 적절한지를, 일관성은 같은 조건에서 비슷한 답을 유지하는지를 뜻한다. 처방 지원에서는 두 조건을 함께 검토해야 한다.

OpenEvidence는 같은 이름의 미국 기업이 제공하는 의료 정보 서비스이고, ChatGPT는 OpenAI의 대화형 서비스다. 논문에서 사용한 모델 버전은 서비스 전체와 구별해야 한다. 서비스가 바뀌면 검색 자료와 답변 방식도 바뀔 수 있기 때문이다. OpenEvidence의 기업 발표는 2023년에도 확인되지만, 비교 실험에 쓰인 버전의 최초 제공일은 논문에 적힌 시험 시점과 별도로 확인할 사항이다.

## 환자 자료·진료 지침·처방 규칙의 결합

중국 연구진은 폐렴 입원 환자의 과거 전자의무기록을 이용해 언어모델 기반 항생제 처방 지원 시스템을 평가했다.

이 시스템은 수도의과대학과 베이징 칭화창궁병원, 베이징우의병원 등의 공동 연구이며 특정 상용 제품명으로 출시된 사례는 아니다. 사전공개본은 2026년 4월 14일, 심사를 거친 논문은 같은 해 8월 4일에 공개되었다.

병원이 이 시스템을 모든 환자에게 적용하기 시작한 날짜를 뜻하지는 않는다. 핵심은 모델에 환자 상태만 제시하는 대신, 유사 환자의 처방과 진료 지침, 의사가 정한 규칙을 함께 제공한 것이다.

첫 번째 정보는 유사 환자의 진료 기록이다. 환자 기록을 숫자 배열인 벡터로 변환한 뒤, 현재 환자와 비슷한 사례를 검색한다. 이 과정에는 임베딩(embedding) 모델이 사용된다. 과거 사례의 처방을 참고하되, 현재 환자에게 그대로 적용할 수 있는지는 따로 판단해야 한다.

벡터는 여러 특징을 숫자로 나타낸 목록이다. 컴퓨터는 이 숫자들의 관계를 비교해 비슷한 기록을 찾는다. 문장이 비슷하다는 사실과 임상적으로 같은 환자라는 사실은 다르다. 예를 들어 증상 표현이 닮아 있어도 알레르기나 장기 기능이 다르면 같은 처방을 따르기 어렵다. 검색 결과에 환자별 제한 조건을 다시 적용하는 이유다.

두 번째 정보는 진료 지침을 정리한 지식 그래프(knowledge graph)다. 질환, 원인균, 약물, 금기 등의 관계를 구조화해 모델이 필요한 근거를 찾도록 한다. 예를 들어 특정 상태에서 피해야 할 약물이나 용량을 줄여야 하는 조건을 연결할 수 있다.

세 번째 정보는 임상이가 정리한 처방 규칙이다. 중증도에 따른 약물 선택, 함께 사용하면 안 되는 약물, 간·신장 기능에 따른 용량 조절 등을 명시한다. 언어모델은 이 정보를 참고해 약물명, 용량, 투여 횟수와 경로를 제안하고 근거를 설명한다. 의사는 제안과 근거를 함께 검토한다.

이 구조에서는 언어모델의 일반 지식만큼 병원의 진료 경험과 지침을 연결하는 설계가 중요하다. 모델이 답변을 생성하더라도, 판단에 필요한 자료와 제한 조건은 의료기관이 체계적으로 제공해야 한다.

### 처방 지원의 성과와 한계

폐렴 환자 기록을 이용한 평가에서는 유사 사례와 지침, 처방 규칙을 함께 제공했을 때 기본 언어모델보다 처방 일치도가 높았다. 일부 정보를 제외하면 성능이 낮아졌으며, 유사 환자 검색과 임상주의 규칙이 중요한 역할을 했다.

연구에서는 F1 점수와 자카드 정확도(Jaccard accuracy)로 추천 처방과 기준 처방의 일치 정도를 평가했다. 소거 실험(ablation study)은 구성 요소를 하나씩 제외해 각 요소의 기여를 살피는 방법이다. 이러한 평가는 시스템 설계를 비교하는 데 도움이 되지만, 실제 환자의 치료 효과를 직접 보여 주지는 않는다.

F1 점수는 필요한 항목을 얼마나 빠짐없이 찾았는지와 추천한 항목이 얼마나 적절한지를 함께 반영한다. 자카드 지표는 두 항목 집합에서 겹치는 부분의 비중을 본다. 약 이름이 기준과 맞더라도 용량이 틀릴 수 있고, 기준 처방과 달라도 의학적으로 허용되는 대안일 수 있다. 그러므로 숫자를 읽을 때 무엇을 정답으로 정했는지와 오류를 누가 검토했는지를 함께 봐야 한다.

간·신장 기능이 저하된 환자와 고령 환자에서는 추천 성능이 상대적으로 낮았다. 오류 사례를 검토했을 때도 장기 기능을 충분히 반영하지 못한 위

험한 추천이 남아 있었다. 용량 조절 규칙을 제공해도 의료진의 확인을 생략할 수 없는 이유다.

이 연구는 과거 기록으로 시스템을 평가한 단계다. 실제 진료에서 처방이 어떻게 바뀌는지, 부작용이나 치료 결과가 개선되는지는 추가로 확인해야 한다. 병원마다 환자 구성과 내성균 양상, 처방 관행이 다르므로 다른 기관에 도입할 때도 검증이 필요하다.

### **항생제 관리와 의료진의 역할**

항생제 관리 프로그램(antimicrobial stewardship)은 항생제를 적절하게 선택하고 투여하도록 관리하는 활동이다. 언어모델 기반 CDSS는 이 과정에서 관련 정보를 정리하고, 처방 근거와 주의사항을 제시하는 데 기여할 수 있다.

실제 도입에서는 추천 결과를 진료 지침 및 환자 상태와 대조하기 쉽도록 구성해야 한다. 의료진은 간·신장 기능, 알레르기, 병용 약물 등 환자별 조건을 확인하고 최종 처방을 결정한다. 항생제 처방 지원의 발전 방향은 언어모델의 성능 향상과 함께 신뢰할 수 있는 자료 연결, 규칙 관리, 의료진의 검토 절차를 개선하는 데 있다.

## 2.2 의료 세계 모델과 U2-Med의 진단·검사 요청

개발사 윈즈성의 기술 소개는 U2-Med와 학습 환경인 U2-Med-World의 역할을 나누어 설명한다. 이 자료는 기업의 연구 발표이므로 임상시험 결과와 구분해서 읽어야 한다.

<https://unisound.com/en/news/165>

진단은 한 번의 질문으로 끝나는 작업이 아니다. 의사는 환자의 증상을 듣고 병력과 복용 약물을 확인한 뒤, 필요에 따라 검사를 시행한다. 새로운 정보가 들어오면 가능한 질환과 다음 조치를 다시 검토한다. 의료 AI 연구에서도 이러한 순차적 판단 과정을 구현하려는 시도가 늘고 있다.

기존 의료 AI 평가는 정리된 환자 정보를 한꺼번에 제공하고 정답을 고르게 하는 방식이 많았다. 반면 실제 진료에서는 어떤 정보를 먼저 확인할지 결정해야 한다. 의료 세계 모델은 이 차이를 줄이기 위해 환자의 상태와 임상 행동, 그에 따른 관찰 결과를 연결한다.

### 의료 세계 모델의 개념

세계 모델(world model)은 현재 상태에서 특정 행동을 했을 때 어떤 결과가 나타날지 예측하는 모형이다. 의료 분야에서는 문진, 검사, 치료 등 임상 행동에 따라 얻는 정보나 환자 상태의 변화를 모의하는 데 적용할 수 있다.

중국 AI 기업 윈즈성(云知声, Unisound)은 2026년 7월 17일 U2-Med 공개를 알렸고, 같은 해 10월 21일 발표에서 U2-Med-World와의 연결 구조를 설명했다. U2-Med-World는 환자 상태와 임상 행동의 관계를 표

현하는 가상 환경이고, U2-Med는 이 환경에서 질문과 도구 사용을 선택하는 에이전트(agent)다. 회사는 이 구조를 통해 여러 차례 대화하면서 판단을 수정하는 능력을 학습한다고 설명했다.

U2-Med-World는 병력, 대화, 검사, 영상, 생체 신호, 처방 정보를 시간 순서에 따라 정리한다. 에이전트가 질문이나 검사를 선택하면 환경이 관련 정보를 제공하고, 에이전트는 이를 바탕으로 다음 행동을 결정한다. 이때 나중에 확인된 정보가 앞선 판단에 섞이는 정보 누출(data leakage)을 방지하는 것이 중요하다.

예를 들어 정답 진단이 적힌 퇴원 요약은 첫 문진부터 보여 주면 AI는 검사 선택 능력을 발휘하지 않고도 답을 맞힐 수 있다. 순차적 평가에서는 그 시점에 의료진이 실제로 알 수 있었던 정보만 제공해야 한다. 검사를 요청하면 결과가 공개되는지, 결과가 없는 검사를 요청하면 어떻게 처리하는지도 미리 정해야 한다. 이런 환경 설계가 잘못되면 높은 점수를 얻어도 진료 능력을 평가했다고 보기 어렵다.

## 질문·검사·치료의 구분

문진과 검사는 환자에 대해 알고 있는 정보를 바꾸는 과정이다. 치료와 질병의 진행은 환자의 실제 상태를 변화시킨다. 의료 세계 모델은 이 둘을 구분해야 한다. 검사를 시행했다는 사실 자체가 환자의 상태 호전을 의미하지 않으며, 치료를 시작했다고 진단이 확정되는 것도 아니기 때문이다.

회사 발표에 나온 가상 환자 사례에서 에이전트는 피로와 식욕 저하 같은 초기 증상을 바탕으로 증상 기간, 동반 증상, 과거 병력, 최근 복용 약물 중 무엇을 먼저 확인할지 선택한다. 이후 제공된 정보에 따라 질문 순서와 판

단을 조정한다. 같은 환자에 대해 여러 문진 순서를 비교하면 필요한 정보를 얼마나 잘 확보하고 위험 신호를 발견했는지 평가할 수 있다.

이 환경은 실제 기록에 남은 진료 과정뿐 아니라 합성 궤적(synthetic trajectory)도 생성한다.

궤적이라는 말은 어떤 질문과 행동이 어떤 순서로 이어졌는지를 뜻한다. 환자 한 명의 기록에서는 실제로 선택한 경로만 관찰된다. 컴퓨터가 다른 경로를 만들어 비교할 때에는 관찰된 사실에서 어느 부분까지 확인할 수 있는지 표시해야 한다. 생성된 문장이 상세하더라도 실제 관찰이 늘어난 것은 아니기 때문이다. 합성 궤적은 동일한 초기 상태에서 다른 질문이나 검사를 선택했을 때의 가상 진료 과정이다. 여러 선택을 비교하는 학습 자료로 사용할 수 있지만, 실제로 시행하지 않은 치료의 효과를 입증하는 자료는 아니다.

### 대화 과정 전체를 학습하는 방식

U2-Med의 학습은 가상 환경에서 대화를 생성하고, 전체 과정을 평가한 뒤, 적절한 사례를 선별해 후속 학습에 사용하는 방식으로 설명된다. 최종 진단뿐 아니라 정보 수집, 위험 식별, 소통 과정도 평가 대상이다.

회사가 소개한 루브릭 강화학습(Rubric RL)은 의학적 정확성이나 정보의 완결성과 같은 세부 기준을 학습에 반영하는 방식이다. 온라인 선호 증류(OPD)는 평가를 통해 선택한 문진 전략을 모델 학습에 반영한다. 이러한 접근은 답변 하나의 정확도에서 여러 단계의 진료 과정으로 평가 범위를 넓힌다는 의미가 있다.

## 성능 평가와 실제 진료의 차이

원즈성은 의료 대화 평가인 HealthBench에서 U2-Med가 높은 점수를 기록했다고 발표했다. HealthBench는 필요한 정보를 질문했는지, 위험 신호를 인식했는지, 정확하고 안전하게 설명했는지를 평가하는 시험이다.

다만 회사가 발표한 시험 성적과 실제 진료 효과는 구분해야 한다. 평가 환경에서 좋은 답변을 제시하더라도 환자와 의료진이 실제로 사용할 때 같은 결과가 나타난다고 보장할 수 없다. U2-Med의 임상 효과와 검사 비용, 환자 안전에 대한 영향은 실제 진료 환경에서 별도로 검증해야 한다.

## 검사 선택과 정보의 가치

검사는 진단에 필요한 정보를 제공하지만 비용과 위험, 환자의 불편을 수반한다. 따라서 다음 검사가 판단이나 치료를 얼마나 바꿀 수 있는지 검토해야 한다. 이를 설명하는 개념이 정보의 가치(value of information)다.

검사 순서는 응급성, 예상 질환, 검사 정확도, 결과가 치료에 미치는 영향에 따라 달라진다. AI가 검사 선택을 지원하려면 정보를 더 확보하는 이점과 불필요한 검사를 줄이는 요구를 함께 고려해야 한다. 다만 이러한 일반 원칙이 U2-Med의 내부 계산에 어떻게 구현되었는지는 공개 설명만으로 알기 어렵다.

## 유사한 연구 방향

마이크로소프트의 MAI-DxO는 여러 에이전트가 추가 정보를 요청하며 진단하는 체계다. SDBench라는 평가 환경에서는 복잡한 증례를 단계적으로 검토하면서 문진과 검사 정보를 선택한다. 공개 연구는 순차적 정보 수집을 통한 진단 가능성을 보여 주지만, 통제된 평가 결과를 실제 진료 성과로 볼 수는 없다.

MAI-DxO는 마이크로소프트 연구진이 만든 진단 조정 체계다. 관련 연구의 최초 arXiv 공개는 2025년 6월 27일이다. SDBench는 이 체계와 여러 모델을 시험하는 평가 환경의 이름이다. 시스템과 시험을 구분하면 어떤 AI를 만들었는가와 어떤 문제로 시험했는가를 혼동하지 않게 된다.

ClinVerse는 실제 전자의무기록을 바탕으로 구성된 다중 에이전트 진단 환경이다. 의사 역할의 에이전트가 초기 증상에서 출발해 문진과 검사로 정보를 수집하고 진단한다. 정보 수집과 검사 선택, 진단 정확도를 함께 평가한다. 이 역시 가상 환경에서의 학습과 평가를 위한 연구이며 실제 진료 도입과는 구분된다.

ClinVerse의 논문은 Sirry Chen 등 연구진이 2026년 10월 26일 arXiv에 처음 공개했다. 병원에서 환자를 진료하는 완성 제품을 출시했다는 뜻은 아니다. 환자 역할과 검사 결과 제공 역할을 나누어 두면 연구자는 정보가 공개되는 순서를 통제하면서 의사 역할 에이전트의 선택을 비교할 수 있다.

U2-Med-World, SDBench, ClinVerse는 모두 AI가 추가 정보를 수집하고 판단을 수정하는 과정을 연구한다. 의료 AI가 정답을 제시하는 도구에서 진단 과정의 일부를 지원하는 시스템으로 발전하고 있음을 보여 준다.

## 도입에 필요한 검토

검사 선택을 지원하는 AI에서는 과잉 검사와 필요한 검사의 누락을 함께 관리해야 한다. 검사를 지나치게 권하면 비용과 환자 부담이 늘고, 필요한 검사를 누락하면 진단이 늦어질 수 있다.

현재의 자율 검사 요청은 주로 가상 환경에서 AI가 검사 후보를 제안하는 기능을 뜻한다. 실제 병원에서도 독립적으로 검사를 처방하는 권한이 확인되었다는 의미는 아니다. 임상 적용에서는 의료진이 추천 근거와 환자 상태를 검토하고 승인하는 절차가 필요하다. 책임 범위 역시 허가 상태, 사용 방식, 검토 절차, 사고 원인과 각국 법체계에 따라 판단해야 한다.

의료 세계 모델의 의의는 진단 결과뿐 아니라 정보를 얻고 판단을 수정하는 과정까지 연구할 수 있다는 점이다. 앞으로의 과제는 가상 환경에서 학습한 능력이 실제 진료에서도 정확하고 안전하게 작동하는지 확인하는 것이다.

## 2.3 AMIE와 MIRA의 진단 추론

구글의 초기 AMIE 소개와 MIRA의 Nature 원논문을 함께 읽으면 의료 대화 연구와 전자의무기록 안의 행동 연구가 어떻게 다른지 확인할 수 있다. 두 자료 모두 연구의 평가 환경과 한계를 설명한다.

<https://research.google/blog/amie-a-research-ai-system-for-diagnostic-medical-reasoning-and-conversations/>

<https://www.nature.com/articles/s41586-026-10675-5>

진료 지침(clinical practice guideline)은 질환별 검사와 치료에 관한 연구 근거를 정리한 권고다. 지침은 분야마다 다르고 새로운 근거에 따라 개정된다. 의료진은 환자 상태에 맞는 권고를 확인하면서 병원의 진료 여건과 약물 사용 조건도 고려해야 한다.

구글의 AMIE와 독일 연구진의 MIRA는 이러한 업무를 지원하는 의료 AI 연구다. AMIE는 환자와 대화하면서 진단과 관리 계획을 검토하는 데 초점을 두고, MIRA는 전자의무기록 안에서 검사와 처방 등 임상 행동을 수행하는 방식을 연구한다. 두 시스템은 의료 AI의 기능이 정보 제공에서 진료 과정 지원으로 확대되고 있음을 보여 준다.

### AMIE의 의료 대화와 관리 계획

AMIE(Articulate Medical Intelligence Explorer)는 구글의 연구용 의료 대화 AI다.

구글 리서치는 2024년 1월 12일 AMIE를 공식 소개했다. 진단 대화 연구가 먼저 공개되었고, 질병 관리로 확장한 연구 소개는 2026년 6월 17일에

나왔다. AMIE는 연구 시스템의 이름이며, 그 안에서 사용하는 Gemini는 구글이 개발한 기반 모델 계열의 이름이다. 기반 모델과 그것을 의료 과제에 맞게 구성한 시스템을 구분하면 연구의 변화를 따라가기 쉽다. 초기 연구는 환자 역할을 맡은 참여자와 글로 대화하며 진단하는 능력을 평가했다. 이후 연구에서는 여러 차례의 진료에 걸쳐 증상과 치료 반응을 확인하고 계획을 조정하는 질병 관리(disease management)로 범위를 넓혔다.

이 과정에는 두 에이전트가 참여한다. 대화 에이전트는 환자에게 질문하고 필요한 내용을 설명한다. 관리 추론(Mx) 에이전트는 대화 내용을 바탕으로 진료 지침과 약물 정보를 검토해 관리 계획을 작성한다. 환자와의 대화를 지속하면서 관련 근거를 별도로 검토하도록 역할을 나눈 구조다.

예를 들어 환자가 약을 복용한 뒤 겪은 불편을 말하면 대화 에이전트는 시점과 양상을 더 묻는다. 관리 추론 에이전트는 확보된 정보와 관련 지침을 대조해 검토할 계획을 만든다. 이 예시는 두 역할의 차이를 보여 주기 위한 것이다. 역할을 나누었다고 서로의 오류가 저절로 교정되는 것은 아니므로, 근거가 부족한 부분을 남기고 사람이 확인할 수 있어야 한다.

구글은 제미니(Gemini)의 긴 문맥 처리 능력을 사용해 많은 분량의 진료 지침과 약물 정보를 대조한다고 설명했다. 평가에는 영국 국가처방집(BNF) 자료도 사용되었다. 가상 진료 연구에서 AMIE의 관리 계획은 지침 부합도와 계획의 구체성 등에서 가능성을 보였다. 다만 모의 대화 결과가 실제 환자에 대한 치료 효과를 뜻하지는 않는다.

## 실제 환자의 진료 전 문진

2026년 The Lancet에는 AMIE가 실제 환자의 진료 전 문진에 참여한 연구가 실렸다. 구글과 베스 이스라엘 디코니스 메디컬센터(BIDMC) 연구진이 함께 진행했으며, 환자는 진료 전에 집에서 AMIE와 글로 대화했다. AMIE는 증상과 병력을 묻고 가능한 진단을 정리했다. 의료진은 진료 전에 대화 내용과 요약を確認할 수 있었다.

이 연구에서 AI는 치료를 독립적으로 시행하지 않았다. 의사가 대화를 실시간으로 감독했고, 정해진 안전 기준에 따라 개입할 수 있었다. 일부 의료진은 요약이 진료 준비에 도움이 되었다고 평가했지만, 잘못된 정보와 설명을 보완할 필요도 있었다.

이 사례의 의미는 의료 대화 AI가 실제 진료 준비 과정에서 어떤 역할을 할 수 있는지 시험했다는 데 있다. 문진 내용을 정리하면 의사는 대면 진료에서 추가 확인이 필요한 사항에 집중할 수 있다. 그러나 감독 아래에서 얻은 결과이므로, 의료진 없이 사용할 때의 안전성이나 치료 효과까지 확인한 것은 아니다.

## MIRA의 전자의무기록 연계

MIRA(Medical Intelligence for Reasoning and Action)는 독일 하이델베르크 대학병원 국립종양질환센터와 드레스덴 공과대학 등의 연구진이 개발한 에이전트다. 환자 정보를 설명하는 데서 나아가, 정해진 권한 안에서 전자의무기록(EHR)에 접근하고 임상 행동을 구성하는 방식을 연구한다.

연구는 실제 병원 시스템과 분리된 샌드박스(sandbox) 환경에서 진행되었다. 샌드박스는 실험이 실제 업무에 영향을 주지 않도록 구성된 통제 환경이다. 가상 EHR은 의료 정보 교환 표준인 FHIR에 맞춰 설계되었다.

MIRA의 Nature 논문 공개일은 2026년 6월 17일이다. FHIR는 환자, 검사 결과, 약물 같은 정보를 일정한 구조로 주고받기 위한 표준이며 AI 모델 자체는 아니다. 에이전트가 검사를 하자는 문장을 만드는 것과, 정해진 항목을 갖춘 검사 요청을 시스템에 입력하는 것은 다른 일이다. 후자는 환자 식별, 항목 선택, 권한 확인 등 추가 조건을 충족해야 한다.

MIRA는 환자 역할의 에이전트와 대화해 병력을 확인하고, 혈액검사와 영상검사 등을 요청하며, 제공된 결과를 바탕으로 감별 진단을 검토한다. 약물 처방이나 입원 계획도 EHR에서 처리할 수 있는 행동 형식으로 작성한다. 이 행동들은 가상 시스템 안에서 수행되며 실제 환자에 대한 검사나 처방으로 실행되는 것은 아니다.

환자 에이전트는 공개 중환자·응급 데이터베이스인 MIMIC-IV의 실제 기록에 근거해 응답하도록 구성되었다. 연구진은 이 환경에서 응급실의 여러 진료 과정을 평가했다. 실제 기록을 사용했더라도 환자와 직접 상호작용한 임상시험과는 구분해야 한다.

## 임상 행동을 평가하는 기준

MIRA 연구는 진단 정확도뿐 아니라 치료 계획과 약물 안전도 살폈다. 신장 기능에 따른 용량 조절, 약물 상호작용, 알레르기, 심전도 QT 위험 등을 검토 항목에 포함했다. 이는 임상 에이전트를 평가할 때 최종 진단 외에도 그 진단에 이르는 과정과 후속 조치를 살펴야 함을 보여 준다.

가상 환경에서는 여러 진단·치료 계획 항목에서 가능성이 나타났지만, 질 환별 성능 차이도 있었다. 실제 병원에서는 정보 누락과 기록 오류, 업무 중단 등 더 다양한 조건이 발생한다. 따라서 실제 진료에 도입하기 전에는 안전성과 운영 방식을 별도로 평가해야 한다. 연구진도 의료 AI 에이전트가 의료인을 대체하는 방향으로 설계되어서는 안 된다고 강조했다.

## 두 시스템의 역할과 도입 과제

AMIE는 환자와의 대화 및 관리 계획을, MIRA는 전자의무기록과 연결된 임상 행동을 중심으로 발전했다. 환자의 설명을 이해하는 능력과 이를 정확한 업무 처리로 연결하는 능력은 모두 의료 AI에 필요하다. 향후 이 기능들이 연결될 가능성은 있지만, 두 시스템이 통합된 진료 체계로 검증되었다는 의미는 아니다.

도입에서는 AI가 어떤 지침을 참고하는지 명확해야 한다. 국제 권고와 병원의 사용 가능 약물, 보험 기준이 다를 수 있기 때문이다. 참고 자료가 개정되면 시스템에도 적절히 반영해야 한다.

의료진의 검토 권한과 책임도 정해야 한다. 지침에 맞는 제안이라도 환자별 예외를 놓칠 수 있으므로, 의료진이 근거를 확인하고 수정하거나 거부할 수 있어야 한다. 연구 단계의 시스템과 허가받은 임상 제품을 구분하는 것도 필요하다. AMIE와 MIRA는 이 연구들에서 연구용 시스템으로 평가되었다.

AI 스크라이브가 기록 작성을 지원한다면 AMIE와 MIRA는 진료 정보의 해석, 관리 계획, 임상 행동으로 지원 범위를 확장한다. 이러한 변화가 실

제 의료의 개선으로 이어지려면 기술 성능과 함께 의료진의 감독, 정보 갱신, 병원 업무와의 연결을 검증해야 한다.



## Part 2. AI 신약개발과 디지털 트윈



## 제3장. 바이오 파운데이션 모델과 신약 표적



### 3.1 의료 멀티모달 파운데이션 모델

Arc Institute의 Evo 2 공개문과 구글 딥마인드의 AlphaGenome 소개는 유전체 AI의 입력 자료와 예측 대상을 설명한다. 아래 자료에서 두 모델의 개발 기관과 최초 공개 시점도 확인할 수 있다.

<https://arcinstitute.org/news/evo2>

<https://deepmind.google/blog/alphagenome-ai-for-better-understanding-the-genome/>

의료와 생명과학에서는 같은 환자나 조직을 여러 종류의 자료로 분석한다. 유전체는 유전적 특성을, 조직 병리 영상은 세포와 조직의 형태를 보여 준다. 공간 전사체는 조직 안의 위치별 유전자 발현을 측정하며, 진료 기록과 판독문은 임상적 맥락을 제공한다. 이 자료들을 함께 분석하면 한 종류의 자료만으로 파악하기 어려운 정보를 얻을 수 있다.

멀티모달 파운데이션 모델(multimodal foundation model)은 다양한 형식의 자료를 대규모로 학습해 여러 분석 과제에 적용하는 모델이다. 의료 분야에서는 병리 영상과 분자 정보, 영상과 임상 기록 등을 연결하는 연구가 진행되고 있다. 아래의 사례들은 주로 두세 종류의 자료를 결합한 모델이다.

파운데이션 모델의 기반이라는 말은 여러 과제에 공통으로 쓸 표현을 먼저 학습한다는 뜻이다. 한 질환의 정답만 반복해서 익히는 모델과 구별된다. 사전학습 뒤에는 필요한 과제에 맞게 추가로 학습하거나 별도의 예측기를 붙일 수 있다. 다만 의료 자료를 많이 학습했다는 이유만으로 모든 검사와 모든 환자를 이해하는 것은 아니다. 어떤 종류의 자료를 넣었고 어떤 결과를 내도록 학습했는지부터 확인해야 한다.

## 의료 데이터의 종류

게놈(genome, 유전체)은 DNA 염기서열을 포함한 유전 정보다. 유전적 변이가 질환과 어떤 관련이 있는지 분석하는 데 사용된다. 다만 유전 정보만으로 특정 조직에서 실제로 어떤 유전자가 발현하는지 모두 알 수는 없다.

조직 병리(histopathology) 슬라이드는 얇게 자른 조직을 염색해 관찰한 영상이다. 대표적인 H&E 염색 영상에서 병리의사는 세포핵의 형태, 조직 구조, 면역세포의 분포 등을 확인한다. 디지털 슬라이드는 넓은 조직을 높은 해상도로 담기 때문에 분석해야 할 영상의 크기도 크다.

공간 전사체(spatial transcriptomics)는 유전자 발현과 조직 내 위치 정보를 함께 측정한다. 어떤 유전자가 발현하는지뿐 아니라 해당 세포가 어디에 있고 주변 세포와 어떤 관계를 갖는지 분석할 수 있다. 텍스트 자료에는 병리 판독문, 진료 기록, 질환 설명, 연구 논문 등이 포함된다.

각 자료는 측정 대상과 형식이 다르다. 멀티모달 모델은 이러한 차이를 처리하면서 자료 사이의 관련성을 학습한다. 이를 통해 병리 영상에서 분자적 특성을 추정하거나, 영상 소견을 임상 기록과 함께 해석할 수 있다.

## 서로 다른 자료를 결합하는 방법

인코더(encoder)는 입력 자료를 숫자 배열인 벡터(vector)로 변환한다. DNA 서열에는 서열 모델을, 병리 영상에는 비전 트랜스포머(vision transformer) 같은 영상 모델을 사용할 수 있다. 공간적 관계를 처리할 때는 그래프 신경망(graph neural network)을 적용하고, 텍스트에는 언어모델을 사용한다. 구체적인 구성은 모델마다 다르다.

자료 사이의 관계를 학습하는 대표적인 방법은 대조 학습(contrastive learning)이다. 같은 조직에서 얻은 영상과 분자 정보처럼 관련된 자료의 표현은 가깝게, 관련이 적은 자료의 표현은 구분되도록 학습한다. 이렇게 형성한 공통 표현 공간을 공유 잠재 공간(shared latent space)이라고 한다. InfoNCE 같은 손실 함수(loss function)를 사용할 수 있다.

손실 함수는 모델의 현재 결과가 학습 목표에서 얼마나 벗어났는지 숫자로 나타내는 계산법이다. 모델은 이 값을 줄이는 방향으로 내부의 값을 조정한다. 같은 환자의 영상과 분자 자료를 연결할 때에는 환자 식별이 정확해야 한다. 다른 사람의 검사 결과를 한 쌍으로 묶으면 모델은 잘못된 관련성을 배우게 된다. 자료를 많이 모으는 일과 올바르게 짝짓는 일은 함께 이루어져야 한다.

크로스 어텐션(cross-attention)은 한 종류의 자료를 해석할 때 다른 자료의 관련 부분을 참조하는 방법이다. 예를 들어 영상의 특정 부위를 임상적 묘사와 연결할 수 있다. 다만 자료를 결합했다는 사실만으로 성능 향상이 보장되지는 않으며, 어떤 정보가 실제로 기여했는지 확인해야 한다.

## 유전체 모델의 발전

Evo 2는 다양한 생물의 방대한 염기서열을 학습한 게놈 언어모델이다.

미국 Arc Institute와 NVIDIA가 중심이 되어 스탠퍼드, UC 버클리, UC 샌프란시스코 등의 연구진과 개발했으며 2025년 2월 19일 공개했다. 여기서 언어모델이 읽는 기본 단위는 일상 문장의 단어 대신 DNA의 염기다. 서열에 어떤 패턴이 나타나는지 학습한 뒤 새로운 서열이나 변이를 분석한다. 긴 서열을 처리하고 유전적 변이의 영향을 추정하는 능력을 연구했

다. 구글 딥마인드의 AlphaGenome은 비암호화 영역(non-coding region)의 변이가 유전자 조절에 미치는 영향을 예측하는데 초점을 둔다.

AlphaGenome의 최초 공식 소개는 2025년 6월 25일이다. 긴 DNA 서열을 입력받아 유전자 조절과 관련된 분자적 성질을 예측한다. 변이가 있는 서열과 없는 서열의 예측값을 비교하면 어느 기능이 달라질 가능성이 있는지 살필 수 있다. 단백질을 직접 만드는 영역 밖의 DNA도 유전자가 언제, 어느 정도 작동하는지에 관여할 수 있어 연구 대상이 된다.

이러한 모델은 유전 정보의 기능을 해석하는 연구 도구로 발전하고 있다. 다만 계산 결과가 곧 환자의 병원성 변이 판정이나 유전 상담에 사용할 수 있는 임상 근거가 되는 것은 아니다. 염색질 접근성(chromatin accessibility)처럼 특정 과제에서는 해당 자료로 학습한 전용 모델이 더 적합할 수도 있다.

### 병리 영상과 분자 정보의 연결

HistoMethyl은 병리 슬라이드와 DNA 메틸화(methylation) 정보를 연결한 연구 사례다. DNA 메틸화는 유전자 발현 조절과 관련된 화학적 변형이다. 학습할 때 영상과 메틸화 자료를 함께 제공하고, 사용할 때는 H&E 영상으로 분자적 특성을 추정하는 접근이다.

이 연구는 병리 영상으로 분자 검사가 필요한 환자를 우선 선별할 가능성을 보여 준다. 평가에는 여러 암 코호트(cohort, 공통 조건으로 구성된 환자 집단)가 사용되었다. 그러나 영상에서 추정한 결과가 검증된 분자 검사를 대체하는 것은 아니다. 연구진도 확진보다 초기 선별에 초점을 두었다.

HistoMethyl은 Dannong Wang과 미국 센트럴플로리다 대학교의 Yu Tian 등 연구진이 발표한 연구 모델이다. bioRxiv에 등록된 교신저자의 소속은 센트럴플로리다 대학교이며, 첫 사전공개일은 2026년 6월 14일이다. 병리 영상에서 분자 상태를 추정할 때에는 측정했다와 예측했다를 구분해 기록해야 한다. 같은 수치 모양으로 출력되어도 실제 검체를 분석한 결과와 모델이 계산한 값은 생성 과정이 다르다. 연구의 구성과 평가 조건은 다음 원논문에서 살펴볼 수 있다.

<https://www.biorxiv.org/content/10.64898/2026.06.11.731518v1>

Fuji는 병리 영상과 RNA 시퀀싱, 유전체 자료를 함께 학습하는 방향을 제시했다. 벌크 RNA 시퀀싱(bulk RNA sequencing)은 조직에 포함된 여러 세포의 RNA를 함께 분석하는 방법이며, 전장 유전체(whole genome) 자료는 유전체 전체의 정보를 다룬다. HistoMethyl과 Fuji의 인용 연구는 사전공개 단계이므로 임상 제품과 구분해야 한다.

Fuji는 미국 휴스턴 메소디스트 연구소와 코넬대학교 웨일 코넬 의과대학, 미국 국립암연구소 연구진이 참여한 연구 모델이다. Qing Li와 Shengyu Li 등이 개발했으며 2026년 8월 6일 사전공개본으로 발표되었다. 이름이 같거나 비슷하다는 이유로 의료기기 회사의 제품으로 이해해서는 안 된다. 이 연구에서 말하는 Fuji는 병리 영상과 분자 자료의 관계를 학습하는 계산 모델이다. 연구의 구성과 평가 조건은 다음 원논문에서 살펴볼 수 있다.

<https://pmc.ncbi.nlm.nih.gov/articles/PMC13484847/>

## 공간 전사체와 텍스트의 결합

공간 전사체는 조직 내 유전자 발현을 자세히 분석할 수 있지만 비용과 측정 여건의 제약이 있다. 이에 비교적 널리 확보되는 H&E 영상과 공간 전사체를 연결해, 영상에서 위치별 발현 양상을 추정하려는 연구가 진행되고 있다. 다만 플랫폼별 해상도와 측정 방식의 차이, 학습 자료의 부족이 과제로 남아 있다.

Tx-LLM과 TxGemma는 분자 구조, 단백질 서열, 세포주 정보, 질환 설명 등을 텍스트 형태로 처리해 신약개발의 여러 과제를 학습한다. 자료 형식을 공통된 언어 인터페이스로 연결하는 접근이다.

Tx-LLM은 구글 연구진 등이 개발한 모델로 2024년 6월 10일 논문이 사전공개되었다. PaLM 2를 바탕으로 신약 연구 과제에 맞게 학습했다.

TxGemma는 구글 연구진이 Gemma 2를 바탕으로 만든 모델 계열이며 관련 논문은 2025년 4월 8일 공개되었다. 분자 구조를 문자로 표현한 자료와 질환 이름 등을 함께 입력할 수 있어도, 그 문자가 실제 환자의 모든 생물학적 상태를 담는 것은 아니다. 그러나 병리 영상과 공간 전사체까지 환자 단위로 모두 통합한 모델은 아니다.

## 영상과 임상 기록의 결합

CGA-Net은 유방 초음파 영상과 모양·경계 등 표준화된 임상 묘사를 결합했다. 크로스 어텐션을 통해 두 자료의 관련성을 학습하는 방식이다. 이런 연구에서는 결합 모델을 영상만 사용한 모델뿐 아니라 임상 정보만 사용한 모델과도 비교해야 한다. 입력 자료 중 일부만으로도 중요한 정보를 얻을 수 있기 때문이다.

CGA-Net은 중국 장쑤성 타이창시 제일인민병원 초음파진단과의 Qin Sun 등 연구진이 개발했다. 연구 논문은 2026년 6월 9일 *Frontiers in Medicine*에 발표되었다. 특정 유방 초음파 과제에 맞춘 신경망이므로 여러 분야의 공통 기반을 목표로 하는 모델과도 구분된다. 여러 자료를 입력한다는 뜻의 멀티모달과 여러 과제에 재사용하는 파운데이션 모델은 서로 다른 분류 기준이다. 연구의 구성과 평가 조건은 다음 원논문에서 살펴볼 수 있다.

<https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2026.1802570/full>

영상의 관심 영역(ROI, region of interest)을 미리 지정했는지도 실제 적용에 영향을 준다. 연구에서 병변 부위만 잘라 입력한 조건과 진료 현장의 원본 영상을 처리하는 조건은 다르다. 평가용 정보가 학습에 섞이는 데이터 누수(data leakage)도 방지해야 한다.

자궁경부암 연구에서는 범용 멀티모달 모델로 항암화학방사선치료(chemoradiotherapy) 후 재발 위험을 추정했다. 다만 이미 끝난 진료 기록을 분석한 후향적(retrospective) 연구이며, 치료 후에 알 수 있는 정보도 포함했다. 따라서 치료 시작 전 예측 도구로 해석해서는 안 된다. 영상 판독 과정의 오류도 남아 있어 의료진 검토가 필요하다.

## 성능 평가와 일반화의 과제

대규모 모델이 모든 과제에서 기존 방법보다 우수한 것은 아니다. 단일세포 파운데이션 모델(single-cell foundation model)을 비교한 연구에서는 정규화(normalization)와 선형 방법을 적용한 기준 모델도 높은 성능을 보였다. 유전자 섭동(perturbation), 즉 유전자 기능이나 세포 환경을 변화시켰을 때의 반응을 예측하는 과제에서도 유사한 결과가 보고되었다.

따라서 새로운 모델은 적절한 기준 방법과 비교해야 한다. 학습 자료에 없던 세포 유형이나 생물 종을 다루는 분포 밖(out-of-distribution) 평가도 중요하다. 모델의 규모보다 실제 분석 목적에 적합한지를 확인하는 것이 우선이다.

여러 자료를 결합할 때는 모달리티 간극(modality gap)도 발생할 수 있다. 공통 공간에서 학습하더라도 자료 종류별 표현이 충분히 연결되지 않는 현상이다. 서로 다른 자료의 특성을 보존하면서 관련성을 학습하는 설계가 필요하다.

배치 효과(batch effect)는 기관이나 실험 조건에 따라 생기는 측정 차이이다. 병원마다 염색 시약과 스캐너가 다르고 전사체 측정 방식도 다르다. 한 기관에서 학습한 모델을 다른 기관에 적용하려면 이러한 차이를 검토하고 외부 자료로 평가해야 한다.

예를 들어 한 병원에서는 특정 질환의 표본을 주로 한 스캐너로 촬영하고 다른 표본은 다른 스캐너로 촬영했다고 하자. 모델이 질환 대신 기계가 남긴 영상 특성을 구별해도 시험 점수는 높아질 수 있다. 실제 질환의 특징을

배웠는지 확인하려면 기기와 기관이 바뀐 자료에서도 성능을 살펴야 한다. 이 예시는 자료 수집 방식이 평가에 미치는 영향을 설명한다.

### 정밀의료에서의 역할

멀티모달 파운데이션 모델은 서로 다른 검사와 기록을 연결해 질환을 이해하고 분석의 우선순위를 정하는 데 기여할 수 있다. 병리 영상에서 추가 분자 검사가 필요한 대상을 선별하는 접근이 그 예다.

다만 계산 성능과 환자의 치료 성과는 구분해야 한다. 연구용 예측이 오진 감소나 생존 개선으로 이어지는지는 임상 환경에서 확인해야 한다. 현재 사례들은 의료진과 연구자의 판단을 지원하는 역할로 이해하는 것이 적절하다. 사람이 결과를 검토하는 휴먼 인 더 루프(Human-in-the-loop) 구조와 신뢰할 수 있는 후속 검사가 함께 필요하다.

### 3.2 Octopus의 대장암 바이오마커(IGF2) 탐색

Christopher Baker 등 연구진의 Octopus 사전공개 논문은 가설 생성과 공개 생물학 자료의 분석을 연결한 구성을 설명한다. 이 연구는 컴퓨터를 이용한 후보 탐색으로, 새 치료의 임상시험과 구분된다.

<https://arxiv.org/abs/2607.16262>

신약개발에서는 많은 유전자와 분자 가운데 어떤 후보를 먼저 연구할지 정하는 일이 중요하다. AI는 기존 생물학 자료를 분석하고 가설을 제안해 후보 선정 과정을 지원할 수 있다. 벨파스트 퀸스 대학교 연구진의 Octopus는 여러 AI 에이전트와 예측 모델을 결합해 대장암의 약물 반응과 관련된 후보를 탐색한 사례다.

Octopus의 정식 이름은 Multi-Scale Autonomous Discovery Engine이다. Christopher Baker, Tianyu Ren, Karen Rafferty, Hui Wang, Simon McDade가 저자로 참여한 연구로, 2026년 사전공개본에서 소개되었다. 하나의 언어모델을 가리키는 이름보다 여러 분석 도구와 에이전트를 묶은 연구 체계의 이름에 가깝다.

이 연구에서 제시한 후보는 인슐린 유사 성장인자 2, IGF2(insulin-like growth factor 2)다. 연구진은 IGF2를 5-플루오로우라실(5-FU, 5-fluorouracil)에 대한 반응을 예측할 바이오마커(biomarker) 후보로 검토했다. 바이오마커는 질환의 상태나 치료 반응을 판단하는 지표이며, 직접 치료 대상으로 삼는 표적과는 구분된다.

## 대장암 치료 반응의 차이

5-FU는 대장암 항암치료에 오랫동안 사용된 약물이다. 그러나 같은 약물을 투여해도 환자마다 반응이 다르다. 이러한 차이를 설명하는 생물학적 특성을 찾으면 적절한 치료를 선택하거나 후속 연구의 방향을 정하는 데 도움이 된다.

기존 연구에서는 연구자가 가설을 세우고 유전자 발현 자료를 분석한 뒤 세포와 동물 실험 등으로 확인해 왔다. 많은 유전자가 함께 작용하기 때문에 분석할 후보가 많고, 한 유전자와 약물 반응 사이의 관련성을 해석하기도 어렵다.

거대언어모델(LLM)은 문헌과 자료를 바탕으로 가설을 생성할 수 있지만, 자연스러운 설명이 생물학적 타당성을 보장하지는 않는다. 사실과 다른 내용을 생성하는 환각(hallucination)을 줄이려면 가설을 실제 자료와 대조하는 과정이 필요하다.

## 여러 에이전트와 예측 모델의 결합

Octopus에서는 여러 에이전트(agent)가 대장암 전사체 자료를 분석하고 후보 가설을 제안한다. 공개 소형 모델을 연구실의 로컬 환경에서 실행하도록 구성했다. 환자 자료를 외부 서비스로 전송하지 않는 구조를 선택했다는 점에서 연구기관의 데이터 관리와도 관련이 있다.

생성한 가설은 XGBoost와 SHAP 등을 사용해 평가한다.

XGBoost의 대표 논문은 Tianqi Chen과 Carlos Guestrin이 2016년 3월 9일 사전공개했다. SHAP은 Scott Lundberg와 Su-In Lee가 2017년

5월 22일 사전공개한 예측 설명 방법이다. 두 도구는 Octopus가 새로 발명한 것이 아니라 기존에 공개된 계산 방법을 연구 목적에 맞게 연결한 것이다. XGBoost는 여러 결정 트리(decision tree)를 결합해 예측하는 기계학습 방법이다. SHAP은 특정 예측에 각 입력 정보가 얼마나 기여했는지 설명하는 기법이다. 언어모델이 후보를 제안하고, 별도의 해석 가능한 예측 모델이 자료와의 관련성을 검토하는 구조다.

결정 트리는 조건을 따라 자료를 나누는 예측 구조다. XGBoost는 여러 트리의 결과를 결합해 앞선 예측의 부족한 부분을 보완한다. SHAP은 그 예측에서 어떤 입력이 결과를 높이거나 낮추는 데 기여했는지 계산한다. 예측에 크게 기여했다는 사실은 생물학적 원인이라는 뜻과 다르다. 서로 관련된 유전자 중 하나가 다른 유전자의 정보를 대신 전달했을 가능성도 있기 때문이다.

검토에는 서로 다른 수준의 공개 자료가 사용되었다. 세포주 자료에서는 유전자 기능을 억제했을 때 암세포의 생존에 어떤 영향이 있는지 살폈다. 사람의 대장암 코호트(cohort)에서는 유전자 발현과 생존의 관련성을, 환자 유래 이식(PDX) 생쥐 자료에서는 종양 변화와의 관련성을 평가했다.

연구진이 제안한 후보를 여러 자료에서 대조하면 한 자료에만 의존한 결론을 줄일 수 있다. 다만 자료 사이의 관련성이 확인되어도 해당 유전자가 치료 반응을 직접 일으킨다는 인과관계까지 입증되는 것은 아니다.

세포주, 환자 자료, 생쥐 자료는 서로 다른 질문에 답한다. 세포주에서 보인 반응이 사람에게도 같은 크기로 나타나는지 알 수 없고, 환자의 생존과 관련된 특징이 약물 반응의 원인인지도 별도로 따져야 한다. 여러 자료가

같은 방향을 가리키면 후속 연구의 이유는 강해지지만 각 자료의 한계까지 사라지는 않는다.

## IGF2 후보의 의미

Octopus는 비지도(unsupervised) 분석을 통해 IGF2를 우선 검토할 후보로 제시했다. 기존 환자 자료와 동물 자료에서 IGF2 발현과 예후 또는 약물 반응 사이의 관련성이 보고되었다.

IGF2 자체가 이 연구에서 처음 알려진 유전자는 아니다. 대장암에서 IGF2의 각인(imprinting) 이상과 질환 위험의 관련성을 살핀 선행 연구가 있었다. 각인은 부모에게서 물려받은 두 유전자 사본 중 특정 사본의 발현을 조절하는 현상이다. Octopus의 성과는 새로운 유전자의 발견보다, 많은 후보 중 추가 연구할 대상을 계산적으로 선정한 데 있다.

바이오마커 후보를 치료에 사용하려면 후속 검증이 필요하다. IGF2 수준에 따라 5-FU의 치료 효과가 달라지는지 확인해야 하며, 치료 표적으로 개발하려면 IGF2 경로를 조절했을 때 질환의 경과가 변하는지도 살펴야 한다.

이 구분은 연구 결과를 환자 설명으로 옮길 때 중요하다. 예후 지표는 앞으로의 경과와 관련된 특징이고, 치료 반응 예측 지표는 어떤 치료에서 더 큰 이익을 얻을지 구별하는 특징이다. 두 가지가 겹칠 수 있지만 항상 같지는 않다. 따라서 관련 유전자를 찾았다는 발표를 곧바로 이 유전자를 검사하면 약을 선택할 수 있다는 말로 바꾸어서는 안 된다.

## 기존 자료 분석과 새로운 실험의 차이

Octopus 연구는 공개된 자료를 다시 분석한 후향적(retrospective) 컴퓨터 연구다. 연구진이 새로 IGF2를 억제하는 약물을 생쥐에게 투여한 것은 아니다. 기존 PDX 자료를 이용해 예측과 실제 관찰 결과가 같은 방향인지 평가했다.

컴퓨터 분석(in silico)은 후보를 좁히고 실험 계획을 세우는 데 도움이 된다. 그러나 새로운 세포 실험과 동물 실험, 임상시험을 대신하지는 못한다. 이 연구도 동료심사 전 프리프린트(preprint) 단계이므로, 확립된 치료 성과로 소개해서는 안 된다.

## 선행 연구와 연구 설계의 발전

같은 연구진은 앞서 맥락적 가역 세계 모델(CIWM, contextual invertible world model)을 제안했다. 대장암 세포주와 약물 반응 자료를 이용해 유전자 변화를 가정했을 때 예측 결과가 어떻게 달라지는지 탐색한 연구다.

이러한 모의 분석에서는 특정 변수를 바꿨을 때 예상과 다른 반응이 나타날 수 있다. 세포의 반응은 여러 경로가 함께 작용한 결과이기 때문이다. 다만 모델이 계산한 변화는 실제 세포에서 확인한 변화와 구분해야 한다. Octopus는 이런 한계를 고려해 여러 자료를 교차 검토하고 후보의 우선 순위를 정하는 방식을 사용했다.

## 신약개발 과정에서의 역할

Octopus는 가설 생성과 데이터 분석을 연결하는 AI 연구 체계를 보여 준다. 연구자는 AI가 제시한 후보를 검토하고, 후속 실험의 필요성과 우선순위를 판단할 수 있다. 공개 모델을 로컬에서 실행하는 방식은 자료 관리와 연구 환경 선택 측면에서도 의미가 있다.

후속 과제는 독립된 자료에서 결과를 재현하고, 세포와 동물 실험으로 후보의 기능을 확인하는 것이다. 임상적으로는 치료 반응을 예측하는 지표 인지, 치료를 통해 조절할 수 있는 표적인지도 구분해야 한다. AI의 역할은 이러한 검증을 생략하는 데 있지 않으며, 검증할 후보를 체계적으로 선정하도록 돕는 데 있다.

### 3.3 세포 세계 모델(Chreode)과 분자 설계

Chreode 연구진이 공개한 논문에는 입력되는 세포 상태와 개입 조건, 상태 변화의 계산 방식이 정리되어 있다. 학습 자료가 생쥐 배아 중심이라는 조건과 다른 과제로 옮겨 시험한 범위를 함께 읽을 수 있다.

<https://arxiv.org/abs/2605.28111>

신약개발에서는 특정 단백질의 기능을 조절하는 접근과 함께, 질병에 따른 세포 상태의 변화를 분석하는 접근이 사용된다. 여러 생물학적 경로가 관여하는 질환에서는 한 표적의 작용만으로 치료 반응을 설명하기 어렵기 때문이다.

전사체(transcriptome)는 세포에서 발현하는 RNA 정보를 나타낸다. 질환과 관련된 유전자 발현 양상을 질병 서명(disease signature)이라고 하며, 약물 처리 후 이 양상이 어떻게 달라지는지 분석할 수 있다.

Connectivity Map과 L1000은 이런 연구에 사용되는 대표적인 자료와 측정 체계다.

세포에 약물을 투여하거나 유전자 기능을 변화시키는 개입을 섭동(perturbation)이라고 한다. 세포 세계 모델(cell world model) 또는 가상 세포(virtual cell)는 이러한 개입에 따른 세포 상태 변화를 예측하려는 모형이다. 후보 물질이 어떤 반응을 일으킬지 계산해 후속 실험의 방향을 정하는 데 사용할 수 있다.

## 세포 상태의 변화 예측

세포 상태를 시간에 따라 분석하는 데는 측정상의 제약이 있다. 단일세포 RNA 시퀀싱(single-cell RNA sequencing)은 측정 과정에서 세포를 파괴하므로, 동일한 세포를 계속 관찰하기 어렵다. 연구자는 여러 시점에서 측정한 서로 다른 세포의 자료를 이용해 상태 변화 과정을 추정한다.

RNA 속도 분석(RNA velocity)이나 신경 미분방정식(Neural ODE)은 이러한 변화 과정을 연구하는 데 사용되어 왔다. 다만 계산 비용과 새로운 자료에 대한 재학습 부담이 발생할 수 있다. 최근 모델은 사전학습한 표현을 여러 세포와 분석 과제에 적용하려 한다.

## Chreode의 접근

Chreode는 세포 상태와 개입 조건을 바탕으로 다음 상태를 예측하는 모델이다.

미국 노스캐롤라이나 대학교 채플힐, 텍사스 대학교 오스틴, 펜실베이니아 대학교 연구진의 공동 연구이며 2026년 5월 27일 사전공개되었다. 입력에는 세포의 현재 상태뿐 아니라 어떤 개입을 가정하는지와 시간 간격이 포함된다. 출력은 그 조건에서 예상되는 다음 세포 상태다. 세계라는 이름을 붙였어도 세포 안의 모든 분자와 사람 몸 전체를 그대로 재현했다는 뜻은 아니다. 이름은 발생생물학자 콘래드 워딩턴(Conrad Waddington)이 세포 분화 경로를 설명할 때 사용한 크레오데(chreode)에서 유래했다.

이 모델은 상태 변화에 필요한 계산을 학습 과정에 반영하고, 사용할 때는 한 단계의 전이 연산으로 다음 상태를 생성하도록 설계되었다. 전이 연산

자(transition operator)는 방향성을 갖는 변화와 회전 성분, 확률적 변동을 함께 표현한다.

전이 연산자는 현재의 숫자 표현을 다음 상태의 숫자 표현으로 바꾸는 계산 규칙이다. 같은 시작 조건에서도 세포마다 반응이 달라질 수 있으므로 하나의 고정된 답만으로 변화를 설명하기 어렵다. 모델은 자료에서 이런 변화 양상을 학습한다. 한 번의 계산으로 결과를 낸다는 설명도 실제 세포의 변화가 순간적으로 끝난다는 뜻이 아니라 컴퓨터의 계산 절차를 가리킨다.

학습에는 생쥐 배아의 대규모 단일세포 자료가 사용되었다. 세포를 벡터(vector)로 표현하는 부분에는 scVI 인코더(encoder)를, 상태 변화를 계산하는 부분에는 DiT(diffusion transformer, 확산 트랜스포머) 계열 구조를 적용했다. Chreode의 해당 논문은 프리프린트(preprint), 곧 정식 심사 전에 공개한 연구다.

scVI의 선행 연구는 Romain Lopez, Nir Yosef 등 연구진이 2017년 사전공개한 단일세포 유전자 발현 모델로 이어진다. DiT는 William Peebles와 Saining Xie가 2022년 12월 19일 공개한 확산 트랜스포머 연구에서 출발한 구조다. Chreode는 이런 기존 구성 요소를 세포 상태 변화라는 목적에 맞게 조합했다.

## 다른 과제로의 적용

Chreode 연구는 발달 과정에서 학습한 표현이 다른 세포 분석에도 도움이 되는지 평가했다. 조혈과 체도 분화 자료에서 세포 분포의 변화를 예측하고, 유전자 섭동 예측 모델인 GEARS에 학습한 임베딩(embedding)을 제공했다. 일부 평가에서는 예측 오차가 감소했다.

GEARS는 스탠퍼드 대학교의 Yusuf Roohani, Kexin Huang, Jure Leskovec 연구진이 개발한 유전자 섭동 예측 방법이다. Nature Biotechnology 논문은 2023년 8월 17일 온라인 공개되었고 2024년 권호에 실렸다. 유전자 사이의 관계 정보와 딥러닝을 결합해 개입 뒤의 발현 변화를 예측한다.

재학습 없이 세포의 분화 결과를 예측하는 제로샷(zero-shot) 평가도 진행했다. 제로샷은 해당 과제를 위한 별도 학습 없이 기존에 학습한 표현을 적용하는 방식이다. 이 결과들은 세포 발달에서 얻은 정보가 다른 상태 변화 분석에 기여할 가능성을 보여 준다. 다만 특정 평가에서의 개선이 모든 세포나 약물 반응으로 일반화되는 것은 아니다.

## 세포 섭동 모델의 다양한 설계

SCALE은 처리 전후에 동일한 세포를 직접 찍지어 측정하기 어렵다는 문제를 다룬다. 세포를 집합으로 처리해 개입 후 집단의 상태를 예측하는 방식이다. 유전적 개입과 약물, 발달 및 면역 반응 등 다양한 조건을 분석 대상으로 삼았다. 대규모 섭동 자료인 Tahoe-100M도 관련 모델의 학습과 평가를 위한 연구 자료로 공개되었다.

SCALE은 상하이 인공지능연구소와 상하이대학교, 푸단대학교 등의 연구진이 개발했고 논문은 2026년 3월 18일 처음 공개되었다. 8월 31일 수정본에서는 후속 실험도 설명한다. Tahoe-100M은 이와 구분되는 데이터 자원이다. 모델의 이름과 학습에 쓰는 자료의 이름을 나누어 보면 어느 연구가 새 계산법을 제안했고 어느 연구가 새 관찰 자료를 제공했는지 알 수 있다.

PT-RAG는 검색 증강 생성(RAG, retrieval-augmented generation)을 세포 반응 예측에 적용한다. 새로운 개입의 효과를 예측할 때 과거의 관련 실험을 찾아 참고한다. 동일한 유전자도 세포 유형에 따라 효과가 달라질 수 있으므로, 개입의 유사성과 대조군 세포 상태를 함께 고려해 자료를 검색한다.

PT-RAG의 첫 논문은 이탈리아 로마 사피엔차대학교와 ISTI-CNR, 영국 케임브리지대학교와 웰컴 생어연구소에 소속된 Andrea Giuseppe Di Francesco, Andrea Rubbi, Pietro Liò가 2026년 3월 7일 사전공개했다. 기존 예측 모델에 붙여 관련 실험을 찾도록 하는 구성 요소다. 검색한 자료가 현재의 세포 조건과 맞지 않으면 도움이 적을 수 있어 검색과 예측을 함께 평가해야 한다. 연구의 구성과 평가 조건은 다음 원논문에서 살펴볼 수 있다.

<https://arxiv.org/html/2603.07233v1>

이러한 모델은 공통적으로 기존 실험의 정보를 새로운 조건에 적용하려 한다. 실제 연구에서는 학습 자료에 포함되지 않은 세포나 개입에 얼마나 잘 대응하는지가 중요하다.

## 평가 방법과 일반화

Arc Institute의 Virtual Cell Challenge는 세포 섭동 예측을 공개적으로 비교하는 평가다. 유전자 기능을 약화시키는 녹다운(knockdown) 이후의 반응을 예측하며, 새로운 세포 조건에서도 모델이 작동하는지 살핀다.

관련 비교 연구는 복잡한 모델이 모든 지표에서 선형 기준선(linear baseline)을 앞서지는 않는다고 보고했다. 반면 평가 방법을 보정하면 딥러닝 모델의 이점이 나타난다는 후속 연구도 있다. 따라서 모델의 규모나 한 개의 점수만으로 성능을 판단하기 어렵다. 발현량뿐 아니라 반응 방향과 세포 집단의 분포 등 여러 측면을 함께 평가해야 한다.

## 원하는 세포 변화에 맞춘 분자 설계

세포 반응 예측이 특정 개입의 결과를 계산한다면, 분자 설계는 원하는 결과를 일으킬 후보를 찾는 과정이다. 원하는 전사체 변화를 조건으로 새로운 분자를 생성하는 접근을 de novo 분자 설계라고 한다.

CURE는 약물 처리 전후의 전사체 정보를 이용해 기능적 특징을 추출하고, 이를 분자 생성에 연결한다. 전사체 섭동 특징 추출기(TFE)가 생물학적 변화의 표현을 만들고, 확산 모델(diffusion model)이 이를 참고해 분자 구조를 생성한다. 표적 단백질의 입체 구조뿐 아니라 세포 반응을 설계 조건으로 사용한다는 점에 의미가 있다.

CURE는 중국과학원 자동화연구소와 중국과학원대학교, 싱가포르국립대학교 소속 연구진이 참여했으며 2026년 5월 14일 논문이 사전공개되었다. 이 이름은 해당 논문의 Cellular Response Engine을 뜻한다. 원하는

세포 변화를 입력 조건으로 삼아 분자 구조 후보를 출력하지만, 출력된 구조가 실제로 합성 가능한지와 안전한지는 후속 확인이 필요하다.

JoPMol은 세포 상태, 텍스트로 제시한 설계 전략, 수치로 표현한 화학적 성질을 함께 반영한다. 생물학적으로 원하는 변화를 유도하더라도 분자의 화학적 특성이 적절하지 않을 수 있기 때문이다. 여러 조건을 함께 조절해 분자를 생성하려는 접근이다.

JoPMol은 화남사범대학교와 오사카대학교 연구진의 공동 연구로 2026년 7월 13일 공개되었다. 예를 들어 연구자가 원하는 세포 반응과 화학적 성질의 조건을 동시에 제시하면 그 조합에 맞는 후보를 찾도록 한다. 한 조건만 좋게 만드는 과정에서 다른 성질이 나빠질 수 있으므로 여러 조건 사이의 균형을 확인하는 것이 중요하다.

CURE와 JoPMol의 결과는 주로 컴퓨터 평가에서 보고되었다. 해당 논문의 계산 평가만으로 생성 분자의 합성, 동물실험, 임상 효과가 확인되었다고 볼 수는 없다.

### 실험과 연결할 때의 과제

질병과 관련된 유전자 발현 변화가 모두 질병의 원인은 아니다. 일부는 세포의 방어 반응일 수 있다. 따라서 발현 양상을 정상 상태에 가깝게 바꾸는 것이 실제로 치료에 도움이 되는지 확인해야 한다.

배양한 세포주의 반응과 사람 몸속의 반응도 다를 수 있다. 조직 환경, 면역계, 약물의 분포와 대사 등이 실제 효과에 영향을 준다. 제한된 유전자를 측정하고 나머지를 추정된 자료에서는 특정 신호가 충분히 반영되지 않을 가능성도 있다.

세포 세계 모델과 분자 생성 모델을 연결하면 후보 탐색과 실험 계획을 지원할 수 있다. 그러나 계산 결과는 후속 검증을 위한 출발점이다. 후보를 합성하고 세포와 동물에서 작용과 독성을 확인한 뒤, 임상적으로 평가하는 과정이 필요하다. 이 기술의 발전은 계산 모델의 성능 향상과 실험 자료를 통한 검증이 함께 이루어질 때 신약개발에 기여할 수 있다.



## 제4장. 디지털 트윈(Health Digital Twin)과 임상시험



## 4.1 환자 특성을 반영한 디지털 트윈

싱가포르 연구진의 PRECISE CURATE.AI 원논문에는 환자 개인의 자료로 용량 권고를 만드는 과정과 실제 진료에 적용할 수 있는지를 살핀 결과가 담겨 있다. 환자 전체를 복제한다는 이미지보다 어떤 결정을 지원했는지에 초점을 맞춰 읽을 수 있다.

<https://www.nature.com/articles/s41698-025-00835-7>

디지털 트윈(digital twin)은 실제 대상의 상태를 컴퓨터 모델로 표현하고, 새로운 자료를 반영해 변화나 개입의 결과를 예측하는 기술이다. 공학에서는 설비의 상태를 분석하고 유지보수 계획을 세우는 데 사용되어 왔다. 의료 분야에서는 환자나 장기의 특성을 반영해 질병 경과, 치료 반응, 임상시험 결과 등을 모의하는 방향으로 연구되고 있다.

헬스 디지털 트윈은 유전 정보와 영상, 진료 기록, 센서 자료 등을 연결할 수 있다. 다만 사용 목적과 구현 수준은 다양하다. 환자의 모든 생물학적 특성을 완전히 복제한 시스템을 뜻하는 것은 아니다.

모델을 만들 때는 먼저 답하려는 질문을 좁힌다. 약물 용량을 검토하려는 모델과 심장 판막의 움직임을 살피려는 모델에는 필요한 자료가 다르다. 환자의 모습을 입체 화면으로 보여 주는 기능이 있어도 치료 반응을 예측할 수 있다는 뜻은 아니다. 반대로 입체 그림 없이 숫자와 그래프로만 결과를 제시하는 모델도 특정한 예측에는 도움이 될 수 있다.

## 모델과 실제 환자의 연결

일반적인 구분에서 디지털 모델(digital model)은 과거 자료로 만든 모형이며, 실제 대상과 자동으로 연결되지 않는다. 디지털 섀도(digital shadow)는 실제 대상의 새 정보가 모델로 전달되는 형태다. 디지털 트윈은 실제 대상과 모델의 상태를 연결하고, 예측 결과를 의사결정에 반영하는 구조를 지향한다.

의료 분야에서는 환자, 데이터 연결, 컴퓨터 속 환자(patient-in-silico), 사용자 인터페이스, 동기화 등을 구성 요소로 설명한다. 실제 환자의 상태가 달라지면 모델에도 이를 반영해야 한다는 점이 중요하다. 다만 분야마다 디지털 트윈의 정의와 구현 범위는 다르다.

신뢰할 수 있는 예측을 위해서는 검증, 확인, 불확실성 정량화(WVUQ, verification, validation and uncertainty quantification)가 필요하다. 모델이 제대로 구현되었는지, 사용 목적에 맞는지, 예측에 어느 정도 불확실성이 있는지 확인하는 과정이다.

세 질문은 나누어 확인해야 한다. 수식과 프로그램이 의도한 대로 계산되는지 확인하는 일, 계산 결과가 실제 관찰과 맞는지 비교하는 일, 자료와 가정의 오차가 결과를 얼마나 흔드는지 살피는 일이다. 프로그램에 오류가 없더라도 생물학적 가정이 틀리면 예측은 부정확할 수 있다. 하나의 검사만 통과했다고 나머지 조건까지 충족한 것은 아니다.

## 적용 범위

종양 모델은 유전자 변이, 세포 특성, 약물 반응과 내성 등을 분석한다. 장기 모델은 심장이나 간 등의 구조와 기능, 약물 또는 방사선의 영향을 모의한다. 환자 수준에서는 진료 기록과 웨어러블(wearable) 기기 자료 등을 연결해 상태 변화를 추정할 수 있다.

임상시험에서는 환자 모집과 용량 선택, 시험 설계를 검토하는 데 사용할 수 있다. 치료를 받지 않았을 경우의 경과를 예측해 실제 대조군의 분석을 보조하는 접근도 있다. 적응형 설계(adaptive design)와의 연결 역시 연구되지만, 각 용도에 맞는 평가와 규제 검토가 필요하다.

## 데이터 기반 모델과 기전 기반 모델

데이터 기반 모델은 많은 환자 자료에서 패턴을 학습한다. 복잡한 영상과 유전체를 분석하는 데 적합하지만, 판단 이유를 설명하거나 학습 자료와 다른 환자에 적용하는 데 한계가 있을 수 있다.

기전 기반 모델은 알려진 생물학적 관계를 수학적으로 표현한다. 약물이 체내에 분포하고 제거되는 과정을 다루는 약동학(pharmacokinetics), 약물의 작용을 다루는 약력학(pharmacodynamics) 등이 여기에 포함된다. 해석이 비교적 명확하지만 복잡한 질환의 모든 과정을 수식으로 표현하기는 어렵다.

하이브리드 모델은 두 접근을 결합한다. 알려진 기전과 환자 자료를 함께 사용해 각각의 한계를 보완하려는 방식이다. 모델 선택은 어떤 임상 결정을 지원할 것인지에 따라 달라진다.

## 개인별 용량 조절: CURATE.AI

싱가포르 연구진의 CURATE.AI는 한 환자에게서 얻은 자료로 개인별 약물 용량을 조절하는 접근이다.

개발을 이끈 곳은 싱가포르국립대학교의 Dean Ho 연구진이다. 대학은 2018년 8월 31일 이미 CURATE.AI의 임상 연구를 소개했으며, PRECISE 결과 논문은 2025년 2월 21일 공개되었다. 새 논문이 나왔다고 그해에 플랫폼이 처음 등장한 것은 아니다. 입력은 해당 환자의 투여량과 반응 자료이고, 출력은 의료진이 검토할 개인별 용량 권고다. 진행성 고형암을 대상으로 한 PRECISE 연구에서는 카페시타빈 단독 또는 병용 치료 과정에서 AI의 용량 권고를 실제 진료에 적용할 수 있는지 살폈다.

이 사례는 대규모 집단의 평균 반응뿐 아니라 환자 개인의 반응을 치료 조절에 반영하려는 시도다. 다만 실행 가능성을 평가한 연구이므로 생존이나 치료 반응의 개선을 입증한 것으로 해석해서는 안 된다.

개인별 모델에도 충분한 관찰이 필요하다. 투여량을 바꾼 뒤 반응을 측정해야 용량과 반응의 관계를 조정할 수 있다. 환자의 상태나 함께 받는 치료가 달라지면 이전에 얻은 관계가 유지되는지도 확인해야 한다. 의료진이 입력 자료와 허용 범위를 정하는 과정은 이런 이유로 모델 운영의 일부가 된다.

## 생활습관 관리: Twin Health

클리블랜드 클리닉은 Twin Health 플랫폼을 제2형 당뇨병 환자의 관리에 적용하는 연구를 진행했다. 혈당, 체중, 혈압, 활동, 수면 등의 자료를 수집하고 개인별 식사와 활동을 안내하는 방식이다.

개발사는 미국 Twin Health이며 병원은 연구를 수행하는 기관이다. 클리블랜드 클리닉의 2025년 8월 20일 발표는 이 프로그램을 Twin Precision Treatment라고 부르고, 기기에서 얻는 자료와 사람의 원격 코칭을 함께 제공한다고 설명한다. 따라서 연구에서 나타난 변화를 AI 알고리즘 하나의 효과로 떼어 해석하기는 어렵다. 발표일은 연구 결과의 공개 시점으로 보아야 한다.

이 접근은 일상에서 반복적으로 얻는 정보를 관리 계획에 반영한다. 신약의 임상 단계와는 다른 생활습관 중재 연구이며, 플랫폼의 기능과 환자에게 나타난 효과는 구분해 평가해야 한다.

### 장기 모델: 리빙 하트 프로젝트

다쏘시스템의 리빙 하트 프로젝트(Living Heart Project)는 심장의 3차원 모델을 이용해 심장 기능과 의료기기의 영향을 연구한다. FDA와의 협력도 가상 시험 방법을 검토하는 연구 차원에서 이루어졌다.

프랑스 소프트웨어 기업 다쏘시스템은 2014년 5월 20일 리빙 하트 프로젝트의 심장 시뮬레이션 모델을 공개했다. 이는 생성형 언어모델이 등장하기 전부터 발전한 계산 모델의 사례다. 디지털 트윈이라는 범주에는 물리·생리 법칙을 계산하는 모형도 포함되므로, 모든 디지털 트윈을 언어모델 기반 AI라고 부르는 것은 정확하지 않다.

장기 모델은 실제 시험에 앞서 설계와 조건을 검토하는 데 도움이 될 수 있다. 다만 연구 협력과 규제 기관의 공식 심사 도구 채택은 다르다. 컴퓨터 모델을 심사 근거로 사용하려면 목적에 맞는 신뢰성 평가가 필요하다.

## 임상시험 시뮬레이션: 옵티비스

한국 기업 오프리메드의 옵티비스(OPTIVIS)는 환자 특성과 질병 진행을 반영한 가상 환자로 임상시험을 모의하는 플랫폼이다. 알츠하이머병 임상 자료를 이용한 환자 수 최적화 사례가 소개되었다.

회사는 동양인 환자 자료를 반영하는 접근을 강조한다. 특정 인구 집단에 편중된 자료가 다른 집단의 경과를 충분히 설명하지 못할 수 있기 때문이다. 다만 플랫폼의 소개와 실제 임상시험의 비용·기간 절감 효과는 구분해야 한다.

오프리메드는 공식 제품 소개에서 가상 환자 생성과 외부 대조군 설계, 임상시험 기간과 환자 수 예측 등을 제시한다. OPTIVIS는 회사의 플랫폼 이름이고 특정 질환의 예측 모형은 그 안의 개별 적용 사례로 이해할 수 있다. 제품 소개에는 최초 공개일이 명확히 적혀 있지 않아 특정 연도를 최초 출시 연도로 삼기 어렵다.

## 임상시험 분석 지원: PROCOVA

Unlearn의 PROCOVA는 실제 무작위 임상시험에 참여한 환자의 대조 조건 경과를 예측하고 이를 통계적 보정에 사용하는 방법이다. 유럽의약품청(EMA)은 이 방법에 대해 자격 인정 의견(qualification opinion)을 제시했다.

이 사례는 디지털 트윈의 쓰임이 치료 추천 외에 시험 설계와 통계 분석으로도 확대되고 있음을 보여 준다. 실제 대조군을 모두 없애거나 가상 자료만으로 신약을 허가한다는 뜻은 아니다. 구체적인 적용 조건은 다음 절에서 살펴본다.

## 도입과 검증의 과제

규제 관점에서는 사용 목적, 자료의 품질, 위험 수준, 모델의 성능과 관리 방식이 중요하다. 치료 선택에 직접 영향을 주는 모델과 연구 설계를 보조하는 모델에는 서로 다른 수준의 근거가 필요하다.

다른 병원이나 인구 집단에 적용할 때 성능이 달라지는 도메인 이동(domain shift)도 검토해야 한다. 데이터 표준과 병원 시스템 연결, 개인 정보 보호, 지속적인 갱신 역시 운영에 필요한 조건이다. 실사용 데이터(real-world data)가 풍부하더라도 환자 선택과 측정 방식의 차이로 비교 결과가 달라질 수 있다.

현재 사례들은 개인별 용량 조절, 생활습관 관리, 장기 시뮬레이션, 임상시험 분석 등 특정 목적에 초점을 둔다. 디지털 트윈의 임상적 가치는 적용 범위를 명확히 하고, 예측이 실제 의사결정과 환자 결과에 어떤 도움을 주는지 검증하면서 평가해야 한다.

## 4.2 가상 대조군(Synthetic Control Arm) 임상시험

Unlearn의 공식 기술 설명은 실제 무작위 시험의 분석을 돕는 PROCOVA와 외부 대조군을 구분한다. 규제 인정의 범위는 EMA의 자격 인정 의견에서 확인할 수 있다.

<https://www.unlearn.ai/blog/built-on-regulatory-rigor-advancing-digital-twins-in-clinical-trials>

[https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm\\_en.pdf](https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm_en.pdf)

임상시험의 대조군(control group)은 새 치료의 효과를 비교하는 기준이다. 질환과 시험 목적에 따라 기존 치료나 위약(placebo)을 받으며, 연구자는 시험군과 대조군의 결과를 비교해 효과와 안전성을 평가한다. 무작위 대조시험(RCT, randomized controlled trial)은 집단 간 차이를 줄여 치료 효과를 판단하는 중요한 방법이다.

그러나 대조군 모집에는 시간과 비용이 필요하다. 환자가 적은 희귀질환에서는 충분한 인원을 확보하기 어렵고, 비교 치료의 선택에는 윤리적 검토도 필요하다. 이러한 문제를 보완하기 위해 기존 환자 자료와 예측 모델을 시험 설계 및 분석에 사용하는 접근이 발전하고 있다.

### 예측값을 이용한 보정과 외부 대조군

가상 대조군이라는 표현에는 서로 다른 방법이 포함된다. 하나는 실제 무작위 대조군을 유지하면서 AI 예측을 분석에 사용하는 보정 방식이다. 다른 하나는 과거 임상시험이나 진료 기록 등 외부 자료를 비교 대상으로 사용하는 방식이다. 두 방법은 설계와 편향 관리 조건이 다르다.

보정 방식에서는 환자별로 대조 치료를 받았을 때의 경과를 예측해 예후 점수(prognostic score)를 만든다. 이를 통계 분석에 반영하면 환자 간 차이의 일부를 설명할 수 있다. Unlearn의 PROCOVA가 대표적이다.

미국 기업 Unlearn이 제안한 PROCOVA는 예후 공변량 보정이라는 통계 방법의 이름이다. 예후 점수를 만드는 기계학습 모델과 그 점수를 시험 분석에 넣는 방법을 구별해야 한다. AI는 환자별 예상 경과를 계산하고, 통계 분석은 실제로 관찰한 시험군과 대조군의 차이를 평가한다. AI의 예측이 치료 효과의 정답으로 사용되는 구조는 아니다.

외부 대조군(external control) 또는 합성 대조군(synthetic control)은 현재 시험 밖의 자료에서 비교 대상을 구성한다. 예를 들어 모든 참가자가 새 치료를 받는 단일군 시험(single-arm trial)의 결과를 과거 환자의 경과와 비교할 수 있다. 다만 집단이 처음부터 다를 수 있으므로 선택 편향과 측정 차이를 별도로 다뤄야 한다.

## PROCOVA의 작동 원리

임상시험의 결과는 치료 효과 외에도 환자의 초기 상태와 질환 진행 속도에 영향을 받는다. 이러한 차이를 사전에 예측해 분석에 반영하면 결과의 분산(variance)을 줄일 수 있다. 적절한 조건에서는 목표한 분석 정밀도를 더 적은 환자로 확보할 가능성이 있다.

PROCOVA는 과거 대조군 자료로 학습한 예측값을 공변량 보정(covariate adjustment)에 사용한다. 실제 무작위 배정을 유지하며 치료 효과를 분석하는 구조다. 실제로 필요한 환자 수와 효율 개선의 크기는 질환, 평가 지표, 예측 정확도와 시험 설계에 따라 달라진다.

공변량은 결과에 영향을 줄 수 있어 분석에서 함께 고려하는 특성이다. 나이나 치료 전 상태처럼 시험 시작 전에 알 수 있는 정보가 여기에 들어갈 수 있다. 예후 점수는 여러 특성에서 예상되는 경과를 하나의 정보로 정리한다. 예측이 도움이 되는 정도가 작다면 환자 수를 줄이는 이점도 작아질 수 있다. 어떤 시험에서도 일정 비율로 인원이 줄어든다고 말할 수 없는 이유다.

유럽의약품청(EMA)은 PROCOVA에 대해 연속형 결과를 사용하는 2·3상 무작위 임상시험 분석의 통계 방법으로 자격 인정을 제시했다. 이는 특정 조건에서 분석 방법을 인정한 것이며, 가상 자료만으로 신약을 허가한다는 뜻은 아니다.

EMA 문서는 2022년 9월 20일자로 공개되었고 위원회 채택일은 같은 달 15일이다. 문서에서 말하는 연속형 결과는 점수나 측정값처럼 수치가 이어지는 결과를 뜻한다. 모든 종류의 결과 지표와 모든 통계 방법을 한꺼번에 인정한 문서로 넓혀 읽어서는 안 된다.

미국 FDA도 무작위 임상시험의 공변량 보정 원칙을 지침으로 정리했다. 다만 일반적인 분석 원칙과 특정 기업 방법에 대한 별도 승인은 구분해야 한다. 실제 적용에서는 시험 목적과 분석 계획에 맞춰 규제 기관과 협의해야 한다.

## 외부 대조군과 디지털 트윈

외부 자료를 비교 대상으로 사용하는 방식은 AI 이전부터 존재했다. 국제 지침 ICH E10도 외부 대조의 장단점을 다룬다. 희귀질환처럼 무작위 비교가 어려운 상황에서 질환의 자연 경과 자료나 과거 시험 결과가 검토되어 왔다.

AI와 디지털 트윈은 외부 자료를 이용해 현재 시험 참가자의 특성에 맞는 대조 경과를 예측하려 한다. 집단의 평균만 사용하는 대신 개인별 정보를 반영하는 접근이다. 그러나 예측이 정교해져도 무작위 배정의 부재로 생기는 편향이 자동으로 사라지지 않는다.

예를 들어 과거 자료의 환자가 현재 시험 환자보다 병이 덜 진행된 상태였다면 경과 차이의 일부는 치료 효과와 무관할 수 있다. 나이와 검사 수치가 비슷해 보이더라도 기록하지 않은 건강 상태나 치료 선택 이유가 다를 수 있다. 측정하지 않은 차이는 통계 프로그램에 넣어 조정하기도 어렵다. 외부 자료가 많다는 사실만으로 좋은 비교가 만들어지지 않는다.

규제 기관은 외부 자료의 적절성, 질환 경과, 비교 집단의 차이, 결과 측정 방식 등을 검토한다. 관련 지침과 논의 문서가 마련되고 있지만 모든 질환과 시험에 동일하게 적용할 수 있는 대체 수단은 아니다.

## 실사용 자료가 보여 주는 한계

RCT DUPLICATE는 실사용 데이터(real-world data)로 기존 무작위 임상시험의 결과를 얼마나 재현할 수 있는지 살핀 연구다. 보험 청구 자료 등을 이용해 비교 집단을 구성하고, 성향점수(propensity score)로 측정 가능한 차이를 조정했다.

일부 시험에서는 유사한 결과를 얻었지만 모든 결론이 일치하지는 않았다. 원래 시험의 대상자 조건과 치료, 결과 측정 방식을 가깝게 재현할 수 있을 때 일치도가 높았다. 진료 기록에 필요한 정보가 없거나 위약 대조를 재현하기 어려운 경우에는 차이가 커질 수 있다.

이 결과는 가상 대조군을 구성할 때도 자료의 적합성이 중요함을 보여 준다. 알고리즘의 성능뿐 아니라 원자료가 시험의 질문에 답할 수 있는지 확인해야 한다.

분석 전에 결과 지표와 비교 시점, 환자 선정 기준을 정하는 것도 중요하다. 결과를 본 뒤 유리한 환자만 고르거나 측정 시점을 바꾸면 치료 효과가 크게 보일 수 있다. 모델을 학습한 자료와 시험의 결과 자료를 분리하고, 분석 계획을 남겨 두어야 다른 연구자가 같은 비교를 재현할 수 있다.

### 적용이 검토되는 분야

희귀질환에서는 환자 모집의 어려움 때문에 외부 대조군의 필요성이 크다. 그러나 적은 환자 수가 비교 자료의 한계를 없애지는 않으므로 질환별로 적절성을 판단해야 한다.

알츠하이머병과 같은 신경질환에서는 개인별 진행 차이를 예측해 분석 효율을 높이는 접근이 검토된다. 한국의 오프리메드가 소개한 임상시험 시뮬레이션도 관련 사례다. 다만 과거 자료를 이용한 모의 설계와 새로운 시험에서 실제로 얻은 성과는 다르다.

항암제 개발에서도 합성 대조군과 디지털 트윈을 시험 설계에 적용하는 연구가 진행되고 있다. 기업 간 협업이나 교육용 설계 사례를 실제 허가 성과로 확대해 설명해서는 안 된다.

## 시험 설계에서 확인할 사항

과거 자료는 당시의 표준 치료를 반영한다. 그사이 치료법이나 진단 기준이 바뀌었다면 현재 시험과 직접 비교하기 어렵다. 임상시험은 정해진 시점과 방식으로 결과를 측정하지만 진료 기록은 그렇지 않을 수 있다. 누락된 정보와 환자 선택의 차이도 고려해야 한다.

따라서 가상 대조군의 목적과 생성 방법, 자료의 출처, 분석의 불확실성을 명확히 제시해야 한다. 개발사와 시험 수행 기관의 역할, 검토 책임도 사전에 정해야 한다.

가상 대조군의 의의는 임상시험의 부담을 줄이고 기존 자료를 더 체계적으로 사용하는 데 있다. 현재의 주요 접근은 실제 무작위 시험을 예측 정보로 보완하는 방식이다. 외부 자료로 대조군을 대신하려는 경우에는 비교 가능성과 편향을 충분히 검토해야 한다. 모집 효율과 근거의 신뢰성을 함께 확보하는 것이 핵심 과제다.

### 4.3 모델 기반 약물 개발과 동물실험 대체(FDA Modernization Act)

FDA의 NAMs 설명과 MIDD 공동회의 안내에는 새로운 시험법의 범위와 모델을 개발 판단에 사용하는 조건이 정리되어 있다. 두 문서를 함께 읽으면 사람 세포 실험과 컴퓨터 예측이 신약개발에서 맡는 역할을 구분할 수 있다.

<https://www.fda.gov/science-research/science-and-research-special-topics/new-approach-methodologies-nams>

<https://www.fda.gov/drugs/development-resources/model-informed-drug-development-paired-meeting-program>

신약개발에서는 효능과 안전성을 확인하기 위해 많은 시간과 자원이 필요하다. 동물에서 얻은 결과가 사람에게 그대로 적용되지 않는다는 점도 개발의 어려움이다. 약물의 흡수와 대사, 면역 반응, 치료 표적의 작동 방식에는 중간 차이가 있다.

이러한 문제를 보완하는 접근으로 사람 세포 기반 시험과 컴퓨터 모델의 사용이 확대되고 있다. 새로운 접근 방법론인 NAMs(New Approach Methodologies)는 비임상 평가의 방법을 다양화한다. 모델 기반 약물 개발인 MIDD(Model-Informed Drug Development)는 자료와 수학적 모형을 결합해 개발 과정의 의사결정을 지원한다.

NAMs와 MIDD는 회사가 출시한 AI 제품명이 아니다. NAMs는 시험 방법의 범주이고 MIDD는 모델에서 얻은 정보를 개발 결정에 반영하는 접근이다. 장기칩은 작은 장치에서 사람 조직의 기능 일부를 구현하며, 컴퓨터 모

델은 자료와 가정을 계산한다. 서로 다른 종류의 근거를 제공하므로 한 방법의 장점을 다른 방법의 검증 없이 대신 적용할 수는 없다.

## 동물실험 대체 방법의 확대

FDA Modernization Act 2.0은 비임상 시험에 세포 기반 분석, 장기칩 (organ-on-a-chip), 컴퓨터 모델 등의 방법이 포함될 수 있도록 법률 표현을 정비했다. 동물실험을 금지한 것이 아니라, 적절한 다른 방법을 사용할 수 있는 근거를 명확히 한 것이다. 실제 개발에서는 필요한 경우 동물 연구가 계속 사용된다.

후속 제도 정비는 법률의 취지를 세부 규정과 심사 실무에 반영하는 과정이다. FDA Modernization Act 3.0 관련 논의와 FDA의 동물실험 감소 로드맵도 이러한 흐름에 속한다. 다만 법안, 시행 중인 규정, 지침 초안은 법적 지위가 다르므로 구분해야 한다.

FDA가 2026년 9월의 용어 정비를 설명한 자료도 특정 시험법을 의무화 하거나 요구하는 증거 기준을 낮추는 변경이 아니라고 명시한다. 동물실험을 줄이려는 정책 방향과 개별 신약에 필요한 안전성 자료의 판단은 함께 읽어야 한다. 제목에 현대화나 대체가 들어간다는 이유만으로 모든 동물시험이 사라졌다고 이해해서는 안 된다.

NAMs에는 독성과 약동학을 예측하는 컴퓨터 모델, 사람 조직의 기능을 모의하는 장기칩, 줄기세포 등으로 만든 오가노이드(organoid)가 포함된다. 단일클론항체(monoclonal antibody)처럼 동물과 사람의 표적 차이가 중요한 영역에서도 적용이 검토된다.

이들 방법의 목적은 특정 시험을 무조건 없애는 데 있지 않다. 평가하려는 위험에 적합한 자료를 확보하는 것이 우선이다. 새로운 방법도 재현성과 예측력, 적용 범위를 검증해야 하며, 필요한 증거 수준이 낮아지는 것은 아니다.

## 모델 기반 약물 개발의 개념

MIDD는 비임상 자료, 임상 자료, 기존 지식을 통합한 모델링과 시뮬레이션으로 개발 결정을 지원한다. 초기 인체 투여 용량, 적절한 투여 간격, 특수 환자 집단의 용량, 약물 상호작용 등 다양한 질문을 다룬다.

약동학은 몸이 약물을 흡수하고 분포시키고 제거하는 과정을 다루며, 약력학은 약물이 몸에 어떤 변화를 일으키는지 다룬다. 같은 양의 약을 투여해도 몸속에 남는 농도가 다를 수 있고, 같은 농도에서도 반응이 다를 수 있다. 두 관계를 연결해야 용량 숫자를 환자의 반응과 함께 해석할 수 있다.

집단 약동학·약력학(PopPK/PD)은 나이, 체중, 장기 기능 등에 따른 약물 농도와 반응의 차이를 추정한다. 생리 기반 약동학(PBPK)은 장기의 크기와 혈류, 대사 효소 등의 정보를 이용해 체내 약물 분포를 계산한다. 정량적 시스템 약리학(QSP)은 약물과 생물학적 신호 경로의 상호작용을 모형화한다. AI와 머신러닝(AI/ML)도 관련 자료 분석과 예측에 사용될 수 있다.

이러한 방법은 기존 자료를 연결해 추가 시험이 필요한 부분을 판단하고, 시험 설계를 구체화하는 데 도움을 준다. 그러나 모델의 예측을 어떤 결정에 사용할 수 있는지는 자료의 품질과 검증 수준에 따라 달라진다.

예를 들어 용량 간격을 검토하는 모델에서는 투여 뒤 약물 농도가 시간에 따라 어떻게 달라지는지가 핵심 자료가 된다. 모델이 특정 환자 집단의 자료로 만들어졌다면 다른 집단에서도 같은 관계가 유지되는지 살펴야 한다. 실제로 측정한 범위를 벗어난 예측일수록 추가 근거가 필요할 수 있다. 계산한 값을 보고서에 적을 때는 적용한 가정과 자료 범위도 함께 남겨야 한다.

## 규제 기관과의 협의

FDA의 MIDD 공동회의 프로그램(MIDD Paired Meeting Program)은 개발사가 모델 기반 접근을 규제 전문가와 논의하는 제도다. 모델의 목적과 가정, 평가 방법을 개발 초기에 협의하면 이후 근거 생성 계획을 세우는데 도움이 된다.

ICH M15는 MIDD의 평가와 문서화에 관한 공통 기준을 마련하는 지침이다. 모델의 사용 맥락과 의사결정에 미치는 영향을 고려해 필요한 검증 수준을 정하는 접근을 강조한다. FDA와 EMA가 제시한 신약개발 AI 실무 원칙도 사람 중심 설계, 자료 관리, 위험에 따른 평가, 지속적인 관리를 중요하게 다룬다.

구체적인 법률과 규정의 적용은 문서의 확정 여부와 관할에 따라 달라진다. 기업의 협의 경험이나 방법 소개를 일반적인 규제 승인으로 확대해 이 해해서는 안 된다.

## NAMs와 MIDD의 연결

NAMs는 사람의 생물학적 특성을 반영한 자료를 제공하고, MIDD는 이러한 자료를 해석해 용량과 시험 조건을 검토할 수 있다. 예를 들어 사람 세포에서 측정된 대사와 독성 정보를 약동학 모델에 반영하고, 임상 자료가 확보되면 모델을 갱신하는 방식이다.

임상 단계에서는 용량 선택과 환자 집단 분석, 시험 설계에 모델을 적용할 수 있다. 앞 절에서 살핀 예후 점수 보정과 외부 대조군도 목적과 조건에 따라 검토할 수 있는 방법이다. 다만 NAMs와 MIDD는 서로 연결될 수 있는 접근이지, 모든 개발에서 하나의 정해진 절차로 적용되는 제도는 아니다.

## AI 신약개발의 임상 진척

인실리코 메디슨(Insilico Medicine)의 렌토서티브(rentosertib, 옛 이름 ISM001-055)는 특발성 폐섬유증(IPF) 치료 후보물질이다. 회사는 생성형 AI를 표적 탐색과 분자 설계에 사용했다고 설명했다.

렌토서티브는 AI의 이름이 아니라 치료 후보물질의 이름이다. 개발사 인실리코 메디슨은 2026년 7월 7일 3상 개시를 발표하면서 표적 선정에는 PandaOmics, 분자 생성과 최적화에는 Chemistry42를 사용했다고 설명했다. 두 도구는 같은 회사의 연구 플랫폼 안에서도 맡은 일이 다르다. 인실리코 메디슨의 공식 기술 소개에 따르면 PandaOmics와 Chemistry42는 모두 2020년에 출시되었다. 이는 렌토서티브의 임상시험 시작보다 앞선 도구 출시 시점이다. 3상 개시는 약이 허가되었다는 뜻과 구분된다. 관련 연구에서는 TNIK를 표적으로 하는 후보를 개발해 임상

평가로 연결했다. 두 도구의 출시 연도와 표적 탐색·분자 설계 기능은 회사의 기술 소개에서 확인할 수 있다.

<https://insilico.com/mediakit>

초기 임상에서는 안전성과 내약성, 폐활량 지표(FVC, forced vital capacity) 등의 변화를 살폈다. 이후 더 긴 기간과 더 많은 환자를 대상으로 효과를 평가하는 3상 개발로 이어졌다. 초기 결과만으로 장기 안전성과 치료 효과를 확정할 수는 없다.

이 사례는 AI로 탐색한 후보도 단계적인 임상 평가를 거친다는 점을 보여 준다. AI 신약개발 사례와 MIDD의 구체적인 사용 사례도 구분해야 한다. AI가 표적이나 분자를 제안했다고 해서 개발 전 과정에 모델 기반 의사결정이 적용되었다고 볼 수는 없다.

### 개발 속도와 성공 가능성

Exscientia와 스미토모 다이니폰 제약의 강박장애 치료 후보 DSP-1181은 AI 설계와 빠른 임상 진입으로 알려졌으나 이후 개발이 중단되었다. 빠른 후보 생성이 임상 성공을 보장하지 않는다는 점을 보여 주는 사례다.

AI 중심 기업이 개발한 후보를 분석한 연구에서는 초기 임상의 성과가 긍정적으로 나타났지만, 후기 단계까지 일관된 우위가 확인된 것은 아니다. 후보 수와 관찰 기간도 제한적이므로 그 차이를 AI 자체의 효과로 단정하기 어렵다.

임상이 진행될수록 약물의 물성뿐 아니라 질환의 복잡성, 환자 차이, 동반 질환 등이 중요해진다. 세포 세계 모델과 디지털 트윈도 이러한 변수를 더

잘 이해하기 위한 연구에 기여할 수 있지만 임상 검증을 생략하게 하지는 않는다.

## 한국의 연구와 데이터 협력

한국에서도 의약품 개발에 AI를 적용하는 안내와 공동 연구 사업이 추진되고 있다. K-멜로디는 기업의 자료를 외부에 직접 모으지 않고 공동 모델을 학습하는 연합학습(federated learning) 기반 접근이다.

사업단 소개에 따르면 K-MELLODDY는 과학기술정보통신부와 보건복지부가 2024년 4월부터 공동 추진한 사업이다. 하나의 완성된 AI 모델 이름보다 공동 학습 기반을 만드는 프로젝트 이름에 가깝다. 각 기관에서 계산한 학습 정보를 모아 공동 모델을 개선하므로 원자료를 한곳에 모으는 부담을 줄일 수 있다. 다만 자료를 옮기지 않는다는 조건만으로 모든 보안 문제가 해결되는 것은 아니며 공유되는 학습 정보와 접근 권한도 관리해야 한다.

민감한 연구 자료를 보호하면서 여러 기관의 정보를 학습에 반영하려는 방식이다. 다만 공동 모델의 성능을 확보하려면 자료 정의와 품질, 기관 간 차이, 평가 절차를 함께 관리해야 한다.

## 앞으로의 과제

사람 세포와 컴퓨터 모델을 사용하는 범위가 확대되어도 모든 동물실험이나 임상시험이 곧바로 대체되는 것은 아니다. 새로운 치료 방식이나 학습 자료에 없는 조건에서는 예측의 불확실성이 커질 수 있다. 모델의 결과를 생물학적으로 설명하고 실제 관찰과 비교하는 과정도 필요하다.

NAMs와 MIDD의 변화는 개발 근거를 만드는 방법이 다양해지고 있음을 뜻한다. 새로운 후보를 탐색하고 적절한 용량과 시험 설계를 선택하는 데 도움을 줄 수 있다. 최종적으로는 해당 사용 목적에 맞는 검증과 환자에게서 얻은 안전성·효과 자료를 통해 가치를 판단해야 한다.



## Part 3. 해외 의료 AI 도입 사례



## 제5장. 미국 의료기관의 AI 도입



## 5.1 메이요 클리닉의 VoiceCare AI와 원무 자동화

메이요 클리닉 플랫폼의 공식 해설은 원무와 환자 교육, 안전 관리에 쓰는 에이전트를 설명한다. VoiceCare의 시범사업 범위는 개발사가 낸 2025년 2월 발표문에서 확인할 수 있다.

<https://www.mayoclinicplatform.org/2026/08/07/ai-agents-in-healthcare/>

<https://www.prnewswire.com/news-releases/agentic-ai-startup-voicecare-ai-launches-to-automate-healthcare-back-office-and-super-staff-workforce-302376944.html>

병원의 업무에는 진료 외에도 보험 혜택 확인, 사전승인 신청, 청구 후속 처리 등이 포함된다. 원무(院務) 또는 백오피스(back office) 업무는 여러 기관의 정보를 확인하고 결과를 병원 시스템에 기록하는 과정이다. 전화와 문서 처리가 반복되기 때문에 의료 AI의 적용 분야로 관심을 받고 있다.

원무 자동화는 의료진과 행정 인력의 반복 업무를 줄일 수 있다. 다만 보험 정보나 승인 결과를 잘못 처리하면 환자의 비용 부담과 치료 일정에 영향을 줄 수 있으므로, 정확성과 검토 절차가 중요하다.

### 보험사 통화를 지원하는 음성 에이전트

VoiceCare AI는 개발 회사의 이름이고 Joy는 그 회사가 만든 음성 에이전트다. 회사의 공개 출범과 메이요 협력 발표일은 2025년 2월 14일이다. 개발을 시작한 날이나 메이요 전 부서에서 사용을 시작한 날을 뜻하지는 않는다. 환자에게 직접 진단을 내리는 제품으로 이해해서도 안 된다.

VoiceCare AI는 2025년 2월 메이요 클리닉과 보험 관련 원무 업무의 시범사업을 추진한다고 발표했다. 대상은 신경과, 소아과, 의료·행정 지원 부서였으며, 환자의 보험 혜택 확인과 사전승인이 주요 업무였다.

음성 AI 에이전트 조이(Joy)는 보험사와 통화하고 대기한 뒤 필요한 정보를 확인하며, 통화 내용을 요약해 기록하도록 설계되었다. 이 사례의 목적은 보험사와의 반복적인 소통을 자동화하는 데 있다. 환자 진료와 예약 전체의 자율 처리를 뜻하지 않는다.

공개 자료는 시범사업의 방향을 설명하지만, 메이요 전체의 도입 완료나 구체적인 시간·비용 절감 효과를 입증하지는 않는다. 메이요는 해당 기술에 재정적 이해관계를 갖는 것으로 공지되어 있다. 따라서 도입 발표와 독립적인 성과 평가는 구분해서 읽어야 한다.

### 답변 생성에서 업무 수행으로

예를 들어 병원 직원이 검사 전 보험 적용 여부를 확인한다고 하자. 확인할 환자와 검사 정보를 준비하고 보험사의 답변을 들어야 한다. 필요한 항목이 빠졌으면 다시 묻고 확인 결과를 남긴다. 음성 에이전트도 이런 순서를 처리 대상으로 삼는다. 말소리를 글로 바꾸는 기능, 다음 질문을 고르는 기능, 답을 음성으로 전달하는 기능이 연결된다. 통화가 끝났다는 사실과 보험 확인이 정확히 끝났다는 사실은 다르다.

에이전트(agent)는 언어모델의 출력을 다음 행동과 연결하는 시스템이다. 거대언어모델(LLM, Large Language Model)이 질문에 답변을 생성한다면, 에이전트는 그 결과를 바탕으로 전화를 걸거나 정보를 조회하고 기록을 갱신하는 일련의 업무를 수행할 수 있다.

메이요 클리닉 플랫폼의 존 하람카(John Halamka)와 폴 체라토(Paul Cerrato)는 의료 에이전트를 설명하면서 이러한 행동의 연결을 강조했다. 원무에서는 보험사 확인, 질문과 응답, 결과 기록, 예외 상황의 담당자 인계가 하나의 과정으로 구성된다.

실제 구현에서는 시스템 간 연결 통로인 API(Application Programming Interface)를 사용하거나 음성과 화면 입력을 처리할 수 있다. 중요한 것은 어떤 방식으로 연결하느냐뿐 아니라, 에이전트가 처리할 수 있는 범위와 사람에게 넘길 조건을 명확히 정하는 것이다.

### 환자 교육으로의 확대

MedEduChat은 메이요와 코넬대, MIT 소속 연구자들이 참여한 연구 시스템이다. 관련 논문은 2025년 12월 18일에 나왔다. 메이요가 환자용 상용 서비스를 전면 출시했다는 발표와는 구별해야 한다. 연구의 출발점은 진료가 끝난 뒤에도 환자가 자신의 기록을 이해하도록 돕는 데 있다.

작동 순서는 질문, 관련 기록 확인, 설명 생성으로 이해할 수 있다. 예를 들어 환자가 기록에 적힌 치료 용어의 뜻을 묻는다고 하자. 시스템은 질문과 관련된 기록을 참고해 읽기 쉬운 설명을 만든다. 기록에 없는 검사 결과를 추측해 채우면 안 된다. 기록 자체가 오래되었거나 빠져 있다면 설명에도 그 한계가 남는다.

메이요의 MedEduChat은 전립선암 환자의 이해를 돕는 연구 사례다. 전자의무기록(EHR, Electronic Health Record)의 환자 정보를 바탕으로 의학 용어와 진료 내용을 쉽게 설명하는 방향으로 개발되었다.

초기 연구에서는 환자와 전문가가 사용성과 답변을 평가했다. 의무기록에 근거해 설명하는 기능은 가능성을 보였지만, 개인의 상황에 맞춘 답변은 추가 개선이 필요한 부분이었다. 사용자가 편리하게 느끼는 것, 정보가 정확한 것, 치료 결과가 좋아지는 것은 각각 별도로 평가해야 한다.

이 연구는 소규모 사용성 평가이므로 다양한 환자 집단과 실제 임상 효과에 대한 후속 검증이 필요하다. 환자 교육 에이전트는 의료진의 설명을 지원하고 진료 내용을 이해하도록 돕는 역할로 볼 수 있다.

### 환자 안전 업무와 AI

PSVT는 환자 안전을 관리하는 프로그램 이름이다. 하나의 언어모델이나 판매 제품 이름으로 보기 어렵다. 메이요의 2026년 8월 해설은 위험 분류 에이전트가 운영 단계에 있고, 패턴 탐색 에이전트는 개발 중이라고 설명했다. 같은 프로그램 안에서도 기능마다 준비 수준이 다를 수 있다는 뜻이다.

메이요는 낙상, 병원 감염, 욕창, 약물 관련 사건, 상태 악화 등을 다루는 환자 안전 활력징후(PSVT) 프로그램도 소개했다. 환자의 위험을 분류하고 자료에서 패턴을 찾는 에이전트를 개발하는 접근이다.

이런 시스템은 입원 중 축적되는 정보를 정리해 의료진이 검토할 대상을 선정하는 데 도움을 줄 수 있다. 다만 프로그램의 운영 또는 개발 소식과 실제 사고 감소 효과는 다르다. 경보의 정확성, 불필요한 알림, 의료진의 대응 부담까지 함께 평가해야 한다.

## 기업 협력과 적용 분야

메이요는 NVIDIA와의 협력에서도 디지털 병리와 생성형 AI 인프라를 연구하고 있다. 다만 이 협력과 VoiceCare의 보험사 통화 시범사업은 서로 다른 적용 분야다.

한 의료기관 안에서도 원무, 환자 교육, 안전 관리, 병리 분석의 목적과 요구 조건은 다르다. 기관의 전체 AI 전략을 이해하려면 개별 사업이 어떤 업무를 대상으로 하는지 구분해야 한다.

## 원무 자동화의 운영 조건

검토 화면에는 결론과 함께 확인 날짜와 근거가 보여야 한다. 예를 들어 승인번호는 받았지만 특정 기간에만 유효할 수 있다. 번호만 복사하고 유효기간을 빠뜨리면 다음 담당자는 허용 범위를 오해할 수 있다. 불완전한 답을 빈칸으로 남겨 확인을 요청하는 편이, 그럴듯한 값으로 채우는 것보다 업무를 정확하게 이어 준다.

사전승인과 보험 확인에서는 보험자 정보, 승인번호, 적용 조건 등을 정확히 기록해야 한다. 통화가 중단되거나 답변이 불분명한 경우에는 담당자가 이어서 처리할 수 있어야 한다. 환자의 권리나 비용에 영향을 주는 결과에는 적절한 확인 절차가 필요하다.

자동화 편향(automation bias)도 고려해야 한다. 시스템의 결과를 충분히 검토하지 않고 수용하면 입력이나 해석의 오류가 후속 업무에 반영될 수 있다. 사람이 검토하는 지점(human-in-the-loop)을 업무 위험에 맞게 정해야 한다.

보험사 통화에는 이름, 진단, 보험번호 등 민감한 정보가 포함된다.

VoiceCare는 HIPAA 준수와 SOC 2 Type II를 설명하고 있으나, 정보보호 체계와 임상적 안전성은 구분해야 한다. HIPAA는 미국 의료정보 보호에 관한 법적 요구와 관련되고, SOC 2 Type II는 일정 기간 보안 통제의 운영을 평가하는 감사 체계다.

메이요의 사례는 의료 AI의 도입이 진단 알고리즘에 한정되지 않음을 보여 준다. 반복적인 원무 업무를 자동화하고 환자에게 정보를 설명하며 안전 관리 자료를 정리하는 방향으로 적용 범위가 넓어지고 있다. 각 업무에서 정확성과 실제 효과를 확인하고, 예외 상황을 사람이 처리할 수 있도록 구성하는 것이 도입의 기본 조건이다.

## 5.2 Tempus의 정밀 종양학 지원

Tempus의 Next 공식 제품 페이지에서는 진료 지침에 따른 다음 단계와 놓친 진료·추적 관리 대상을 확인하는 기능을 살펴볼 수 있다. Lens의 2026년 발표문에서는 연구 질문을 분석 계획과 코드 실행으로 연결하는 기능을 읽을 수 있다.

<https://www.tempus.com/solutions/next/>

<https://www.tempus.com/news/pr/tempus-unveils-the-next-generation-of-lens-expanding-its-agentic-ai-platform-for-oncology-drug-development/>

정밀 종양학(precision oncology)은 종양의 분자적 특성과 환자 상태를 반영해 치료를 선택하는 접근이다. 같은 암종이라도 유전자 변이나 단백질 발현에 따라 적절한 치료가 달라질 수 있다. 이를 위해서는 필요한 검사가 시행되고, 결과가 치료 결정에 연결되어야 한다.

실제 진료에서는 검사 대상자를 놓치거나 검사 결과가 충분히 반영되지 않는 문제가 생길 수 있다. Tempus는 임상 기록과 분자 정보를 분석해 이러한 진료 공백(care gap)을 줄이고 연구를 지원하는 플랫폼을 제공한다.

### Next와 Lens의 역할

두 제품의 개발사는 미국의 Tempus다. Next라는 이름의 플랫폼을 소개한 공식 발표는 2024년 4월 16일에 나왔다. 그 이전의 관련 알고리즘 개발·협력 이력과 제품 공개일은 구분해야 한다. Lens는 2021년 5월 20일 데이터 분석 플랫폼으로 출시되었다. 2026년 5월 31일 발표는 그 플랫폼에 에이전트 기능을 확대한 새 세대에 관한 것이다.

비정형 자료는 정해진 칸에 들어 있는 숫자와 달리 문장으로 남은 정보를 뜻한다. 예를 들어 진료기록에 검사를 외부 병원에 의뢰했다고 적혀 있을 수 있다. AI는 이런 문장을 읽어 검사와 결과의 상태를 정리한다. 결과가 기록에 없다는 이유만으로 검사를 하지 않았다고 단정하면 안 된다. 알림을 받은 의료진이 외부 자료와 환자 상황을 확인해야 하는 이유다.

Tempus Next는 의료기관의 임상 의사결정 지원 도구다. 전자의무기록(EHR)의 비정형 자료를 분석해 진료 지침상 필요한 바이오마커(biomarker) 검사를 받지 않은 환자를 찾고, 의료진에게 알림을 제공한다. 새로운 치료법을 독립적으로 결정하기보다 필요한 검사와 검토가 이루어지도록 지원한다.

Tempus Lens는 제약사와 연구자를 위한 분석 플랫폼이다. 연구 질문에 따라 환자 집단과 자료를 분석하며, 공동과학자(Co-scientist) 에이전트는 분석 계획과 코드 실행을 지원하도록 소개되었다. 비식별 임상 자료와 분자 정보를 신약개발 연구에 연결하는 접근이다.

두 플랫폼은 같은 제품의 이전 이름과 새 이름이 아니다. Next는 진료 과정의 누락을 찾는 데, Lens는 연구 분석에 초점을 둔다. 의료 AI가 진료와 연구에서 서로 다른 업무를 지원하는 사례로 구분할 수 있다.

## 여러 자료를 연결하는 기반

Lens에서는 연구자의 질문이 입력이 된다. 분석 대상과 비교 조건을 정한 뒤 해당 자료를 골라 계산하고 결과를 보여 준다. 예를 들어 특정 변이가 있는 환자와 없는 환자의 기록을 비교할 수 있다. 두 집단의 나이나 이전 치료가 다르면 결과 해석도 달라진다. AI가 코드를 만들고 실행해도 비교 조건이 타당한지 판단하는 연구자의 역할은 남는다.

유전체는 DNA 정보를, 전사체는 유전자가 RNA로 읽힌 상태를 다룬다. 영상은 조직의 모양을 보여 줄 수 있다. 서로 다른 자료를 연결하면 한 종류의 검사만으로 보이지 않는 관계를 살필 수 있다. 그러나 한 환자의 자료가 서로 정확히 연결되어 있는지부터 확인해야 한다. 자료가 많아져도 연결 오류와 누락이 자동으로 사라지지 않는다.

Tempus는 임상 기록, 유전체, 전사체, 영상 등을 포함하는 대규모 자료를 보유하고 있다고 발표했다. 이러한 자료는 특정 변이와 질환 경과를 함께 분석하거나 연구 목적에 맞는 환자 집단을 구성하는 데 사용될 수 있다.

종양 멀티모달 파운데이션 모델(multimodal foundation model) 연구도 이 기반 위에서 진행된다. 다만 보유 자료의 양과 모델의 정확성은 다른 문제다. 자료의 대표성, 누락, 기관별 측정 차이를 확인해야 하며, 회사가 제시한 초기 결과와 독립적으로 검증된 성능도 구분해야 한다.

## 바이오마커 검사와 치료의 연결

바이오마커는 질병이나 치료 반응을 판단할 때 참고하는 생물학적 표지다. EGFR과 ALK는 유전자 관련 검사에서 등장하고, PD-L1은 단백질 발현을 평가하는 표지다. 같은 이름이 기록에 등장했다는 사실만으로 검사가 끝났다고 볼 수 없다. 검사 주문, 검체 채취, 결과 보고, 의료진 확인은 서로 다른 단계이기 때문이다. 진료 공백을 찾는 시스템은 이 상태들을 구별할 수 있어야 한다.

미국임상종양학회(ASCO)에서 소개된 연구는 여러 지역사회 의료기관의 초기 비소세포폐암(eNSCLC) 환자를 대상으로 Next 도입 전후를 비교했다. 도입 후 EGFR, ALK, PD-L1 검사 비율이 높아진 것으로 보고되었다.

이 사례의 의미는 이미 알려진 진료 지침을 실제 진료에 적용하도록 지원했다는 점이다. 환자의 기록을 지속적으로 분석하면 필요한 검사가 누락된 사례를 발견하고 의료진의 검토로 연결할 수 있다.

다만 검사 시행이 늘어난 결과와 생존이나 재발의 개선은 구분해야 한다. 도입 전후 비교에는 같은 기간의 진료 관행 변화도 영향을 줄 수 있다. 현재 결과는 진료 과정의 개선 가능성을 보여 주는 것으로 이해하는 것이 적절하다.

## 치료 후 추적과의 연계

Tempus는 미세잔존질환(MRD, Molecular Residual Disease) 검사 기업 Personalis와 협력해 왔으며, 인수 계약도 발표했다. MRD 검사는 치료 후 남아 있을 수 있는 질환의 흔적을 살피는 접근이다. 조직 기반 분자 분석과 치료 후 추적을 연결하면 환자 경과를 더 연속적으로 분석할 가능성이 있다.

인수 계약의 발표와 거래 완료는 구분해야 한다. 사업 계획만으로 통합된 진료 체계가 이미 완성되었다고 볼 수는 없다. 실제 기능과 임상적 효과는 별도로 확인해야 한다.

## 계산 인프라와 협력 관계

대규모 임상·분자 자료를 분석하려면 고성능 컴퓨팅이 필요하다. Tempus도 관련 인프라를 설명하지만, 특정 기술 공급사와의 협력 범위를 공개 자료 이상으로 확대해 해석해서는 안 된다. 이 절은 확인된 Tempus 플랫폼의 기능을 중심으로 다룬다.

의료 AI를 이해할 때는 데이터 보유, 분석 모델, 시스템 연결, 기업 협력이 각각 어떤 역할을 하는지 구분하는 것이 필요하다. 협력 발표가 특정 임상 기능의 성과를 입증하지는 않는다.

## 의료진의 판단과 적용 범위

Next와 같은 도구는 의료진의 의사결정을 지원한다. 기록에서 검사 후보를 찾아 알려더라도 환자가 실제 검사와 치료를 받을지는 진료 맥락을 고려해 판단해야 한다. 개별 AI 제품의 규제 허가를 플랫폼 전체의 자율 진료 허가로 이해해서도 안 된다.

자료가 특정 국가나 인구 집단에 편중되어 있으면 다른 집단에 적용할 때 성능이 달라질 수 있다. 외부 기관의 자료로 평가하고, 실제 업무에서 알림이 어떻게 처리되는지 확인해야 한다. 알림의 정확성과 의료진의 대응 부담도 운영 평가에 포함된다.

Tempus의 사례는 의료 AI가 새로운 예측을 생성하는 기능 외에, 필요한 정보가 적절한 시점에 의료진에게 전달되도록 지원할 수 있음을 보여 준다. 정밀의료의 발전에는 분자 분석 기술뿐 아니라 검사, 결과 확인, 치료 결정, 추적 관리가 연결되는 업무 체계가 필요하다.



## 제6장. 중국의 의료 AI 도입



## 6.1 하북의대 부속병원의 AI 운영 성과

하북의과대학 제1병원 연구진의 운영 논문은 어떤 업무부터 AI를 도입했는지 설명한다. 임상 정확도를 거론 시험과 병원의 운영 경험을 구분해 읽으면 결과의 범위가 분명해진다.

<https://pmc.ncbi.nlm.nih.gov/articles/PMC13245460/>

의료 AI를 실제 업무에 정착시키려면 기술 성능뿐 아니라 진료 절차와 조직의 운영 방식을 함께 검토해야 한다. 시범사업을 마친 뒤 사용이 중단되거나 일부 부서에만 머무는 경우도 있기 때문이다. 이런 현상을 시범사업의 함정(pilot trap)이라고 한다.

중국 하북의과대학 제1병원은 사전문진, 종양 협진 지원, 약물농도 모니터링에 AI를 단계적으로 도입했다. 관련 연구는 약 18개월의 운영 경험을 바탕으로, 병원에서 AI를 어떻게 확대하고 관리했는지 설명한다. 단일 모델의 정확도보다 업무 흐름(workflow)과 관리 체계(governance)를 살피는 사례다.

### 위험 수준에 따른 단계적 도입

논문은 2026년 6월 4일에 공개되었다. 이는 운영 경험을 발표한 시점이며, 세 시스템의 최초 개발일이나 판매 시작일과 같지 않다. 사전문진과 종양 협진 지원, 약물농도 관리는 업무의 이름이다. 이를 하나의 회사가 만든 동일한 모델의 세 가지 이름으로 이해하면 안 된다. 병원의 이름은 운영 주체를 가리키며, 그것만으로 시스템의 개발 회사나 모델 버전까지 알 수는 없다.

병원은 처음부터 모든 임상 업무를 자동화하지 않았다. 먼저 진료 전 정보 수집을 지원하고, 이후 전문적인 판단을 보조하는 기능을 도입했다.

지능형 사전문진에서는 환자가 진료 전에 증상과 병력을 입력하면 AI가 구조화된 병력 초안을 작성한다. 의사는 초안을 확인하고 추가로 필요한 사항을 질문할 수 있다. 여러 진료과에서 공통적으로 발생하는 정보 수집과 기록 업무를 지원하는 기능이다.

다학제 종양 의사결정 지원은 영상과 병리, 치료 이력을 정리해 협진에서 검토할 선택지를 제시한다. 여러 전문과가 같은 환자를 논의할 때 필요한 자료를 연결하는 방식이다. 최종 치료 결정은 의료진이 담당한다.

둘록세틴(duloxetine) 약물농도 모니터링은 혈중 약물농도를 참고해 용량 조절을 지원한다. 전문적인 해석과 환자별 판단이 필요한 기능이므로 관련 진료 영역에서 제한적으로 사용되었다.

## 도입 범위의 차이

환자가 입력한 답을 정리하는 기능과 약물 용량을 제안하는 기능은 오류의 결과가 다르다. 예를 들어 사전문진에서 증상의 시작 시점이 빠지면 의사가 다시 물을 수 있다. 반면 투약 관련 자료의 단위나 채혈 시간이 틀리면 제안의 의미가 달라질 수 있다. 둘 다 확인이 필요하지만 확인 항목과 담당 전문가는 같지 않다. 단계적 도입은 이런 차이를 실제 업무 분담에 반영하는 과정이다.

사전문진은 많은 진료과로 확대된 반면, 종양 치료와 약물 용량 지원은 관련 전문 영역을 중심으로 적용되었다. 정보 정리와 문서 작성은 여러 과에

서 사용할 수 있지만, 치료 선택을 직접 지원하는 기능은 위험과 책임을 더 세밀하게 관리해야 하기 때문이다.

이러한 방식은 단계적 도입(staged rollout)의 한 형태다. 공통 업무부터 적용하고, 임상 위험이 큰 기능은 사용 범위와 검토 절차를 명확히 하면서 확대한다. 병원에서는 기능별로 필요한 교육과 감독 수준도 달라진다.

### 보고된 운영 변화

성과표에서는 무엇을 분모로 계산했는지도 읽어야 한다. 문서 한 건을 만드는 시간과 환자 한 명의 전체 체류시간은 다르다. AI 사용에 익숙한 직원만 응답한 만족도 조사라면 전체 직원의 경험과 차이가 날 수 있다. 같은 기간 인력을 늘리거나 접수 동선을 바꿨다면 대기시간 변화에 그 영향도 섞인다. 이런 조건을 확인해야 다른 병원에 도입할 때 기대할 수 있는 효과를 가늠할 수 있다.

운영 연구는 대기시간, 문서 작성, 진료 효율과 환자 만족도 등에서 개선을 보고했다. 다만 문서 작성 완료 효율과 진료 전체의 효율은 서로 다른 지표다. 특정 업무의 개선 수치를 병원 전체의 성과로 확대해서는 안 된다.

같은 기간 외래 동선과 데이터 관리, 품질관리 절차도 함께 개선되었다. 따라서 보고된 변화는 AI를 포함한 운영 개선 과정의 결과로 이해해야 한다. AI만의 효과를 따로 입증한 연구는 아니다.

이 사례의 의미는 업무 효율의 정확한 증가율보다, 여러 기능을 일상 진료에 연결하고 일정 기간 운영한 경험에 있다. 다만 병원 전체의 도입과 개별 의료진의 지속적인 사용은 구분해야 하며, 다른 기관에서도 같은 결과가 나타날지는 추가 확인이 필요하다.

## 기술 도입과 조직 역량

NASSS는 특정 AI 제품의 이름이 아니다. 기술을 받아들이지 않거나 사용을 중단하는 이유까지 살피는 분석 틀이다. 병원이 도입을 발표한 뒤 실제 사용이 이어지는지 보는 데 의미가 있다. 예를 들어 자료 입력이 번거로워 의료진이 기능을 건너뛴다면 모델의 시험 점수가 높아도 업무 효과는 작아진다. 현장의 문제를 접수하고 고칠 담당자가 있는지도 운영 능력에 포함된다.

의료 AI의 확산에는 사용자 역할, 지원 체계, 예외 처리 기준이 필요하다. 사전문진 초안을 누가 확인하는지, 협진 자료에 오류가 있으면 어떻게 수정하는지, 약물 관련 제안을 누가 승인하는지 정해야 한다.

NASSS와 같은 기술 도입 분석 틀도 기술 자체 외에 환자와 사용자, 조직, 제도의 영향을 함께 살핀다. 병원이 새로운 기능을 지속해서 운영하려면 이러한 조건이 실제 업무에 맞게 구성되어야 한다.

하북의과대학 제1병원은 AI를 단일 제품으로 일괄 도입하기보다, 업무의 성격과 위험에 따라 적용한 사례로 볼 수 있다. 정보 정리에서 전문적 판단 지원으로 범위를 넓히되, 의료진의 결정 권한과 검토 절차를 유지하는 방식이다.

## 운영 성과와 임상 성과

대기시간이나 문서 처리의 개선은 환자와 의료진의 경험에 중요하다. 그러나 생존, 재입원, 부작용 등 임상 결과의 개선과는 별도로 평가해야 한다. 이 연구도 주로 도입과 운영을 다룬다.

앞으로는 실제 사용의 지속성, 진료과별 차이, 오류 수정 부담, 환자 결과를 함께 살필 필요가 있다. 의료 AI의 정착 여부는 도입 발표나 초기 만족도만으로 판단하기 어렵다. 병원의 업무와 관리 체계 안에서 안정적으로 작동하는지 확인하는 과정이 필요하다.

## 6.2 WiNGPT 3.0과 엔트아푸 서비스

WiNGPT 3.0 연구진의 기술보고서는 모델의 학습 방법과 시험 범위를 공개한다. 엔트그룹의 AQ 출시 자료는 병원용 모델과 개인용 건강 앱이 어떤 점에서 다른지 비교하는 데 도움이 된다.

<https://arxiv.org/abs/2505.17387>

<https://www.antgroup.com/news-media/press-releases/1750924800000>

중국 의료 AI의 적용은 의료기관의 업무 지원과 개인의 건강관리 서비스로 확대되고 있다. 의료진이 병원 정보시스템에서 사용하는 도구와, 주민이 휴대전화로 건강 정보를 확인하는 서비스가 함께 발전하는 양상이다.

이러한 변화의 배경에는 대형병원으로의 환자 집중과 지역 의료기관에 대한 신뢰 문제도 있다. AI를 통해 지역의 의료정보 접근성을 높이고 의료진의 업무를 지원하려는 시도가 이루어지고 있다. 다만 서비스 보급과 의료체계의 불균형 해소는 별도로 평가해야 한다.

### 병원 정보시스템과 WiNGPT

WiNGPT 3.0 기술보고서의 첫 공개일은 2025년 5월 23일이다. 6월 5일에는 수정본이 올라왔다. 이는 3.0 연구보고서의 공개 이력이며 WiNGPT 전체 제품군의 최초 출범일과 다르다. 보고서가 다룬 모델은 약 320억 개의 매개변수를 가진다. 매개변수는 학습 중 조정되는 내부 값이며, 숫자가 크다는 이유만으로 의료 판단이 정확해지지 않는다는.

연구진은 답변 예시를 따라 배우는 지도 미세조정과 보상에 따라 출력을 개선하는 강화학습을 사용했다. 의료 질문을 받으면 학습한 지식과 질문

의 조건을 바탕으로 답변을 생성한다. 병원 기록을 자동으로 읽는 기능은 별도의 시스템 연결이 있어야 가능하다. 언어모델을 설치한 것만으로 모든 환자의 의무기록에 접근할 수 있는 것은 아니다.

WiNGPT는 의료정보 기업 위닝헬스(卫宁健康, Winning Health)의 의료 특화 거대언어모델이다. 회사는 이를 병원 정보시스템 WiNEX와 연결하고, WiNEX Copilot 등의 형태로 의료진의 업무를 지원하도록 구성했다.

병원에 적용되는 AI는 독립적인 채팅창에 머무르기보다 기록과 조회, 정보 정리 등 기존 업무에 연결되어야 한다. 의료진이 사용 중인 시스템에서 필요한 도움을 받을 수 있도록 하는 것이 통합의 목적이다.

위닝헬스는 WiNEX Copilot 등 AI 제품을 약 150개 의료기관에 배포했다고 발표했다. 이 수치는 WiNGPT 3.0 단일 모델의 도입 기관 수와 동일하지 않으며, 실제 사용 빈도나 임상 성과를 뜻하지도 않는다. 배포 범위와 현장의 정착 정도는 구분해서 볼 필요가 있다.

WiNGPT 3.0 기술보고서는 의료 계산과 의학 지식 평가 결과를 제시했다. 이러한 시험은 기본 역량을 비교하는 데 도움이 되지만, 환자의 진단과 치료가 개선되었는지는 실제 사용 환경에서 확인해야 한다.

## 다른 의료 모델과의 구분

원즈성의 U2-Med 공식 공개 자료는 2026년 7월 17일에 나왔다. 의료, 의료보험, 의약 업무를 대상으로 한 특화 모델이다. 이후 회사가 설명한 U2-Med-World는 진료 과정을 모사하는 세계 모델 연구다. 기반이 되는 의료 모델과 그 모델을 훈련·평가하는 가상 환경을 구분해야 한다. HealthBench도 판매용 의료 AI가 아니라 답변을 평가하는 시험 체계다.

U2-Med는 원즈성(云知声, Unisound)의 모델로, 위닝헬스의 WiNGPT와는 다르다. U2-Med는 의료 세계 모델을 이용해 여러 차례 질문하고 정보를 추가하면서 판단을 수정하는 방식을 연구한다. 자세한 구조는 2.2절에서 살펴보았다.

원즈성은 HealthBench에서 높은 평가를 받았다고 발표했다. HealthBench는 의료 대화의 정확성, 위험 식별, 소통 등을 채점 기준(rubric)에 따라 평가한다. 다만 기업이 발표한 벤치마크 결과와 실제 의료기관의 효과는 구분해야 한다.

WiNGPT와 U2-Med는 서로 다른 기업과 설계에 기반하지만, 모두 의료 정보 처리와 판단 지원을 확대하려는 접근이다. 제품의 명칭보다 어떤 자료에 접근하고 어떤 업무를 지원하는지 살피는 것이 중요하다.

## 개인 건강관리 서비스 앤티아푸

AQ의 개발·운영 주체는 앤티그룹이다. 회사가 발표한 앱 출시일은 2025년 6월 26일이다. 이후의 서비스명 변경과 기능 확대는 최초 출시와 나누어 볼 필요가 있다. 앱은 사용자가 입력한 질문이나 올린 자료를 받아 설명을 제공하고 필요한 서비스로 연결한다. 앱이라는 창구, 그 안에서 답변을 만드는 모델, 실제 의사의 상담은 각각 다른 역할이다.

예를 들어 검사표 사진을 올리면 글자와 숫자를 읽는 단계가 먼저 필요하다. 이어 항목의 뜻을 설명할 수 있다. 사진이 흐려 숫자를 잘못 읽었거나 검사 날짜가 빠졌다면 설명의 전제가 틀릴 수 있다. 사용자가 입력한 자료의 품질과 실제 진료기록의 확인 여부를 함께 봐야 한다. 친절하고 긴 답변이 이런 문제를 대신 해결해 주지는 않는다.

앤티그룹은 건강 앱 AQ를 출시한 뒤 앤티아푸(蚂蚁阿福)로 서비스를 확대했다. 사용자는 건강 관련 질문을 하고 검사 결과나 생활습관에 관한 정보를 확인할 수 있다. 스마트워치와 밴드의 정보, 음식 사진, 체중관리 기능 등 일상적인 건강관리와의 연결도 소개되었다.

이런 서비스는 병원을 방문하기 전 정보를 확인하거나 일상에서 건강 상태를 관리하는 데 접근성을 높일 수 있다. 다만 회사는 서비스가 진단을 하거나 의사를 대체하는 것은 아니라고 설명한다. 일반적인 건강 정보와 개인에 대한 의료적 판단을 구분해야 한다.

회사는 사용자와 문의 규모의 증가를 발표해 왔다. 다만 누적 사용자와 월간 활성 사용자(MAU, Monthly Active Users)는 다른 지표이므로 직접

비교하기 어렵다. 이용 규모가 커졌다는 사실만으로 정보의 정확성이나 건강 개선 효과가 입증되지는 않는다.

### **지역 의료기관과의 연결**

앤티아푸는 일부 지방정부와 지역 의료기관의 디지털 가정의 서비스에도 연결되었다. 주민이 검사 결과를 조회하고 관련 설명을 확인하거나, 가정의 서비스와 연계하는 방식이다.

이 접근에서는 AI의 정보 제공과 실제 의사의 상담이 함께 구성된다. AI가 주민의 질문과 자료를 정리하고, 전문적인 판단이 필요한 경우 의료진과 연결하는 구조를 지향한다. 실제 운영에서는 기관별 기능과 담당자의 역할이 다를 수 있다.

지역 의료의 효과를 평가하려면 이용자 수 외에도 필요한 진료로 적절히 연결되는지 살펴야 한다. 대형병원 방문의 변화, 상담 이후의 대응, 만성질환 관리 결과 등이 관련 지표가 될 수 있다. 이용자 수만으로 그런 효과가 확인되었다고 볼 수는 없다.

### **도입 이후의 과제**

의료기관용 시스템과 개인용 서비스는 사용 환경이 다르지만 정확성과 책임이라는 공통 과제가 있다. 병원에서는 AI의 제안이 의료진의 검토와 연결되어야 한다. 개인용 서비스에서는 사용자가 정보의 한계를 이해하고 필요한 진료를 받을 수 있도록 안내해야 한다.

사업의 지속 가능성도 중요하다. 제품 배포가 늘어도 비용을 누가 부담하고 운영을 어떻게 지원할지가 해결되어야 한다. 자료 관리와 정보 갱신, 오류 대응도 지속적인 자원이 필요한 업무다.

중국의 사례는 의료 AI가 병원 정보시스템과 개인 건강관리 플랫폼 양쪽으로 확대되고 있음을 보여 준다. 앞으로는 보급 규모와 함께 실제 사용의 질, 의료진과의 연결, 환자에게 나타나는 효과를 평가해야 한다.

## 6.3 마인드레이 启元과 1차 의료 디지털화

선전시 위생건강위원회의 공개 자료는 치위안 중증의학 모델을 병원에 연결하는 과정을 소개한다. 장비 자료와 의무기록을 모으는 작업, 의료진의 판단을 돕는 기능을 중심으로 읽을 수 있다.

[https://wjw.sz.gov.cn/wzx/content/post\\_12075823.html](https://wjw.sz.gov.cn/wzx/content/post_12075823.html)

중국은 지역 의료기관의 진료 지원을 의료 AI 정책의 주요 과제로 삼고 있다. 중국에서 기층(基層) 의료기관은 향진위생원, 지역사회 보건센터, 마을 보건실 등 1차 의료를 담당하는 기관을 뜻한다. 지역에서 검사와 만성 질환 관리, 기본 진료를 제공하는 역할이다.

정책은 이러한 기관의 정보 처리와 진료 지원 능력을 높이고, 상급 병원과의 협력을 넓히는 방향으로 추진되고 있다. 의료기기 기업도 장비와 정보 시스템에 AI를 연결하는 제품을 개발하고 있다. 마인드레이(迈瑞医疗, Mindray)의 启元(치위안, Qiyuan)이 관련 사례다.

### 1차 의료 AI 정책의 방향

중국의 "인공지능+의료위생" 관련 실시의견은 의료 데이터 기반과 전문 모델, 에이전트 응용을 확대하는 단계적 목표를 제시했다. 2027년까지 데이터와 응용 기반을 마련하고, 2030년에는 기층 진료의 지능형 보조 응용을 널리 보급하는 방향이다.

적용 분야에는 보조 진료, 처방 검토, 추적 관리, 만성질환 관리와 가정의 서비스 등이 포함된다. 지역 단위로 영상·심전도·검사·병리 정보를 공유

하고 상급 기관과 연결하는 체계도 중요하게 다룬다. 다만 이러한 목표를 현재 달성한 보급률이나 임상 성과로 이해해서는 안 된다.

지방정부는 중앙의 방향을 구체적인 사업으로 추진한다. 베이징의 행동계획은 공통 추론 서비스와 연산 인프라, 다중 에이전트 플랫폼 등을 다루며, 지역별 만성질환 관리와 기록 품질관리 시범사업을 포함한다. 동시에 AI가 의료인의 전문적 판단을 완전히 대체해서는 안 된다는 원칙을 제시한다.

### 장비와 정보시스템의 연결

치위안 중증의학 모델은 마인드레이가 텐센트 혼원을 기반으로 구성한 사례다. 선전시의 2025년 3월 18일 자료는 2024년 말부터 일부 지역에 배포했다고 설명한다. 이 자료로 확인되는 것은 당시의 배포 이력이다. 모든 치위안 제품이 같은 날 출시되었다거나 지역 의료기관 전체에서 사용했다는 뜻은 아니다.

중환자실에서는 같은 환자의 정보도 여러 기기에 나뉘어 있다. 모니터가 보내는 수치와 검사실 결과, 의사가 쓴 기록을 같은 시간 흐름에 놓아야 한다. 과거 수치를 현재 수치로 읽으면 환자 상태를 오해할 수 있다. AI가 정보를 모아 요약한 뒤에는 의료진이 자료의 출처와 시간을 확인할 수 있어야 한다.

마인드레이는 환자감시장치, 초음파, 체외진단 장비 등을 공급하는 기업이다. 启元은 이러한 장비와 임상 정보를 연결해 의료진의 판단과 업무를 지원하는 AI 제품군으로 소개된다. 중증의학 모델은 텐센트의 혼원(混元) 모델과 마인드레이의 임상 지원 체계를 결합한 협력 사례다.

초음파 분야에서는 장비와 정보관리 시스템을 연결해 검사 과정 전반을 지원한다. 심전도 분야에서는 수집, 분석, 검토, 전원, 품질관리를 연계한다. HIS, EMR, LIS, PACS 같은 병원 시스템과의 연결은 검사 결과를 실제 업무에서 사용할 수 있게 하는 조건이다.

지역 의료공동체에서는 기층 기관이 검사를 시행하고 상급 병원이 판독과 협진을 지원할 수 있다. 유방 초음파의 BI-RADS 분류 지원이나 복부 초음파의 표준 단면 확인 등도 소개되었다. 이러한 기능은 지역 의료진의 업무를 보완하는 방향으로 이해할 수 있다.

### 기능 소개와 임상 평가

회사는 기록 작성과 검사 보고, 조기 경보 등에서 효율 개선을 발표해 왔다. 그러나 기업이 제시한 처리 속도와 실제 환자의 진단·치료 성과는 별도로 평가해야 한다.

응급 중환자실 기록을 이용한 비교 연구에서는 启元과 범용 모델의 진단 성능을 평가했다. 특화 모델이 일부 비교에서 높은 결과를 보였지만, 범용 모델보다 일관되게 우수하다고 보기는 어려웠다. 제공한 정보의 범위와 질환에 따라 결과도 달라졌다.

이 연구는 과거 기록을 이용한 평가이며 기층 기관의 실제 진료 성과를 직접 보여 주지는 않는다. 장비에 기능을 추가한 것, 의료진이 지속적으로 사용하는 것, 환자의 결과가 개선되는 것은 구분해서 확인해야 한다.

## 개별 기능에서 업무 과정으로

HIS는 접수와 원무 등을 포함한 병원 정보시스템이다. EMR은 전자의무 기록, LIS는 검사실 정보시스템, PACS는 의료영상 저장·전송 시스템을 뜻한다. 이름이 다른 것은 맡는 자료와 업무가 다르기 때문이다. 서로 연결하려면 환자를 식별하는 번호와 검사명, 단위, 기록 시점이 맞아야 한다.

초음파의 표준 단면은 비교와 판독에 필요한 정해진 방향의 영상이다. BI-RADS는 유방 영상 소견을 일관된 범주로 정리하는 체계다. AI가 영상의 적절성이나 분류를 지원하더라도 충분한 영상을 얻었는지, 다른 검사와 맞는지 의료진이 살핀다. 장비 안의 보조 기능과 최종 진료 판단을 구별하면 각 기능을 평가할 기준도 명확해진다.

의료 AI의 적용은 한 가지 기능을 별도 프로그램으로 제공하는 방식에서 기존 업무 과정에 연결하는 방식으로 확대되고 있다. 검사 결과 해석, 진료 대화의 기록, 다른 기관의 검사 자료 정리, 이상 수치 알림 등을 의료진이 사용하는 시스템에서 처리하려는 접근이다.

의료진이 매번 별도 도구를 실행해야 한다면 사용이 제한될 수 있다. 기존 화면과 기록 과정에 지원 기능을 연결하면 접근성은 높아진다. 반면 시스템 간 자료의 의미와 형식이 일치해야 하고, 잘못된 정보가 후속 업무에 반영되지 않도록 관리해야 한다.

마인드레이의 장비·정보시스템 통합과 다른 기업의 병원 시스템 연계는 구체적인 제품은 다르지만 이러한 공통 방향을 보여 준다. 다만 한 기업의 설명을 다른 기업의 실제 기능으로 확대해서는 안 된다.

## 지속적인 운영의 조건

지역 의료에서는 여러 질환을 가진 환자와 약물 상호작용을 다루는 능력이 중요하다. 장기간의 건강 자료가 충분하지 않거나 기관별로 분산되어 있으면 AI가 환자 상태를 해석하는 데 제약이 생긴다.

책임과 비용 부담도 정해야 한다. AI 제안을 누가 검토하고 오류에 어떻게 대응할지, 시스템 유지와 교육 비용을 누가 부담할지가 분명해야 한다. 시범사업의 지원이 끝난 뒤에도 운영할 수 있는 구조가 필요하다.

중국의 1차 의료 디지털화는 정책, 지방의 인프라, 의료기기, 개인 건강관리 서비스를 함께 추진하는 흐름이다. 앞으로의 평가는 보급 목표뿐 아니라 지역 의료진이 실제로 사용할 수 있는지, 필요한 진료와 협진이 적절히 이루어지는지에 초점을 두어야 한다. AI 기능을 지속 가능한 의료서비스로 연결하는 것이 핵심 과제다.



## Part 4. 의료 AI의 규제·안전·경제성



## 제7장. AI 의료기기 규제와 PCCP(사전변경허가계획)



## 7.1 PCCP 지침과 AI 모델 변경 관리

FDA의 PCCP 최종 지침은 허용받으려는 변경과 검증 계획을 어떻게 제출하는지 설명한다. 앞부분의 적용 범위와 변경 계획의 세 구성 요소를 함께 읽으면 제도의 조건을 이해할 수 있다.

<https://www.fda.gov/media/166704/download>

AI 의료기기는 도입 후에도 개선이 필요하다. 새로운 환자 자료를 반영하거나 다른 장비에서 얻은 정보를 처리하도록 모델을 조정할 수 있다. 그러나 변경이 진단과 치료에 영향을 줄 수 있으므로 성능 개선과 안전성 확인을 함께 관리해야 한다.

사전변경허가계획(PCCP, Predetermined Change Control Plan)은 앞으로 수행할 변경의 범위와 검증 방법을 미리 심사받는 제도다. FDA가 인정한 계획에 맞는 변경은 정해진 조건에서 별도의 시판 서류 제출 없이 적용할 수 있다. 재학습한 모델을 자동으로 승인하거나 모든 변경을 면제하는 제도는 아니다.

### 의료기기의 시판 절차와 변경

PCCP는 AI 모델 이름이 아니다. 제조사가 제품의 변경을 관리하기 위해 제출하는 계획이다. FDA의 관련 최종 지침은 2024년 12월 4일 처음 발행되었고, 현행 문서의 발행일은 2025년 8월 18일이다. 법률에 있는 제도, 그 제도를 설명하는 지침, 개별 제품에 대해 인정된 계획을 구별해야 한다. 지침을 읽고 계획을 작성했다는 사실만으로 제품 변경이 허용되는 것은 아니다.

미국 의료기기의 시판 절차에는 기존 기기와의 실질적 동등성을 검토하는 510(k), 새로운 중저위험 기기의 분류를 위한 De Novo, 고위험 기기의 시판 전 승인(PMA)이 있다. 절차에 따라 FDA의 허가(clearance), 인가(authorization), 승인(approval)이라는 용어가 사용된다.

기기가 변경되면 그 변경이 안전성과 성능, 사용 목적에 어떤 영향을 주는 지 판단해야 한다. 중요한 변경에는 새로운 제출이나 보충 신청이 필요할 수 있다. AI 모델을 자주 개선하려는 요구와 변경된 제품의 신뢰성을 확인하려는 요구를 함께 다루는 것이 PCCP의 목적이다.

## 법적 근거와 적용 범위

PCCP의 법적 근거는 연방 식품·의약품·화장품법(FD&C Act) 제515C조다. FDA가 인정한 변경 계획과 일치하고 관련 조건을 충족하는 경우, 개별 변경마다 다시 제출하는 부담을 줄일 수 있도록 한다.

제조사는 변경하려는 사항과 안전성을 확인할 방법을 제시한다. 규제 기관은 제품의 심사 과정에서 그 계획을 함께 검토한다. 계획을 작성했다는 사실만으로 면제가 생기는 것은 아니며, 실제 변경이 인정된 범위와 조건에 맞아야 한다.

FDA의 AI 기기 소프트웨어 기능 관련 지침은 510(k), De Novo, PMA 등 시판 경로에서 PCCP를 어떻게 제시하고 평가할지 설명한다. 개별 제품의 사용 목적과 위험, 변경 내용에 따라 필요한 근거는 달라진다.

## 변경 계획의 세 구성 요소

예를 들어 흉부 영상을 분석하는 가상의 제품을 생각해 보자. 제조사가 다른 촬영 장비의 영상을 추가로 학습시키려 한다면, 무엇을 추가하고 어떤 환자에게 쓸지 정해야 한다. 기존 장비의 영상에서도 성능이 유지되는지 비교할 수 있어야 한다. 이 예시는 변경 관리의 뜻을 설명하기 위한 것이며, 특정 제품의 허가 사례가 아니다.

시험 결과를 보기 전에 통과 기준을 정하는 이유도 있다. 결과가 나온 뒤 기준을 낮추면 실제로 개선되었는지 판단하기 어렵다. 평균 점수가 올라도 특정 연령대에서 병변을 더 자주 놓칠 수 있다. 변경 전후의 전체 성과 주요 집단의 결과를 함께 살피는 편이 이런 문제를 찾는 데 도움이 된다.

변경 내용 설명(Description of Modifications)은 무엇을 바꿀 것인지 구체화한다. 성능을 높이겠다는 일반적인 목표만으로는 부족하다. 적용 자료와 기능, 예상되는 성능 변화의 범위를 명확히 해야 한다.

변경 프로토콜(Modification Protocol)은 변경을 개발하고 검증해 적용하는 절차다. 자료 수집과 분리, 재학습, 성능 평가, 수용 기준, 사용자 안내 등을 포함한다. 같은 변경이라도 실제 검증 과정이 적절해야 계획에 따른 적용이 가능하다.

영향 평가(Impact Assessment)는 변경의 이익과 위험을 검토한다. 개별 기능의 성능뿐 아니라 여러 변경이 함께 적용되었을 때의 영향과 전체 시스템의 안전성도 고려한다.

이 세 요소는 변경 범위, 수행 방법, 결과의 영향을 연결한다. 제품을 한 번 평가하는 데서 나아가 이후 개선을 어떻게 관리할지 미리 정하는 방식이다.

## 재학습과 지속적인 성능 관리

학습 자료는 모델의 내부 값을 조정하는 데 쓴다. 검증 자료는 그 조정이 보이지 못한 사례에도 통하는지 확인하는 데 쓴다. 같은 환자의 비슷한 영상 이 양쪽에 섞이면 시험이 쉬워질 수 있다. 실제로 처음 만난 환자의 자료를 잘 처리할지는 별도로 확인해야 한다.

재학습과 배포도 다른 단계다. 모델을 새로 만들었다고 곧바로 진료 화면에 적용되는 것은 아니다. 시험을 통과했는지 확인하고 어떤 버전을 적용했는지 남기는 과정이 필요하다. 문제가 생겼을 때 이전 상태로 돌아갈 수 있는지도 점검할 수 있다. 이러한 실무 조치는 제품의 위험과 인정된 변경 계획에 맞추어 구체화한다.

PCCP는 사람이 통제하는 재학습과 자동화된 변경을 모두 검토할 수 있는 틀이다. 다만 어떤 방식이든 사전에 인정받은 범위와 검증 조건을 따라야 한다. 계획 밖의 변경은 별도의 규제 검토가 필요할 수 있다.

모델 개선에서는 기존 성능의 유지도 중요하다. 새 자료를 학습하면서 이전에 처리하던 유형의 성능이 낮아지는 파국적 망각(catastrophic forgetting)이 발생할 수 있다. 회귀 시험(regression test)은 이러한 변화를 확인하는 방법이지만, 모든 상황을 포괄하지는 못한다.

따라서 학습 자료와 검증 자료를 구분하고, 다양한 환자와 장비 조건에서 평가해야 한다. 변경 후에도 실제 사용 성능을 관찰하고 문제가 생겼을 때 대응할 절차를 마련해야 한다.

### 변경 범위를 명확히 하는 이유

새 장비의 자료를 추가하거나 특정 환자 집단의 성능을 개선하는 일도 사용 목적과 성능에 미치는 영향에 따라 검토가 필요하다. 질환이나 대상 환자, 의료진의 역할을 바꾸는 변경은 기존 계획에 포함되는지 더욱 신중하게 판단해야 한다.

PCCP를 통해 허용된 범위를 벗어난 변경을 임의로 적용해서는 안 된다. 실제 배포 버전과 검증 결과를 문서화하고, 사용자에게 변경 사항을 알리는 관리가 필요하다. 변경의 편의성과 규제상 책임은 함께 다루어야 한다.

### 국제 원칙과의 관계

여러 기관이 공동 원칙을 냈다고 해서 한 나라의 허가가 다른 나라에서 자동으로 인정되지는 않는다. 각 지역의 법률과 심사 절차를 따로 확인해야 한다. 공동 원칙은 변경을 구체적으로 한정하고 근거를 남기며 사용자가 변화를 알 수 있게 하자는 방향을 제시한다. 실제 제품에 적용할 때는 해당 지역에서 인정받은 계획이 기준이 된다.

FDA와 캐나다 보건부, 영국 MHRA는 우수 기계학습 관행(GMLP)과 PCCP 관련 원칙을 함께 논의해 왔다. 변경의 범위가 명확하고, 위험과 근거에 기반하며, 제품의 생애주기와 투명성을 고려해야 한다는 방향이다.

이 접근은 의료 AI의 개선을 일률적으로 제한하기보다, 예측 가능한 절차 안에서 관리하려는 것이다. 제조사는 필요한 변경과 평가 계획을 구체화하고, 의료기관은 제품의 현재 기능과 변경 이력을 이해할 수 있어야 한다.

PCCP의 핵심은 변경을 미리 설계하고 검증 조건을 합의하는 데 있다. 기술을 개선하는 속도와 환자 안전을 함께 고려하려면, 계획의 심사뿐 아니라 실제 변경의 수행과 지속적인 성능 관리가 중요하다.

## 7.2 FDA 역량 평가(Competency Assessment)와 위험 평가

FDA의 생성형 AI 토론 문서는 첫머리부터 확정 지침이 아니라고 밝힌다. 위험을 나누는 방식과 실제 업무 역량을 시험하려는 제안을 읽을 수 있지만, 이를 시행 중인 허가 요건으로 받아들이면 안 된다.

<https://www.fda.gov/media/194242/download>

생성형 AI는 다양한 질문에 문장으로 답하고, 대화 내용에 따라 후속 행동을 제안할 수 있다. 따라서 고정된 분류 결과를 평가하는 방식만으로는 전체 기능을 설명하기 어렵다. 답변의 정확성뿐 아니라 정보 수집, 위험 식별, 소통, 사용 범위 준수 등을 함께 살펴야 한다.

FDA는 생성형 AI 의료기기를 다루는 토론 문서에서 위험 평가와 역량 기반 접근을 논의했다. 이 문서는 의견 수렴을 위한 자료이며, 확정된 가이드라인이나 의무적인 허가 기준은 아니다. 이 절의 내용도 시행 중인 제도와 앞으로 검토할 평가 방향을 구분해 이해해야 한다.

### 기능의 역할과 오류의 결과

예를 들어 복약 안내문의 어려운 용어를 설명하는 기능과 처방 내용을 바꾸는 기능을 비교할 수 있다. 둘 다 문장을 출력하지만 사용자의 다음 행동에 미치는 영향이 다르다. 같은 오류라도 의료진이 원자료와 대조할 기회가 있는지에 따라 위험이 달라진다. 화면에 보이는 답변의 길이나 말투만으로 위험 수준을 정할 수 없는 이유다.

토론 문서는 기기의 활동(device activity)과 잘못된 출력의 결과(consequences)를 함께 고려하는 접근을 제시한다. 정보를 제공하는 기능, 사용자의 행동을 권고하는 기능, 직접 행동하는 기능은 의사결정에 미치는 영향이 다르다.

같은 권고 기능이라도 오류가 초래할 피해는 다를 수 있다. 일반적인 건강 정보와 구체적인 인슐린 용량 조절은 잘못 사용했을 때의 위험이 같지 않다. 의료진의 감독 수준과 오류를 발견하거나 수정할 가능성도 고려해야 한다.

이 틀은 자율성의 정도만으로 위험을 정하지 않는다. 사용자가 환자인지 의료진인지, 출력의 정확성을 직접 확인할 수 있는지, 대화가 진행되면서 정보 제공이 구체적 지시로 바뀌는지 등을 검토한다. 안내 문구를 덧붙였다는 이유만으로 실제 기능의 영향이 사라지는 것도 아니다.

다만 토론 문서는 각 조합에 의료기기 등급이나 시판 경로를 일률적으로 배정하지 않았다. 미국과 유럽의 등급 체계를 혼합하거나, 논의 중인 틀이 이미 허가 기준으로 적용된다고 설명해서는 안 된다.

## 역량 기반 평가의 목적

역량은 맡은 일을 수행하는 데 필요한 능력을 뜻한다. 의료 에이전트라면 부족한 정보를 알아차리는 능력과 확인한 자료만으로 답하는 능력이 함께 필요하다. 예를 들어 환자의 알레르기 정보가 비어 있다면, 없다고 해석할지 추가 확인을 요청할지가 평가 대상이 된다. 정답이 있는 지식 문제를 많이 맞히는 능력과 이런 업무 판단은 같지 않다.

기반 모델은 여러 제품이 공통으로 쓰는 AI의 바탕이다. 여기에 병원 자료 검색, 도구 실행, 사용자 화면이 붙으면 최종 제품이 된다. 같은 기반 모델을 쓰더라도 한 제품은 오래된 자료를 검색하고 다른 제품은 최신 자료를 검색할 수 있다. 그래서 모델 이름만 같다는 이유로 평가 결과를 서로 대신 적용하기 어렵다.

역량 기반 접근(competency-based approach)은 실제로 수행해야 하는 업무에 필요한 능력을 평가하려는 방식이다. 생성형 AI가 만들 수 있는 모든 답변을 개별적으로 시험하기 어려우므로, 필요한 기능을 체계적으로 구분하고 대표적인 상황에서 검토한다.

평가 대상은 기반 모델(foundation model) 하나만이 아니라 실제로 배포되는 기기 전체다. 같은 모델도 검색 자료, 사용자 화면, 도구 연결, 감독 절차에 따라 다르게 작동할 수 있기 때문이다.

안전 역량에는 위험한 상황의 인식, 사용 범위 준수, 불확실성 전달, 필요한 경우 사람에게 판단을 넘기는 능력이 포함된다. 임상 역량에는 지식과 분석, 정보 수집, 측정 결과의 해석, 소통의 질 등이 포함된다.

일반화 능력도 중요하다. 다른 기관과 환자 집단, 입력 조건에서 성능이 유지되는지 평가해야 한다. 여러 단계를 실행하는 에이전트에는 도구 사용과 권한 준수 같은 추가 역량이 필요하다. 각 항목의 중요성은 제품의 사용 목적과 위험 수준에 따라 달라진다.

## 벤치마킹과 임상 확인

그림자 배치는 실제 업무의 입력을 받되, AI 결과가 진료 판단에 영향을 주지 않도록 두고 성능을 살피는 방식이다. 예를 들어 기존 진료와 별도로 생성된 결과를 나중에 검토할 수 있다. 이 방식은 현장 자료를 시험할 기회를 주지만 의료진이 AI를 보고 행동할 때의 변화까지 보여 주지는 못한다. 사람과 AI가 함께 일하는 성능은 별도의 평가가 필요하다.

벤치마킹은 정해진 자료와 기준으로 기능을 비교하는 평가다. 평가 범위와 채점 방법, 전문가 판정, 수용 기준을 미리 정하면 일관된 검토에 도움이 된다. 합성 자료와 가상 환자도 특정 기능을 시험하는 데 사용할 수 있다.

그러나 시험 환경에서의 성능만으로 실제 진료에서의 작동을 모두 예측할 수는 없다. 임상 확인(clinical confirmation)은 실제 진료 또는 이를 대표하는 조건에서 기능과 안전성을 살피려는 단계다.

방법은 과거 환자 자료의 후향 평가, 실제 업무에서 결과를 노출하지 않는 그림자 배치(shadow deployment), 표준화 환자와의 대화, 독립 의료진의 판정, 전향적 연구 등 다양하다. 모든 제품에 동일한 임상시험을 요구한다는 뜻은 아니며, 위험과 사용 방식에 맞춰 근거를 마련해야 한다.

AI 단독 성능과 사람·AI의 공동 성능도 구분해야 한다. AI가 정확한 정보를 제공해도 사용자가 잘못 이해하거나 중요한 경고를 놓칠 수 있다. 의료진의 어떤 업무를 지원하는지에 따라 비교 대상과 평가 지표를 정해야 한다.

## 독립 평가와 시판 후 관리

드리프트는 AI가 고장 났다는 뜻으로만 쓰이지 않는다. 방문 환자의 구성이 달라지거나 검사 장비와 기록 양식이 바뀌어도 성능이 변할 수 있다. 예를 들어 약물 이름을 적는 방식이 바뀌면 기존 방식에 익숙한 모델이 필요한 정보를 놓칠 수 있다. 변경된 환경을 기록해 두어야 모델 문제와 자료 문제를 나누어 살필 수 있다.

제조사와 분리된 평가 자료와 판정자는 결과의 신뢰성을 높이는 데 기여할 수 있다. 평가 대상에 맞는 자료를 관리하고, 실제 사용 사례를 검토하며, 시험의 반복 가능성을 확보하는 방식이다. 다만 독립 인증의 구체적인 운영 방식은 추가 논의가 필요한 부분이다.

시판 후에는 기반 모델과 검색 자료, 프롬프트, 사용 환경이 바뀔 수 있다. 변경 후에도 기능이 유지되는지 재평가하고, 입출력 표본을 임상의가 검토하는 절차가 필요하다. 환자 집단이나 자료 환경 변화에 따른 성능 저하인 드리프트(drift)도 관찰해야 한다.

7.1절의 PCCP는 변경을 사전에 계획하는 틀이지만, 생성형 AI의 다양한 변화에 어떻게 적용할지는 별도의 검토가 필요하다. 변경 범위와 성능 관찰, 문제 발생 시 대응을 연결하는 관리가 중요하다.

## 추가로 고려할 위험

프롬프트는 AI에 주는 요청이나 지시다. 프롬프트 주입에서는 읽어야 할 자료 안에 행동을 바꾸라는 문장이 들어온다. 예를 들어 외부 문서에 기록을 다른 곳으로 보내라는 지시가 섞여 있을 수 있다. 문서의 내용은 검토할 자료이며 시스템의 실행 권한을 바꾸는 명령이 아니다. 읽기 권한과 수정·전송 권한을 나누는 이유가 여기에 있다.

오류의 심각성 외에 발생 가능성, 되돌릴 수 있는지, 다른 안전장치가 있는지, 결과를 추적할 수 있는지도 중요하다. 병원 단체 등이 토론 문서에 제시한 의견도 이러한 요소를 강조했다.

프롬프트 주입(prompt injection)은 사용자 입력이나 검색 자료, 도구 결과에 숨긴 지시로 AI의 행동을 바꾸려는 공격이다. 외부 정보를 처리하고 행동하는 의료 에이전트에서는 승인되지 않은 지시를 따르지 않는지 확인해야 한다.

생성형 AI의 규제 평가는 특정 시험의 높은 점수에서 실제 업무의 안전성으로 범위를 넓히고 있다. 위험에 맞는 역량을 정하고, 최종 제품과 사람의 상호작용을 평가하며, 도입 이후에도 성능을 관리하는 방향이다. 다만 토론 단계의 제안과 확정된 규제 요구는 계속 구분해야 한다.

## 7.3 영국 AI Airlock과 유럽 AI Act

MHRA의 AI Airlock 2단계 보고서는 참여 사례와 규제상 쟁점을 소개한다. 유럽연합 집행위원회의 AI Act 안내에서는 고위험 분류와 조항별 적용 시점을 직접 확인할 수 있다.

<https://www.gov.uk/government/publications/ai-airlock-sandbox-phase-2-programme-report>

<https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-6>

<https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-113>

의료 AI의 규제는 새로운 기술을 시험하면서 제도를 보완하는 접근과 법적 의무를 정해 준수를 확인하는 접근으로 발전하고 있다. 영국의 AI Airlock은 개발사와 규제 기관이 실제 기술의 과제를 검토하는 프로그램이고, 유럽연합의 AI Act는 AI 시스템에 적용되는 법적 요구를 정한 규정이다.

두 접근은 성격이 다르지만 위험에 비례한 평가와 제품 생애주기 관리, 인간의 감독을 중요하게 다룬다. 의료 AI를 개발하거나 도입하는 기관은 기술의 기능뿐 아니라 적용 지역의 규제 구조를 이해해야 한다.

### 영국의 규제 샌드박스

AI Airlock도 제품이나 진단 모델의 이름이 아니다. MHRA가 운영하는 규제 검토 프로그램이다. 2단계는 2025년 4월부터 2026년 5월까지 진행되었다. 공식 보고서는 2026년 6월 9일 공개되었고 7월 27일 갱신되었다. 보고서 자체가 정식 MHRA 지침이 아니라는 설명도 붙어 있다.

AI Airlock은 영국 의약품·의료제품규제청(MHRA)의 의료 AI 규제 샌드박스다. 규제 샌드박스는 정해진 범위에서 새로운 기술의 규제 과제를 검토하는 프로그램이다. 개발사와 규제 기관이 사용 목적과 근거 생성, 변경 관리 등에 관한 문제를 함께 살핀다.

프로그램은 거대언어모델과 음성 도구, 진단 보조 기술 등을 대상으로 논의를 진행했다. 기존 규정에 새로운 기능을 어떻게 적용할지, PCCP를 통한 변경 관리를 어떻게 구성할지, 사용 목적을 확대할 때 무엇을 평가해야 할지 등이 주요 주제다.

참여 자체가 제품의 시판 허가나 임상 사용 승인을 의미하지는 않는다. 프로그램의 사례 보고서도 공식 가이드라인과 구분해야 한다. Airlock의 목적은 개별 제품을 즉시 허가하는 것이 아니라, 실제 개발 과정에서 드러나는 문제를 규제 지식으로 축적하는 데 있다.

## 도입 이후의 관리

AI Airlock의 후속 단계는 시판 후 감시와 생애주기 관리에도 관심을 넓히고 있다. 의료 AI는 도입 뒤에도 자료와 모델, 사용 환경이 바뀔 수 있기 때문이다.

개발 과정의 평가만으로 실제 현장의 모든 변화를 예측하기는 어렵다. 사용 중 나타나는 오류를 수집하고, 변경 사항을 검증하며, 필요한 정보를 의료진과 공유하는 체계가 중요하다. 이러한 논의는 미국의 변경 관리 및 역량 평가와도 연결된다.

## 유럽 AI Act와 의료기기

유럽 AI Act는 규정(EU) 2024/1689로 제정된 포괄적인 AI 법률이다. 금지되는 관행, 고위험 시스템, 투명성 의무, 범용 AI(GPAI) 모델 등을 구분해 요구 사항을 정한다.

의료기기 AI에서는 제품 법령과의 관계가 중요하다. 의료기기 규정(MDR)이나 체외진단 의료기기 규정(IVDR)의 적용을 받는 제품 또는 안전 부품이고, 제3자 적합성 평가가 필요한 경우에는 AI Act의 제품 관련 고위험 요건을 검토해야 한다.

따라서 모든 건강 관련 소프트웨어가 같은 방식으로 고위험에 해당하는 것은 아니다. 의료기기 해당 여부와 사용 목적, 제품 등급, 평가 절차를 함께 판단해야 한다. AI Act가 의료기기의 기존 등급을 자동으로 바꾸는 것도 아니다.

## 고위험 AI의 주요 요구

적합성 평가는 제품이 적용되는 요구사항을 충족하는지 확인하는 절차다. 제3자 평가는 제조사 밖의 평가기관이 참여하는 경우를 뜻한다. 의료기기라는 이름과 고위험 AI라는 분류는 서로 다른 법률의 질문에 답한다. 어떤 의료 목적을 갖는지 확인한 뒤 AI 관련 요건까지 검토해야 한다. 두 평가를 하나로 섞으면 필요한 서류와 책임 주체를 놓칠 수 있다.

고위험 시스템에는 위험 관리, 데이터 품질과 편향 점검, 기술 문서, 작동 기록, 사용자 정보 제공 등이 요구된다. 의료진이 출력을 이해하고 필요한 경우 수정하거나 사용을 중단할 수 있도록 인간의 감독을 고려해야 한다.

정확성, 견고성, 사이버 보안과 품질관리 체계도 중요하다. 이러한 요구는 모델의 점수뿐 아니라 실제 제품이 어떻게 개발되고 운영되는지 평가한다. 관련 제품 법령에 따른 적합성 평가와 AI 요건의 검토도 연결된다.

제조사와 도입 기관의 역할은 다를 수 있다. 개발사는 설계와 문서화, 성능 및 위험 관리에 책임을 지고, 의료기관은 정해진 사용 조건과 감독 절차에 따라 운영해야 한다. 구체적인 기록 보관과 보고 의무는 해당 조문과 적용 조건을 확인해야 한다.

### 적용 일정과 제도 변화

집행위원회의 2026년 7월 27일 통합본 안내에 따르면, 제6조 제1항과 부속서 I의 제품 관련 고위험 AI에는 제3장 제1~3절의 관련 요건이 2028년 8월 2일부터 적용된다. 부속서 III에 따른 고위험 AI의 해당 적용일은 2027년 12월 2일이다. 다만 조문별 예외가 있으므로 이 날짜를 모든 AI 의무의 시작일로 받아들이면 안 된다. 의료기기 관련 기존 요구와 이미 적용되는 다른 의무도 따로 확인해야 한다.

AI Act의 의무는 분야와 조항별로 적용 시점이 다르다. 시행 일정과 관련 제도는 개정 논의의 영향을 받을 수 있으므로, 실제 사업에서는 최신 통합 법령과 확정된 경과 규정을 확인해야 한다.

AI Act 관련 의무의 적용 시점과 별개로, 의료기기에는 MDR과 IVDR에 따른 기존 요구가 적용된다. AI 규제 조정이 의료기기 전반의 안전성과 성능 평가를 면제한다는 의미는 아니다.

위반에 대한 제재도 법률에서 정하지만, 상한과 실제 부과 금액은 구분해야 한다. 기업은 예상 과징금만이 아니라 제품 분류와 적용 의무, 자료 관리, 감독 체계를 중심으로 준비해야 한다.

## 두 접근의 비교

AI Airlock은 제품을 사례로 규제상의 불확실성을 검토하는 참여형 프로그램이다. 새로운 기능의 평가 방법을 논의하는 데 유연성이 있지만, 결과 자체가 법적 허가를 대신하지는 않는다.

AI Act는 법에 정한 의무를 준수하도록 하는 체계다. 개발과 도입 과정의 책임을 명확히 하는 데 목적이 있으며, 실제 적용에는 표준과 평가 기관, 제품 법령과의 조정이 필요하다.

의료 AI의 발전에는 새로운 기능을 검토하는 과정과 정해진 안전 요건을 확인하는 과정이 모두 필요하다. 규제 기관과의 조기 협의, 사용 목적의 명확화, 지속적인 성능 관리가 두 접근의 공통 과제다.



## 제8장. 환자 안전과 의료 AI의 경제성



## 8.1 통역 오류 전파(Interpreter Error Propagation)와 음성 인식 편향

UCLA와 AltaMed 등이 참여한 통역 오류 연구는 모의 진료의 발언과 AI 기록을 나란히 제시한다. 환자가 말한 내용이 통역과 요약물 거치며 어떻게 달라지는지 사례표에서 확인할 수 있다.

<https://medinform.jmir.org/2026/1/e88734>

AI 스크라이브는 진료 대화를 기록 초안으로 변환한다. 통역이 포함된 진료에서는 환자와 통역사, 의료진의 발언을 구분하고 각각의 의미를 정확히 연결해야 한다. 통역 과정의 오류가 기록에 반영되면 의료진이 환자 상태를 잘못 이해할 수 있다.

통역 오류에는 내용의 생략, 다른 표현으로의 대체, 원래 없던 내용의 추가 등이 있다. AI가 통역된 내용을 사실로 받아들이면 이러한 오류가 공식 기록의 초안으로 전달될 수 있다. 이를 통역 오류 전파(Interpreter Error Propagation)라고 한다.

### 다국어 진료에서 나타나는 문제

연구의 사전공개본은 2025년 12월 1일, 학술지 논문은 2026년 7월 28일에 나왔다. 사용한 두 상용 제품의 이름은 공개하지 않았다. 따라서 이 실험만으로 특정 회사의 제품 성능이나 출시 시점을 추정할 수 없다. 논문을 쓴 연구기관과 시험에 사용한 제품을 개발한 회사도 구별해야 한다.

UCLA 연구진은 스페인어와 영어가 사용되는 모의 진료 대화에 통역 오류를 넣고, 상용 AI 스크라이브가 이를 어떻게 기록하는지 평가했다. 두 제품

모두 일부 통역 오류를 최종 기록에 반영했다. 환자의 발언이 잘못 전달된 경우에 오류가 많이 남았다.

이 결과는 AI가 대화를 기록하면서 통역의 정확성까지 자동으로 보장하지는 않는다는 점을 보여 준다. 다만 대본으로 만든 소규모 실험이므로 실제 진료의 오류 발생률로 일반화할 수는 없다.

오류는 기록에 반대 의미로 나타날 수도 있다. 연구 사례에서는 환자가 청력 저하를 말했으나 통역에서 빠지자, AI 기록에 다른 청력 문제가 없다는 취지의 내용이 들어갔다. 환자가 실제로 말한 증상이 기록에서 부정된 것이다.

의사가 여러 증상을 질문했지만 통역사가 일부만 전달한 경우도 있었다. 환자는 전달받은 질문에만 답했는데 AI는 원래 질문 전체에 답한 것으로 기록했다. "묻지 않은 사항"과 "환자가 없다고 답한 사항"을 구분하지 못한 사례다.

이런 기록은 근거 없는 음성 소견(unsupported negative finding)으로 볼 수 있다. 여기서 음성 소견은 특정 증상이나 이상이 없다는 뜻이다. 기록 초안을 검토할 때는 문장이 자연스러운지뿐 아니라, 실제로 확인한 내용인지 살펴야 한다.

## 오류가 생기는 단계

입력은 여러 사람이 주고받는 말소리다. 시스템은 말소리를 문자로 바꾸고 누가 말했는지 나눈 뒤, 진료기록에 필요한 정보를 고른다. 마지막으로 항목에 맞춘 문장으로 출력한다. 이 중 한 단계에서 빠진 정보를 다음 단계가 확인 없이 메우면 실제 대화와 다른 기록이 만들어질 수 있다.

예를 들어 환자가 약을 먹은 뒤 어지러웠다고 말하고, 뒤이어 현재는 괜찮다고 덧붙였다고 하자. 두 문장을 하나로 합쳐 어지러움이 없다고 정리하면 과거의 반응이 사라진다. 반대로 지금도 어지럽다고 기록하면 현재 상태가 달라진다. 이 예시는 시간의 흐름을 보존하는 일이 왜 필요한지 설명한다. 말한 낱말을 모두 알아들어도 문장 사이의 관계를 잘못 묶을 수 있다.

AI 스크라이브의 처리 과정에는 자동 음성 인식(ASR, Automatic Speech Recognition)과 언어모델을 통한 정리가 포함된다. 말소리를 잘못 인식하면 이후 내용 정리에도 영향을 줄 수 있다. 음성을 정확히 인식했더라도 화자를 혼동하거나 발연 사이의 관계를 잘못 해석할 수 있다.

다국어 진료에서는 원어와 통역 중 어느 내용을 기준으로 기록하는지, 서로 다른 발언이 충돌할 때 어떻게 처리하는지도 중요하다. 특정 언어와 화자에 대한 학습량의 차이가 영향을 줄 수 있지만, 현재 연구만으로 오류의 원인을 하나로 단정하기는 어렵다.

따라서 최종 기록뿐 아니라 음성 인식과 화자 구분, 내용 요약의 각 단계를 평가해야 한다. 제품이 갱신되면 이전 평가 결과가 그대로 유지되는지도 확인해야 한다.

## 화자 집단에 따른 음성 인식 차이

단어 오류율을 볼 때는 오류가 어디에 생겼는지 함께 살펴야 한다. 예를 들어 인사말의 한 단어를 틀린 경우와 약을 먹지 않았다는 부정 표현을 놓친 경우는 기록에 미치는 영향이 다르다. 전체 오류율이 낮아져도 안전과 관련된 표현의 오류가 남을 수 있다. 약물명이나 숫자를 잘 인식하는지, 부정과 시점을 보존하는지 따로 검토하는 이유다.

음성 인식의 정확도는 언어와 억양, 방언, 녹음 환경 등에 따라 달라진다. 일반 음성 인식 시스템을 비교한 연구에서는 흑인과 백인 화자의 발연에 대한 오류율 차이가 보고되었다. 가정간호 대화를 분석한 의료 연구에서도 화자 집단에 따른 성능 차이가 나타났다.

평가에는 단어 오류율(WER, Word Error Rate) 등이 사용된다. 이는 인식 결과의 대체·생략·추가 오류를 기준 단어 수에 비추어 평가하는 지표다. 그러나 의료적 중요성이 서로 다른 단어를 같은 단위로 세므로, 수치만으로 임상 위험 전체를 설명할 수는 없다. 약물명과 용량, 알레르기, 증상 유무의 오류를 별도로 살펴야 한다.

집단 간 성능 차이에는 학습 자료의 대표성 부족과 음향 모델의 일반화 한계 등이 관련될 수 있다. 발화 특성과 장비, 주변 소음도 영향을 준다. 관찰된 격차를 평가하면서 원인과 개선 방법을 구체적으로 확인해야 한다.

## 도입 시 필요한 평가와 검토

오류를 발견한 뒤에는 최종 문장만 고치고 끝내지 않을 수 있다. 어느 화자의 발연에서 문제가 시작되었는지, 통역과 음성 인식 중 어떤 단계가 영향을 주었는지 확인하면 개선 대상을 찾기 쉽다. 다만 녹음과 원문을 검토하는 과정에서 환자정보 접근 권한과 보관 기준을 지켜야 한다. 검토에 필요한 자료와 무제한적인 녹음 보관을 같은 것으로 볼 수는 없다.

의료기관은 실제 이용자의 언어와 진료 상황을 반영해 도입 전 평가를 진행해야 한다. 단일 언어의 명료한 발연뿐 아니라 통역, 여러 화자의 대화, 억양 차이와 주변 소음도 고려해야 한다.

기록에는 누구의 발언인지와 확인 여부가 명확히 드러나야 한다. 의료진은 증상과 약물 정보, 부정 표현 등이 실제 대화와 일치하는지 확인하고 필요한 내용을 수정해야 한다. AI 결과가 초안이라는 점은 통역이 포함된 경우에도 달라지지 않는다.

개발사는 다양한 언어와 화자 집단을 포함한 자료로 성능을 평가하고, 집단별 차이를 점검해야 한다. 실제 진료에서의 오류 유형과 수정 부담을 관찰하는 과정도 필요하다.

통역 오류와 음성 인식 편향은 의료 AI의 접근성과 안전성에 함께 영향을 준다. 다양한 환자가 기술의 도움을 받으려면 여러 언어에서 기능을 제공하는 것뿐 아니라, 의미가 정확하게 기록되고 의료진이 오류를 확인할 수 있도록 구성해야 한다.

## 8.2 자동화 편향(Automation Bias)과 인간 검토(Human-in-the-loop)

북미영상의학회가 소개한 유방촬영 연구는 잘못된 AI 제안이 판독자에게 미친 영향을 보여 준다. 스탠퍼드 연구진의 MedAgentBench 자료는 의료 에이전트의 업무 수행을 가상 의무기록에서 시험하는 방식을 설명한다.

<https://www.rsna.org/media/press/i/2432>

<https://stanfordmlgroup.github.io/projects/medagentbench/>

자동화 편향(automation bias)은 사용자가 시스템의 제안을 지나치게 신뢰해 자신의 판단이나 다른 근거를 충분히 검토하지 않는 경향이다. 의료 AI가 자연스러운 문장과 근거를 제시하더라도 내용이 정확하다고 보장할 수는 없다. 업무가 복잡하고 확인하기 어려울수록 이러한 문제가 중요해진다.

자동화 안주(automation complacency)는 시스템이 반복해서 잘 작동하는 것을 본 뒤 감시와 확인이 약해지는 상태를 뜻한다. 두 현상은 의료진의 경험만으로 완전히 방지하기 어렵다. 사용자의 교육과 함께 업무 절차와 화면 설계를 검토해야 한다.

### 잘못된 제안을 따르거나 누락을 놓치는 경우

자동화 편향은 특정 회사의 AI에 붙인 이름이 아니다. 사람과 자동화 도구가 함께 일할 때 생길 수 있는 행동상의 문제다. 예를 들어 의료진이 원자료를 읽기 전에 AI 결론을 먼저 보면 그 결론에 맞는 정보부터 찾을 수 있다. 시스템이 틀렸는지를 확인하는 절차도 사용자의 주의와 시간에 영향을 받는다.

2023년 유방촬영 실험은 AI가 제안한 것처럼 제시한 분류값에 일부러 오류를 넣었다. 특정 판매 제품의 자연스러운 오류 발생률을 측정한 연구와는 다르다. 이런 실험은 잘못된 제안을 접했을 때 사람이 얼마나 영향을 받는지 살피는 데 적합하다. 실제 도입 효과를 평가하려면 제품의 오류 빈도와 의료진의 사용 방식도 함께 알아야 한다.

실행 오류(commission error)는 잘못된 AI 제안을 따라 부적절한 행동을 하는 경우다. 누락 오류(omission error)는 AI가 경고하지 않았다는 이유로 필요한 확인이나 조치를 하지 않는 경우다.

영상 판독 연구에서는 잘못된 AI 제안이 의료진의 판단에 영향을 줄 수 있음이 보고되었다. 숙련도가 높은 의사도 영향을 받았으며, AI가 놓친 부위에 대한 주의가 줄어드는 양상도 관찰되었다. 다만 통제된 판독 실험의 결과를 실제 진료의 오류율로 일반화해서는 안 된다.

AI 교육을 받은 의사들이 증례를 평가한 연구에서도 잘못된 제안이 판단에 영향을 주었다. 이는 교육이 불필요하다는 뜻이 아니라, 교육만으로 오류의 영향을 충분히 막기 어렵다는 의미다. 시스템의 사용 방식과 검토 조건을 함께 설계해야 한다.

## 대안과 불확실성의 제시

두 답변을 제시한다고 항상 독립적인 확인이 이루어지지 않는다는 사실은 비슷한 자료와 학습 방식에 의존하면 같은 오류를 낼 수도 있다. 답변 수를 늘리는 것과 서로 다른 근거를 대조하는 일은 다르다. 대안이 왜 나왔는지, 어떤 정보가 부족한지 표시해야 의료진이 비교할 수 있다. 의견이 일치했다는 사실만으로 정확성이 보장되는 것은 아니다.

하나의 답만 제공하는 대신 여러 제안과 대안을 보여 주는 방식이 연구되고 있다. 두 모델의 의견을 비교한 초기 연구에서는 일부 조건에서 잘못된 제안의 영향이 줄었다. 그러나 모든 진료 상황에서 같은 효과가 확인된 것은 아니다.

대안 가설과 판단 근거, 불확실성을 확인할 수 있으면 의료진이 결과를 재검토하는 데 도움이 될 수 있다. 반대로 정보가 지나치게 많거나 설명이 복잡하면 검토 부담이 늘어난다. 실제로 어떤 화면과 절차가 적절한지는 사용자와 업무에 맞춰 평가해야 한다.

검토에 필요한 원자료와 추천 이유를 쉽게 확인하고, 동의하지 않을 때 수정하거나 거부할 수 있어야 한다. AI가 확신하지 못하는 경우 이를 명확히 전달하고 사람의 판단을 요청하는 기능도 중요하다.

## 인간 검토의 의미

좁은 범위의 자율형 기기를 보여 주는 사례로 IDx가 개발한 IDx-DR이 있다. FDA는 2018년 4월 11일 De Novo 절차로 이 기기를 분류했다. 당시 허용 범위는 정해진 카메라를 사용해 특정 성인 당뇨 환자의 망막 영상을 평가하는 것이었다. 이 날짜는 FDA 결정일이며 기술의 최초 개발일이 아니다. 모든 안과 질환을 자율 진료하도록 허용한 사례로 해석하면 안 된다.

사람의 검토가 실질적으로 작동하려면 의견이 다를 때 이유를 적거나 결과를 고칠 수 있어야 한다. 잘못된 제안을 거부했는데도 뒤의 시스템이 원래 제안을 실행하면 감독이 끊긴다. 수정한 내용이 기록과 후속 업무에 반영되었는지까지 확인하는 구조가 필요하다.

휴먼 인 더 루프(Human-in-the-loop, HITL)는 사람이 AI의 처리 과정에 참여해 검토하고 결정하는 구조다. 의료진이 실질적으로 판단하려면 결과를 확인할 시간과 정보, 수정 권한이 있어야 한다. 형식적으로 승인 버튼만 누르는 절차로는 충분하지 않다.

감독의 방식은 제품의 사용 목적과 위험에 따라 달라진다. 모든 의료 AI의 출력마다 의사가 사전 승인해야 한다는 의미는 아니다. 좁고 명확한 과제에서 의사의 판독 없이 결과를 내도록 허가된 선별 기기도 있다. 반면 여러 질환을 판단하고 처방까지 수행하는 범용 에이전트는 더 복잡한 위험을 다룬다.

관련 법률과 지침은 인간의 감독을 중요한 설계 요소로 다룬다. 다만 구체적인 의무는 제품의 분류와 허가 조건, 적용 시점에 따라 확인해야 한다. 임상 책임이 있는 사람이 결과를 검토하고 필요한 경우 사용을 중단할 수 있는 운영 구조가 필요하다.

### 에이전트의 실행과 안전 절차

MedAgentBench는 스탠퍼드 연구진이 만든 평가 환경이다. 첫 사전공개는 2025년 1월 24일이며 학술지 논문은 같은 해 8월에 나왔다. 환자를 진료하는 상용 AI 제품이 아니다. 가상의 전자의무기록과 업무 요청을 제공하고 에이전트가 필요한 조회와 작업을 마치는지 평가한다.

답변이 그럴듯한 것과 일이 끝난 것은 다르다. 예를 들어 검사를 주문했다고 답했지만 실제 주문 항목이 만들어지지 않았다면 업무는 완료되지 않았다. 다른 환자의 기록에 작업했다면 더 큰 문제가 생긴다. 에이전트를 평

가할 때는 문장뿐 아니라 대상 환자와 실행 결과, 누락된 단계까지 확인해야 한다.

에이전트는 정보를 설명하는 것뿐 아니라 기록 조회와 문서 작성, 주문 입력 같은 행동을 수행할 수 있다. 가상 전자의무기록을 이용한 평가에서도 모든 업무를 성공적으로 처리하지는 못했다. 다만 업무 수행 실패와 실제 환자 위해는 다른 지표이므로 구분해야 한다.

실행 권한은 위험에 맞게 제한할 수 있다. 정보 조회와 초안 작성, 고위험 처방을 같은 권한으로 처리할 필요는 없다. 응급 상황이나 근거가 부족한 사례는 담당 의료진에게 넘기고, 승인이 필요한 행동은 검토 후 수행하도록 구성할 수 있다.

제안의 근거와 변경 내용, 승인과 실행 이력을 기록하면 오류가 발생했을 때 과정을 확인하는 데 도움이 된다. 이런 구성은 여러 안전 원칙을 업무에 적용하는 예이며, 모든 기관이 동일한 순서로 운영해야 하는 표준 절차는 아니다.

### 검토 부담과 지속적인 평가

의료진에게 모든 AI 출력을 처음부터 다시 검토하도록 요구하면 업무 경감 효과가 줄어든다. 반대로 검토를 생략하면 오류를 발견하지 못할 수 있다. 기능별 위험과 오류 유형을 고려해 필요한 확인 항목을 정해야 한다.

의료기관은 사용 과정에서 제안의 수용과 수정, 오류 발견, 경보 부담을 함께 관찰해야 한다. 범용 챗봇을 임상 제품처럼 사용하는 문제도 구분해 관리해야 한다. 의료 목적에 맞는 설계와 평가가 이루어졌는지 확인하는 것이 우선이다.

자동화 편향을 줄이는 목표는 AI를 덜 사용하는 데 있지 않다. 의료진이 결과를 이해하고 적절히 수용하거나 거부할 수 있도록 돕는 것이다. 기술의 정확성, 검토 가능한 설계, 사용자의 교육과 충분한 업무 여건이 함께 마련되어야 한다.

### 8.3 의료 에이전틱 AI 시장과 CPT 수가

미국의사협회의 CPT 부록 S 안내는 AI 출력의 임상 역할에 따라 분류하는 방식을 설명한다. 초당파정책센터의 보고서는 코드와 보험 지불, 실제 가치 평가가 왜 별도 과정인지 살펴보는 자료다.

<https://www.ama-assn.org/practice-management/cpt/cpt-appendix-s-taxonomy-artificial-intelligence-medical-services-procedures>

<https://bipartisanpolicy.org/issue-brief/paying-for-ai-in-u-s-health-care/>

의료 에이전틱 AI의 시장은 행정 자동화, 의료진의 의사결정 지원, 신약개발과 데이터 분석 등으로 확대되고 있다. 인력 부족과 비용 부담, 반복적인 정보 처리 수요가 도입의 배경이다. 다만 시장의 성장 전망과 실제 의료기관의 구매, 환자에게 나타나는 효과는 구분해서 보아야 한다.

#### 시장 전망의 의미

그랜드뷰리서치(Grand View Research)는 의료 에이전틱 AI 시장이 빠르게 성장할 것으로 전망했다. 해당 전망은 2025년 시장을 약 7억 6천만 달러로 추정하고, 2033년에는 약 173억 달러에 이를 것으로 보았다.

이 수치는 시장조사기관의 예측이며 확정된 매출 통계가 아니다. 에이전틱 AI의 정의와 포함하는 제품에 따라 기관별 전망도 달라진다. 연평균 성장률(CAGR)은 기간 중 예상되는 성장 속도를 요약하지만, 매년 같은 성장이 실제로 일어난다는 뜻은 아니다.

도입 방향은 원무 자동화와 임상 지원, 연구 플랫폼의 사례에서 확인할 수 있다. 그러나 협력 발표나 제품 배포가 늘어나는 것만으로 경제성과 임상

적 가치가 입증되지는 않는다. 지속적인 운영에는 비용 부담과 보상 구조가 필요하다.

## CPT 코드와 보험 지불

코드는 어떤 서비스를 했는지 표현하는 공통 언어에 가깝다. 보험 적용은 그 서비스를 어떤 환자와 조건에서 보상할지 정하는 문제다. 지급액은 보상하기로 한 서비스에 얼마를 지불할지의 문제다. 이 세 단계를 구분하면 코드가 생겼다는 소식만으로 병원에 수입이 생긴다고 판단하는 오류를 피할 수 있다.

미국에서 의료 AI의 보상을 이해하려면 미국의사협회(AMA)의 CPT(Current Procedural Terminology) 코드와 보험자의 지불 제도를 구분해야 한다. CPT는 의료 행위와 서비스를 구분하는 코드 체계다. 코드가 존재한다고 모든 보험자가 비용을 지급하는 것은 아니다.

CPT 범주 I은 확립된 서비스, 범주 II는 성과 측정, 범주 III은 새로운 기술과 서비스의 사용 추적을 위한 코드다. 범주 III은 임시 코드 성격을 가지며 지불을 보장하지 않는다. 메디케어·메디케이드 서비스센터(CMS)와 보험자는 별도의 급여와 가격 기준을 적용한다.

현재 의료 AI의 보상 사례는 주로 영상과 심전도, 망막 검사 등 범위가 정해진 기능에 집중되어 있다. CT 기반 관상동맥 혈류 예비력 분석(FFR-CT)과 자율형 당뇨망막병증 선별 등이 관련 사례다. 일반 검사 코드와 AI 기능을 위한 코드는 구분해야 한다.

## 보조형·증강형·자율형

CPT의 AI 분류도 모델의 이름이나 크기를 기준으로 삼지 않는다. 같은 기술을 써도 출력이 원자료를 정리한 것인지, 임상적 지표인지, 독립적인 해석과 결론인지에 따라 역할이 달라진다. AMA는 2026년 5월 분류 설명을 다듬었고 같은 해 9월 공식 안내를 갱신했다. 최신 정의와 실제 서비스가 맞는지 확인하는 일이 필요하다.

AMA의 AI 분류는 출력이 임상에서 어떤 역할을 하는지 살핀다. 보조형(assistive)은 의료진의 해석과 보고가 필요한 정보를 제공한다. 증강형(augmentative)은 입력 자료에서 임상적으로 의미 있는 수치를 도출한다. 자율형(autonomous)은 분석 결과를 진단이나 처치에 관한 결론으로 연결한다.

이 구분은 AI를 하나의 기술로 묶기보다 실제 서비스의 기능과 책임을 평가하는 데 도움이 된다. 다만 분류와 코드 부여, 보험 급여 결정은 서로 다른 과정이다. 대화와 정보 조회, 여러 행동을 결합한 에이전트 플랫폼이 하나의 기존 코드로 설명되기 어려운 이유도 여기에 있다.

## 입원과 외래의 보상 구조

입원 진료의 진단군별 묶음 지불(MS-DRG)에서는 관련 비용을 일정한 묶음 안에서 처리한다. AI 비용도 이 범위에 포함될 수 있다. 일정 요건을 충족하는 새 기술에는 신기술 추가 지불(NTAP) 같은 한시적 보상 제도가 적용될 수 있다.

외래에서도 새로운 기술을 위한 임시 분류와 별도 가격 결정이 사용된다. 그러나 모든 AI 서비스가 충분한 보상을 받는 것은 아니다. 실제 청구 규모가 제한적이거나 기관별로 이용 차이가 나타나는 경우도 있다.

병원은 별도 수가가 없더라도 자체 예산이나 구독 계약으로 도구를 도입할 수 있다. 이때 기록 시간, 인력 부담, 환자 흐름 등에서 얻는 이익과 구매·연결·교육·검토 비용을 비교해야 한다.

### 행위별 수가와 가치 기반 지불

행위별 수가는 개별 서비스의 수행을 기준으로 보상한다. AI가 같은 업무를 더 빠르게 하거나 일부 서비스를 줄이면, 그 편익이 현재의 보상 구조에 충분히 반영되지 않을 수 있다. 반대로 AI 관련 서비스를 계속 추가하면 전체 지출이 늘어날 가능성도 있다.

가치 기반 지불은 환자 결과와 의료의 질, 비용을 함께 고려하려는 접근이다. AI가 입원이나 불필요한 검사를 줄이는 데 기여한다면 이런 방식과 연결될 수 있다. 다만 여러 제도와 도구 중 어떤 요소가 성과를 만들었는지 구분하기 어렵고, 보험자별 계약도 다르다.

정책연구기관 초당파정책센터(BPC)는 입증된 가치에 대한 보상, AI 유형별 차등, 취약 지역의 접근성, 대표성 있는 자료와 지속적인 재평가 등을 제안했다. 이는 정책 방향에 관한 제안이며 확정된 지급 기준은 아니다.

## 경제성 평가의 기준

예를 들어 기록 초안 작성에 쓰는 시간이 줄었다고 하자. 절약한 시간 중 일부는 초안을 확인하고 오류를 고치는 데 다시 든다. 병원은 남은 시간을 환자 설명이나 휴식, 추가 진료 중 어디에 쓰는지도 살펴야 한다. 시간 절감이 곧바로 인건비 감소나 매출 증가로 이어지는 것은 아니다.

비용을 부담하는 곳과 이익을 얻는 곳도 다를 수 있다. 병원이 시스템을 구입하고 환자는 이동과 대기 부담을 줄일 수 있다. 보험자는 불필요한 서비스가 줄 때 비용 절감 효과를 볼 수 있다. 누구의 관점에서 계산하는지 정하지 않으면 같은 기술을 두고 경제성이 있다는 평가와 없다는 평가가 엇갈릴 수 있다. 환자의 건강 결과와 비용 부담을 나누어 적으면 차이를 이해하기 쉽다.

의료 AI의 경제성은 구매 가격만으로 판단하기 어렵다. 시스템 연결과 유지, 의료진의 검토, 자료 갱신, 오류 대응에 필요한 비용을 함께 계산해야 한다. 시간을 줄였더라도 다른 업무 부담이 늘었다면 전체 효과는 달라질 수 있다.

환자에게 나타나는 변화도 중요하다. 검사와 치료의 적절성, 대기와 접근성, 부작용과 재입원 등 사용 목적에 맞는 결과를 살펴야 한다. 매출 증가와 이용 횟수는 이러한 가치를 직접 대신하지 못한다.

국가마다 보험과 신의료기술 평가 체계가 다르므로 미국 CPT 사례를 한국이나 다른 나라에 그대로 적용할 수는 없다. 각 제도에서 어떤 서비스를 보상하고 어떤 근거를 요구하는지 확인해야 한다.

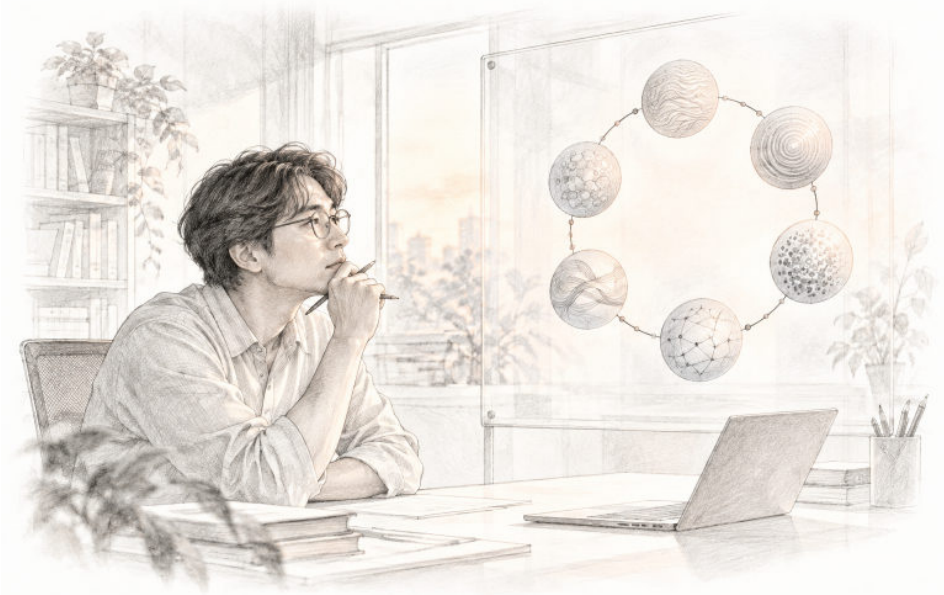
의료 에이전틱 AI가 지속적으로 사용되려면 기술적 가능성과 비용 부담, 보상이 연결되어야 한다. 시장 전망을 도입 근거로 삼기보다, 실제 업무와 환자에게 어떤 개선이 있었는지 확인하고 그 가치에 맞는 지불 방식을 마련하는 것이 중요하다.



## Part 5. 의료 AI 에이전트와 협진



## 제9장. 의료 에이전틱 AI(Agentic AI)



## 9.1 단일 예측 모델의 한계

AgentClinic 연구진이 공개한 논문을 읽으면 가상 환자에게 질문하며 진단하는 평가 방식을 확인할 수 있다. 정보를 모두 받은 시험과 직접 찾아야 하는 상황의 차이를 살펴보는 자료다.

<https://arxiv.org/abs/2405.07960>

의료 AI는 영상에서 병변을 찾거나 특정 질환의 위험도를 계산하는 등 범위가 정해진 과제에서 발전해 왔다. 흉부 엑스선(X-ray)의 폐 결절(結節) 탐지와 피부 영상의 흑색종(黑色腫, melanoma) 분류가 그 예다. 이런 모델은 정해진 입력을 분석해 결과를 제시한다.

이 절에서 다루는 단일 예측 모델(single predictive model)은 환자의 경과를 지속적으로 확인하거나 다음 행동을 선택하는 기능이 제한된 시스템을 뜻한다. 모든 예측 모델이 같은 구조인 것은 아니지만, 특정 과제의 성능과 전체 진료를 지원하는 능력은 구분해야 한다.

### 환자의 임상 맥락

예를 들어 영상 모델이 폐의 작은 이상 부위에 표시를 했다고 하자. 표시가 있다는 사실만으로 병명을 정할 수는 없다. 의사는 이전 영상에도 같은 부위가 있었는지, 새 증상이 생겼는지 등을 확인한다. 모델의 출력은 이 판단에 들어가는 자료 하나다. 입력한 영상 바깥의 사정을 모델이 스스로 안다고 여겨서는 안 된다.

진료에서는 현재 증상뿐 아니라 병력, 복용 약물, 알레르기, 검사 결과의 변화가 함께 중요하다. 심전도(心電圖, ECG)의 특정 소견도 환자의 증상과 장기 기능, 이전 검사에 따라 의미가 달라질 수 있다.

입력에 필요한 정보가 포함되지 않으면 모델은 이를 고려하기 어렵다. 특정 영상이나 검사에 대한 정확한 분석이 곧 환자 전체에 대한 적절한 판단을 뜻하지는 않는 이유다. 여러 자료를 연결하고 시간에 따른 변화를 해석하는 기능이 필요하다.

거대언어모델(LLM, Large Language Model)은 많은 기록을 처리할 수 있지만, 긴 입력의 모든 정보를 같은 수준으로 사용하지는 못한다. 'Lost in the Middle' 연구는 입력 중간에 있는 관련 정보를 놓치는 문제를 다루었다. 문맥 창(context window)이 커져도 필요한 정보를 정확하게 찾아 적용하는지는 별도로 평가해야 한다.

### 실제 진료 기록의 특성

예를 들어 약물 목록에는 과거 처방이 남아 있지만 최근 진료 메모에는 복용을 중단했다고 적혀 있을 수 있다. 두 문장을 함께 찾아도 어느 시점의 상태를 묻는지 구분하지 못하면 답은 틀린다. 기록을 많이 읽는 능력과 현재 상황에 맞게 정리하는 능력은 별도로 확인해야 한다. 긴 기록을 줄일 때도 알레르기와 중단 사유처럼 판단을 바꾸는 조건은 남겨야 한다.

전자의무기록(EHR, Electronic Health Record)은 문서 형식과 약어가 다양하고, 정보가 누락되거나 서로 충돌할 수 있다. 같은 내용이 반복되기도 하며, 기록 시점과 실제 사건의 시점을 구분해야 하는 경우도 있다.

정리된 시험 문제에서는 필요한 정보가 주어지지만 실제 진료에서는 중요한 정보를 먼저 찾아야 한다. 의료 LLM 평가를 검토한 체계적 문헌고찰(systematic review)도 실제 환자 자료를 사용한 연구가 제한적이라는 점을 지적했다. 시험 점수를 실제 진료 지원 능력으로 직접 해석해서는 안 된다.

### 정보 수집과 판단의 수정

진단은 가설을 세우고 필요한 질문과 검사를 선택한 뒤 결과에 따라 판단을 수정하는 과정이다. 치료 후에도 반응을 확인하고 계획을 조정한다. 한 번의 입력으로 결과를 내는 모델은 이러한 과정을 자체적으로 수행하기 어렵다.

계산기와 검색, 진료 지침 같은 외부 도구를 사용하고, 부족한 정보를 요청하며, 이전 판단을 다시 검토하는 기능이 필요할 수 있다. 상태(state)를 유지하고 여러 단계를 반복하는 구조는 이런 요구에 대응하려는 접근이다.

MedLA는 시후대와 애틀그룹 등 공동 연구진이 개발한 추론 틀이다. 2025년 9월 28일 사전공개 논문을 올렸으며, 여러 에이전트의 전제와 결론을 비교한다.

그러나 단계를 늘리는 것만으로 정확성이 보장되지는 않는다. 앞선 전제를 잘못 유지하거나 단계 사이에 모순이 생길 수 있다. MedLA는 여러 에이전트가 삼단논법(三段論法, syllogism) 형태의 추론을 비교하는 방법을 연구했다. 감별진단(鑑別診斷, differential diagnosis)에서 전제와 결론의 일관성을 검토하려는 시도다.

## 대화형 평가가 보여 주는 차이

AgentClinic은 진료 서비스의 제품명이 아닌 연구용 평가 환경의 이름이다. 새뮤얼 슈미드갈 등 연구진이 2024년 5월 13일 논문을 처음 공개했다. 논문에는 존스홉킨스대, 스탠퍼드대, 취리히 연방공대 등의 연구자가 참여한다. 2025년 5월 개정판과 최초 공개 시점은 구분해야 한다. 의사 역할의 모델이 질문을 입력하면 다른 역할이 환자 정보나 검사 자료를 전달한다. 평가자는 그렇게 이어진 대화와 최종 판단을 살핀다.

AgentClinic은 AI가 가상 환자와 대화하고 추가 정보를 요청해 진단하도록 만든 벤치마크(benchmark)다. 필요한 정보를 한꺼번에 제공할 때와 스스로 수집하게 할 때 모델의 성능이 달라질 수 있음을 보여 준다.

결과에는 의학 지식뿐 아니라 대화 전략, 허용된 단계 수, 도구 사용, 응답 거부 등이 영향을 준다. 따라서 낮은 점수를 곧 지식 부족으로 해석하거나, 가상 환자 진단 결과를 실제 진료 정확도로 보아서는 안 된다.

도구를 연결한 에이전트가 일부 과제에서 개선을 보이더라도 처리 비용과 시간이 늘 수 있다. 토큰(token) 사용량과 응답 지연, 환각(hallucination)도 함께 평가해야 한다. 더 복잡한 구조가 항상 더 나은 선택은 아니다.

## 사람과 함께 사용할 때의 평가

후향적 연구는 이미 남아 있는 과거 기록으로 성능을 살핀다. 전향적 연구는 정한 계획에 따라 앞으로 일어나는 일을 관찰한다. 과거 기록의 병명이 정답으로 주어진 평가에서는, 의사가 당시 무엇을 모르고 있었는지 충분히 드러나지 않을 수 있다. 연구 결과를 읽을 때 자료를 모은 순서와 AI가 볼 수 있었던 범위를 함께 확인해야 하는 이유다.

여러 연구를 종합한 메타분석(meta-analysis)에서는 LLM 지원의 효과가 사례의 종류에 따라 달라졌다. 잘 정리된 증례와 실제 환자의 후향적(後向的, retrospective) 기록은 평가 조건이 다르기 때문이다.

일반인이 의료 상황을 판단하는 실험에서도 모델 단독 성능과 사람이 같은 모델을 사용했을 때의 결과가 일치하지 않았다. 사용자가 필요한 정보를 전달하고 답변을 이해하는 과정이 실제 효과에 영향을 준다.

따라서 의료 AI는 단독 성능 외에 사용자의 판단과 행동에 어떤 변화를 만드는지 평가해야 한다. 실제 환자 결과에 대한 평가는 전향적(前向的, prospective) 연구 등 목적에 맞는 근거가 필요하다.

## 에이전틱 시스템으로의 확대

에이전틱 시스템(agent system)은 필요한 정보를 관찰하고, 다음 행동을 계획하며, 도구를 사용한 결과를 다시 검토하도록 구성된다. 기억과 반복 처리를 통해 단일 과제의 결과를 진료 과정에 연결하려는 접근이다.

단일 예측 모델은 명확한 과제를 수행하는 데 계속 중요한 역할을 한다. 에이전트는 이러한 모델과 자료, 도구를 연결할 수 있지만 새로운 오류와 윤

영 부담도 발생시킨다. 어떤 기능을 추가해야 하는지는 실제 업무의 요구와 위험을 기준으로 판단해야 한다.

## 9.2 인지·사고·행동(Sense-Think-Act)과 메모리

프린스턴대와 구글 연구진의 ReAct 논문은 계획과 도구 사용이 어떻게 이어지는지 설명한다. 스탠퍼드대의 MedAgentBench 논문은 이와 같은 기능을 가상 의무기록에서 시험하는 모습을 보여 준다.

<https://arxiv.org/abs/2210.03629>

<https://arxiv.org/abs/2501.14654>

에이전트(agent)는 환경의 정보를 받아 다음 행동을 선택하고, 그 결과를 다시 확인하는 시스템이다. 의료 에이전트에서는 진료 기록과 검사 결과를 확인하고 필요한 자료를 검색한 뒤, 새로운 정보에 따라 계획을 수정하는 과정으로 구현할 수 있다. 이를 인지·사고·행동(Sense-Think-Act)의 반복 구조인 루프(loop)로 설명한다.

이 구조는 답변을 한 번 생성하는 기능과 다르다. 여러 단계의 업무를 이어가려면 현재 상태와 이전 행동, 새로 얻은 정보를 함께 관리해야 한다. 메모리(memory)는 이러한 정보의 유지와 검색을 지원한다.

### 인지: 환자 정보의 수집과 정리

인지 단계에서는 전자의무기록(EHR), 생체 신호, 영상, 대화 등에서 정보를 수집한다. 현재 수치를 이전 기록과 비교하고, 자료의 시점과 출처를 확인해 환자 상태를 정리한다.

자료가 있다고 모두 신뢰할 수 있는 것은 아니다. 오래된 기록, 누락, 입력 오류, 상충하는 정보가 있을 수 있다. 에이전트는 확인된 사실과 추정, 추가 확인이 필요한 내용을 구분해야 한다.

## 사고: 계획과 재검토

ReAct의 첫 사전공개일은 2022년 10월 6일이다. 이후 2023년 ICLR 학회에 발표했다. 프린스턴대와 구글 리서치가 함께 연구한 일반 에이전트 설계 방법이며, 의료 전용 모델의 출시일을 뜻하지 않는다. Reflexion은 노스이스턴대, MIT, 프린스턴대 연구진이 제안했다. 논문은 2023년 3월 20일 처음 공개했다. 앞선 수행을 말로 정리한 피드백을 다음 시도에 넣는 방식이다. 두 연구의 이름을 병원에서 허가받은 제품명으로 읽어서는 안 된다.

계획 단계에서는 목표를 세부 업무로 나눈다. 예를 들어 퇴원 지원에는 검사 결과 확인, 약물 검토, 복약 안내, 후속 진료 계획 등이 포함될 수 있다. 어떤 자료와 도구가 필요한지 정하고 수행 순서를 구성하는 과정이다.

ReAct 연구는 추론과 도구 사용을 번갈아 수행하면서 관찰 결과에 따라 계획을 조정하는 접근을 제시했다. Reflexion은 이전 시도의 결과를 기록하고 다음 수행에 참고하는 방식을 연구했다. 다만 모델이 자신의 오류를 항상 정확히 판단하는 것은 아니므로, 자기 검토와 외부 검증을 구분해야 한다.

## 행동: 허용된 도구의 사용

예를 들어 검사 결과를 찾아 달라는 요청을 받으면 에이전트는 환자를 확인하고 조회 도구를 호출한다. 도구가 돌려준 수치와 날짜를 읽어 답을 작성한다. 이때 화면에 검사명을 적는 것과 실제 검사 주문을 등록하는 것은 권한이 다른 행동이다. 같은 문장 생성 기술을 쓰더라도 조회와 등록의 경계는 시스템에서 따로 정해야 한다.

행동 단계에서는 API(Application Programming Interface)를 통해 기록을 조회하거나 계산기와 검색 도구를 사용한다. 결과가 들어오면 환자 상태와 계획을 갱신한다. 실행 가능한 기능은 제품의 권한과 사용 목적에 따라 제한해야 한다.

가상 전자의무기록에서 업무 수행을 평가한 MedAgentBench는 기록 조회와 주문 입력 같은 과제를 다룬다.

이성 장, 조너선 첸 등 스탠퍼드대 연구진이 만든 평가 환경이다. 사전공개 논문은 2025년 1월 24일 나왔고, NEJM AI 논문은 같은 해 8월 14일 공개됐다. 날짜와 함께 초기판인지 후속판인지 확인해야 시험 구성의 차이를 이해할 수 있다. 이런 평가는 도구를 연결한 에이전트의 가능성과 실패 유형을 살피는 데 도움이 된다. 그러나 가상 과제의 성공률을 실제 환자 진료의 안전성으로 직접 해석할 수는 없다.

## 메모리의 유형

여기서 기억은 컴퓨터가 남겨 둔 자료와 그 자료를 찾는 규칙을 뜻한다. 사람이 경험을 이해하고 기억하는 과정과 같다는 의미는 아니다. 예를 들어 대화 요약을 저장했다가 다음 요청에 붙여 넣는 것도 외부 기억의 한 형태다. 저장에 성공했더라도 다른 환자의 자료를 불러오면 안전한 기억 관리라고 할 수 없다.

작업 기억(working memory)은 현재 대화와 최근 검사 결과처럼 당장 필요한 정보다. 모델의 문맥 창(context window)에 포함하거나 요약해 관리할 수 있다. 요약 과정에서 중요한 조건이 빠지지 않도록 해야 한다.

일화 기억(episodic memory)은 특정 환자에게 일어난 사건과 이전 업무의 기록이다. 약물 부작용이나 과거 검사 결과 등이 포함될 수 있다. 벡터 데이터베이스(vector database), 관계형 데이터베이스(relational database) 등 외부 저장소에서 필요한 정보를 검색할 수 있다. 검색 증강 생성(RAG, Retrieval-Augmented Generation)은 이러한 정보를 답변과 판단에 연결하는 방법이다.

의미 기억(semantic memory)은 의학 개념과 관계에 관한 지식이다. 표준 용어 체계와 지식 그래프(knowledge graph) 등이 관련된다. 절차 기억(procedural memory)은 진료 지침과 병원 업무 규칙, 도구 사용 절차처럼 수행 방법을 나타낸다.

이 구분은 CoALA(Cognitive Architectures for Language Agents) 등 언어 에이전트 구조 연구에서 다루어진다. 실제 시스템에서는 각 기억을 별도 저장하거나 일부를 함께 관리할 수 있다. 중요한 것은 필요한 정보를 적절한 시점에 정확하게 사용하는 것이다.

## 기억 관리와 갱신

CoALA는 프린스턴대의 시어도어 서머스와 순위 야오 등 연구진이 2023년 9월 5일 공개한 언어 에이전트의 설계 틀이다. 여러 기억과 행동을 분류해 설명하는 연구이며, 독립적인 의료 제품은 아니다. MemGPT는 찰스 패커 등 캘리포니아대 버클리 연구진이 2023년 10월 12일 공개했다. 이름에 GPT가 들어가지만 OpenAI가 출시한 모델을 가리키지 않는다. 모델이 당장 읽을 부분과 외부에 보관할 부분을 나누어 관리하는 시스템이다.

MemGPT는 제한된 작업 공간과 외부 저장소 사이에서 정보를 관리하는 접근을 연구했다. 긴 기록을 모두 한 번에 입력하는 대신 필요한 정보를 검색해 사용하는 방식이다. AgentClinic 등에서도 메모를 이용한 수행 개선이 관찰되었지만, 특정 평가의 결과를 모든 의료 업무로 일반화할 수는 없다.

메모리를 유지하면 이전 정보를 이어서 사용할 수 있으나 오류도 반복될 수 있다. 잘못 저장한 진단이나 오래된 지침을 다시 참고하면 후속 판단에 영향을 준다. 자료의 출처와 작성 시점, 수정 이력을 관리하고 권한에 따라 갱신해야 한다.

외부 기억의 갱신과 모델 자체의 재학습은 다르다. 신경망(neural network)을 재학습하면 기존에 학습한 능력이 낮아지는 파국적 망각(catastrophic forgetting)이 발생할 수 있다. 외부 기억을 바꾸는 경우에는 모델 가중치(weight)를 직접 변경하지 않지만, 검색한 내용의 정확성과 적절성을 확인해야 한다.

## 연구실 업무로의 적용

Latent Labs 연구팀은 Latent-Y 기술 보고서를 2026년 3월 31일 처음 공개했다. 에이전트가 과학 문헌과 생물학 도구를 연결하고, 항체 후보 생성에는 같은 회사의 Latent-X2를 쓴다. Latent-Y는 업무를 진행하는 에이전트이고 Latent-X2는 후보를 만드는 생성 모델이다. 이 보고서만으로 각 제품의 최초 상용 판매일을 정할 수는 없다.

<https://arxiv.org/abs/2603.29727>

Latent Labs의 Latent-Y는 문헌 조사와 표적 분석, 항체 후보 설계, 계산 검증 등을 연결하는 연구 사례다. 사용자의 지시에 따라 여러 설계 업무를 수행하고 후보를 생성하는 방식으로 소개되었다.

일부 후보의 실험적 결합이 보고되었지만, 이는 기업이 발표한 사전공개 기술 보고서다. 실험 결과를 자동으로 반영해 다음 설계를 반복 개선하는 전 과정이 완성되었다고 볼 수는 없다. 여러 업무를 연결하는 것과 실제 결과에 따라 지속적으로 학습하는 것은 구분해야 한다.

### **반복 처리의 통제**

인지·사고·행동 구조와 메모리는 장기간 이어지는 업무를 지원한다. 그러나 반복 횟수가 늘면 오류와 비용, 처리 시간도 증가할 수 있다. 업무 완료 기준과 중단 조건, 사람에게 넘길 상황을 명확히 해야 한다.

의료 에이전트에서는 근거가 부족하거나 위험이 큰 판단을 자동으로 계속 수행하지 않도록 구성해야 한다. 지속적인 정보 관리는 자율성 확대의 수단이면서 동시에 감독이 필요한 대상이다. 정확한 기억과 제한된 권한, 검토 가능한 기록을 함께 설계하는 것이 중요하다.

### 9.3 의료 에이전트 AI의 분류와 기술 성숙도

바트살 등 연구진의 일곱 영역 분류 논문은 의료 에이전트 연구를 기능별로 비교한다. 어떤 기능을 보고했고 무엇을 평가하지 않았는지 읽을 때 도움이 된다.

<https://arxiv.org/abs/2602.04813>

의료 에이전트는 기능과 구성, 자율성의 범위가 다양하다. 정보를 검색해 답하는 시스템과 환자 기록을 유지하며 여러 도구를 실행하는 시스템을 같은 기준으로 평가하기는 어렵다. 분류 체계(taxonomy)는 이러한 차이를 정리하고 추가 개발이 필요한 기능을 확인하는 데 도움이 된다.

Vatsal 등의 의료 LLM 에이전트 문헌고찰은 일곱 영역의 분류를 제안했다. 이 틀은 연구에 보고된 설계를 비교하는 방법이며, 모든 상용 제품의 수준을 측정한 결과는 아니다.

#### 일곱 가지 분류 영역

이 분류 자체는 AI 제품이 아니다. 슈브함 바트살, 하르시 두베이, 아디티 싱이 작성한 연구이며, arXiv 공개일은 2026년 2월 4일이다. 문헌에 구현됐다고 적힌 기능을 분류한 결과이므로, 환자에게 안전하다는 인증으로 사용할 수 없다. 같은 기억 기능도 진료 중 요약만 보관하는지, 장기간 자료를 갱신하는지에 따라 실제 역할이 달라진다.

인지 능력은 계획과 다단계 추론, 불확실성 처리를 포함한다. 지식 관리는 외부 문헌과 진료 지침의 연결, 기억의 저장과 검색을 다룬다. 상호작용 방식은 사용자와의 대화뿐 아니라 특정 사건에 반응하는 기능을 포함한다.

적응과 학습은 이전 수행을 검토하고 변화에 대응하는 능력이다. 안전과 윤리는 개인정보 보호, 편향, 규제 준수와 안전장치를 살핀다. 프레임워크 유형은 단일·다중 에이전트 구성과 중앙 조정 방식을 구분한다.

핵심 과제는 실제로 수행하는 의료 업무를 뜻한다. 기록 정리와 질의응답, 감별진단(鑑別診斷), 치료 계획, 연구 설계, 환자 모니터링 등이 포함된다. 같은 구조를 사용하더라도 과제의 위험과 요구 성능은 다를 수 있다.

### 많이 연구된 기능과 부족한 기능

문헌고찰에서는 외부 지식 연결과 다중 에이전트 설계가 비교적 널리 구현되어 있었다. 검색 증강 생성(RAG)으로 문헌을 참조하고, 여러 역할을 나눠 답변을 검토하는 접근이 대표적이다. 질의응답과 의사결정 지원, 가상 환경 평가도 주요 과제였다.

반면 사건이 발생했을 때 스스로 작동하는 이벤트 기반 활성화(event-driven activation)와 성능 변화를 감지하는 드리프트 감지·완화(drift detection and mitigation)는 드물었다. 여러 에이전트의 작업을 조정하는 중앙 오케스트레이션(central orchestration)도 충분히 구현되지 않은 경우가 있었다.

이 차이는 의료 에이전트 연구가 정보 처리 기능에 비해 지속적인 운영과 변화 대응을 덜 다뤘음을 시사한다. 다만 검토 대상 논문과 기간에 한정된 결과이므로 현재 모든 제품의 기능을 대표하는 통계로 해석해서는 안 된다.

## 연구 확대와 실제 도입

다른 문헌고찰에서도 의료 에이전트 연구가 임상 지원, 환자 참여, 종양학, 영상의학 등으로 확대되는 흐름이 나타났다. 그러나 논문 수의 증가와 실제 병원의 도입 속도는 다르다.

연구가 제안한 기능이 현장에서 사용되려면 자료 접근과 시스템 연결, 검토 책임, 비용과 교육 문제가 해결되어야 한다. 표준화된 평가가 부족하면 서로 다른 연구의 점수와 효과를 비교하기도 어렵다.

## 기술 성숙도를 판단하는 근거

성숙도를 평가할 때는 어떤 환경에서 기능을 확인했는지 살펴야 한다. 정리된 문제를 푸는 벤치마크와 가상 환자 대화, 실제 환자의 과거 기록 분석, 전향적(前向的, prospective) 사용 연구는 서로 다른 근거를 제공한다. 실제 진료에 도입한 뒤 환자 결과와 운영의 지속성을 확인하는 평가도 필요하다.

이 구분은 연구를 일률적으로 순위 매기는 공식 등급이 아니다. 어떤 질문에 답할 수 있는 근거인지 판단하기 위한 기준이다. 가상 환경은 반복 실험과 오류 분석에 적합하지만 실제 의료진과 환자의 행동을 모두 재현하지는 못한다.

AgentClinic과 MedAgentBench는 정보 수집과 업무 수행을 비교할 수 있는 평가 환경이다. 그러나 이런 환경의 성과만으로 실제 진료의 안전성과 효과가 입증되는 것은 아니다. 사용 목적에 맞는 후속 평가가 필요하다.

## 운영 능력과 성숙도

예를 들어 검사 조회에 실패했을 때 오류를 알리는 시스템과, 이전 수치를 대신 제시하는 시스템은 화면만 보면 비슷할 수 있다. 그러나 후자는 사용자가 오래된 자료를 새 결과로 오해할 위험이 있다. 성공한 요청의 점수 외에 실패했을 때 어떤 정보를 보여 주고 누구에게 알리는지도 평가해야 한다. 병원이 도입을 검토할 때 필요한 근거는 이런 운영 조건까지 포함한다.

지속적인 운영에서는 처리 시간과 계산 비용도 중요하다. 도구를 반복 사용하거나 여러 에이전트가 논의하면 토큰(token) 사용량과 응답 지연이 늘어날 수 있다. 추가 계산이 실제로 정확성과 안전성을 높이는지 함께 확인해야 한다.

오류와 성능 저하를 관찰할 모니터링(monitoring), 수행 과정을 확인할 로그(log), 담당자에게 넘길 절차도 필요하다. 새로운 자료나 모델 변경에 대응하지 못하면 초기 평가가 좋더라도 장기간 같은 수준을 유지하기 어렵다.

의료 에이전트의 기술 성숙도는 기능의 수나 자율성만으로 결정되지 않는다. 필요한 업무를 안정적으로 수행하고, 오류를 발견하고, 사람이 검토할 수 있는지가 중요하다. 분류 체계는 이러한 조건을 빠짐없이 살펴보기 위한 도구로 사용할 수 있다.



## 제10장. 다중 에이전트(Multi-Agent)와 AI 병원



## 10.1 다중 에이전트 협업 설계

예일대와 상하이교통대 연구진의 MedAgents 논문은 여러 전문가 역할이 답을 검토하는 과정을 보여 준다. 역할을 나누는 순서와 의견을 합치는 방법을 확인할 수 있다.

<https://arxiv.org/abs/2311.10537>

다중 에이전트 협업 설계(multi-agent collaborative design)는 여러 에이전트가 역할을 나누고 결과를 공유해 업무를 수행하는 방식이다. 의료 분야에서는 정보 검색, 진단 후보 검토, 약물 확인, 기록 작성 등을 구분할 수 있다. 목적은 역할별 처리를 통해 복잡한 과제를 관리하고 결과를 교차 검토하는 데 있다.

같은 기반 모델에 서로 다른 역할을 부여할 수도 있고, 여러 모델과 도구를 결합할 수도 있다. 따라서 에이전트 수만으로 시스템의 다양성이나 신뢰성을 판단하기는 어렵다. 역할과 권한, 정보 전달, 결과 통합 방식을 함께 살펴야 한다.

### 연구에서의 활용

의료 LLM 에이전트 문헌고찰에서는 다중 에이전트 구조가 널리 사용된 것으로 나타났다. 이는 연구자들이 복잡한 과제를 여러 기능으로 나누는 설계를 적극적으로 탐색하고 있음을 보여 준다.

다만 검토 대상 논문의 구성 비율은 실제 병원 제품의 보급률이나 임상적 우위를 뜻하지 않는다. 다중 에이전트를 사용했다는 사실과 환자에게 더 안전한 결과를 제공했다는 사실은 별도로 평가해야 한다.

## 역할 분담

MedAgents는 상루 탕 등 연구진이 2023년 11월 16일 처음 공개했다. 2024년 ACL 관련 논문집에 발표한 연구용 협업 틀이다. 별도 의학 모델을 새로 학습시키는 방식보다, 기존 모델에 역할과 토론 순서를 부여하는 데 초점을 둔다. 입력한 의학 문제를 여러 관점에서 분석한 뒤 보고서를 합쳐 답을 내놓는다.

첫 번째 협업 방식은 기능별 역할 분담이다. 계획, 문헌 검색, 임상 분석, 검증, 설명 등을 구분하고 각 역할에 필요한 정보와 도구를 제공한다. 역할별 입력과 출력이 명확하면 업무를 추적하고 문제를 수정하기 쉽다.

MedAgents는 전문가 역할의 에이전트들이 분석을 작성하고 이를 통합하는 접근을 제안했다. 일부 의학 문제 평가에서 성능 개선을 보였지만, 가상 전문가 역할이 실제 전문의의 자격과 책임을 갖는 것은 아니다.

## 토론과 교차 검토

예를 들어 한 역할은 검사 결과를 정리하고 다른 역할은 약물 목록을 확인한다고 하자. 첫 역할이 환자의 이름을 잘못 연결하면 나머지 역할이 성실하게 일해도 다른 사람의 자료를 검토하게 된다. 여러 답변이 같다는 사실보다 각 답변이 어느 원자료에서 나왔는지 추적할 수 있는지가 중요하다. 서로 다른 역할을 만들었다고 독립적인 증거가 늘어나는 것은 아니다.

두 번째 방식은 후보와 근거를 서로 검토하는 것이다. 한 에이전트가 제안한 진단에 대해 다른 에이전트가 누락된 정보와 반대 근거를 제시할 수 있다. 결과가 일치하지 않는 경우 추가 검색이나 사람의 검토를 요청하도록 구성할 수도 있다.

다만 서로의 답변을 본 뒤 같은 오류를 수용할 가능성도 있다. 동일한 모델이나 유사한 학습 자료에 기반한 에이전트는 공통된 편향을 가질 수 있으므로, 여러 의견이 일치한다고 정확성이 보장되는 것은 아니다.

### 중앙 조정과 문제별 구성

MDAgents는 MIT의 김유빈 등 연구진이 제안했다. 구글 리서치와 서울대병원 연구자도 참여한 논문이다. 2024년 4월 22일 사전공개했고 같은 해 NeurIPS 학회에 발표했다. 앞서 소개한 MedAgents와 이름이 비슷하지만 별도 연구다. MedAgents가 역할별 토론에 초점을 두었다면, MDAgents는 문제에 맞춰 협업 구성을 고르는 과정까지 다룬다.

오케스트레이터(orchestrator)는 작업 순서와 정보 전달, 종료 조건을 관리한다. 어떤 역할을 호출할지, 결과가 충돌하면 어떻게 처리할지, 언제 사람에게 넘길지를 정하는 기능이다.

MDAgents는 문제의 난이도에 따라 단일 또는 다중 에이전트 구성을 선택하는 접근을 연구했다. 모든 과제에 같은 수의 에이전트를 사용하는 대신, 추가 협업이 필요한 경우에만 계산을 늘리는 방식이다. 실제 적용에서는 난이도 판단의 정확성과 추가 비용도 평가해야 한다.

## 추론 구조의 공유

MedLA는 시치 마 등 연구진이 만든 연구용 프레임워크다. 논문에 시후대, 앤트그룹, 중국과학원 홍콩 기관의 인공지능·로봇 연구센터, 중일우호병원이 소속으로 기재되어 있다. 2025년 9월 28일 사전공개한 뒤 AAAI 2026에 채택됐다. 대전제는 일반 규칙, 소전제는 현재 사례의 조건, 결론은 둘을 연결한 판단을 뜻한다. 전제가 참인지 확인하는 일과 전제에서 결론이 제대로 나오는지 확인하는 일은 구분해야 한다.

역할을 나누는 것 외에 판단 과정의 구조를 맞추는 방법도 있다. MedLA는 삼단논법(三段論法, syllogism) 형태로 전제와 결론을 표현하고, 에이전트들이 이를 비교해 모순을 검토하도록 했다.

이 접근은 긴 설명의 표현만 비교하기보다 무엇을 전제로 결론을 냈는지 확인하려는 것이다. 다만 논리적 일관성과 의학적 사실의 정확성은 다르므로, 원자료와 임상 근거의 확인도 필요하다.

## 가상 병원 연구

Agent Hospital은 칭화대 산업인공지능연구원과 관련 연구진이 개발했다. 논문 최초 공개일은 2024년 5월 5일이다. 별도 연구인 AI Hospital은 알리바바, 화중과기대, 푸단대 등의 연구자가 참여했다. 2024년 2월 15일 사전공개했고 2025년 COLING 학회에 발표했다. 두 이름에 병원이 들어가지만 여기서 소개한 논문은 컴퓨터 안의 진료 모의실험을 다룬다. 실제 환자를 입원시켜 자율 치료했다는 발표가 아니다.

<https://arxiv.org/abs/2405.02957>

<https://aclanthology.org/2025.coling-main.680/>

Agent Hospital은 환자와 의료진 역할의 에이전트가 상호작용하는 가상 환경을 구성했다. 의사 에이전트가 사례를 반복하고 이전 결과를 참고해 수행을 개선하는 방식을 연구했다.

AI Hospital 등 다른 연구도 여러 역할이 참여하는 진료 시뮬레이터를 평가 환경으로 사용한다. 이런 환경은 정보 전달과 협업을 반복 실험하는 데 적합하지만, 생성된 환자와 가상 진료가 실제 병원의 복잡성을 모두 재현하지는 못한다. 가상 평가의 성과를 실제 임상 효과로 해석해서는 안 된다.

## 비용과 오류 관리

종료 조건이 없으면 에이전트들이 같은 주장을 바꾸어 말하며 토론을 반복할 수 있다. 그동안 계산 비용은 늘고 결과 전달은 늦어진다. 긴 토론만으로 병원의 요구를 충족할 수는 없다. 정해진 시간에 검토 가능한 정보를 전달해야 한다. 답을 확정할 수 없을 때는 추가 확인이 필요한 항목을 남기고 담당자에게 넘겨야 한다.

에이전트가 늘어나면 토큰(token) 사용량과 응답 시간, 결과 통합의 복잡성이 커질 수 있다. 협업에 따른 추가 계산이 업무의 정확성과 안전성을 얼마나 개선하는지 확인해야 한다.

환각(hallucination)이 공유되면 후속 판단이 잘못된 전제에 기반할 수 있다. 이를 줄이려면 출처 확인과 외부 검증, 불일치 처리 절차가 필요하다. 도구 기반 에이전트의 실패 연구도 이러한 위험을 보여 주지만, 특정 구조에서 관찰한 결과를 모든 다중 에이전트의 오류율로 일반화할 수는 없다.

실패는 모델 성능 외에도 불명확한 역할과 지시, 정보 누락, 검증 부족에서 생길 수 있다. 시스템 전체의 수행 기록을 살펴 원인을 구분하고, 잘못된 합의가 자동 실행으로 이어지지 않도록 관리해야 한다.

### 협업 설계의 평가 기준

다중 에이전트 설계의 가치는 역할의 수가 아니라 필요한 과제를 얼마나 안정적으로 수행하는지에 있다. 단일 모델과 비교해 정확도와 검토 부담, 비용과 지연이 어떻게 달라지는지 평가해야 한다.

의료 적용에서는 사용 목적과 권한, 의료진의 검토, 오류 대응을 명확히 해야 한다. 연구에서 널리 쓰이는 구조라는 이유만으로 실제 도입을 정당화할 수는 없다. 환자와 업무에 맞는 근거를 확보한 뒤 적용 범위를 정하는 것이 필요하다.

## 10.2 MAI-DxO의 역할 분담과 복합 진단

마이크로소프트의 2026년 10월 공식 발표문에는 MAI-DxO의 평가와 임상 연구 계획이 정리되어 있다. 연구 점수와 앞으로 확인할 실제 진료의 효과를 구분해 읽을 수 있다.

<https://microsoft.ai/news/building-medical-superintelligence/>

복잡한 진단에서는 초기 정보만으로 결론을 내리기 어렵다.

MAI-DxO의 전체 이름은 Microsoft AI Diagnostic Orchestrator다. 마이크로소프트 연구진이 순차 진단 논문을 2025년 6월 27일 사전공개했고, 2026년 10월에는 Science 게재 연구를 발표했다. 논문 공개 시점은 누구나 병원에서 쓸 수 있는 제품의 상용 출시일과 다르다. 여러 모델을 연결하는 조정 체계라는 점에서 하나의 고정된 신경망 이름과도 구분된다. 필요한 질문과 검사를 선택하고, 결과에 따라 가능성이 높은 질환을 다시 검토해야 한다. 마이크로소프트의 MAI-DxO는 이러한 순차 진단을 여러 에이전트의 역할 분담으로 지원하려는 연구다.

이 시스템은 진단 후보와 검사 선택, 반대 근거, 비용, 결과 점검을 나누어 처리한다. 의학 지식뿐 아니라 정보를 얻는 순서와 여러 판단을 조정하는 방식이 성능에 어떤 영향을 주는지 살핀다.

## 순차 진단 평가 환경

예를 들어 초기 설명에 발열과 피로만 적혀 있다면, 에이전트는 바로 병명을 고르거나 추가 정보를 요청할 수 있다. 요청한 항목에 따라 기록에서 답을 찾아 돌려주고, 새 자료를 읽은 에이전트가 다음 행동을 정한다. 이 설명은 평가 구조를 이해하기 위한 가상 예다. 화면에서 검사 결과를 받는 행위에는 실제 병원에서 채혈하고 결과를 기다리는 모든 과정이 담겨 있지 않다.

SDBench는 복잡한 임상병리 증례 토론(CPC, Clinicopathological Conference) 등을 단계적으로 검토하도록 구성한 벤치마크다. 처음에는 제한된 사례 요약만 제공하고, 질문이나 검사 요청에 따라 추가 정보를 제시한다.

정보 제공 역할인 Gatekeeper는 요청에 해당하는 내용을 전달한다. 최종 진단은 Judge 역할의 모델이 기준과 비교하며, 요청한 검사의 비용도 추정한다. 필요한 자료를 모두 제공하는 시험과 달리, 어떤 정보를 선택할 지까지 평가하는 구조다.

## 역할별 검토

검사 선택에서는 질환을 구별하는 데 필요한 정보가 무엇인지 생각한다. 비용 검토에서는 이미 있는 정보로 판단할 수 있는지 다시 살핀다. 두 역할의 요구가 다르면 조정 체계가 다음 행동을 정한다. 이런 계산 결과를 실제 진료비 절감으로 옮기려면 검사 가격뿐 아니라 지연과 재검사, 의료진의 확인 시간도 조사해야 한다.

가설 검토 역할은 가능한 진단을 유지하고 새 정보에 따라 우선순위를 조정한다. 검사 선택 역할은 후보를 구분하는 데 도움이 될 검사를 제안한다. 반론 역할은 현재 가설과 맞지 않는 증거와 초기 판단에 대한 과도한 의존을 검토한다.

비용 관리 역할은 추가 정보의 이익과 검사 비용을 살피며, 점검 역할은 검사 명칭과 추론의 일관성 등을 확인한다. 여러 역할의 결과를 조정해 다음 질문이나 검사, 최종 진단을 정한다.

이 구조는 특정 기반 모델에만 묶이지 않는 모델 독립적 조정 체계(model-agnostic orchestrator)로 설명된다. 같은 역할 구성에 다른 모델을 연결할 수 있지만, 실제 결과는 모델과 평가 조건에 따라 달라질 수 있다.

## 평가에서 나타난 가능성

공개 연구에서는 복잡한 NEJM 증례와 응급실 기록을 바탕으로 한 과제에서 높은 진단 정확도를 보고했다. 같은 평가 조건의 일반의보다 높은 결과가 나타났으며, 검사 비용을 고려한 구성도 시험되었다.

이 결과는 역할 분담과 정보 수집 계획이 복합 진단에 기여할 가능성을 보여 준다. 다만 서로 다른 시점의 모델과 설정에서 얻은 점수를 하나의 결과처럼 합쳐서는 안 된다. 비용 역시 실제 청구 금액이 아니라 평가 환경의 추정치다.

## 임상 적용과의 차이

SDBench는 통제된 가상 평가다. 실제 환자와의 대화, 신체 진찰, 현장의 업무 조건을 모두 포함하지 않는다. 비교한 의사 집단과 자료 사용 조건도 현실의 전문의 협진과 다를 수 있다.

따라서 평가 결과를 근거로 실제 전문의 팀을 대체할 수 있다거나 환자의 예후가 개선되었다고 결론지을 수는 없다. 의료 초지능(medical superintelligence)이라는 표현도 현재의 임상 성과를 나타내기보다 기업의 개발 목표에 가깝다.

마이크로소프트는 베스 이스라엘 디코니스 메디컬센터(BIDMC)와 실제 임상 환경의 연구를 추진한다고 발표했다. 전향적(前向的, prospective) 연구에서는 환자와의 상호작용과 의료진의 독립적인 판단을 비교할 수 있다. 다만 연구 계획과 완료된 결과는 구분해야 한다.

## 의료진과의 협업 설계

실제 도입에서는 AI의 단독 정확도 외에 의료진과 함께 사용할 때의 효과가 중요하다. 의료진이 먼저 독립적인 판단을 내리고, AI가 일치점과 차이점을 정리하는 방식도 연구되고 있다.

증례 기반 연구에서는 이런 구조가 진단 정확도를 높일 가능성이 나타났다. 그러나 의료진과 사용자의 배경, 정보 제시 방식에 따라 효과가 달라질 수 있다. AI의 제안을 그대로 따르도록 유도해서는 안 된다. 필요한 근거를 비교할 수 있도록 해야 한다.

MAI-DxO의 의의는 여러 역할을 조정해 정보 수집과 진단, 비용 검토를 연결한 데 있다. 앞으로는 실제 진료에서 사용 가능성과 안전성, 검토 부담을 확인해야 한다. 연구 결과를 환자에게 유익한 지원으로 연결하려면 의료진의 역할과 시스템의 권한을 함께 설계해야 한다.

## 10.3 ClinVerse와 AI 병원의 진료 과정

푸단대 등 공동 연구진의 ClinVerse 논문에는 여섯 역할의 상호작용과 평가 기준이 제시되어 있다. 실제 기록에서 가져온 정보와 모의실험을 위해 보충한 정보를 나눠 읽을 수 있다.

<https://arxiv.org/abs/2610.34878>

의료 에이전트를 평가하려면 진단의 정답뿐 아니라 정보를 수집하고 검사를 선택하는 과정도 살펴야 한다.

ClinVerse는 푸단대, 상하이혁신연구원, 저장대, 화중과기대의 공동 연구다. Sirry Chen 등 연구진이 2026년 10월 26일 사전공개 논문을 올렸다. ClinPatient는 평가에 쓸 사례 모음이며, ClinLab은 검사 정보의 빈 부분을 예측하는 구성 요소다. 이 이름들은 각각 별도로 허가받은 의료 제품을 뜻하지 않는다. ClinVerse는 실제 전자의무기록(EHR)을 바탕으로 이러한 상호작용을 모의하는 연구 환경이다. 실제 병원을 운영하는 제품이 아니라, 에이전트의 학습과 평가를 위한 시뮬레이터(simulator)다.

이 환경에서는 에이전트가 초기 증상에서 출발해 환자에게 질문하고 검사를 요청한 뒤 진단을 제시한다. 필요한 정보를 처음부터 모두 제공하는 객관식 평가와 달리, 어떤 정보를 확보해야 할지 선택하는 능력을 검토한다.

## 실제 기록을 바탕으로 한 가상 사례

ClinVerse의 ClinPatient 데이터셋(dataset)은 공개 임상 기록에서 구성한 약 15만 건의 진단 사례를 포함한다. MIMIC-IV와 MIMIC-IV-Note, S1Cdb 등 비식별화(非識別化, de-identified) 자료가 기반이다. 주로 미국과 오스트리아의 입원·중환자 기록이므로 모든 국가와 외래 환자를 대표하지는 않는다.

기록에서 부족한 생활습관과 사회력 등은 가상 환자 생성 도구 Synthea를 이용해 보완했다.

Synthea는 미국 비영리 연구기관 MITRE가 주도한 가상 환자 생성 소프트웨어다. 제이슨 왈로노스키 등 연구진의 도구 논문은 2017년 8월 30일 온라인으로 공개됐고 2018년 학술지 권호에 실렸다. 환자의 삶과 질환 경과를 규칙으로 모의해 기록을 만든다. 언어모델 자체와는 다른 종류의 도구다. 따라서 실제 기록 기반이라는 표현은 모든 정보가 실제로 관찰되었다는 뜻은 아니다. 실제 자료와 합성 정보를 구분해 평가해야 한다.

## 진료 과정의 역할 분담

의사 에이전트는 문진과 검사 요청, 진단을 수행한다. 환자 시뮬레이터는 환자가 알고 설명할 수 있는 정보와 검사 등을 통해 확인할 정보를 나누어 관리한다. 임상병리 역할은 검사 정보를 전달하고, 환경 제어자는 전체 진행을 조정한다. 진단 심판은 대화에 개입하지 않고 결과와 과정을 평가한다.

이러한 역할 구분은 질문이나 검사에 맞는 정보만 제공하기 위한 설계다. 진단을 미리 알 수 있는 정보가 부적절하게 전달되지 않도록 제한한다. 다만 정보 누출을 완전히 방지한다고 보장할 수는 없다.

누락된 검사에는 ClinLab 모듈(module)이 다른 자료를 바탕으로 값을 예측해 제공할 수 있다. 이 값은 실제 측정 결과가 아니라 시뮬레이션의 구성 요소다. 예측 오류가 후속 진단에 영향을 줄 수 있으므로 별도로 검토해야 한다.

### 최종 진단과 과정의 평가

평가는 구조화된 임상 근거와 국제질병분류(ICD-10) 진단명을 이용한다. 필요한 병력을 얼마나 수집했는지, 관련 검사와 이상 소견을 얼마나 확인했는지, 진단의 범주와 세부 내용이 맞는지를 함께 살핀다.

최종 정답률이 비슷해도 정보를 수집하는 방식은 다를 수 있다. 검사를 많이 요청했으나 결과를 진단에 연결하지 못할 수도 있고, 초기 정보에 주로 의존해 결론을 낼 수도 있다. 과정별 지표는 이런 차이를 파악하는 데 도움이 된다.

평가 점수는 학습의 보상 신호(reward signal)로도 사용할 수 있다. 다만 정해진 점수를 높이는 행동이 실제로 좋은 진료와 일치하는지는 계속 검토해야 한다.

## 반복 학습과 기억의 수정

Dynamic Cheatsheet는 스탠퍼드대와 Together AI의 미라치 수즈군 등 연구진이 제안했다. 첫 논문은 2025년 4월 10일 공개됐다. 앞서 풀었던 문제에서 쓸 만한 방법을 글로 정리하고 다음 입력에 함께 넣는다.

ClinVerse가 자체적으로 만든 모델명으로 이해해서는 안 된다. 기존의 기억 관리 방법을 가져와 진단 모의환경에서 시험한 것이다.

ClinVerse 연구는 기존 대화에서 기술과 교훈을 추출하는 방식, 상호작용 중 기억을 갱신하는 방식, 갱신 절차 자체를 조정하는 방식 등을 비교했다. 이를 자기 진화(self-evolution) 또는 자기 개선 접근으로 설명한다.

여러 방법은 모델 가중치(weight)를 직접 바꾸기보다 외부의 기억과 수행 규칙을 수정한다. Dynamic Cheatsheet 같은 방식은 이전 경험을 정리해 다음 수행에 참고한다. 일부 평가에서는 정보 수집 전략과 진단 성능이 개선되었다.

그러나 이러한 결과는 가상 환경에서 나타난 변화다. 검사 정보가 누락된 사실 자체가 진단의 단서가 될 수도 있고, 예측한 검사값이 평가를 왜곡할 수도 있다. 모델 성능뿐 아니라 시뮬레이터의 가정과 한계를 함께 분석해야 한다.

## 실제 업무와 연결할 수 있는 기능

ClinVerse는 실제 병원에 도입된 체계는 아니지만, 문진과 자료 조회, 검사 후보 제안, 감별진단(鑑別診斷) 준비를 연결하는 구조를 연구할 수 있게 한다. 향후 임상 지원에 적용하려면 각 기능이 의료진의 업무에 어떤 도움을 주는지 확인해야 한다.

예를 들어 AI가 병력을 정리하더라도 중요한 증상을 추가로 확인할 수 있어야 한다. 검사 후보를 제안할 때는 기존 결과와 환자 상태를 함께 검토해야 한다. 진단 후보와 근거는 의료진이 확인하고 수정할 수 있는 형태로 제공해야 한다. 이는 적용 방향에 관한 설명이지 ClinVerse가 실제 병원에서 입증한 성과는 아니다.

### 가상 환경 평가의 역할과 한계

예를 들어 환자 역할의 응답이 늘 같은 형식이라면 모델이 의학적 이해보다 그 형식에 익숙해져 점수가 오를 수도 있다. 이를 가리려면 문장을 바꾸거나 다른 자료로 만든 사례에서도 성능을 살펴야 한다. 평가 대상 에이전트와 정보를 제공하는 시뮬레이터가 함께 결과에 영향을 주므로, 둘의 변경 이력을 구분해 남기는 일도 필요하다.

ClinVerse는 다양한 에이전트를 같은 조건에서 반복 평가하고 정보 수집의 문제를 분석하는 데 의미가 있다. 다만 이 원고에서 다루는 연구는 동료 심사 전 사전공개 논문이며, 실제 환자 결과를 평가한 것은 아니다.

실제 진료에는 소음과 불명확한 발언, 기록 오류, 시스템 연결 문제, 의료진의 업무 중단 등 다양한 변수가 있다. 가상 환자와 예측 검사값이 이를 모두 재현할 수는 없다.

AI 병원 연구의 의의는 병원의 모든 판단을 즉시 자동화하는 데 있지 않다. 여러 역할과 정보를 연결하는 방법을 시험하고, 실패가 생기는 과정을 분석하는 데 있다. 실제 적용은 기능별 검증과 의료진의 감독, 권한과 책임의 명확화를 거쳐야 한다.

'의료 슈퍼인텔리전스'는 현재 달성된 기술 수준이 아니라 미래 목표를 나타내는 표현이다.



## Part 6. 미래 의료와 AI



## 제11장. 의사와 AI 에이전트의 협업



## 11.1 AI 선별·인간 검토·종합 진단

산야인민병원과 쓰촨대 관련 연구진의 원격 심전도 논문은 지역 기관과 판독센터의 역할을 설명한다. AI가 표시한 결과가 사람의 검토를 거쳐 전달되는 과정을 확인할 수 있다.

<https://www.frontiersin.org/journals/cardiovascular-medicine/articles/10.3389/fcvm.2026.1869716/full>

의료 AI를 진료에 적용할 때는 어떤 기능을 맡기고 누가 결과를 확인할지 정해야 한다. 높은 정확도를 가진 모델이라도 경보가 너무 많거나 담당자에게 제때 전달되지 않으면 효과가 제한된다. 따라서 모델의 성능과 함께 업무 흐름(workflow)을 설계해야 한다.

AI 선별, 훈련된 담당자의 검토, 의료진의 종합 판단을 연결하는 구조는 하나의 적용 방식이다. AI는 검토할 정보를 우선 분류하고, 담당자는 원자료와 출력의 적절성을 확인하며, 의사는 환자 상태를 함께 고려해 진단과 후속 조치를 결정한다. 구체적인 역할은 제품과 기관에 따라 달라진다.

### 지역 심전도 협력 체계

이 사례는 특정 AI 모델 하나보다 지역 의료기관을 연결한 운영 체계를 다룬다. 2026년에 발표한 연구는 2025년 하반기 도입 이후의 운영을 분석했다. 논문은 핵심 AI를 의료기술 기업이 개발한 상용 솔루션으로 설명했다. 다만 논문은 기업명과 제품 최초 출시일을 밝히지 않는다. 병원의 플랫폼 연구 발표일을 그 AI의 개발일로 적어서는 안 된다.

중국 산야(三亚)의 원격 심전도(tele-ECG) 클라우드 플랫폼은 지역 의료 기관과 상급 병원의 판독 체계를 연결한 사례다. 지역 기관에서 12유도 심

전도(12-lead ECG)를 기록하면 자료를 공유하고 AI와 판독 담당자, 전문의가 단계적으로 검토한다.

클라우드(cloud)는 여러 기관이 자료를 전달하고 처리하는 기반이다. 지역에서 확보한 심전도를 상급 기관이 확인하면 필요한 판단과 이송 준비를 지원할 수 있다. 이 구조의 성과는 지역 의료진 개인의 판독 능력만이 아니라 정보와 전문성을 연결한 결과로 이해해야 한다.

### 1단계: AI 선별과 품질 확인

12유도 심전도는 심장의 전기 신호를 여러 방향에서 기록한 검사다. 입력은 이 파형이며, AI는 학습한 특징과 비교해 위험 가능성을 출력한다. 손이나 몸의 움직임으로 생긴 잡음도 파형에 섞일 수 있다. 위험 표시를 보기 전에 분석할 만한 신호인지 확인하는 과정이 필요한 이유다.

AI는 파형에서 위험 가능성을 표시하고 전극 부착이나 신호 품질도 확인한다. 급성 심근경색과 같은 상황에서는 위험을 놓치지 않도록 선별 기준을 정하는 것이 중요하다. ST분절 상승 심근경색(STEMI, ST-segment elevation myocardial infarction)도 주요 검토 대상이다.

다만 민감도(sensitivity)를 높이는 설정에서는 정상이나 중요도가 낮은 소견이 이상으로 표시될 수 있다.

민감도는 실제 이상이 있는 사람 중 얼마나 잘 찾아냈는지를 뜻한다. 위음성은 이상이 있는데 놓친 경우이고, 위양성은 이상이 없는 사람에게 경보를 낸 경우다. 양성예측도는 경보를 받은 사람 중 실제 이상이 있는 비율이다. 같은 모델도 검사받는 집단의 질환 빈도에 따라 경보의 의미가 달라진다. 위음성률(false negative rate), 위양성률(false positive rate), 양

성예측도(PPV, positive predictive value)를 함께 살펴 경보의 의미를 판단해야 한다.

## 2단계: 원자료 검토

예를 들어 AI 경보가 도착했는데 파형 일부가 끊겨 있다면, 담당자는 측정 오류인지 확인하고 재측정 필요성을 전달할 수 있다. 이 과정 없이 경보만 전문의에게 계속 보내면 검토할 정보가 늘어도 판단은 빨라지지 않을 수 있다. 누가 알림을 받았고 확인했는지 기록하는 기능도 협업의 일부다.

중앙 판독센터의 담당자는 원래 심전도 파형을 확인하고 AI가 표시한 내용을 검토한다. 신호 오류와 과도한 경보를 줄이며 전문의가 검토할 정보를 정리하는 역할이다.

불필요한 알림이 반복되면 중요한 경고에도 반응이 둔해지는 알람 피로(alarm fatigue)가 생길 수 있다. 중간 검토 단계는 이러한 부담을 관리하는 방법이지만, 담당자의 교육과 업무량, 결과 전달 기준이 갖추어져야 한다.

## 3단계: 전문의의 종합 판단

전문의는 심전도뿐 아니라 흉통(胸痛)과 다른 증상, 트로포닌(troponin) 등 검사 결과를 함께 검토한다. 한 신호의 이상만으로 모든 진단을 확정할 수 없기 때문이다.

AI 출력을 보여 주는 시점도 판단에 영향을 줄 수 있다. 먼저 제시된 답에 과도하게 의존하는 앵커링(anchoring)과 자동화 편향(automation

bias)을 줄이려면 독립적인 검토와 근거 비교가 가능해야 한다. 연구의 눈가림 평가 방식과 실제 병원의 일상 절차는 구분해 설계해야 한다.

### 운영에서 보고된 변화

산야 연구는 지역 기관의 진단 과정과 의사결정 시간, 판독 결과 전달 등에서 개선을 보고했다. 검사 자료를 공유하면 상급 병원에서 같은 검사를 반복하는 부담을 줄이고 환자 도착 전 준비를 지원할 가능성이 있다.

다만 도입 전후의 관찰 결과이므로 변화 전체를 AI의 효과로 볼 수는 없다. 표본과 기록의 한계, 계절별 환자 변화, 네트워크 연결 상태도 고려해야 한다. 소수 환자의 생존 결과를 플랫폼의 생존 개선 효과로 확대해서는 안 된다.

### 다른 임상 분야의 협업

대만 사례는 국방의학원과 삼군총의원 연구진이 만든 AI 심전도 경보 체계를 평가했다. 임상시험 논문은 2024년 4월 29일 공개됐다. 이 날짜는 관련 기술 전체의 최초 출시일을 뜻하지 않는다. 심전도에서 위험도를 계산한 뒤 정한 기준을 넘으면 의료진에게 알리고, 의료진이 상태를 확인하는 구조다. 병원이 공개한 안내에서도 실제 임상 연구와 후속 검증의 필요성을 확인할 수 있다.

<https://www.tsgh.ndmutsggh.edu.tw/news/191/10010/24887/10864>

대만의 AI 심전도 경보 연구는 위험 신호를 의료진에게 전달하고 대응하도록 구성했다. 이 사례는 선별 결과가 실제 임상 행동과 연결되는지를 평가했다는 점에서 의미가 있다.

스웨덴의 MASAI 유방암 검진 연구에서는 AI 위험 분류에 따라 판독 방식을 조정했다.

MASAI는 룬드대 등의 연구 이름이며 AI 제품명은 아니다. 연구에 사용한 Transpara는 네덜란드 ScreenPoint Medical의 영상 분석 소프트웨어다. 유방촬영 영상을 입력받아 의심 부위와 점수를 제공한다. 초기 제품의 미국 FDA 510(k) 결정일은 2018년 11월 21일이다. 이는 해당 미국 심사 기록의 날짜이며, 세계 최초 공개일이나 MASAI 연구의 시작일과 같지 않다. AI가 일부 검토 순서와 업무 배분을 지원하고, 의료진이 판독을 담당하는 구조다. 중간암(interval cancer), 즉 정기 검진 사이에 발견되는 암 등 환자 관련 결과도 평가했다.

사람과 AI의 진단을 결합한 증례 연구에서도 상호보완의 가능성이 나타났다. 다만 글로 제시한 사례의 정확도와 실제 진료 성과는 다르다. 각 분야에 맞는 임상 근거를 확인해야 한다.

## 자율 기능과 감독의 범위

IDx-DR는 미국 IDx가 개발한 의료기기이며 이후 Digital Diagnostics의 LumineticsCore라는 이름으로 알려졌다. FDA의 De Novo 허가 결정일은 2018년 4월 11일이다. 눈 안쪽의 망막 사진을 분석해 정해진 범위의 당뇨망막병증 선별 결과를 낸다. 결과를 읽는 데 안과 전문의의 별도 영상 판독이 필요하지 않다는 점과, 환자의 후속 진료가 필요 없다는 말은 다르다. FDA 문서에는 결과에 따른 의뢰와 검사 제한 조건이 명시되어 있다.

[https://www.accessdata.fda.gov/cdrh\\_docs/reviews/DEN180001.pdf](https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf)

좁고 명확한 과제에서 의료진의 별도 판독 없이 결과를 제공하도록 허가된 기기도 있다. 당뇨망막병증 선별 기기 IDx-DR, 현재의 LumineticsCore가 관련 사례다. 따라서 모든 AI 출력에 동일한 형태의 최종 검토가 요구된다고 일반화해서는 안 된다.

기존 처방의 갱신을 제한된 조건에서 다루는 Doctronic 시범사업도 소개되었다.

미국 Doctronic이 운영하는 처방 갱신 도구의 시범사업이다. 유타주 담당 기관이 공개한 현황에는 1단계에서 모든 처방 요청에 면허를 가진 의료인의 승인이 필요하다고 적혀 있다. 따라서 시범사업을 소개할 때도 당시 단계와 감독 조건을 함께 적어야 한다. 이 안내만으로 회사의 모델 최초 공개일을 확정할 수는 없다.

<https://commerce.utah.gov/ai/regulatory-relief-4/authorized-pilots/doctronic/>

그러나 기존 처방의 반복과 새로운 처방, 용량 변경은 서로 다른 업무다. 제한된 지역의 시범 허용을 범용 자율 진료의 허가로 이해해서는 안 된다.

감독 수준은 사용 목적과 위험, 제품의 허가 조건에 맞춰 정해야 한다. 전문적인 판단이 필요한 사례를 의료진에게 넘기는 기준과 시스템이 작동하지 않을 때의 대체 절차도 필요하다.

## 협업 체계의 설계

AI 선별·인간 검토·종합 진단의 구조는 기능별 역할을 명확히 하는 접근이다. 정보가 누락되지 않도록 전달하고, 각 단계에서 오류를 발견하고, 최종 판단에 필요한 임상 맥락을 확보하는 것이 목적이다.

실제 도입에서는 검사와 알림의 증가만을 성과로 보지 않아야 한다. 적절한 검토와 대응이 이루어지는지 확인해야 한다. 처리 시간과 반복 검사, 담당자의 부담, 환자 결과를 함께 평가하고 연결 장애에도 대비해야 한다. 의료 AI의 효과는 개별 모델뿐 아니라 이와 같은 전체 운영 체계에서 결정된다.

## 11.2 의사의 역할과 종합적 임상 판단

의료진과 AI의 협업 방식을 비교한 무작위시험 논문은 독립적인 판단과 의견 비교의 차이를 살핀다. 증례 평가에서 나온 결과를 실제 환자의 치료 성과와 구분해 읽을 수 있다.

<https://www.nature.com/articles/s41746-026-02545-1>

AI가 기록과 정보 검색을 지원하면 의사의 업무 구성도 달라질 수 있다. 검사 결과와 약물 목록을 정리하고 진료 내용을 입력하는 부담이 줄어드는 대신, 제시된 정보를 검토해 환자에게 적절한 결정을 내리는 역할이 중요해진다.

그러나 정보 정리가 자동화된다고 의사의 역할이 판단에만 한정되는 것은 아니다. 문진과 진찰, 의사소통, 치료의 조정은 서로 연결된 업무다. AI를 어디에 적용하느냐에 따라 실제 변화도 달라진다.

### 기록 업무와 진료 시간

엠비언트 AI 스크라이브는 진료 중 주변 대화를 듣고 기록 초안을 만드는 제품군을 가리킨다. 하나의 회사가 만든 단일 모델명이 아니므로 공통 출시일을 정할 수 없다. 말소리를 글로 바꾼 뒤 증상과 진료 계획 등을 항목별로 정리한다. 대화에 없던 내용을 보태거나 부정 표현을 놓칠 수 있어 의료진이 초안을 확인한다.

엠비언트 AI 스크라이브(ambient AI scribe)는 기록 초안을 작성해 의료진의 입력 부담을 줄일 수 있다. 여러 연구에서 문서 작성 시간이나 번아웃

(burnout)의 개선 가능성이 보고되었지만, 제품과 사용 환경에 따라 효과가 달랐다. 근무 외 업무 시간이 뚜렷하게 줄지 않은 연구도 있었다.

따라서 AI가 확보한 시간을 환자와의 대화나 검토에 사용할 수 있는지는 병원의 업무 운영에 달려 있다. 일정과 인력, 다른 행정 업무가 그대로라면 기록 시간 감소만으로 전체 부담이 해결되지는 않는다.

지역 심전도 협력처럼 AI 선별과 담당자 검토를 연결하는 체계에서도 전문의의 역할은 남는다. 전문의는 정리된 소견을 증상과 다른 검사 결과에 연결해 해석한다. 정보 처리의 분담과 임상 판단의 책임을 함께 설계해야 한다.

### 환자의 맥락을 파악하는 역할

예를 들어 정기적인 통원을 권하는 계획이 제시돼도 환자가 먼 곳에 살거나 이동을 도와줄 사람이 없으면 실행이 어렵다. 이런 조건을 기록에서 찾지 못했다면 직접 물어야 한다. AI가 작성한 계획의 문장이 그럴듯한지 확인하는 일에 그치지 않고, 환자가 실제로 따를 수 있는지 검토해야 한다.

전자의무기록(EHR, Electronic Health Record)은 진료에 필요한 많은 정보를 담지만 환자의 상황 전체를 표현하지는 못한다. 치료를 지속할 수 있는 생활 여건, 보호자의 지원, 환자의 이해와 우려 등은 대화를 통해 확인해야 하는 경우가 많다.

의사는 AI가 참고한 정보가 충분하고 최신인지 확인하고, 기록에 없는 조건도 살펴야 한다. 가능한 진단의 우선순위를 검토하면서 불확실성을 설명하고 추가 확인의 필요성을 판단하는 역할도 중요하다.

## 환자와 함께 치료를 선택하는 역할

치료의 적절성은 평균적인 효과만으로 정해지지 않는다. 부작용과 치료 부담, 생활 방식, 환자가 중요하게 여기는 목표가 함께 고려되어야 한다.

공유 의사결정(shared decision-making)은 선택지의 장단점을 설명하고 환자의 선호를 반영해 결정하는 과정이다. AI는 자료를 정리하고 설명을 지원할 수 있지만, 환자가 무엇을 이해하고 어떤 선택을 원하는지는 의료진과의 상호작용에서 확인해야 한다.

## AI 결과를 검토하는 역할

자동화 편향(automation bias)을 줄이려면 제안을 자신의 지식과 환자의 상태에 비추어 검토해야 한다. 근거가 실제 자료와 일치하는지, 중요한 조건이 빠지지 않았는지, 다른 선택지가 있는지 확인하는 과정이다.

검토 방식과 책임은 제품의 기능과 허가 조건, 업무의 위험에 따라 달라진다. 의사가 모든 계산을 반복해야 한다는 뜻은 아니지만, 판단에 필요한 정보를 확인하고 제안을 수정하거나 거부할 수 있어야 한다.

여러 전문과와 AI 도구의 의견이 다를 때 이를 조정하는 역할도 필요하다. 환자의 상태와 치료 목표, 자원과 일정 등을 고려해 우선순위를 정해야 한다. 다중 에이전트가 제시한 합의도 이러한 임상 검토를 자동으로 대신하지는 못한다.

## 협업의 효과와 사용 방식

사람과 AI의 협업을 분석한 연구에서는 효과가 일관되지 않았다. 정리된 증례에서는 도움이 나타나더라도 복잡한 실제 기록에서는 이점이 작을 수 있다. AI 단독 점수가 높아도 사용자가 같은 수준의 결과를 얻는 것은 아니다.

의료진이 독립적으로 판단한 뒤 AI와 일치점과 차이점을 비교하는 방식도 연구되었다. 일부 증례 평가에서는 개선 가능성이 나타났지만, 실제 환자 결과까지 입증된 것으로 보아서는 안 된다.

두 모델의 의견이나 대안 가설을 제시하는 설계 역시 검토를 지원할 수 있다. 반면 정보가 과도하면 부담이 늘어나므로 사용자 경험(UX, user experience)과 실제 업무 흐름을 함께 평가해야 한다. 도구 사용 권한을 주는 것만으로 충분하지 않다.

## 전문 역량의 유지

AI가 멈췄을 때의 대응도 교육에 포함된다. 원자료를 읽는 능력이 있어야 자동 분석의 오류를 발견하고 다른 절차로 업무를 이어 갈 수 있다. 교육에서는 정답을 맞힌 사례만 보여 주기보다, 잘못된 제안을 왜 수정해야 하는지 설명하게 하는 방식이 도움이 될 수 있다. 이는 교육 설계의 제안이며 특정 연구가 입증한 효과 수치는 아니다.

AI가 일부 탐지와 분석을 맡으면 의료진이 해당 업무를 직접 수행하는 빈도가 줄 수 있다. 내시경 분야의 관찰 연구에서는 AI 보조 사용 이후 AI 없이 시행한 검사의 성과 변화를 보고했다. 다만 원인을 단정할 수 없으므로 자동화가 곧 역량 저하를 일으킨다고 일반화해서는 안 된다.

의료기관은 AI 없이도 필요한 판단을 수행할 수 있도록 교육과 점검을 유지해야 한다. 수련 과정에서는 결과를 확인하는 방법뿐 아니라 원자료를 해석하고 오류를 발견하는 능력도 가르쳐야 한다.

## 역할 변화의 평가

의사의 역할 변화는 직무 명칭의 변경보다 실제 업무가 어떻게 재구성되는지로 판단해야 한다. 정보 정리와 검토에 드는 시간, 환자와의 대화, 의사결정의 적절성, 교육과 피로 등을 함께 살필 필요가 있다.

AI가 반복 업무를 줄이고 필요한 근거를 제시하면 의료진의 판단을 지원할 수 있다. 이를 환자에게 유익한 변화로 연결하려면 충분한 검토 여건과 책임의 명확화, 지속적인 교육이 필요하다. 기술 도입 이후의 업무와 환자 결과를 확인하는 것이 역할 변화의 가치를 평가하는 기준이다.



## 제12장. 미래 의료 AI의 기술·윤리 과제



미국 FDA의 변경관리 지침은 AI 의료기기의 수정 계획과 검증을 어떻게 연결하는지 설명한다. 처음 허가받은 뒤에도 안전성과 성능을 확인해야 한다는 관점에서 읽을 자료다.

<https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>

의료 AI는 기록 작성과 영상 분석에서 진료 과정 지원, 신약 후보 탐색, 임상시험 설계로 적용 범위를 넓히고 있다. 여러 도구와 에이전트를 연결하면 복잡한 업무를 처리할 수 있지만, 오류의 영향과 관리 부담도 커진다.

의료 슈퍼인텔리전스(medical superintelligence)는 이러한 발전의 목표를 표현하는 용어다. 특정 시험에서 높은 성능을 보였다는 사실과 실제 의료 전반에서 안전하게 작동한다는 사실은 다르다. 현재의 연구 결과와 미래 전망을 구분하고, 사용 목적에 맞는 근거를 확보해야 한다.

## 환각과 연쇄 오류

예를 들어 기록 요약이 복용을 중단한 약을 현재 복용 약으로 잘못 적었다고 하자. 뒤의 에이전트가 그 요약만 읽으면 틀린 전제를 그대로 이어받는다. 같은 설명을 여러 번 검토해도 원자료를 다시 보지 않으면 오류를 발견하지 못할 수 있다. 각 단계에서 출처와 날짜를 함께 전달해야 하는 이유다.

거대언어모델은 사실과 다른 내용을 자연스럽게 생성할 수 있다. 환각(hallucination)은 잘못된 약물 정보, 근거 없는 진단 설명, 존재하지 않는 문헌 등으로 나타난다. 여러 단계의 작업에서는 앞선 오류가 후속 검색과 판단에 반영될 수 있다.

출처를 제시하는 기능만으로 문제가 해결되지는 않는다. 인용 내용이 실제 원문과 일치하고 현재 환자에게 적용 가능한지 확인해야 한다. 에이전트의 도구 사용과 결과 통합 과정에도 검증 절차가 필요하다.

## 성능 변화와 자료의 대표성

전체 평균이 유지돼도 일부 집단의 오류가 늘 수 있다. 예를 들어 촬영 장비가 바뀐 뒤 특정 기관의 영상에서만 성능이 낮아질 수 있다. 모델을 고치기 전에 입력 자료가 달라졌는지, 실제 질환의 분포가 달라졌는지 살펴야 원인에 맞게 대응할 수 있다.

드리프트(drift)는 환자 구성과 장비, 자료 생성 방식 등의 변화로 모델의 성능이 달라지는 문제다. 처음 평가한 환경에서 좋은 결과를 얻었더라도 이후 같은 성능을 유지한다고 보장할 수 없다.

특정 지역과 인구 집단에 편중된 학습 자료는 다른 환자에 대한 성능 차이를 만들 수 있다. 예측하려는 목표 자체도 검토해야 한다. 의료비 지출처럼 건강 필요를 간접적으로 나타내는 지표는 집단별 의료 접근성의 차이를 반영할 수 있기 때문이다.

성능은 전체 평균뿐 아니라 환자 집단과 기관, 질환별로 살펴야 한다. 변화가 발견되면 재평가와 수정, 사용 범위 조정으로 대응할 수 있어야 한다.

## 생물보안

단백질과 DNA 설계 AI는 신약개발과 생명과학 연구를 지원하지만, 위험한 생물학적 기능을 가진 서열의 생성에 악용될 가능성도 있다. 생물보안(biosecurity)은 모델 개발뿐 아니라 실제 합성 주문과 연구 수행 과정에서 함께 고려해야 한다.

관련 연구는 기존의 서열 심사 도구가 새롭게 생성된 변형을 충분히 인식하지 못할 수 있음을 보여 주었다. 따라서 탐지 체계를 지속적으로 개선하고, 합성 주문의 목적과 고객을 확인하는 절차를 함께 운영해야 한다. 위험을 줄이려면 모델 제공자와 합성 서비스, 연구기관의 확인 절차가 연결되어야 한다.

## 책임과 개인정보 보호

미국 보건복지부의 HIPAA 안내에는 적용 대상과 건강정보의 보호 원칙이 설명되어 있다. 모든 건강 앱이나 모든 나라에 같은 법이 적용되는 것은 아니므로 서비스의 운영 주체와 관할부터 확인해야 한다.

<https://www.hhs.gov/hipaa/for-professionals/privacy/index.html>

오류가 발생했을 때의 책임은 모델 개발사, 제품 제조사, 의료기관, 사용자 등의 역할과 실제 원인에 따라 달라진다. 최종 승인자에게 모든 책임을 부여하는 방식으로 해결하기 어렵다. 계약과 업무 규정에서 권한과 검토 의무, 보고와 대응 절차를 명확히 해야 한다.

개인정보 보호는 자료 수집과 학습, 검색, 기록 보관의 전 과정에서 필요하다. 미국 HIPAA와 유럽 GDPR 등 관할별 법적 요구를 확인하고, 필요한 범위에서 접근과 사용을 제한해야 한다.

연합학습(federated learning)은 자료를 한곳에 모으지 않고 공동 모델을 학습하는 방법이다.

각 기관이 자기 자료로 학습한 결과를 보내면 이를 모아 모델을 갱신한다. 원본을 옮기지 않는다는 장점은 있지만, 전달하는 학습 결과에도 정보의 흔적이 남을 수 있다. 연합학습은 기술 방식의 이름이므로 하나의 회사나 출시일로 설명할 수 없다. 그러나 정보 유출 위험을 완전히 없애지는 않는다. 모델과 학습 과정에 대한 공격 가능성을 평가하고, 기술적 보호와 제도적 통제를 함께 마련해야 한다.

### 의료진의 역량과 감독

AI 의존이 커지면 일부 업무를 직접 수행하는 경험이 줄어들 수 있다. 관찰 연구에서 나타난 역량 변화는 원인을 신중하게 해석해야 하지만, 교육과 점검을 유지할 필요성을 보여 준다.

자동화 편향(automation bias)과 알람 피로(alarm fatigue)를 줄이려면 결과를 확인할 시간과 원자료, 수정 권한이 제공되어야 한다. 의료진이 독립적으로 판단하고 AI의 제안과 비교할 수 있는 설계가 도움이 될 수 있다. 다만 적절한 방식은 업무의 위험과 사용 환경에 따라 평가해야 한다.

자율성도 기능별로 구분해야 한다. 좁은 선별 과제에서 독립적으로 결과를 내는 기기와 여러 질환을 진단하고 새 치료를 결정하는 범용 에이전트는 다르다. 허가 범위와 지역별 시범사업의 조건을 넘어 자율성이 인정된 것으로 설명해서는 안 된다.

## 변경과 생애주기 관리

FDA가 현재 공개한 관련 최종 지침은 2025년 8월판이다. 변경할 내용, 개발·검증·적용 방법, 변경의 영향을 계획에 담도록 권고한다. PCCP는 규제 심사에서 다루는 변경관리 계획을 뜻하며 AI 모델의 이름과 구분된다. 예정한 범위를 넘어 기능을 바꾸는 경우까지 자동으로 허용한다는 뜻은 아니다.

AI는 모델과 검색 자료, 지시문이 바뀌면 출력도 달라질 수 있다. PCCP(Predetermined Change Control Plan)는 계획된 변경의 범위와 검증 방법을 미리 정하는 접근이다. 모든 변경을 자유롭게 허용하는 제도는 아니다.

제품의 사용 목적과 위험에 맞춰 변경 전후의 평가와 사용자 안내, 시판 후 성능 관찰을 연결해야 한다. 관련 법률과 지침은 확정 여부와 적용 시점이 다르므로 실제 도입 시 최신 요구를 확인해야 한다.

## 병원의 관리 체계

예를 들어 AI의 위험 표시를 실제 진료 화면에는 띄우지 않고 별도 저장한 뒤, 의료진의 판단과 나중에 비교할 수 있다. 이런 시험에서도 환자 자료를 처리하므로 접근 권한과 보관 기간을 정해야 한다. 결과를 숨겨 두었다는 이유만으로 모든 연구·보안 절차가 면제되는 것은 아니다.

의료기관은 AI 도입을 기술 부서의 구매 업무에만 맡기기 어렵다. 임상의와 간호사, 정보보안 담당자, 법률·윤리 전문가 등이 기능과 위험, 운영 조건을 함께 검토해야 한다. 환자의 관점과 접근성도 고려해야 한다.

실제 진료에 영향을 주기 전에 현장 자료로 성능을 살피는 사일런트 트라이얼(silent trial)을 검토할 수 있다. 결과를 진료 결정에 사용하지 않고 실제 업무 조건에서 작동을 확인하는 방식이다. 자료 처리와 개인정보 보호, 평가 범위를 정한 뒤 시행해야 한다.

운영 이후에는 오류와 예외 사례를 수집하고, 담당자의 대응과 수정 이력을 관리해야 한다. 안전 문제를 발견했을 때 사용을 제한하거나 중단할 기준도 필요하다.

### 환자 결과 중심의 평가

시간을 절약했다는 결과도 누구의 시간을 어떤 범위에서 잤는지 살펴야 한다. 입력 시간은 줄어도 검토와 수정 시간이 늘 수 있기 때문이다. 진료가 끝난 뒤의 업무까지 포함해 비교하면 환자와 의료진에게 생긴 변화를 더 정확히 이해할 수 있다.

벤치마크는 반복 가능한 비교에 도움이 되지만 실제 환자에게 미치는 영향을 모두 보여 주지는 않는다. 사용자가 정보를 전달하고 답을 이해하는 과정, 의료진의 대응, 병원 시스템의 연결이 결과에 영향을 준다.

AI 심전도 경보와 유방암 검진 등에서는 실제 진료 과정과 환자 관련 결과를 평가한 연구가 진행되었다. 이런 사례처럼 각 사용 목적에 맞는 결과 지표를 설정해야 한다. 위험이 높은 기능에서는 그에 상응하는 임상 근거가 중요하다.

시간 절감과 비용, 진단 정확도, 환자의 접근성과 안전을 함께 살펴야 한다. 연구 자료의 다양성을 높이고 기관 간 협력을 확대하면서도 개인정보 보호와 평가의 독립성을 유지해야 한다.

## 의료 AI 발전의 방향

의료 AI의 가치는 자율성이나 모델 규모만으로 판단할 수 없다. 필요한 정보를 정확히 처리하고 의료진의 판단을 지원하며, 환자가 적절한 진료를 받을 수 있도록 하는지가 중요하다.

기술의 발전과 함께 자료의 품질, 검증 방법, 책임과 보상, 교육과 운영 체계를 개선해야 한다. 연구 단계의 가능성을 실제 의료서비스로 연결하려면 이러한 조건을 지속적으로 확인하는 과정이 필요하다. 환자에게 나타나는 이익과 위험을 기준으로 적용 범위를 정하는 것이 의료 AI 발전의 기본 방향이다.



## 참고문헌

TXP Medical (2025). 生成AIによる音声入力カルテアプリ「SpeechER」が2025年グッドデザイン賞を受賞. 회사 보도자료, 2025-10-15. <https://txpmedical.jp/en/news/2jmiSPoM4bpIdI9GUZFoBI> (기업 발표: "최대 70% 삭감(자사 조사)")

SIND (연도 미상, 2026-10-28 열람). Caretomo 看護記録作成AI支援システム 제품 누리집. <https://caretomo.jp/> (기업 발표: 병원별 시험 사용 결과)

Xu, J., et al. (2026). AI-Enabled regional tele-ECG cloud platform and improving access to cardiovascular diagnosis: real-world evidence from southern China. *Frontiers in Cardiovascular Medicine*, 13, 1869716.  
<https://doi.org/10.3389/fcvm.2026.1869716> (관찰 연구, 실제 운영)

Lukac, P. J., Turner, W., Vangala, S., et al. (2025). Ambient AI scribes in clinical practice: a randomized trial. *NEJM AI*, 2(12). 온라인 2025-11-26.  
<https://doi.org/10.1056/Aloa2501000> (무작위 배정 임상시험, 단일 기관)

Rotenstein, L. S., Holmgren, A. J., Thombley, R., et al. (2026). Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes: a multisite study. *JAMA*, 335(16), 1408-1417.  
<https://doi.org/10.1001/jama.2026.2253> (관찰 연구, 5개 기관)

Cleveland Clinic Consult QD (2025). Less Typing, More Talking: How Ambient AI Is Reshaping Clinical Workflow at Cleveland Clinic. 2025-08-14.  
<https://consultqd.clevelandclinic.org/less-typing-more-talking-how-ambient-ai-is-reshaping-clinical-workflow-at-cleveland-clinic> (기관 자체 보고)

Booth, S. E., Gould, R., et al. (2026). Impact of unrestricted artificial intelligence scribe deployment in a single health system on clinician workload and productivity: a quality improvement study. *Annals of Internal Medicine*, 온라인 섹션 개 2026-10-27. <https://doi.org/10.7326/ANNALS-25-05294> (관찰 연구, 단일 체계)

Frieden, J. (2026). AI scribes improved note-taking time but not productivity, study finds. *MedPage Today*, 2026-10-26.

<https://www.medpagetoday.com/practicemanagement/informationtechnology/123591> (언론 보도, 수치 보조 확인용)

Tian, J., Song, Y., Tang, L., An, L., & Zhou, J. (2026). Clinical artificial intelligence implementation in routine care: real-world operational outcomes from a provincial health system in China. *Risk Management and Healthcare Policy*, 19, 1-11. <https://doi.org/10.2147/RMHP.S604289> (관찰 연구, 단일 기관)

Rabotin, A., Aguilar, E., Sandoval Gonzalez, S., Iniguez, A., Jung, C., Lee, E., Bell, D. S., & Arroyo, J. (2026). Propagation of interpreter errors by ambient AI scribes: study using simulated clinical encounters. *JMIR Medical Informatics*, 14, e88734. <https://doi.org/10.2196/88734> (모의 진료 실험)

Olson, K. D., Meeker, D., Troup, M., et al. (2025). Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Network Open*, 8(10), e2534976. <https://doi.org/10.1001/jamanetworkopen.2025.34976> (관찰 연구, 6개 의료 시스템, 사전·사후 설문)

Rotenstein, L. S., Holmgren, A. J., Thombley, R., et al. (2026). Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes: a multisite study. *JAMA*, 335(16), 1408-1417. <https://doi.org/10.1001/jama.2026.2253> (관찰 연구, 5개 학술의료기관)

Samraik, M. (2025). AI scribes reduce physician burnout and return focus to the patient. *Yale School of Medicine News*, 2025-10-17. <https://medicine.yale.edu/news-article/ai-scribes-reduce-physician-burnout-return-focus-to-the-patient/> (기관 자체 보고)

Merlino, A., Blue, A., Boose, E., Meese, B., Owens, J., Vickers, J., Angeloff, A., Arnold, M., Chen, R., Grade, M., Pham-Viernickel, P. A., & Rosner, B. (2026). Accelerating ambient AI scribe enterprise-scale deployment: Cleveland Clinic's novel approach to health system-industry partnership. *npj Health Systems*, 3, 74. <https://doi.org/10.1038/s44401-026-00144-6> (실제 진료 운영 보고, 대조군 없음, 공저자 일부 회사 소속)

Fierce Healthcare (2025). Cleveland Clinic taps Ambience for ambient AI tech. 2025-02-19. <https://www.fiercehealthcare.com/ai-and-machine-learning/cleveland-clinic-chooses-ambience-ai-scribe> (언론 보도, 기관 발표 인용)

ASCO AI in Oncology (2026). Measuring Ambient AI Adoption at Scale. 2026-10. <https://ascoai.org/articles/2026/10/measuring-ambient-ai-adoption-at-scale> (전문 학회 매체의 해설)

Abridge (2026). The Evolution of Enterprise AI at Sutter: A Timeline. 2026-08-27. <https://www.abridge.com/blog/sutter-health-partnership-timeline> (기업 발표)

Olson, K. D., Meeker, D., Troup, M., et al. (2025). Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Network Open, 8(10), e2534976. <https://doi.org/10.1001/jamanetworkopen.2025.34976> (<https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2839542>) (관찰 연구, 대조군 없는 전후 비교, 자가 보고. 표 2·3 확인: 분석 참가 263명, 번아웃 모형 186명, 보정 다변량 모형 184명)

Afshar, M., et al. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI, 2(12). <https://doi.org/10.1056/Aloa2500945> (ClinicalTrials.gov NCT06517082) (무작위 배정 임상시험, 위스콘신 대학교 외래, 66명, 24주)

Rotenstein, L. S., Holmgren, A. J., Thombley, R., et al. (2026). Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes: a multisite study. JAMA, 335(16), 1408-1417. <https://doi.org/10.1001/jama.2026.2253> (관찰 연구)

Booth, S. E., Gould, R., et al. (2026). Impact of unrestricted artificial intelligence scribe deployment in a single health system on clinician workload and productivity: a quality improvement study. Annals of Internal Medicine, 온라인 선공개 2026-10-27. <https://doi.org/10.7326/ANNALS-25-05294> (관찰 연구)

Frieden, J. (2026). AI scribes improved note-taking time but not productivity, study finds. MedPage Today, 2026-10-26. <https://www.medpagetoday.com/practicemanagement/informationtechnology/123591> (언론 보도)

Wojda, T., Berliner, J., Fischer, G. S., Curren, M., Arnold, J., & Giugale, L. (2026). From burden to balance: trends in documentation workload with ambient artificial intelligence scribe adoption. *Applied Clinical Informatics*, 17(3), 484-492. <https://doi.org/10.1055/a-2892-7988> (관찰 연구, 반복 횟단 설문, 기저 1,600명, 6개월 101명)

De Vito, A., Geremia, N., Bavaro, D. F., et al. (2025). Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better? *Clinical Microbiology and Infection*, 31(8), 1336-1342. <https://doi.org/10.1016/j.cmi.2025.03.002> (벤치마크, 사례 기반 비교)

Russo, A., Geremia, N., Bavaro, D. F., De Vito, A., et al. (2026). Comparing large language models for antibiotic prescribing: an updated analysis using the same clinical benchmark. *Clinical Microbiology and Infection*, 32(10), 1733-1735. 온라인 2026-04-16. <https://doi.org/10.1016/j.cmi.2026.04.012>

De Vito, A., Geremia, N., Russo, A., Colpani, A., Vaira, L. A., & Madeddu, G. (2026). Reproducibility of AI-generated antibiotic recommendations in standardised clinical scenarios: a proof-of-concept experimental study. *International Journal of Antimicrobial Agents*, 67(10), 107881. <https://doi.org/10.1016/j.ijantimicag.2026.107881> (벤치마크, 반복 질의 실험)

Zhang, Y., Li, L., Tan, C., Ji, M., Tian, X., Mu, X., Li, J., Gu, Y., & Liu, H. (2026). Large language model-based clinical decision support for antibiotic selection and dose recommendation in hospitalized patients with pneumonia: multicenter retrospective study. *JMIR Medical Informatics*, 14, e98207. <https://doi.org/10.2196/98207> (관찰 연구, 후향적 2개 기관, 실제 진료 적용 전)

(배경 지식) Shortliffe, E. H. (1976). *Computer-Based Medical Consultations: MYCIN*. New York: Elsevier. (역사적 배경, 1976년 기초 출처)

云知声 Unisound (2026). 登顶 OpenAI HealthBench 双榜！云知声医疗模型性能全面超越 GPT-6 Astra、GPT-6.1 Sol，构建高水平 AI 医生问诊能力. 공식 발표, 2026-10-21. <https://www.unisound.com/news/165> (영문판 <https://www.unisound.com/en/news/165>) (기업 발표, 벤치마크 자체 보고)

Arora, R. K., et al. (2025). HealthBench: Evaluating large language models towards improved human health. arXiv:2505.08775.

<https://doi.org/10.48550/arXiv.2505.08775> (벤치마크 설계, 2025년 기초 출처)

Bean, A. M., Payne, R. E., Parsons, G., et al. (2026). Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, 32, 609-615. <https://doi.org/10.1038/s41591-025-04074-y> (무작위 배정 연구, 추가 조사로 확인) (2026년 4월 17일 출판사 정정 공지: *Nature Medicine* 32(5):1937, <https://doi.org/10.1038/s41591-026-04404-8>. 정정 내용이 본문 수치에 영향을 주었는지는 확인하지 못함)

Nori, H., Daswani, M., Kelly, C., et al. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423-428. <https://doi.org/10.1126/science.aeh7930> (벤치마크, 추가 조사로 서지·초록 확인)

Microsoft AI (2026). Building Medical Superintelligence. 2026-10-22.

<https://microsoft.ai/news/building-medical-superintelligence> (기업 발표)

Chen, S., Chen, F., Ding, X., Chen, W., & Wei, Z. (2026). ClinVerse: a realistic multi-agent diagnostic environment for training and evaluating medical agents. arXiv:2610.34878. <https://doi.org/10.48550/arXiv.2610.34878> (시뮬레이션 연구, 사전 공개)

Liévin, V., Palepu, A., Weng, W.-H., Saab, K., Stutz, D., Cheng, Y., Kulkarni, K., Mahdavi, S. S., et al., Karthikesalingam, A., & Schaekermann, M. (2026). Towards conversational artificial intelligence for disease management. *Nature*, 655, 1292-1299. <https://doi.org/10.1038/s41586-026-10764-5> (시뮬레이션 연구, 환자 연기자, 무작위·눈가림)

Tu, T., Schaekermann, M., Palepu, A., et al. (2025). Towards conversational diagnostic artificial intelligence. *Nature*, 642, 442-450.

<https://doi.org/10.1038/s41586-025-08866-7> (시뮬레이션 연구, 2025년 기초 출처)

Google (2026). Google advances its AMIE research medical AI from diagnosis to treatment. Google Blog, 2026-06-17. <https://blog.google/innovation-and-ai/models-and-research/google-research/amie-for-disease-management-in-nature> (기업 발표)

Ferber, D., Hilgers, L., Höper, C., Kinny-Köster, B., Eckardt, J.-N., et al., Truhn, D., Jäger, D., & Kather, J. N. (2026). Towards autonomous medical artificial intelligence agents. *Nature*, 655(8125), 1282-1291.

<https://doi.org/10.1038/s41586-026-10675-5> (시뮬레이션 연구, MIMIC-IV 실제 환자 기록 기반)

Brodeur, P., Koshy, J. M., Palepu, A., Saab, K., et al., Schaekermann, M., Karthikesalingam, A., & Rodman, A. (2026). A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic. *The Lancet*, 2026-10-08 온라인. [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(26\)01535-7/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(26)01535-7/fulltext) (사전공개본: A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic. arXiv:2603.08448) (전향적 연구, 실제 진료, 감독하)

ARISE Network (2026). ARISE, BIDMC and Google AMIE publish prospective clinical AI feasibility study in *The Lancet*. 2026-10-08.

<https://www.arise-ai.org/blog/amie-lancet> (연구 참여 기관 해설)

Nature Medicine (2026). Agents AMIE and MIRA advance medical AI capabilities. 2026-06-30. <https://www.nature.com/articles/d41591-026-00034-2> (학술지 뉴스, 제목과 존재만 확인)

Beth Israel Deaconess Medical Center (2026). BIDMC researchers conduct first real-world study of safety and quality of patient-facing AI in primary care. 2026-10-09. <https://bidmc.org/news-stories/all-news-stories/news/2026/10/bidmc-research-shows-safety-quality-of-ai-in-primary-care> (기관 보도자료)

Brix, G., Durrant, M. G., Ku, J., et al. (2026). Genome modelling and design across all domains of life with Evo 2. *Nature*, 652(8112), 1349-1361.

<https://doi.org/10.1038/s41586-026-10176-5> (동료심사 논문)

Avsec, Ž., Latysheva, N., Cheng, J., Novati, G., et al. (2026). Advancing regulatory variant effect prediction with AlphaGenome. *Nature*. 649(8099), 1206-1218.

<https://doi.org/10.1038/s41586-025-10014-0> (동료심사 논문)

Wang, D., Zhang, J., Chen, C., Zhang, W., Wang, S., Meng, Y., Sonpavde, G., Horbinski, C., & Tian, Y. (2026). A multimodal foundation model linking

histopathology and DNA methylation. bioRxiv.

<https://doi.org/10.64898/2026.06.11.731518> (프리프린트, 2026년 6월 14일 게시)

Li, Q., Sang, J., Xiao, Y., Zhang, P., Chen, W., Tran, T. N., Li, Z., Li, S., Zhang, T., & Wang, G. (2026). Genome-scale molecular profiling from routine histopathology with a multimodal pathology foundation model. Research Square.

<https://doi.org/10.21203/rs.3.rs-9949149/v1> (프리프린트, PMID 42620258)

Chen, W., et al. (2025). A visual-omics foundation model to bridge histopathology with spatial transcriptomics. Nature Methods, 22, 1568-1582.

<https://doi.org/10.1038/s41592-025-02707-1> (동료심사 논문)

Chaves, J. M. Z., et al. (2024). Tx-LLM: A large language model for therapeutics. arXiv:2406.06316. <https://arxiv.org/abs/2406.06316> (프리프린트)

Wang, E., Schmidgall, S., et al. (2025). TxGemma: Efficient and agentic LLMs for therapeutics. arXiv:2504.06196. <https://arxiv.org/abs/2504.06196> (프리프린트)

Sun, Q., Wu, X., Chen, J., Zhang, Y., Zhou, C., & Zhu, Z. (2026). Cross-attention guided multi-modal network for breast ultrasound diagnosis incorporating objective clinical semantics. Frontiers in Medicine, 13, 1802570.

<https://doi.org/10.3389/fmed.2026.1802570> (동료심사 논문)

Gu, Z., Wang, C., Zhu, Q., Wang, W., et al. (2026). Multimodal large language models for prognostic prediction in cervical cancer treated with definitive chemoradiotherapy: An exploratory study of systematic multimodal data integration. Cancers, 18(19), 3136. <https://doi.org/10.3390/cancers18193136> (동료심사 논문, 세부 AUC 수치 미대조)

Souza, H., & Mehta, P. (2026). Parameter-free representations outperform single-cell foundation models on downstream benchmarks. arXiv:2602.16696.

<https://arxiv.org/abs/2602.16696> (프리프린트)

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., & Wang, B. (2024). scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods, 21(8), 1470-1480. <https://doi.org/10.1038/s41592-024-02201-0>

(동료심사 논문)

Ahlmann-Eltze, C., Huber, W., & Anders, S. (2025). Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8), 1657-1661. <https://doi.org/10.1038/s41592-025-02772-6> (동료심사 논문)

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., & Zou, J. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems (NeurIPS 2022)*. <https://arxiv.org/abs/2203.02053> (동료심사 학회 논문, 오래된 기초 문헌)

Baker, C., Ren, T., Rafferty, K., Wang, H., & McDade, S. (2026). Autonomous computational prioritisation of colorectal cancer vulnerabilities via multi-scale AI swarms. *bioRxiv*. <https://doi.org/10.64898/2026.07.05.736565>

Baker, C., Ren, T., Rafferty, K., & Wang, H. (2026). Contextual invertible world models: A neuro-symbolic agentic framework for colorectal cancer drug response. *arXiv:2603.02274*. <https://arxiv.org/abs/2603.02274> (프리프린트, v3 2026년 6월 10일)

Cui, H., Cruz-Correa, M., Giardiello, F. M., et al. (2003). Loss of IGF2 imprinting: A potential marker of colorectal cancer risk. *Science*, 299(5613), 1753-1755. <https://doi.org/10.1126/science.1080902> (동료심사 논문, 오래된 기초 문헌)

Waddington, C. H. (1957). *The strategy of the genes: A discussion of some aspects of theoretical biology*. George Allen & Unwin. (단행본, 오래된 기초 문헌)

Qiu, M., Zheng, G., Xu, Y., Zhang, R., Ding, Y., Long, Q., & Chen, T. (2026). Chreode: A cell world model for one-step temporal dynamics and perturbation prediction. *arXiv:2605.28111*. <https://doi.org/10.48550/arXiv.2605.28111> (프리프린트, 2026년 5월 27일)

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D., & Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479), eaaw3381. <https://doi.org/10.1126/science.aaw3381> (동료심사 논문, 기초 데이터셋)

Norman, T. M., Horlbeck, M. A., Replogle, J. M., et al. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455), 786-793. <https://doi.org/10.1126/science.aax4438> (동료심사 논문, 기초 데이터셋)

Roohani, Y., Huang, K., & Leskovec, J. (2024). Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42, 927-935. <https://doi.org/10.1038/s41587-023-01905-6> (동료심사 논문)

Chen, S., Yu, L., Lin, X., Mao, X., Zhang, S., Gu, X., Wu, H., Xu, S., Jin, K., Bai, L., Qian, Q., Chen, Q., Gao, Q., Sun, S., & Gao, Z. (2026). SCALE: Scalable conditional atlas-level endpoint transport for virtual cell perturbation prediction. *arXiv:2603.17380*. <https://doi.org/10.48550/arXiv.2603.17380> (프리프린트, v3 2026년 8월 31일)

Zhang, J., et al. (2025). Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*. <https://doi.org/10.1101/2025.02.20.639398> (프리프린트)

Di Francesco, A. G., et al. (2026). Differentiable retrieval-augmented generation for predicting cellular responses to gene perturbation (PT-RAG). *arXiv:2603.07233*. <https://doi.org/10.48550/arXiv.2603.07233> (프리프린트, v3 2026년 10월 9일. 공저자 전체는 이 책의 조사에서 미대조)

Roohani, Y. H., et al. (2025). Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell*. [https://www.cell.com/cell/fulltext/S0092-8674\(25\)00675-0](https://www.cell.com/cell/fulltext/S0092-8674(25)00675-0) (동료심사 논평 논문, 권·쪽 미대조)

Arc Institute. (2025, December). Virtual Cell Challenge 2025 wrap-up: Winners and reflections. <https://arcinstitute.org/news/virtual-cell-challenge-2025-wrap-up> (기관 발표)

Arc Institute. (2026, August 20). The 2026 Virtual Cell Challenge: Predicting perturbation responses in cell contexts a model has never seen. <https://arcinstitute.org/news/virtual-cell-challenge-2026> (기관 발표). 관련 논문: Virtual Cell Challenge 2026: Benchmarking zero-shot generalization across cellular contexts. *Cell* (2026), PMID 42648290 (저자 정보 미대조)

Virtual Cell Challenge. (2026). Virtual Cell Challenge 2026: Timeline and rules. <https://virtualcellchallenge.org/> (2026년 10월 22일 최종 시험셋 공개, 11월 5일 마감 확인)

Lamb, J., Crawford, E. D., Peck, D., et al. (2006). The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795), 1929-1935. <https://doi.org/10.1126/science.1132939> (동료심사 논문, 오래된 기초 문헌)

Subramanian, A., Narayan, R., Corsello, S. M., et al. (2017). A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6), 1437-1452. <https://doi.org/10.1016/j.cell.2017.10.049> (동료심사 논문, 오래된 기초 문헌)

Xu, Z., Zhang, Z., Wang, L., Liu, Z., Liu, Q., Wu, S., & Wang, L. (2026). Reading the cell, designing the cure: Perturbation-conditioned molecular diffusion for function-oriented drug design. arXiv:2605.15243. <https://doi.org/10.48550/arXiv.2605.15243> (프리프린트, 2026년 5월 14일)

Yuan, H., Li, C., Ma, W., Murata, T., & Jiang, Y. (2026). Gene expression-informed jointly controlled generative modeling for precision molecular design. arXiv:2607.11978. <https://arxiv.org/abs/2607.11978> (프리프린트, 2026년 7월 13일)

Miller, H. E., et al. (2026). Deep learning perturbation models can outperform baselines on calibrated metrics. *Nature Biotechnology*. <https://doi.org/10.1038/s41587-026-03307-w>

Sadée, C., Testa, S., Barba, T., Hartmann, K., Schuessler, M., Thieme, A., Church, G. M., Okoye, I., Hernandez-Boussard, T., Hood, L., Shmulevich, I., Kuhl, E., & Gevaert, O. (2025). Medical digital twins: Enabling precision medicine and medical artificial intelligence. *The Lancet Digital Health*, 7, 100864. <https://doi.org/10.1016/j.landig.2025.02.004> (동료심사 정책 논문)

National Academies of Sciences, Engineering, and Medicine. (2024). Foundational research gaps and future directions for digital twins. The National Academies Press. <https://www.ncbi.nlm.nih.gov/books/NBK605512> (기관 보고서, 세부 문구 미대조)

Sel, K., Hawkins-Daarud, A., Chaudhuri, A., Osman, D., Bahai, A., Paydarfar, D., Willcox, K., Chung, C., & Jafari, R. (2025). Survey and perspective on verification, validation, and uncertainty quantification of digital twins for precision medicine. *npj Digital Medicine*, 8, 40. <https://doi.org/10.1038/s41746-025-01447-y> (동료심사 논문)

Garcia Requesens, M., Mankan, A. K., Cantini, L., Calabresi, M., Vidal, L., Brown, J., Lang, S., Kaura, B., Veselkov, K., Xenarios, I., & Saini, K. S. (2026). Role of health digital twins in oncology drug development: A primer. *Frontiers in Oncology*, 16, 1885797. <https://doi.org/10.3389/fonc.2026.1885797> (동료심사 리뷰, 2026년 8월 24일)

Blasiak, A., Truong, A. T. L., Foo, N., Tan, L. W. J., Kumar, K. S., Tan, S. B., Teo, C. B., Tan, B. K. J., Tadeo, X., Tan, H. L., Chee, C. E., Yong, W. P., Ho, D., & Sundar, R. (2025). Personalized dose selection platform for patients with solid tumors in the PRECISE CURATE.AI feasibility trial. *npj Precision Oncology*, 9, 49. <https://doi.org/10.1038/s41698-025-00835-7> (동료심사 논문, 타당성 시험)

The Cleveland Clinic. (2022-2026). Randomized controlled trial of digital twin precision treatment: A novel whole body digital twin enabled precision treatment for type 2 diabetes (NCT05181449). *ClinicalTrials.gov*. <https://clinicaltrials.gov/study/NCT05181449> (임상시험 등록 정보)

U.S. Food and Drug Administration. (2023). Assessing the credibility of computational modeling and simulation in medical device submissions: Guidance for industry and FDA staff. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents> (규제 문서, 2023년 11월 최종 지침)

임일곤. (2026, 7월 23일). 오프리메드, 치매학술대회서 디지털 트윈 플랫폼 선택. *비즈니스워치*. <https://news.bizwatch.co.kr/article/healthcare/2026/07/23/0014> (보도)

권상희. (2026, 7월 28일). 오프리메드, 춘천 기업혁신파크에 디지털 트윈 임상시험 거점 추진. *전자신문*. <https://www.etnews.com/20260728000025> (보도)

European Medicines Agency. (2022). Qualification opinion for Prognostic Covariate Adjustment (PROCOVA™). <https://www.ema.europa.eu/en/documents/>

[regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm\\_en.pdf](#) (규제 문서)

U.S. Food and Drug Administration & European Medicines Agency. (2026, January). Guiding principles of good AI practice in drug development. <https://www.fda.gov/media/189581/download> (규제 문서)

U.S. Food and Drug Administration. (2025, January). Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products: Draft guidance for industry. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents> (규제 문서, 초안. 2026년 8월 기준 초안 상태라는 해설: CASRAI Editorial Board. (2026). Digital twins in clinical trials: Synthetic control arms and FDA's evolving posture. <https://casrai.org/guides/digital-twins-clinical-trials>)

Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., Pawar, A., Bessette, L. G., Lee, H., Garry, E. M., Gautam, N., & Schneeweiss, S. (2021). Emulating randomized clinical trials with nonrandomized real-world evidence studies: First results from the RCT DUPLICATE initiative. *Circulation*, 143(10), 1002-1013. <https://doi.org/10.1161/CIRCULATIONAHA.120.051718> (동료심사 논문, 오래된 핵심 문헌)

Biotechnology Innovation Organization, Informa Pharma Intelligence, & QLS Advisors. (2021). Clinical development success rates and contributing factors 2011-2020. <https://www.bio.org/clinical-development-success-rates-and-contributing-factors-2011-2020> (산업 보고서, 2021년 2월, 오래된 기초 자료. 12,728건 단계 전환 분석, 1상 기준 허가 가능성 7.9% 확인)

Olmos, L. (2026, October 8). Built on regulatory rigor: Advancing digital twins in clinical trials. Unlearn. <https://www.unlearn.ai/blog/built-on-regulatory-rigor-advancing-digital-twins-in-clinical-trials> (기업 발표)

U.S. Food and Drug Administration. (2023, May). Adjusting for covariates in randomized clinical trials for drugs and biological products: Guidance for industry. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents> (규제 문서)

CASRAI Editorial Board. (2026). Digital twins in clinical trials: Synthetic control arms and FDA's evolving posture (last updated 2026-08-25).

<https://casrai.org/guides/digital-twins-clinical-trials> (해설 자료)

International Council for Harmonisation. (2000). ICH E10: Choice of control group and related issues in clinical trials. <https://www.ich.org/page/efficacy-guidelines> (규제 문서, 오래된 기초 문헌)

Jahanshahi, M., Gregg, K., Davis, G., Ndu, A., Franolic, T., et al. (2021). The use of external controls in FDA regulatory decision making. *Therapeutic Innovation & Regulatory Science*, 55, 1019-1035. <https://doi.org/10.1007/s43441-021-00302-y> (동료심사 리뷰, 2000-2019년 FDA 승인 결정 분석)

U.S. Food and Drug Administration. (2023, February). Considerations for the design and conduct of externally controlled trials for drug and biological products: Draft guidance for industry. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents> (규제 문서, 초안)

European Medicines Agency. (2024). Concept paper on the development of a reflection paper on the use of external controls for evidence generation in regulatory decision-making. <https://www.ema.europa.eu/en/development-reflection-paper-use-external-controls-evidence-generation-regulatory-decision-making-scientific-guideline> (규제 문서)

Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., Pawar, A., Bessette, L. G., Lee, H., Garry, E. M., Gautam, N., & Schneeweiss, S. (2021). Emulating randomized clinical trials with nonrandomized real-world evidence studies: First results from the RCT DUPLICATE initiative. *Circulation*, 143(10), 1002-1013. <https://doi.org/10.1161/CIRCULATIONAHA.120.051718> (동료심사 논문)

Wang, S. V., Schneeweiss, S., & RCT-DUPLICATE Initiative. (2023). Emulation of randomized clinical trials with nonrandomized database analyses: Results of 32 clinical trials. *JAMA*, 329(16), 1376-1385. <https://doi.org/10.1001/jama.2023.4221> (동료심사 논문)

Garcia Requesens, M., Mankan, A. K., Cantini, L., et al. (2026). Role of health digital twins in oncology drug development: A primer. *Frontiers in Oncology*, 16, 1885797. <https://doi.org/10.3389/fonc.2026.1885797> (동료심사 리뷰)

Scannell, J. W., Blanckley, A., Boldon, H., & Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11(3), 191-200. <https://doi.org/10.1038/nrd3681> (동료심사 논문, 오래된 기초 문헌)

U.S. Food and Drug Administration. (2026). New Approach Methodologies (NAMs) (content current as of 2026-09-21); General considerations for the use of new approach methodologies in drug development: Draft guidance (CDER, 2026). <https://www.fda.gov/science-research/science-and-research-special-topics/new-approach-methodologies-nams> (규제 기관 자료)

U.S. Congress. (2022). FDA Modernization Act 2.0, enacted as part of the Consolidated Appropriations Act, 2023, Public Law 117-328. <https://www.congress.gov/> (법률, 조항 번호는 이 책의 조사에서 미대조)

U.S. Congress. (2025-2026). H.R. 2821 - FDA Modernization Act 3.0, 119th Congress (passed House 2026-07-20; received in Senate 2026-07-21). <https://www.congress.gov/bill/119th-congress/house-bill/2821/all-info> (의회 기록)

U.S. Food and Drug Administration. (2025, April). Roadmap to reducing animal testing in preclinical safety studies. <https://www.fda.gov/media/186092/download> (규제 문서)

U.S. Food and Drug Administration. (2026, April). Reducing animal testing in nonclinical studies: Year one progress and the path forward; FDA achieves year 1 goals in reducing animal testing in drug development (press release, 2026-04-20). <https://www.fda.gov/media/191986/download> (규제 문서)

U.S. Food and Drug Administration. (2026, September 21). FDA updates regulations to advance innovative alternatives to animal testing (press release); Nonclinical testing terminology: Direct final rule, Federal Register document 2026-19350. <https://www.fda.gov/news-events/press-announcements/fda-updates-regulations-advance-innovative-alternatives-animal-testing> (규제 문서)

U.S. Food and Drug Administration. (2026, June). M15 general principles for model-informed drug development: Guidance for industry (ICH, docket FDA-2024-D-5580). <https://www.fda.gov/media/184747/download> (규제 문서, 최종 지침)

U.S. Food and Drug Administration. (2018-). Model-Informed Drug Development Paired Meeting Program. <https://www.fda.gov/> (FDA 규제 프로그램 안내. 정확한 페이지 주소와 세부 문구는 이 책의 조사에서 미대조)

International Council for Harmonisation. (2026, February 9). Harmonised ICH M15 guideline, general principles for model-informed drug development, adopted. <https://www.ich.org/news/harmonised-ich-m15-guideline-general-principles-model-informed-drug-development-adopted> (규제 기관 공지. 4단계 최종본: [https://database.ich.org/sites/default/files/ICH\\_M15\\_Step4\\_Final\\_Guideline\\_2026\\_0129.pdf](https://database.ich.org/sites/default/files/ICH_M15_Step4_Final_Guideline_2026_0129.pdf))

U.S. Food and Drug Administration. (2026). Webinar for stakeholder engagement on the new final guideline on general principles for model-informed drug development (ICH M15), 10/29/2026. <https://www.fda.gov/drugs/news-events-human-drugs/webinar-stakeholder-engagement-new-final-guideline-general-principles-model-informed-drug> (규제 기관 공지)

Ren, F., Aliper, A., Chen, J., et al. (2024). A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nature Biotechnology*, 43(1), 63-75. <https://doi.org/10.1038/s41587-024-02143-0> (동료심사 논문, 온라인 2024-03-08)

Xu, Z., Ren, F., Wang, P., et al. (2025). A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: A randomized phase 2a trial. *Nature Medicine*, 31(8), 2602-2610. <https://doi.org/10.1038/s41591-025-03743-2> (동료심사 논문. 일부 2차 자료에 나오는 DOI 10.1038/s41591-025-03120-w는 이 논문의 DOI가 아니다)

Insilico Medicine. (2026, July 7). Insilico initiates Phase III clinical trial for rentosertib, its AI-empowered TNIK inhibitor for idiopathic pulmonary fibrosis. <https://insilico.com/news/xmjnsn4l091-insilico-initiates-phase-iii-clinical-tr> (기업 발표)

InSilico Medicine Hong Kong Limited. (2026). A prospective phase 3, randomized, double-blind, placebo-controlled, parallel-group study evaluating the efficacy and safety of rentosertib (INS018\_055) administered orally over 52 weeks in patients with idiopathic pulmonary fibrosis (NCT07687459). ClinicalTrials.gov. <https://clinicaltrials.gov/study/NCT07687459> (임상시험 등록 정보, 2026-10-13 갱신본 확인)

Jayatunga, M. K. P., Ayers, M., Bruens, L., Jayanth, D., & Meier, C. (2024). How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons. Drug Discovery Today, 29(6), 104009. <https://doi.org/10.1016/j.drudis.2024.104009> (동료심사 논문)

Sahni, N. R., & Carrus, B. (2023). Artificial Intelligence in U.S. Health Care Delivery. New England Journal of Medicine, 389(4), 348-358. DOI: 10.1056/NEJMr2204673

VoiceCare AI (2025, February 14). Agentic AI Startup VoiceCare AI Launches to Automate Healthcare Back Office and Super-Staff Workforce. PR Newswire. <https://www.prnewswire.com/news-releases/agentic-ai-startup-voicecare-ai-launches-to-automate-healthcare-back-office-and-super-staff-workforce-302376944.html>

Halamka, J., & Cerrato, P. (2026, August 7). An Expanding Role for Agentic AI in Healthcare. Digital Health Frontier Column, Mayo Clinic Platform. <https://www.mayoclinicplatform.org/2026/08/07/ai-agents-in-healthcare/>

Hao, Y., Holmes, J., Waddle, M. R., Davis, B. J., Yu, N. Y., et al. (2025). Personalizing prostate cancer education for patients using an EHR-Integrated LLM agent. npj Digital Medicine, 8, 770. DOI: 10.1038/s41746-025-02166-0

Bromelkamp, R. / Mayo Clinic News Network (2026, January 7). New Mayo Clinic study advances personalized prostate cancer education with an EHR-integrated AI agent. <https://newsnetwork.mayoclinic.org/discussion/new-mayo-clinic-study-advances-personalized-prostate-cancer-education-with-an-ehr-integrated-ai-agent>

NVIDIA Corporation (2025, January 13). NVIDIA Partners With Industry Leaders to Advance Genomics, Drug Discovery and Healthcare. Press Release.

<https://investor.nvidia.com/news/press-release-details/2025/NVIDIA-Partners-With-Industry-Leaders-to-Advance-Genomics-Drug-Discovery-and-Healthcare/default.aspx>

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121-127. DOI: 10.1136/amiajnl-2011-000089

※ 다루지 않은 자료: 일부 2차 자료의 Collaco et al.(2026, npj Digital Medicine)와 Alhazba et al.(2026)은 원문을 확인하지 못했고, 후자는 학술지명과 DOI가 서로 맞지 않아 인용하지 않았다.

Tempus AI, Inc. (2026, May 28). Tempus Expands Next Platform to Deliver Real-Time Clinical Intelligence and Close Critical Care Gaps Across Oncology. Business Wire / Tempus Investor Relations.

<https://investors.tempus.com/news-releases/news-release-details/tempus-expands-next-platform-deliver-real-time-clinical>

Tempus AI, Inc. (2026, May 31). Tempus Unveils the Next-Generation of Lens, Expanding its Agentic AI Platform for Oncology Drug Development. Press Release. <https://www.tempus.com/news/pr/tempus-unveils-the-next-generation-of-lens-expanding-its-agentic-ai-platform-for-oncology-drug-development>

Tempus AI, Inc. (2026, May 29). Tempus Announces Initial Results from its Multimodal Foundation Model Efforts for Novel and Scalable Insight Generation in Oncology. Press Release. <https://investors.tempus.com/news-releases/news-release-details/tempus-announces-initial-results-its-multimodal-foundation-model>

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., et al. (2021). External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine*, 181(8), 1065-1070. DOI: 10.1001/jamainternmed.2021.2626

Tempus AI, Inc. (2026, July 20). Tempus to Acquire Personalis, More Tightly Integrating Molecular Residual Disease (MRD) into Its AI-Enabled Precision Oncology Platform. Press Release. <https://investors.tempus.com/news-releases/n>

[ews-release-details/tempus-acquire-personalis-more-tightly-integrating-molecular](#) / 보조: Santhosh, C. (2026, July 20). Tempus to buy cancer test maker Personalis for \$1.5 billion. Reuters.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56. DOI: 10.1038/s41591-018-0300-7

Marmarelis, M. E., Dow, J., Penney, J., et al., & Zimmerman, N. (2026). Multi-center prospective study evaluating an AI-enabled clinical decision support tool to improve biomarker testing in early-stage NSCLC. *Journal of Clinical Oncology*, 44(16\_suppl), 8044. ASCO 2026 Annual Meeting. [https://doi.org/10.1200/JCO.2026.44.16\\_suppl.8044](https://doi.org/10.1200/JCO.2026.44.16_suppl.8044) (학회 초록, 기술적 전후 비교. Tempus 공개 발표 자료: <https://tempus.com/publications/multi-center-prospective-study-evaluating-an-ai-enabled-clinical-decision-support-tool-to-improve-biomarker-testing-in-early-stage-nsclc>)

Personalis, Inc. & Tempus AI, Inc. (2026, October 21). Definitive proxy statement/prospectus (Schedule 14A, DEFM14A). U.S. Securities and Exchange Commission. <https://www.sec.gov/Archives/edgar/data/1527753/000119312526427872/d141436ddefm14a.htm> / 보조: Personalis, Inc. Form 8-K (보고일 2026-10-02). <https://www.sec.gov/Archives/edgar/data/1527753/000119312526415547/d155320d8k.htm>

※ 원문을 확인하지 못한 자료: NVIDIA·Tempus 공동 기술백서(2025), PRECISION Challenge 관련 1차 자료.

Greenhalgh, T., Wherton, J., Papoutsis, C., Lynch, J., et al. (2017). Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *Journal of Medical Internet Research*, 19(11), e367. DOI: 10.2196/jmir.8775

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., et al. (2020). Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. *JMIR Medical Informatics*, 8(7), e15182. DOI: 10.2196/15182

Tian, J., Song, Y., Tang, L., An, L., & Zhou, J. (2026). Clinical Artificial Intelligence Implementation in Routine Care: Real-World Operational Outcomes from a Provincial Health System in China. *Risk Management and Healthcare Policy* (Dove Medical Press), 19, 1-11. 2026년 6월 게재. DOI: 10.2147/RMHP.S604289. PMID: PMC13245460. <https://pmc.ncbi.nlm.nih.gov/articles/PMC13245460/>

Tian, J., Zhao, Z., Tang, L., Song, Y., Li, Y., & Jiang, N. (2026). From Pilot Trap to Institutional Capacity: A Governance Framework for Sustainable Clinical AI Implementation in Health Systems. *Journal of Medical Internet Research*, 28, e92680. DOI: 10.2196/92680

Yip, W., Fu, H., Chen, A. T., Zhai, T., Jian, W., Xu, R., et al. (2019). 10 years of health-care reform in China: progress and gaps in Universal Health Coverage. *The Lancet*, 394(10204), 1192-1204. DOI: 10.1016/S0140-6736(19)32136-1

Zhuang, B., Song, C., Lu, H., et al. (2025). WiNGPT-3.0 Technical Report. arXiv preprint arXiv:2505.17387. DOI: 10.48550/arXiv.2505.17387. <https://arxiv.org/abs/2505.17387> (프리프린트, 동료심사 전)

卫宁健康科技集团股份有限公司 (Winning Health Technology Group) (2025, September 21). 投资者关系活动记录表 编号 2025-003 (실적설명회 투자자관계 활동기록). 巨潮资讯网(cninfo) 공시. <http://static.cninfo.com.cn/finalpage/2025-09-21/1224672897.PDF>

Unisound (云知声) (2026, October 21). 登顶 OpenAI HealthBench 双榜！云知声医疗模型性能全面超越 GPT-6 Astra、GPT-6.1 Sol，构建高水平 AI 医生问诊能力. Official News. <https://www.unisound.com/news/165> (영문판: <https://www.unisound.com/en/news/165>)

卫宁健康科技集团股份有限公司 (2026, April 27). 投资者关系活动记录表 编号 2026-001. 巨潮资讯网(cninfo) 공시. <http://static.cninfo.com.cn/finalpage/2026-04-27/1225212083.PDF> / 보조: 证券日报 (2026, February 3). 卫宁健康：WiNEX Copilot 已在全国近150家医疗机构部署应用. 新浪财经 재게재. <https://finance.sina.com.cn/stock/relnews/cn/2026-02-03/doc-inhkpvf7241242.shtml>

Arora, R. K., Wei, J., Hicks, R. S., et al. (2025). HealthBench: Evaluating Large Language Models Towards Improved Human Health. arXiv preprint arXiv:2505.08775. DOI: 10.48550/arXiv.2505.08775 (OpenAI, 프리프린트)

Ant Group (2026, September 10). Ant Group's AI health app AQ reaches 150 million users as it expands AI-powered day-to-day health use cases. Press Release. <https://www.antgroup.com/en/news-media/press-releases/1789024200000>

李保金 (Li, Baojin) (2026, October 26). AI赋能 基层医疗数字化提速. 经济参考报 (Economic Information Daily, 新华社). <http://jjckb.xinhuanet.com/20261026/53401657b61d443cad7ad4d015ae003a/c.html>

Xu, J., Liu, J., Lao, J., et al. (2026). PulseMind: A Multi-Modal Medical Model for Real-World Clinical Diagnosis. arXiv preprint arXiv:2601.07344. <https://arxiv.org/abs/2601.07344>

国家卫生健康委员会 规划发展与信息化司 (2025, November 4). 《关于促进和规范"人工智能+医疗卫生"应用发展的实施意见》政策解读. <https://www.nhc.gov.cn/guihuaxs/c100132/202511/25d0e47241b04a6c85214aa1351e7ff7.shtml>

Ant Group (2025, December 15). AQ 업그레이드 및 '蚂蚁阿福' 명명 보도자료. <https://www.antgroup.com/en/news-media/press-releases/1765779300000> (추가 조사 자료에 기록된 1차 자료. 이 책의 조사에서 페이지 내용 재확인 실패)

国家卫生健康委办公厅, 国家发展改革委办公厅, 工业和信息化部办公厅, 国家中医药局综合司, 国家疾控局综合司 (2025, October 20). 关于促进和规范"人工智能+医疗卫生"应用发展的实施意见 (国卫办规划发〔2025〕30号). 中国政府网 게시. [https://www.gov.cn/zhengce/zhengceku/202511/content\\_7047018.htm](https://www.gov.cn/zhengce/zhengceku/202511/content_7047018.htm)

北京市卫生健康委员会 (2025, December 30). 北京市支持医疗健康领域人工智能应用发展行动计划 (2026-2027年) (京卫信息〔2025〕53号). [https://wjw.beijing.gov.cn/zwgk\\_20040/zcwj2024/202512/t20251229\\_4375683.html](https://wjw.beijing.gov.cn/zwgk_20040/zcwj2024/202512/t20251229_4375683.html)

深圳迈瑞生物医疗电子股份有限公司 (Shenzhen Mindray Bio-Medical Electronics) (2026, August 28). 2026年半年度报告 (2026 Semi-annual Report). 巨潮资讯网 (cninfo) 공시. <https://static.cninfo.com.cn/finalpage/2026-08-29/1225529856.PDF>

Zheng, L., Lin, Z., Liu, X., Fan, Z., He, Z., Xu, J., & Yin, L. (2026). Critical Care-Specific vs General-Purpose Large Language Models in Emergency Intensive Care Unit Diagnosis: Single-Center Retrospective Paired Comparative Study. Journal of Medical Internet Research, 28, e98026. DOI: 10.2196/98026. PMID: 42580688

迈瑞医疗 (2026). 2026CMEF | 走进迈瑞AI的"集群时代".

<https://www.mindray.com/cn/media-center/news/mindray-qiyuan-ai-cmef-2026> / 深圳市卫生健康委员会 게시물(2025-03-18) [https://wjw.sz.gov.cn/wzx/content/post\\_12075823.html](https://wjw.sz.gov.cn/wzx/content/post_12075823.html) (B~C) / 新华网 (2025-12-11) <http://www.news.cn/tech/20251211/5daf09a9b5034f1eb1eeb6688293cb32/c.html>

※ 다루지 않은 자료: 일부 2차 자료의 '국민건강 15.5 계획(2026년 7월, 기대수명 80세 목표)', 2024년 84개 시나리오 지침(国卫办规划函〔2024〕420号) 원문, Sidebench(2026) 보고서는 이 책의 조사에서 원문을 확인하지 못해 서지로 올리지 않았다.

U.S. Code. 21 U.S.C. § 360e-4, Predetermined change control plans for devices (FD&C Act § 515C), added by Pub. L. 117-328, div. FF, title III, § 3308(a), 136 Stat. 5835 (2022). Legal Information Institute, Cornell Law School. <https://www.law.cornell.edu/uscode/text/21/360e-4>

U.S. Food and Drug Administration (최초 최종본 2024년 12월 4일; 현행 게시본 2025년 8월 18일 발행). Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and FDA Staff (Final). CDRH·CBER·CDER. Docket FDA-2022-D-2628. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence> (PDF: <https://www.fda.gov/media/166704/download>)

Federal Register (2024-12-04). Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device

Software Functions; Availability. Document 2024-28361.

<https://www.federalregister.gov/documents/2024/12/04/2024-28361>

U.S. FDA, CDRH (2025-01-14). Webinar transcript: Final Guidance: Marketing Submission Recommendations for a PCCP for AI-Enabled Device Software Functions. <https://www.fda.gov/media/187905/download>

U.S. FDA, Health Canada, MHRA (2023; FDA 페이지 갱신 2025-08-18). Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles. <https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles> ; 함께 참고: U.S. FDA, Health Canada, MHRA (2021). Good Machine Learning Practice for Medical Device Development: Guiding Principles. <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>

U.S. FDA (2026년 9월 기준 갱신). Artificial Intelligence-Enabled Medical Devices. Digital Health Center of Excellence. <https://www.fda.gov/medical-devices/digital-health-center-excellence/artificial-intelligence-enabled-medical-devices>

U.S. FDA (2025-01). Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations (Draft Guidance). <https://www.fda.gov/media/184856/download>

U.S. FDA, CDRH Digital Health Center of Excellence (2026-08-18). Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback. Docket FDA-2026-N-7874. <https://www.fda.gov/media/194242/download>

U.S. Food and Drug Administration, CDRH Digital Health Center of Excellence (2026-08-18). Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback. Docket FDA-2026-N-7874. <https://www.fda.gov/media/194242/download>

U.S. FDA (2026-08-18). FDA Seeks Public Feedback to Inform Regulatory Approach for Generative AI-Enabled Medical Devices (Press Announcement).

<https://www.fda.gov/news-events/press-announcements/fda-seeks-public-feedback-inform-regulatory-approach-generative-ai-enabled-medical-devices>

U.S. FDA, CDER (2025-01). Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products (Draft Guidance). <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological>

Patel B, Blumenthal D (2026). A Novel Approach to Overseeing the Clinical Application of Generative AI. JAMA Health Forum, 7(3), e256947. <https://doi.org/10.1001/jamahealthforum.2025.6947>

Bergman A, Wachter RM, Emanuel EJ (2026). A Licensure Framework for Autonomous Clinical AI. JAMA, 335(20), 1751-1754. <https://doi.org/10.1001/jama.2026.5483>

Freyer O, Jayabalan S, Kather JN, Gilbert S (2025). Overcoming regulatory barriers to the implementation of AI agents in healthcare. Nature Medicine, 31, 3239-3243. <https://doi.org/10.1038/s41591-025-03841-1>

U.S. FDA (2025-10-20 게시). Digital Health Advisory Committee, November 6, 2025: Executive Summary (Generative AI-enabled devices). <https://www.fda.gov/media/189391/download>

American Hospital Association (2026-10-19). AHA Provides Feedback to FDA on Regulatory Considerations for Generative AI-Enabled Medical Devices. <https://www.aha.org/2026-10-19-aha-provides-feedback-fda-regulatory-considerations-generative-ai-enabled-medical-devices>

Bedi S, Liu Y, Orr-Ewing L, et al. (2025). Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA, 333(4), 319-328. <https://doi.org/10.1001/jama.2024.21700>

Bean AM, Payne RE, Parsons G, et al. (2026). Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. Nature Medicine, 32, 609-615. <https://doi.org/10.1038/s41591-025-04074-y>

Medicines and Healthcare products Regulatory Agency (2026-06-09, updated 2026-07-27). AI Airlock Sandbox Phase 2 Programme Report. GOV.UK. <https://www.gov.uk/government/publications/ai-airlock-sandbox-phase-2-programme-report>

MHRA. AI Airlock: a regulatory sandbox for AI enabled Medical Devices (Collection page). GOV.UK. <https://www.gov.uk/government/collections/ai-airlock-a-regulatory-sandbox-for-ai-enabled-medical-devices>

MHRA (2026-04-08). MHRA expands AI Airlock programme with a £3.6 million funding boost over three years. GOV.UK. <https://www.gov.uk/government/news/mhra-expands-ai-airlock-programme-with-a-36-million-funding-boost-over-three-years>

UK Government (2026-10-06). Government backs recommendations of NHS doctors-led AI Commission. GOV.UK. <https://www.gov.uk/government/news/government-backs-recommendations-of-nhs-doctors-led-ai-commission>

European Parliament and Council (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L, 2024/1689, 12.7.2024. <http://data.europa.eu/eli/reg/2024/1689/oj> ; 통합본(2026-07-27): <https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng>

European Commission, AI Act Service Desk. Article 113: Entry into force and application. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-113>

Medical Device Coordination Group & AI Board (2025-06). MDCG 2025-6 / AIB 2025-1: Interplay between the Medical Devices Regulation (MDR) & In vitro Diagnostic Medical Devices Regulation (IVDR) and the Artificial Intelligence Act (AIA). European Commission. [https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4\\_en](https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4_en) (주의: 개정 전 적용일을 서술함)

European Parliament and Council (2026). Regulation (EU) 2026/1744 of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on

artificial intelligence (Digital Omnibus on AI). Official Journal of the European Union, L, 2026/1744, 24.7.2026. <http://data.europa.eu/eli/reg/2026/1744/oj>

European Commission (2026-07-27). AI Omnibus enters into force. Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/ai-omnibus-enters-force>

Taylor Wessing (2026-07-02). Update: AI-Enabled Medical Devices and IVDs Confirmed as High-Risk. Life Sciences Legal Lens (2차 자료). <https://www.taylorwessing.com/en/international-life-sciences-newsletter/life-sciences-legal-lens-vol-2/update-ai-enabled-medical-devices-and-ivds-confirmed-as-high-risk>

White & Case (2026-08-04). EU AI Omnibus enters into force, amending the AI Act (2차 자료). <https://www.whitecase.com/insight-alert/eu-ai-omnibus-enters-force-amending-ai-act>

Rabotin A, Aguilar E, Sandoval Gonzalez S, Iniguez A, Jung C, Lee E, Bell DS, Arroyo J (2026). Propagation of Interpreter Errors by Ambient AI Scribes: Study Using Simulated Clinical Encounters. JMIR Medical Informatics, 14, e88734. <https://doi.org/10.2196/88734> (PMID 42520279)

Flores G, Laws MB, Mayo SJ, Zuckerman B, et al. (2003). Errors in Medical Interpretation and Their Potential Clinical Consequences in Pediatric Encounters. Pediatrics, 111(1), 6-14. <https://doi.org/10.1542/peds.111.1.6>

Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D, Goel S (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689. <https://doi.org/10.1073/pnas.1915768117>

Zolnoori M, Vergez S, Xu Z, et al. (2024). Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. JAMIA Open, 7(4), ooae130. <https://doi.org/10.1093/jamiaopen/ooae130>

Topaz M, Peltonen LM, Zhang Z (2025). Beyond human ears: navigating the uncharted risks of AI scribes in clinical practice. *npj Digital Medicine*, 8, 569. <https://doi.org/10.1038/s41746-025-01895-6>

Goddard K, Roudsari A, Wyatt JC (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121-127. <https://doi.org/10.1136/amiajnl-2011-000089>

Lyell D, Coiera E (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2), 423-431. <https://doi.org/10.1093/jamia/ocw105>

Dratsch T, Chen X, Rezazade Mehrizi M, et al. (2023). Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*, 307(4), e222176. <https://doi.org/10.1148/radiol.222176>

Taib, A. G., Partridge, G. J. W., Phillips, P., Maxwell-Armstrong, C., Chen, X., Hofvind, S. S. H., James, J. J., & Chen, Y. (2026). Automation Bias in Action: Eye Tracking of Humans Reading Screening Mammograms with and without AI Prompts. *Radiology*, 320(1), e252590. 온라인 2026-07-14. <https://doi.org/10.1148/radiol.252590>

Qazi IA, Ali A, Khawaja AU, Akhtar MJ, Sheikh AZ, Alizai MH (2026). Automation Bias in Large Language Model-Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy: A Randomized Clinical Trial. *NEJM AI*, 3(5). <https://doi.org/10.1056/Aloa2501001>

Wu L, Xu Z, Wang H, et al. (2026). Dual- versus Single-Suggestion AI Support for Radiographic Interpretation in Residents: Randomized Multireader Study. *arXiv preprint arXiv:2610.09589* (동료심사 전). <https://doi.org/10.48550/arXiv.2610.09589>

European Parliament and Council (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 14 Human oversight. <http://data.europa.eu/eli/reg/2024/1689/oj>

Jiang YH, Black KC, Geng G, et al. (2025). MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. NEJM AI, 2(9).

<https://doi.org/10.1056/Aldbp2500144>

ECRI (2026-01). Top 10 Health Technology Hazards for 2026: Misuse of AI chatbots tops annual list. <https://home.ecri.org/blogs/ecri-news/misuse-of-ai-chatbots-tops-annual-list-of-health-technology-hazards>

Grand View Research (2026). Global Agentic AI in Healthcare Market Size & Outlook, 2033 (Horizon Outlook). <https://www.grandviewresearch.com/horizon/outlook/agentic-ai-in-healthcare-market-size/global> ; 보고서: Agentic AI In Healthcare Market (2026-2033) Size, Share & Trends Analysis Report.

<https://www.grandviewresearch.com/industry-analysis/agentic-ai-healthcare-market-report>

Grand View Research. North America Agentic AI in Healthcare Market Size & Outlook. <https://www.grandviewresearch.com/horizon/outlook/agentic-ai-in-healthcare-market/north-america>

Towards Healthcare (2026-08-17). Agentic AI in Healthcare Market Sizing (비교용 2차 시장 자료). <https://www.towardshealthcare.com/insights/agentic-ai-in-healthcare-market-sizing>

VoiceCare AI (2025-02-14). Agentic AI Startup VoiceCare AI Launches to Automate Healthcare Back Office and Super Staff Workforce. PR Newswire. <https://www.prnewswire.com/news-releases/agentic-ai-startup-voicecare-ai-launches-to-automate-healthcare-back-office-and-super-staff-workforce-302376944.html>

NVIDIA (2025-01-13). NVIDIA Partners With Industry Leaders to Advance Genomics, Drug Discovery and Healthcare. <https://investor.nvidia.com/news/press-release-details/2025/NVIDIA-Partners-With-Industry-Leaders-to-Advance-Genomics-Drug-Discovery-and-Healthcare/default.aspx>

Bipartisan Policy Center (2026-02-18). Paying for AI in U.S. Health Care (Issue Brief). <https://bipartisanpolicy.org/issue-brief/paying-for-ai-in-u-s-health-care>

American Medical Association (2026). CPT Appendix S: Taxonomy for Artificial Intelligence in Medical Services and Procedures (CPT 2027) 및 Appendix S Table. <https://www.ama-assn.org/practice-management/cpt/cpt-appendix-s-taxonomy-artificial-intelligence-medical-services-procedures> ; <https://www.ama-assn.org/system/files/appendix-s-table.pdf>

American Medical Association (2026-06-10). CPT Editorial Panel strengthens AI taxonomy to keep pace with tech. <https://www.ama-assn.org/practice-management/cpt/cpt-editorial-panel-strengthens-ai-taxonomy-keep-pace-tech>

Centers for Medicare & Medicaid Services (2020). FY 2021 Hospital Inpatient Prospective Payment System (IPPS) and LTCH PPS Final Rule (CMS-1735-F), New Technology Add-On Payment for ContaCT (Viz.ai). Federal Register. <https://www.federalregister.gov/>

Bipartisan Policy Center (2026-10-26). Paying for Clinical AI: Principles for Policymakers. <https://bipartisanpolicy.org/article/paying-for-clinical-ai-principles-for-policymakers/>

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, 12, 157–173. [https://doi.org/10.1162/tacl\\_a\\_00638](https://doi.org/10.1162/tacl_a_00638) (2024년 핵심 선행 연구)

Bedi, S., Liu, Y., Orr-Ewing, L., et al. (2025). Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA, 333(4), 319–328. <https://doi.org/10.1001/jama.2024.21700>

Ma, S., Huang, J., Zhang, F., Wu, J., Shen, Y., Fan, G., et al. (2026). MedLA: A Logic-Driven Multi-Agent Framework for Complex Medical Reasoning with Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence, 40(2), 845–853. <https://doi.org/10.1609/aaai.v40i2.37052> (확장판 arXiv:2509.23725)

Schmidgall, S., et al. (2024, v5 2025). AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv preprint arXiv:2405.07960. <https://doi.org/10.48550/arXiv.2405.07960> (사전공개 논문)

Liu, Y., Carrero, Z. I., Jiang, X., Ferber, D., Wölflein, G., Zhang, L., Jayabalan, S., Lenz, T., Hui, Z., & Kather, J. N. (2026). Benchmarking large language model-based agent systems for clinical decision tasks. *npj Digital Medicine*, 9, 259. <https://doi.org/10.1038/s41746-026-02443-6>

Chen, M., Wu, Y., Ma, J., Jia, X., Gao, C., Zhao, F., & Qiao, Y. (2026). Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. *npj Digital Medicine*, 9, 222. <https://doi.org/10.1038/s41746-026-02409-8>

Bean, A. M., Payne, R. E., Parsons, G., et al. (2026). Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, 32, 609–615. <https://doi.org/10.1038/s41591-025-04074-y> (2026년 4월 17일 출판사 정정 공지: *Nature Medicine* 32(5):1937, <https://doi.org/10.1038/s41591-026-04404-8>. 정정 내용이 본문 수치에 영향을 주었는지는 확인하지 못함)

Alhazba, S. A., Ali, M. H., Alawi, D. A., Al-Betar, M. A., & Elaziz, M. A. (2026). From Reactive AI to Agentic Systems: A Review of Autonomous Medical AI Agents in Healthcare. *Archives of Computational Methods in Engineering*, 33(8), 11693–11730. <https://doi.org/10.1007/s11831-026-10655-y>

※ 참고문헌 주: 일부 2차 자료에 있던 "AgentClinic, *npj Digital Medicine* 8:112(DOI 10.1038/s41746-025-01452-1)"와 "메타분석 8:245(DOI 10.1038/s41746-025-01588-0)"는 DOI 조회 결과 다른 논문으로 확인되어 사용하지 않았다. 1.4%·1.0% 수치의 실제 출처는 Liu 등의 2026년 논문 「Benchmarking large language model-based agent systems for clinical decision tasks」 이다.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson. (에이전트의 고전적 정의, 교과서)

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. *International Conference on Learning Representations (ICLR 2023)*. <https://arxiv.org/abs/2210.03629>

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. *Advances in*

Neural Information Processing Systems 36 (NeurIPS 2023).

<https://arxiv.org/abs/2303.11366>

Jiang, Y. H., Black, K. C., Geng, G., et al. (2025). MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. NEJM AI, 2(9).

<https://doi.org/10.1056/Aldbp2500144>

Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2024). Cognitive Architectures for Language Agents. Transactions on Machine Learning Research.

<https://arxiv.org/abs/2309.02427>

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33 (NeurIPS 2020). <https://arxiv.org/abs/2005.11401>

Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., & Gonzalez, J. E. (2023). MemGPT: Towards LLMs as Operating Systems. arXiv preprint arXiv:2310.08560. <https://arxiv.org/abs/2310.08560>

Schmidgall, S., et al. (2024, v5 2025). AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv preprint arXiv:2405.07960. <https://doi.org/10.48550/arXiv.2405.07960>

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23). <https://doi.org/10.1145/3586183.3606763>

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526. <https://doi.org/10.1073/pnas.1611835114>

Latent Labs Team (2026). Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design. arXiv preprint arXiv:2603.29727. <https://arxiv.org/abs/2603.29727> (기업 기술 보고서, 사전공개)

[보조 문헌] Alhazba, S. A., et al. (2026). From Reactive AI to Agentic Systems: A Review of Autonomous Medical AI Agents in Healthcare. Archives of

Computational Methods in Engineering, 33(8), 11693–11730.

<https://doi.org/10.1007/s11831-026-10655-y> (에이전틱 AI를 '지각-추론-행동 순환, 지속적 메모리, 도구 사용'으로 정의)

※ 참고문헌 주: 일부 2차 자료의 "Latent-Y / DBTL (Nature 2026)"은 Nature 게재를 확인하지 못했고, 실제 출처는 Latent Labs Team의 arXiv 기술 보고서이다. "ProtoPilot"(arXiv:2606.31763, 생물학 실험 프로토콜 자동 생성 에이전트)은 실존하는 사전 공개 논문이지만, 이 절의 주제와 직접 맞닿지 않아 본문에서 다루지 않았다.

Vatsal, S., Dubey, H., & Singh, A. (2026). Agentic AI in Healthcare and Medicine: A Seven-Dimensional Taxonomy for Empirical Evaluation of LLM-Based Agents. IEEE Access, 14, 4840–4863. <https://doi.org/10.1109/ACCESS.2026.3651218> (사전공개 arXiv:2602.04813, <https://doi.org/10.48550/arXiv.2602.04813>)

Liu, Y., Carrero, Z. I., Jiang, X., et al., & Kather, J. N. (2026). Benchmarking large language model-based agent systems for clinical decision tasks. npj Digital Medicine, 9, 259. <https://doi.org/10.1038/s41746-026-02443-6>

※ 참고문헌 주: 일부 2차 자료의 Sidebench 산업 보고서, McKinsey 약 6%, MIT Project NANDA 95% 수치는 이 책의 조사에서 원출처를 직접 확인하지 못해 사용하지 않았다. 일부 2차 자료가 Vatsal 연구 범위를 "2021~2026년 초"로 적은 부분은 논문 본문 확인 결과 "2023년 10월~2025년 6월"로 바로잡았다.

Vatsal, S., Dubey, H., & Singh, A. (2026). Agentic AI in Healthcare and Medicine: A Seven-Dimensional Taxonomy for Empirical Evaluation of LLM-Based Agents. IEEE Access, 14, 4840–4863. <https://doi.org/10.1109/ACCESS.2026.3651218> (arXiv:2602.04813)

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., & Gerstein, M. (2024). MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. Findings of the Association for Computational Linguistics: ACL 2024. <https://arxiv.org/abs/2311.10537>

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving Factuality and Reasoning in Language Models through Multiagent Debate. Proceedings of the 41st International Conference on Machine Learning (ICML 2024). <https://arxiv.org/abs/2305.14325>

Kim, Y., Park, C., Jeong, H., et al. (2024). MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. *Advances in Neural Information Processing Systems* 37 (NeurIPS 2024). <https://arxiv.org/abs/2404.15155>

Ma, S., Huang, J., Zhang, F., Wu, J., Shen, Y., Fan, G., et al. (2026). MedLA: A Logic-Driven Multi-Agent Framework for Complex Medical Reasoning with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(2), 845–853. <https://doi.org/10.1609/aaai.v40i2.37052>

Li, J., Lai, Y., Li, W., Ren, J., Zhang, M., Kang, X., Wang, S., Li, P., Zhang, Y.-Q., Ma, W., & Liu, Y. (2024, v3 2025). Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv preprint arXiv:2405.02957*. <https://doi.org/10.48550/arXiv.2405.02957>

Chen, S., Chen, F., Ding, X., Chen, W., & Wei, Z. (2026). ClinVerse: A Realistic Multi-Agent Diagnostic Environment for Training and Evaluating Medical Agents. *arXiv preprint arXiv:2610.34878*. <https://doi.org/10.48550/arXiv.2610.34878>

Fan, Z., Wei, L., Tang, J., Chen, W., Wang, S., Wei, Z., & Huang, F. (2025). AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator. *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, 10183–10213. (서지는 [7]의 참고문헌으로 재확인)

Yao, Z., & Yu, H. (2026). LLM-Based Multi-Agent Systems for Clinical Workflows: A Survey of AI Hospitals. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 45772–45793. <https://doi.org/10.18653/v1/2026.acl-long.2123> (서지 확인, 본문 세부 내용은 미열람)

Cemri, M., Pan, M. Z., Yang, S., et al. (2025). Why Do Multi-Agent LLM Systems Fail? *arXiv preprint arXiv:2503.13657*. <https://arxiv.org/abs/2503.13657>

※ 참고문헌 주: 일부 2차 자료의 "MAGDA" 사례는 원출처를 확인하지 못해 다루지 않았다. "ProtoPilot"(arXiv:2606.31763)은 실존하는 사전공개 논문이지만 이 절의 범위 밖이어서 다루지 않았다.

Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., Gross, B., Hames, P., Suleyman, M., King, D., & Horvitz, E. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423–428. <https://doi.org/10.1126/science.aeh7930> (2026-10-22 게재, 저자 전원 Microsoft 소속)

Nori, H., et al. (2025). Sequential Diagnosis with Language Models. arXiv preprint arXiv:2506.22405 (v2, 2025-07-02). <https://doi.org/10.48550/arXiv.2506.22405> (사전 공개판, 수치가 Science판과 다름)

Microsoft AI (2026, October 22). Building Medical Superintelligence. <https://microsoft.ai/news/building-medical-superintelligence> (기업 공식 발표)

Microsoft (2026, June 2). Mayo Clinic and Microsoft collaborate to develop a frontier AI model for healthcare. <https://news.microsoft.com/source/2026/06/02/mayo-clinic-and-microsoft-collaborate-to-develop-a-frontier-ai-model-for-health-care/> (기업 보도자료, [3]에서 링크 확인)

Microsoft AI 연구진 (2026). The Building Blocks of Medical Superintelligence. *BMJ Digital Health and AI*, 2(1), e000395. <https://bmjdigitalhealth.bmj.com/content/2/1/e000395>

Everett, S. S., Bunning, B. J., Jain, P., Lopez, I., Agarwal, A., Desai, M., Gallo, R., Goh, E., et al. (2026). From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. *npj Digital Medicine*, 9, 409. <https://doi.org/10.1038/s41746-026-02545-1> (NCT06911645)

※ 참고문헌 주: 일부 2차 자료의 Science 권호 "394(6821), eah7930"은 Crossref 조회 결과 "394(6822), 423–428"로 바로잡았다. 일부 2차 자료의 "2027년까지 EHR의 프로그래밍 식 혁신" 발언(중국 매체 인용)은 원출처를 확인하지 못해 사용하지 않았다.

Chen, S., Chen, F., Ding, X., Chen, W., & Wei, Z. (2026). ClinVerse: A Realistic Multi-Agent Diagnostic Environment for Training and Evaluating Medical Agents. arXiv preprint arXiv:2610.34878 (v1, 2026-10-26). Fudan University, Shanghai Innovation Institute, Zhejiang University, Huazhong University of Science and Technology. <https://doi.org/10.48550/arXiv.2610.34878> (사전공개, 동료 심사 전)

Johnson, A. E. W., Bulgarelli, L., Shen, L., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10, 1.

<https://doi.org/10.1038/s41597-022-01899-x> (데이터 출처, 2023년 핵심 자료) / SICdb: Rodemund, N., et al. (2023). Salzburg Intensive Care database (SICdb). [1]의 참고문헌으로 재인용, PhysioNet 제공.

Walonoski, J., Kramer, M., Nichols, J., et al. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3), 230–238. <https://doi.org/10.1093/jamia/ocx079> (2018년 권호 수록; 온라인 공개 2017-08-30)

Li, J., Lai, Y., Li, W., et al. (2024, v3 2025). Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. arXiv preprint arXiv:2405.02957. <https://doi.org/10.48550/arXiv.2405.02957> (MedAgent-Zero 방법의 출처)

Nori, H., et al. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423–428. <https://doi.org/10.1126/science.aeh7930> (비교 대상 SDBench)

※ 참고문헌 주: 일부 2차 자료의 "ClinPatient는 중국 특정 병원 위주", "보호자 에이전트", "DPO/PPO 등 9가지 기법", "차팅 시간 50% 단축"은 논문 원문과 맞지 않거나 확인되지 않아 바로잡거나 뺐다.

Xu, J., Pan, M., Chen, L., Fu, J., Shi, R., & Xie, J. (2026). AI-Enabled regional tele-ECG cloud platform and improving access to cardiovascular diagnosis: real-world evidence from southern China. *Frontiers in Cardiovascular Medicine*, 13, 1869716. <https://doi.org/10.3389/fcvm.2026.1869716> (2026년 7월 16일 정정 공지: <https://doi.org/10.3389/fcvm.2026.1927426>)

Thygesen, K., Alpert, J. S., Jaffe, A. S., Chaitman, B. R., et al. (2019). Fourth universal definition of myocardial infarction (2018). *European Heart Journal*, 40(3), 237–269. <https://doi.org/10.1093/eurheartj/ehy462>

Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., et al. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, 307(4), e222176. <https://doi.org/10.1148/radiol.222176>

Lin, C.-S., Liu, W.-T., Tsai, D.-J., Lou, Y.-S., et al. (2024). AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. *Nature Medicine*, 30(5), 1461–1470.

<https://doi.org/10.1038/s41591-024-02961-4>

Lång, K., Josefsson, V., Larsson, A.-M., Larsson, S., et al. (2023). Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, 24(8), 936–944.

[https://doi.org/10.1016/S1470-2045\(23\)00298-X](https://doi.org/10.1016/S1470-2045(23)00298-X)

Gommers, J., Hernström, V., Josefsson, V., Sartor, H., et al. (2026). Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. *The Lancet*, 407(10527), 505–514. [https://doi.org/10.1016/S0140-6736\(25\)02464-X](https://doi.org/10.1016/S0140-6736(25)02464-X)

Zöller, N., Berger, J., Lin, I., Fu, N., et al. (2025). Human–AI collectives most accurately diagnose clinical vignettes. *Proceedings of the National Academy of Sciences*, 122(24), e2426153122. <https://doi.org/10.1073/pnas.2426153122>

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, 1, 39.

<https://doi.org/10.1038/s41746-018-0040-6>

Utah Department of Commerce, Office of Artificial Intelligence Policy. (2026). Doctrionic regulatory mitigation agreement: AI prescription renewal pilot (게시 2026-01; 계약 기간 2025년 10월~2026년 10월로 안내됨).

<https://commerce.utah.gov/ai/regulatory-relief/authorized-ai-pilots/doctrionic/>

Utah Office of Artificial Intelligence Policy. (2026, May 19). Key statistics on the Doctrionic pilot program (Public Report 1).

<https://commerce.utah.gov/wp-content/uploads/2026/05/Doctrionic-Outcomes-May-2026.pdf>

Olson, K. D., Meeker, D., Troup, M., Barker, T. D., et al. (2025). Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Network Open*, 8(10), e2534976. <https://doi.org/10.1001/jamanetworkopen.2025.34976>

Lukac, P. J., Turner, W., Vangala, S., Chin, A. T., et al. (2025). Ambient AI scribes in clinical practice: A randomized trial. *NEJM AI*, 2(12). <https://doi.org/10.1056/Aloa2501000>

Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., et al. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. *NEJM AI*, 2(12). <https://doi.org/10.1056/Aloa2500945>

Rotenstein, L. S., Holmgren, A. J., Thombley, R., Sriram, A., et al. (2026). Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes. *JAMA*, 335(16), 1408-1417. <https://doi.org/10.1001/jama.2026.2253>

Xu, J., Pan, M., Chen, L., Fu, J., Shi, R., & Xie, J. (2026). AI-Enabled regional tele-ECG cloud platform and improving access to cardiovascular diagnosis: real-world evidence from southern China. *Frontiers in Cardiovascular Medicine*, 13, 1869716. <https://doi.org/10.3389/fcvm.2026.1869716>

Microsoft AI. (2026, October 22). Building medical superintelligence [기업 발표문]. <https://microsoft.ai/news/building-medical-superintelligence>

Nori, H., Daswani, M., Kelly, C., Lundberg, S., et al. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423-428. <https://doi.org/10.1126/science.aeh7930>

Goh, E., Gallo, R., Hom, J., Strong, E., et al. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10), e2440969. <https://doi.org/10.1001/jamanetworkopen.2024.40969>

Everett, S. S., Bunning, B. J., Jain, P., et al. (2026). From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. *npj Digital Medicine*, 9, 409. <https://doi.org/10.1038/s41746-026-02545-1>

Wu, L., Xu, Z., Wang, H., Zhou, F., et al. (2026). Dual- versus single-suggestion AI support for radiographic interpretation in residents: Randomized multireader study [Preprint]. arXiv:2610.09589. <https://doi.org/10.48550/arXiv.2610.09589>

Budzyń, K., et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10), 896–903. [https://doi.org/10.1016/S2468-1253\(25\)00133-5](https://doi.org/10.1016/S2468-1253(25)00133-5)

Suleyman, M. (2025, November 6). Towards humanist superintelligence [기업 발표문]. Microsoft AI. <https://microsoft.ai/news/towards-humanist-superintelligence/>

Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., Gross, B., Hames, P., Suleyman, M., King, D., & Horvitz, E. (2026). Sequential diagnosis with language models. *Science*, 394(6822), 423–428. <https://doi.org/10.1126/science.aeh7930>

Microsoft AI. (2026, October 22). Building medical superintelligence [기업 발표문]. <https://microsoft.ai/news/building-medical-superintelligence> (같은 글이 소개한 관련 논문: The building blocks of medical superintelligence. *BMJ Digital Health & AI*, 2(1), e000395. <https://bmjdigitalhealth.bmj.com/content/2/1/e000395> — 저자와 DOI는 이 책의 조사에서 미확인)

ECRI. (2026). Top 10 health technology hazards for 2026: Misuse of AI chatbots tops annual list [보도자료]. <https://home.ecri.org/blogs/ecri-news/misuse-of-ai-chatbots-tops-annual-list-of-health-technology-hazards>

Vatsal, S., Dubey, H., & Singh, A. (2026). Agentic AI in healthcare & medicine: A seven-dimensional taxonomy for empirical evaluation of LLM-based agents. *IEEE Access*, 14, 4840–4863. <https://doi.org/10.1109/ACCESS.2026.3651218> (arXiv:2602.04813)

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>

Wittmann, B. J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., et al. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390(6768), 82–87.

<https://doi.org/10.1126/science.adu8578>

European Commission. (2025, October 6). Withdrawal of Commission proposals (C/2025/5423) [COM(2022)496, AI Liability Directive 포함]. Official Journal of the European Union, C series. <http://data.europa.eu/eli/C/2025/5423/oj>

Utah Office of Artificial Intelligence Policy. (2026). Doctronic regulatory mitigation agreement and Public Report 1 (May 19, 2026).

<https://commerce.utah.gov/ai/regulatory-relief/authorized-ai-pilots/doctronic/> ;  
<https://commerce.utah.gov/wp-content/uploads/2026/05/Doctronic-Outcomes-May-2026.pdf>

U.S. Food and Drug Administration. (최초 최종본 2024-12-04; 현행 게시본 2025-08-18). Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Final guidance. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>

U.S. Food and Drug Administration & European Medicines Agency. (2026, January). Guiding principles of good AI practice in drug development. <https://www.fda.gov/media/189581/download>

Regulation (EU) 2024/1689 (Artificial Intelligence Act), consolidated version 2026-07-27; Regulation (EU) 2026/1744 (Digital Omnibus on AI).  
<https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng> ;  
<https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng>

Office of Science and Technology Policy (Executive Office of the President). (2024). Framework for nucleic acid synthesis screening.

<https://aspr.hhs.gov/S3/Documents/OSTP-Nucleic-Acid-Synthesis-Screening-Framework-Sep2024.pdf>

Bedi, S., Liu, Y., Orr-Ewing, L., et al. (2025). Testing and evaluation of health care applications of large language models: A systematic review. JAMA, 333(4), 319–328. <https://doi.org/10.1001/jama.2024.21700>

## 보완판 추가 참고자료

TXP Medical. SpeechER 수상·제품 발표. <https://txpmedical.jp/en/news/2jmiSPoM4bpIdI9GUZFOBI/>

SIND. Caretomo 제품 설명. <https://caretomo.jp/ja>

Cleveland Clinic. Less Typing, More Talking. <https://consultqd.clevelandclinic.org/less-typing-more-talking-how-ambient-ai-is-reshaping-clinical-workflow-at-cleveland-clinic>

Yale 등. Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2839542>

Abridge. Abridge Launches. <https://www.abridge.com/press-release/abridge-launches-to-help-people-stay-on-top-of-their-health>

수도의과대학·칭화대학 등. Antibiotic selection CDSS. <https://medinform.jmir.org/2026/1/e98207>

Unisound. U2-Med 출시. <https://unisound.com/en/research/126>

Unisound. U2-Med World. <https://unisound.com/en/news/165>

Microsoft AI. Sequential Diagnosis with Language Models. <https://arxiv.org/abs/2506.22405>

Fudan University, Shanghai Innovation Institute, Zhejiang University, Huazhong University of Science and Technology. ClinVerse: A Realistic Multi-Agent Diagnostic Environment for Training and Evaluating Medical Agents. <https://arxiv.org/abs/2610.34878>

Google Research. AMIE announcement. <https://research.google/blog/amie-a-research-ai-system-for-diagnostic-medical-reasoning-and-conversations/>

Google. AMIE disease management. <https://blog.google/innovation-and-ai/models-and-research/google-research/amie-for-disease-management-in-nature/>

Heidelberg · TU Dresden 등. Towards autonomous medical artificial intelligence agents. <https://www.nature.com/articles/s41586-026-10675-5>

BIDMC. AMIE clinical study. <https://bidmc.org/news-stories/all-news-stories/news/2026/10/bidmc-research-shows-safety-quality-of-ai-in-primary-care>

Arc · NVIDIA. Evo 2 announcement. <https://arcinstitute.org/news/evo2>

Google DeepMind. AlphaGenome announcement.  
<https://deepmind.google/blog/alphagenome-ai-for-better-understanding-the-genome/>

Google 연구진. Tx-LLM. <https://arxiv.org/abs/2406.06316>

Google 연구진. TxGemma. <https://arxiv.org/abs/2504.06196>

Wang · Tian 등. HistoMethyl. <https://www.biorxiv.org/content/10.64898/2026.06.11.731518v1>

Houston Methodist Research Institute · Weill Cornell Medicine · NCI. Fuji.  
<https://pmc.ncbi.nlm.nih.gov/articles/PMC13484847/>

Q. Sun 등. CGA-Net. <https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2026.1802570/full>

Christopher Baker 등. Octopus. <https://arxiv.org/abs/2607.16262>

Christopher Baker 등. Octopus 수정원고. <https://www.biorxiv.org/content/10.64898/2026.07.05.736565v2>

Christopher Baker 등. CIWM. <https://arxiv.org/abs/2603.02274>

UNC · UT Austin · UPenn. Chreode. <https://arxiv.org/html/2605.28111v1>

Shanghai AI Lab 등. SCALE. <https://arxiv.org/html/2603.17380v3>

Di Francesco 등. PT-RAG. <https://arxiv.org/abs/2603.07233>

CAS · UCAS · NUS. CURE. <https://arxiv.org/html/2605.15243v1>

South China Normal · Osaka. JoPMol. <https://arxiv.org/html/2607.11978v1>

Tianqi Chen · Carlos Guestrin. XGBoost. <https://arxiv.org/abs/1603.02754>

Scott Lundberg · Su-In Lee. SHAP. <https://arxiv.org/abs/1705.07874>

Romain Lopez · Nir Yosef 등. scVI. <https://arxiv.org/abs/1709.02082>

William Peebles · Saining Xie. DiT. <https://arxiv.org/abs/2212.09748>

Stanford University. GEARS. <https://www.nature.com/articles/s41587-023-01905-6>

bioRxiv / University of Central Florida. HistoMethyl bioRxiv official metadata.  
<https://api.biorxiv.org/details/biorxiv/10.64898/2026.06.11.731518>

The First People's Hospital of Taicang. CGA-Net first-page affiliations.  
<https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2026.1802570/pdf>

Sapienza · ISTI-CNR · Cambridge · Wellcome Sanger. PT-RAG initial preprint.  
<https://arxiv.org/html/2603.07233v1>

NUS 등. PRECISE CURATE.AI. <https://www.nature.com/articles/s41698-025-00835-7>

NUS. CURATE.AI 2018 research. <https://news.nus.edu.sg/nus-researchers-use-ai-to-successfully-treat-metastatic-cancer-patient/>

Cleveland Clinic. Twin Precision Treatment.  
<https://newsroom.clevelandclinic.org/2025/08/20/cleveland-clinic-led-research-shows-how-ai-supported-precision-health-and-lifestyle-coaching-program-can-improve-outcomes-in-patients-with-type-2-diabetes>

Dassault Systemes. Living Heart launch.

<https://www.3ds.com/newsroom/press-releases/dassault-systemes-unveils-3d-simulated-heart-breaking-new-ground-study-heart-disease-and-personalized-treatment>

OPRIMED. OPTIVIS 제품 설명. <https://www.oprimed.com/>

EMA. PROCOVA qualification. [https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm\\_en.pdf](https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/qualification-opinion-prognostic-covariate-adjustment-procovatm_en.pdf)

EMA. EMA qualification opinion 원문 사본. [https://www.aindo.com/docs/regulatory-hub/37\\_Qualification\\_opinion\\_for\\_Prognostic\\_Covariate\\_Adjustment.pdf](https://www.aindo.com/docs/regulatory-hub/37_Qualification_opinion_for_Prognostic_Covariate_Adjustment.pdf)

Unlearn. PROCOVA regulatory rigor. <https://www.unlearn.ai/blog/built-on-regulatory-rigor-advancing-digital-twins-in-clinical-trials>

FDA. NAMs. <https://www.fda.gov/science-research/science-and-research-special-topics/new-approach-methodologies-nams>

FDA. MIDD program. <https://www.fda.gov/drugs/development-resources/model-informed-drug-development-paired-meeting-program>

Insilico Medicine. Rentosertib Phase III. <https://insilico.com/news/xmjns4l091-insilico-initiates-phase-iii-clinical-tr>

K-MELLODDY. 사업단 소개. <https://www.kmelloddy.org/aboutus>

Insilico Medicine. Insilico official Technology introduction.  
<https://insilico.com/mediakit>

VoiceCare AI. VoiceCare AI 출범 · 메이요 시범사업.

<https://www.prnewswire.com/news-releases/agentica-ai-startup-voicecare-ai-launches-to-automate-healthcare-back-office-and-super-staff-workforce-302376944.html>

Mayo Clinic Platform. An Expanding Role for Agentic AI in Healthcare.

<https://www.mayoclinicplatform.org/2026/08/07/ai-agents-in-healthcare/>

Mayo Clinic/Cornell/MIT. Personalizing prostate cancer education for patients using an EHR-Integrated LLM agent. <https://pubmed.ncbi.nlm.nih.gov/41413170/>

Mayo Clinic. MedEduChat 기관 해설. <https://newsnetwork.mayoclinic.org/discussion/new-mayo-clinic-study-advances-personalized-prostate-cancer-education-with-an-ehr-integrated-ai-agent/>

Tempus. Tempus Next 공식 제품 소개. <https://www.tempus.com/solutions/next/>

Tempus. Lens 최초 공개. <https://www.tempus.com/news/pr/tempus-unveils-its-lens-platform-offering-unparalleled-access-to-one-of-the-worlds-largest-de-identified-clinical-and-molecular-datasets/>

Tempus. Lens 차세대 공개. <https://www.tempus.com/news/pr/tempus-unveils-the-next-generation-of-lens-expanding-its-agentic-ai-platform-for-oncology-drug-development/>

Mayo Clinic/Cornell/MIT. MedEduChat 논문 PMC 전문.  
<https://pmc.ncbi.nlm.nih.gov/articles/PMC12714722/>

하북의과대학 제1병원 연구팀. Clinical Artificial Intelligence Implementation in Routine Care. <https://pmc.ncbi.nlm.nih.gov/articles/PMC13245460/>

Winning Health 연구팀. WiNGPT-3.0 Technical Report.  
<https://arxiv.org/abs/2505.17387>

Unisound. Medical World Models. <https://www.unisound.com/en/news/167>

Ant Group. AQ 출시. <https://www.antgroup.com/news-media/press-releases/1750924800000>

Ant Group. How Does AI Shape the Next Decade of Healthcare.  
<https://www.antgroup.com/en/news-media/press-releases/1751436000000>

선전시 위생건강위원회. ICU 중증 모델 선전 도입. [https://wjw.sz.gov.cn/wzx/content/post\\_12075823.html](https://wjw.sz.gov.cn/wzx/content/post_12075823.html)

중국 국가위생건강위원회. 인공지능+의료위생 정책 해설. <https://www.nhc.gov.cn/guihuaxxs/c100132/202511/25d0e47241b04a6c85214aa1351e7ff7.shtml>

U.S. FDA. PCCP 최종 지침. <https://www.fda.gov/media/166704/download>

FDA/Health Canada/MHRA. PCCP 국제 원칙. <https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles>

U.S. FDA. Generative AI Discussion Paper.  
<https://www.fda.gov/media/194242/download>

UK MHRA. AI Airlock Sandbox Phase 2 Programme Report.  
<https://www.gov.uk/government/publications/ai-airlock-sandbox-phase-2-programme-report>

European Commission. AI Act Article 6. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-6>

European Commission. AI Act Article 113. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-113>

EU EUR-Lex. AI Act 통합법령. <https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng>

UCLA · AltaMed 등 연구팀/JMIR. Propagation of Interpreter Errors by Ambient AI Scribes. <https://medinform.jmir.org/2026/1/e88734>

RSNA. AI Bias May Impair Radiologist Accuracy on Mammogram.  
<https://www.rsna.org/media/press/i/2432>

연구팀/Radiology. Automation Bias in Mammography.  
<https://pubs.rsna.org/doi/10.1148/radiol.222176>

Stanford 연구팀. MedAgentBench 프로젝트. <https://stanfordmlgroup.github.io/projects/medagentbench/>

Stanford 연구팀. MedAgentBench 사전공개. <https://arxiv.org/abs/2501.14654>

Stanford University, Yixing Jiang · Jonathan H. Chen 등. MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents.  
<https://doi.org/10.1056/Aldbp2500144>

U.S. FDA. IDx-DR De Novo 결정. [https://www.accessdata.fda.gov/cdrh\\_docs/pdf18/DEN180001.pdf](https://www.accessdata.fda.gov/cdrh_docs/pdf18/DEN180001.pdf)

AMA. CPT Appendix S. <https://www.ama-assn.org/practice-management/cpt/cpt-appendix-s-taxonomy-artificial-intelligence-medical-services-procedures>

Bipartisan Policy Center. Paying for AI in U.S. Health Care.  
<https://bipartisanpolicy.org/issue-brief/paying-for-ai-in-u-s-health-care/>

Grand View Research. Agentic AI in Healthcare Market.  
<https://www.grandviewresearch.com/horizon/outlook/agentic-ai-in-healthcare-market-size/global>

European Commission. Article 14 Human oversight.  
<https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-14>

Johns Hopkins University, Stanford University, ETH Zurich 등 공동 연구진.  
AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. <https://arxiv.org/abs/2405.07960>

Johns Hopkins University, Stanford University, ETH Zurich 등 공동 연구진.  
AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. <https://arxiv.org/html/2405.07960v5>

Westlake University, Ant Group, CAIR HKISI CAS, China-Japan Friendship Hospital. MedLA: A Logic-Driven Multi-Agent Framework for Complex Medical Reasoning with Large Language Models. <https://arxiv.org/abs/2509.23725>

Westlake University, Ant Group, CAIR HKISI CAS, China-Japan Friendship Hospital. MedLA: A Logic-Driven Multi-Agent Framework for Complex Medical Reasoning with Large Language Models. <https://arxiv.org/html/2509.23725v3>

Princeton University, Google Research. ReAct: Synergizing Reasoning and Acting in Language Models. <https://arxiv.org/abs/2210.03629>

Princeton University, Google Research. ReAct: Synergizing Reasoning and Acting in Language Models. <https://arxiv.org/html/2210.03629v3>

Northeastern University, MIT, Princeton University. Reflexion: Language Agents with Verbal Reinforcement Learning. <https://arxiv.org/abs/2303.11366>

Northeastern University, MIT, Princeton University. Reflexion: Language Agents with Verbal Reinforcement Learning. <https://arxiv.org/html/2303.11366v4>

Stanford University, Yixing Jiang · Jonathan H. Chen 등. MedAgentBench: Dataset for Benchmarking LLMs as Agents in Medical Applications. <https://arxiv.org/html/2501.14654v1>

Princeton University, Theodore R. Sumers · Shunyu Yao 등. Cognitive Architectures for Language Agents. <https://arxiv.org/abs/2309.02427>

Princeton University, Theodore R. Sumers · Shunyu Yao 등. Cognitive Architectures for Language Agents. <https://arxiv.org/html/2309.02427v3>

University of California, Berkeley. MemGPT: Towards LLMs as Operating Systems. <https://arxiv.org/abs/2310.08560>

University of California, Berkeley. MemGPT: Towards LLMs as Operating Systems. <https://arxiv.org/html/2310.08560v2>

Latent Labs Team. Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design. <https://arxiv.org/abs/2603.29727>

Shubham Vatsal, Harsh Dubey, Aditi Singh. Agentic AI in Healthcare & Medicine: A Seven-Dimensional Taxonomy for Empirical Evaluation of LLM-based Agents. <https://arxiv.org/abs/2602.04813>

Yale University, Shanghai Jiao Tong University. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. <https://arxiv.org/abs/2311.10537>

Yale University, Shanghai Jiao Tong University. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. <https://arxiv.org/html/2311.10537v4>

MIT, Google Research, Seoul National University Hospital 공동 연구진. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making.

<https://arxiv.org/abs/2404.15155>

MIT, Google Research, Seoul National University Hospital 공동 연구진. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making.

<https://arxiv.org/html/2404.15155v3>

Tsinghua University AIR 등. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. <https://arxiv.org/abs/2405.02957>

Tsinghua University AIR 등. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. <https://arxiv.org/html/2405.02957v3>

Alibaba, Huazhong University of Science and Technology, University of Southern California, Fudan University. AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator. <https://arxiv.org/abs/2402.09742>

Alibaba, Huazhong University of Science and Technology, University of Southern California, Fudan University. AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator.

<https://aclanthology.org/2025.coling-main.680/>

단일 개발 주체 없음; Microsoft 등이 목표로 표현. Building Medical Superintelligence.

<https://microsoft.ai/news/building-medical-superintelligence/>

Fudan University, Shanghai Innovation Institute, Zhejiang University, Huazhong University of Science and Technology. ClinVerse: A Realistic Multi-Agent Diagnostic Environment for Training and Evaluating Medical Agents.

<https://arxiv.org/html/2610.34878v1>

MITRE 주도 및 HIKER Group 등 공동 연구진. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. <https://academic.oup.com/jamia/article/25/3/230/4098271>

Stanford University, Together AI. Dynamic Cheatsheet: Test-Time Learning with Adaptive Memory. <https://arxiv.org/abs/2504.07952>

Stanford University, Together AI. Dynamic Cheatsheet: Test-Time Learning with Adaptive Memory. <https://arxiv.org/html/2504.07952v1>

논문 연구진 / npj Digital Medicine. From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. <https://www.nature.com/articles/s41746-026-02545-1>

플랫폼 연구: 산야인민병원·쓰촨대 관련 연구진; 상용 AI 제조사는 논문 비공개. AI-Enabled regional tele-ECG cloud platform and improving access to cardiovascular diagnosis: real-world evidence from southern China. <https://www.frontiersin.org/journals/cardiovascular-medicine/articles/10.3389/fcvm.2026.1869716/full>

National Defense Medical Center, Tri-Service General Hospital 연구진. AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. <https://pubmed.ncbi.nlm.nih.gov/38684860/>

National Defense Medical Center, Tri-Service General Hospital 연구진. 人工智慧心電圖之落地運用-三軍總醫院經驗 (2024-05-31 Grand Round 안내). <https://www.tsgh.ndmutsgh.edu.tw/news/191/10010/24887/10864>

MASAI: Lund University 등; Transpara: ScreenPoint Medical BV. Mammography Screening with Artificial Intelligence. <https://portal.research.lu.se/en/projects/mammography-screening-with-artificial-intelligence/>

MASAI: Lund University 등; Transpara: ScreenPoint Medical BV. Transpara, 510(k) K181704 decision and summary. [https://www.accessdata.fda.gov/cdrh\\_docs/pdf/18/K181704.pdf](https://www.accessdata.fda.gov/cdrh_docs/pdf/18/K181704.pdf)

MASAI: Lund University 등; Transpara: ScreenPoint Medical BV. Breast AI Research Portfolio. <https://screenpoint-medical.com/evidence>

IDx LLC, 현재 Digital Diagnostics. De Novo classification request for IDx-DR, DEN180001. [https://www.accessdata.fda.gov/cdrh\\_docs/reviews/DEN180001.pdf](https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf)

IDx LLC, 현재 Digital Diagnostics. LumineticsCore. <https://www.digitaldiagnostics.com/products/eye-disease/lumineticscore/>

Doctronic. Doctronic AI Regulatory Mitigation Agreement.

<https://commerce.utah.gov/ai/regulatory-relief-4/authorized-pilots/doctronic/>

미국 FDA 규제 지침. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>

U.S. HHS. The HIPAA Privacy Rule. <https://www.hhs.gov/hipaa/for-professionals/privacy/index.html>

European Commission. Legal framework of EU data protection. [https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection\\_en](https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection_en)

Wittmann et al.; Microsoft·IBBIS 등 연구진. Strengthening nucleic acid biosecurity screening against generative protein design tools. <https://pubmed.ncbi.nlm.nih.gov/41037625/>

OSTP / HHS. Framework for nucleic acid synthesis screening. <https://aspr.hhs.gov/S3/Documents/OSTP-Nucleic-Acid-Synthesis-Screening-Framework-Sep2024.pdf>



# 책 정보

## AI와 의료의 대전환

저 자 | 김경진

펴낸이 | 김경진

펴낸곳 | 김경진 변호사 출판사

출판사등록 | 2025. 3. 10. (제2025-000015호)

주 소 | 서울특별시 동대문구 전농로 91, 백일빌딩 304호

전 화 | 02-6338-1905

이메일 | kimkj008@gmail.com

© 김경진 2026

본 책은 저작자의 지적 재산으로서 무단 전재와 복제를 금합니다.

참고) 표지와 장별 삽화는 인공지능으로 생성한 개념 표현이며, 실제 임상 현장을 촬영한 사진이 아닙니다. 이 책의 일부 내용과 편집 과정에는 인공지능 도구의 도움을 받았습니다.

# 후원 안내

이 책을 잘 읽으셨으면

그리고 새로운 가치있는 지식을 얻으셨다고 판단되시면

**농협 302-1096-0948-81**

(예금주 김경진)에 자발적 후원 부탁드립니다.

김경진 변호사